Journal of Machine Learning Research 26 (2025) 1-79 Submitted 11/23; Revised 2/25; Published 2/25

Outlier Robust and Sparse Estimation of Linear Regression Coefficients

Takeyuki Sasai[1] TAKEYUKI.SASAI@GMAIL.COM

Hironori Fujisawa[2,1,3] FUJISAWA@ISM.AC.JP

[1] *The Graduate Institute for Advanced Studies, SOKENDAI, Tokyo, Japan*

[2] *The Institute of Statistical Mathematics, Tokyo, Japan*

[3] *Center for Advanced Intelligence Project, RIKEN, Tokyo, Japan*

Editor: Daniel Hsu

Abstract

We consider outlier-robust and sparse estimation of linear regression coefficients, when the covariates and the noises are contaminated by adversarial outliers and noises are sampled from a heavy-tailed distribution. Our results present sharper error bounds under weaker assumptions than prior studies that share similar interests with this study. Our analysis relies on some sharp concentration inequalities resulting from generic chaining.

Keywords: learning theory, robustness, sparsity, tractability, concentration inequality

1. Introduction

This study considers outlier-robust and sparse estimation of linear regression coefficients. Consider the following sparse linear regression model:

$$y_i = \mathbf{x}_i^\top \boldsymbol{\beta}^* + \xi_i, \quad i = 1, \cdots, n, \tag{1}$$

where $\boldsymbol{\beta}^* \in \mathbb{R}^d$ represents the true coefficient vector with s nonzero elements, $\{\mathbf{x}_i\}_{i=1}^n$ denotes a sequence of independent and identically distributed (i.i.d.) random covariate vectors, and $\{\xi_i\}_{i=1}^n$ denotes a sequence of i.i.d. random noises. Throughout the present paper, we assume $s \geq 1$ and $d/s \geq 3$ for simplicity. There are many studies on estimation problems of $\boldsymbol{\beta}^*$ (Tibshirani, 1996; Fan and Li, 2001; Zou and Hastie, 2005; Yuan and Lin, 2006; Candes and Tao, 2007; Bickel et al., 2009; Raskutti et al., 2010; Zhang, 2010; Belloni et al., 2011; Dalalyan and Chen, 2012; Sivakumar et al., 2015; Su and Candes, 2016; Fan et al., 2017; Derumigny, 2018; Bellec et al., 2018; Lecué and Mendelson, 2018; Fan et al., 2021). Let $\|\mathbf{v}\|_2$ denote the ℓ_2 norm for a vector $\mathbf{v}$. Especially, using the method in Bellec et al. (2018), with probability at least $1-\delta$, we can construct an estimator $\hat{\boldsymbol{\beta}}$ such that

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \lesssim \sqrt{\frac{s\log(d/s)}{n}} + \sqrt{\frac{\log(1/\delta)}{n}}, \tag{2}$$

where $\lesssim$ is an inequality up to an absolute constant factor, when, for simplicity, $\{\mathbf{x}_i\}_{i=1}^n$ and $\{\xi_i\}_{i=1}^n$ are the sequences of i.i.d. random covariate vectors sampled from the multivariate Gaussian distribution with $\mathbb{E}\mathbf{x}_i = 0$ and $\mathbb{E}\mathbf{x}_i\mathbf{x}_i^\top = I$, and random noises sampled from the Gaussian distribution with $\mathbb{E}\xi_i = 0$ and $\mathbb{E}\xi_i^2 = 1$, respectively.

This paper considers the situation where $\{\mathbf{x}_i, y_i\}_{i=1}^n$ suffers from malicious outliers. We allow an adversary to inject outliers into (1), yielding

$$y_i = \mathbf{X}_i^\top \boldsymbol{\beta}^* + \xi_i + \sqrt{n}\theta_i, \quad i = 1, \cdots, n, \tag{3}$$

where $\mathbf{X}_i = \mathbf{x}_i + \boldsymbol{\varrho}_i$ for $i = 1, \cdots, n$, and $\{\boldsymbol{\varrho}_i\}_{i=1}^n$ and $\{\theta_i\}_{i=1}^n$ are outliers. Let $\mathcal{O}$ be the index set of outliers. We assume the following.

Assumption 1 *Assume that*

(i) the adversary can freely choose the index set $\mathcal{O}$;

(ii) $\{\boldsymbol{\varrho}_i\}_{i\in\mathcal{O}}$ and $\{\theta_i\}_{i\in\mathcal{O}}$ are allowed to be correlated freely with each other and with $\{\mathbf{x}_i\}_{i=1}^n$ and $\{\xi_i\}_{i=1}^n$;

(iii) $\boldsymbol{\varrho}_i = (0, \cdots, 0)^\top$ and $\theta_i = 0$ for $i \in \mathcal{I} = \{1, 2, \cdots, n\} \setminus \mathcal{O}$.

We note that, under Assumption 1, $\{\mathbf{x}_i\}_{i\in\mathcal{I}}$ and $\{\xi_i\}_{i\in\mathcal{I}}$ are no longer sequences of i.i.d. random variables because $\mathcal{O}$ is freely chosen by an adversary. This type of contamination by outliers is sometimes called strong contamination, in contrast to Huber contamination (Diakonikolas and Kane, 2019). Huber contamination is more manageable to tame than strong contamination because outliers of Huber contamination are not correlated to the inliers and do not destroy the independence of the inliers. We consider a problem to estimate $\boldsymbol{\beta}^*$ in (3), and construct a computationally tractable estimator having a property similar to (2).

We briefly review recent developments in robust and computationally tractable estimators. Chen et al. (2018) derived optimal error bounds for the estimation of means and covariance (scatter) matrices in the presence of outliers and proposed estimators, which achieve the optimal error bounds. However, these estimators are computationally intractable. Subsequently, Lai et al. (2016) and Diakonikolas et al. (2019b) considered tractable estimators for similar problem settings. After Lai et al. (2016) and Diakonikolas et al. (2019b), many outlier-robust tractable estimators have been developed: moment estimation (for example, Diakonikolas et al. (2017a); Kothari et al. (2018); Depersin and Lecué (2022); Cheng et al. (2019b,a); Dong et al. (2019); Lugosi and Mendelson (2021); Dalalyan and Minasyan (2022)), moment estimation with sparsity (for example, Diakonikolas et al. (2019c, 2022b); Zeng and Shen (2022); Diakonikolas et al. (2022a); Cheng et al. (2022); Prasad et al. (2020)), linear regression (for example, Balakrishnan et al. (2017); Dalalyan and Thompson (2019); Diakonikolas et al. (2019d); Chinot (2020); Cherapanamjeri et al. (2020); Lecué and Lerasle (2020); Chinot et al. (2020); Bakshi and Prasad (2021); Liu et al. (2020); Pensia et al. (2020); Minsker et al. (2024); Merad and Gaïffas (2023); Diakonikolas et al. (2024)), half-space estimation (for example, Diakonikolas et al. (2019a); Montasser et al. (2020); Diakonikolas et al. (2020, 2021)), Gaussian mixture models (for example, Diakonikolas et al. (2018); Liu and Moitra (2021)). Their primary interests are deriving sharp error bounds, deriving information-theoretically optimal bounds, and reducing computational complexity. However, there are few studies on combining outlier-robust properties with sparsity in linear regression setting (Nguyen and Tran, 2012; Chen et al., 2013; Balakrishnan et al., 2017; Diakonikolas et al., 2019c; Dalalyan and Thompson, 2019; Liu et al., 2020; Chinot, 2020; Gao, 2020; Lecué and Lerasle, 2020; Chinot et al., 2020; Sasai, 2022; Minsker et al., 2024; Merad

and Gaïffas, 2023; Diakonikolas et al., 2024). Especially, Chen et al. (2013), Balakrishnan et al. (2017) and Liu et al. (2020) dealt with the estimation problem of $\boldsymbol{\beta}^*$ from (3) under the assumption of Gaussian and subGaussian tails of $\{\mathbf{x}_i, \xi_i\}_{i=1}^n$, with computationally tractable estimators. Our study can be considered as an extension of these prior studies from two perspectives: sharpening the error bound and relaxing the assumption with improved analysis. Our estimation method builds on the approach proposed in Pensia et al. (2020). However, since Pensia et al. (2020) focuses on non-sparse settings, we introduce a version of the Hanson–Wright inequality and a refined analysis of the ℓ_1-penalized Huber loss to incorporate sparsity. Further details are provided in Section 3.2.

The present paper is organized as follows. In Section 2, we present our main results in rough statements and describe some relationships to previous studies. In Sections 3 and 5, we describe our estimation methods and main results, without proofs. In Section 6, we describe key propositions, a lemma, and a corollary without proofs. In Section 7, we provide some proofs of the propositions in Sections 3–6. In Section 8, we provide some numerical experiments. In the appendices, we provide the proofs that are omitted in Sections 3–7. In the remainder of this paper, we assume that Assumption 1 holds for outliers, and for simplicity, $0 < \delta \leq 1/4$.

2. Our Results and Relationship to Previous Studies

We state our results in Section 2.1. We compare our results with previous studies in Section 2.2.

2.1 Our Results

Before showing our results, we introduce some definitions. First, we introduce the ψ_α-norm, $\mathfrak{L}$-subGaussian random variable, affine-transformed coordinate-wise independent $\mathfrak{L}$-subGaussian random vector and $\mathfrak{L}$-subGaussian random vector, which are extensions of subGaussian random variables in high-dimensional settings.

Definition 1 (ψ_α-norm) *For a random variable f, let*

$$\|f\|_{\psi_\alpha} := \inf\{\eta > 0 : \mathbb{E}\exp(|f/\eta|^\alpha) \leq 2\} < \infty.$$

Definition 2 ($\mathfrak{L}$-subGaussian random variable) *A random variable $x \in \mathbb{R}$ with mean $\mathbb{E}x = 0$ is said to be an $\mathfrak{L}$-subGaussian random variable if x satisfies $\|x\|_{\psi_\alpha} \leq \mathfrak{L}$, where $\mathfrak{L}$ is a numerical constant such that $\mathfrak{L} \geq 1$.*

Definition 3 (Affine-transformed coordinate-wise independent $\mathfrak{L}$-subGaussian random vector) *A random vector $\mathbf{x} \in \mathbb{R}^d$ with mean $\mathbb{E}\mathbf{x} = 0$ is said to be an affine-transformed coordinate-wise independent $\mathfrak{L}$-subGaussian random vector if $\mathbf{x} = \Sigma^{\frac{1}{2}}\mathbf{z}$, where $\mathbf{z} = (z_1, \cdots, z_d)$ is a random vector whose coordinates are independent $\mathfrak{L}$-subGaussian random variable and Σ is a positive semi-definite matrix.*

Definition 4 ($\mathfrak{L}$-subGaussian random vector) *A random vector $\mathbf{x} \in \mathbb{R}^d$ with mean $\mathbb{E}\mathbf{x} = 0$ is said to be an $\mathfrak{L}$-subGaussian random vector if for any fixed $\mathbf{v} \in \mathbb{R}^d$,*

$$\|\langle \mathbf{x}, \mathbf{v}\rangle\|_{\psi_2} \leq \mathfrak{L}\left(\mathbb{E}|\langle \mathbf{x}, \mathbf{v}\rangle|^2\right)^{\frac{1}{2}}, \tag{4}$$

where the norm $\|\cdot\|_{\psi_2}$ *is defined in Definition 1 and* $\mathfrak{L}$ *is a numerical constant such that* $\mathfrak{L} \geq 1$.

We note that an affine-transformed coordinate-wise independent $\mathfrak{L}$-subGaussian random vector is an $\mathfrak{L}$-subGaussian random vector. Additionally, we note that, for example, the multivariate Gaussian random vector with covariance I is an affine-transformed coordinate-wise independent $\mathfrak{L}$-subGaussian random vector with $\mathfrak{L} = \sqrt{8/3}$. The affine-transformed coordinate-wise independent $\mathfrak{L}$-subGaussian random vector has been used to obtain non-asymptotic estimation bounds (Bartlett et al., 2020; Tsigler and Bartlett, 2023; Cheng and Montanari, 2022).

Second, we introduce the restricted eigenvalue condition for Σ (Bühlmann and Van De Geer, 2011). This allows us to treat covariates with singular covariance, which is typical of high-dimensional settings. For a vector $\mathbf{v} \in \mathbb{R}^d$, define $\mathbf{v}|_i$ as the i-th element of $\mathbf{v}$, and define the ℓ_1 norm of $\mathbf{v}$ as $\|\mathbf{v}\|_1$. For a vector $\mathbf{v} \in \mathbb{R}^d$ and set $\mathcal{J}$, define $\mathbf{v}_{\mathcal{J}}$ as a vector such that $\mathbf{v}_{\mathcal{J}}|_i = \mathbf{v}|_i$ for $i \in \mathcal{J}$ and $\mathbf{v}_{\mathcal{J}}|_i = 0$ for $i \notin \mathcal{J}$. For a set $\mathcal{J}$, define $\mathcal{J}^c$ as the complement set of $\mathcal{J}$. Additionally, for a vector $\mathbf{v}$, define the number of nonzero elements of $\mathbf{v}$ as $\|\mathbf{v}\|_0$. For a set $\mathcal{S}$, let $|\mathcal{S}|$ be the number of elements of $\mathcal{S}$. Let $o = |\mathcal{O}|$.

Definition 5 (Restricted eigenvalue condition for Σ) *The covariance matrix* Σ *is said to satisfy the restricted eigenvalue condition* $\mathrm{RE}(s, c_{\mathrm{RE}}, \mathfrak{r})$ *with some positive constants* $c_{\mathrm{RE}}, \mathfrak{r}$*, if* $\|\Sigma^{\frac{1}{2}}\mathbf{v}\|_2 \geq \mathfrak{r}\|\mathbf{v}\|_2$ *for any vector* $\mathbf{v} \in \mathbb{R}^p$ *and any set* $\mathcal{J}$ *such that* $|\mathcal{J}| \leq s$ *and* $\|\mathbf{v}_{\mathcal{J}^c}\|_1 \leq c_{\mathrm{RE}}\|\mathbf{v}_{\mathcal{J}}\|_1$.

For simplicity, we redefine $\mathfrak{r} = \inf_{\mathbf{v}\in\mathbb{R}^d,\, \|\mathbf{v}\|_0 \leq s,\, \|\mathbf{v}_{\mathcal{J}^c}\|_1 \leq c_{\mathrm{RE}}\|\mathbf{v}_{\mathcal{J}}\|_1} \frac{\|\Sigma^{\frac{1}{2}}\mathbf{v}\|_2}{\|\mathbf{v}\|_2}$. Lastly, we introduce the following two quantities related to the minimum/maximum eigenvalues:

$$\kappa_{\mathrm{l}} = \inf_{\|\mathbf{v}\|_0 \leq s} \frac{\|\Sigma^{\frac{1}{2}}\mathbf{v}\|_2}{\|\mathbf{v}\|_2}, \quad \kappa_{\mathrm{u}} = \sup_{\|\mathbf{v}\|_0 \leq 2s^2} \frac{\|\Sigma^{\frac{1}{2}}\mathbf{v}\|_2}{\|\mathbf{v}\|_2}.$$

We note that, from the definition, we see that the minimum eigenvalue of Σ is smaller than κ_{l}. Define $\rho = \max_{i\in\{1,\cdots,d\}} \sqrt{\Sigma_{ii}}$ and the maximum eigenvalue of Σ as $\Sigma_{\max}$. In the present paper, we present four main results. For the first and second results, we make the following assumption on $\{\mathbf{x}_i, \xi_i\}_{i=1}^n$:

Assumption 2 *Assume that*

(i) $\{\mathbf{x}_i\}_{i=1}^n$ *is a sequence of i.i.d. random vectors sampled from an affine-transformed coordinate-wise independent* $\mathfrak{L}$*-subGaussian random vector with* $\mathbb{E}\mathbf{x}_i = 0$, $\mathbb{E}\mathbf{x}_i^\top \mathbf{x}_i = \Sigma$, $\rho \geq 1$ *and* $\kappa_{\mathrm{l}} > 0$*. Assume that* Σ *satisfies* $\mathrm{RE}(s, c_{\mathrm{RE}}, \mathfrak{r})$ *with* $c_{\mathrm{RE}} > 1$ *and* $\mathfrak{r} \leq 1$*;*

(ii) $\{\xi_i\}_{i=1}^n$ *is a sequence of i.i.d. random variables with* $\mathbb{E}\xi_i^2 \leq \sigma^2$*;*

(iii) $\{\xi_i\}_{i=1}^n$ *and* $\{\mathbf{x}_i\}_{i=1}^n$ *are independent.*

Define $\lesssim_{c_{\mathrm{RE}}}$ as an inequality up to an absolute constant factor and c_{RE}. Define

$$R_{\mathrm{lasso}} = \frac{\rho}{\mathfrak{r}}\sqrt{\frac{s\log(d/s)}{n}} + \sqrt{\frac{\log(1/\delta)}{n}},$$
$$R_{\mathrm{outlier}} = \frac{\kappa_{\mathrm{u}}}{\kappa_{\mathrm{l}}}\left(\sqrt{\frac{o}{n}}\sqrt{s\sqrt{\frac{\log(d/s)}{n}} + \sqrt{\frac{\log(1/\delta)}{n}}} + \frac{o}{n}\sqrt{\log\frac{n}{o}}\right),$$
$$R'_{\mathrm{outlier}} = \frac{\kappa_{\mathrm{u}}}{\kappa_{\mathrm{l}}}\sqrt{\frac{o}{n}}\sqrt{s\sqrt{\frac{\log(d/s)}{n}} + \sqrt{\frac{\log(1/\delta)}{n}}} + \sqrt{\frac{o}{n}}\frac{\Sigma_{\max}}{\kappa_{\mathrm{l}}}.$$

Our first result is as follows (for a precise statement, see Theorem 13 in Section 3.3).

Theorem 6 *Suppose that Assumption 2 holds, and Σ is known to the algorithm. Then, with probability at least $1-4\delta$, we can construct an estimator $\hat{\boldsymbol{\beta}}$ such that*

$$\|\Sigma^{\frac{1}{2}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_2 \lesssim_{c_{\mathrm{RE}}} \mathfrak{L}^3\sigma\left(R_{\mathrm{lasso}} + R_{\mathrm{outlier}}\right), \tag{5}$$

with a computationally tractable method, when R_{lasso} and R_{outlier} are sufficiently small.

Our second result is as follows (for a precise statement, see Theorem 15 in Section 4).

Theorem 7 *Suppose that Assumption 2 holds, and Σ is unknown to the algorithm. Then, with probability at least $1-3\delta$, we can construct an estimator $\hat{\boldsymbol{\beta}}$ such that*

$$\|\Sigma^{\frac{1}{2}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_2 \lesssim_{c_{\mathrm{RE}}} \mathfrak{L}^3\sigma\left(R_{\mathrm{lasso}} + R'_{\mathrm{outlier}}\right), \tag{6}$$

with a computationally tractable method, when R_{lasso} and R'_{outlier} are sufficiently small.

From (5) and (6), we see that the error bounds of our estimators match those of the normal lasso, up to $\mathfrak{L}$ and numerical constant factors, when there are no outliers because R_{lasso} is equivalent to the upper bound in (2) up to constant factors and $R_{\mathrm{outlier}} = 0$ when there are no outliers. We see that R'_{outlier} is larger than R_{outlier}. From the fact that R'_{outlier} is larger than R_{outlier}, we see that there is a deterioration in the error bounds, as a trade-off for not utilizing Σ in the estimation.

Remark 8 *Condition (ii) in Assumption 2 is provided to make the results simple, and condition (ii) can be weakened to a tail probability condition. For details, see Assumption 4.*

Next, we state our third and fourth results. In these results, we relax the assumption on the covariates and we deal with the case where the covariates are $\mathfrak{L}$-subGaussian. We make the following assumption.

Assumption 3 *Assume that $\{\mathbf{x}_i\}_{i=1}^n$ is a sequence of i.i.d. random vectors sampled from an $\mathfrak{L}$-subGaussian random vector with $\mathbb{E}\mathbf{x}_i = 0$, $\mathbb{E}\mathbf{x}_i^\top\mathbf{x}_i = \Sigma$, $\rho \geq 1$ and $\kappa_{\mathrm{l}} > 0$. Assume that Σ satisfies $\mathrm{RE}(s, c_{\mathrm{RE}}, \mathfrak{r})$ with $c_{\mathrm{RE}} > 1$ and $\mathfrak{r} \leq 1$.*

Define

$$R''_{\mathrm{outlier}} = \sqrt{\frac{o}{n}} \times \left(\frac{\mathfrak{L}+\rho}{\kappa_{\mathrm{l}}} \sqrt{s\sqrt{\frac{\log d}{n}} + s\sqrt{\frac{\log(1/\delta)}{n}}} + \mathfrak{L}\frac{\Sigma_{\max}}{\kappa_{\mathrm{l}}} \right).$$

Then, our third result is as follows (for a precise statement, see Corollary 17 in Section 5.1).

Theorem 9 *Suppose that Assumption 3 and conditions (ii) and (iii) of Assumption 2 hold, and Σ is unknown to the algorithm. Then, with probability at least $1-3\delta$, we can construct an estimator $\hat{\boldsymbol{\beta}}$ such that*

$$\|\Sigma^{\frac{1}{2}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_2 \lesssim_{c_{\mathrm{RE}}} \mathfrak{L}^2\sigma \left(R_{\mathrm{lasso}} + R''_{\mathrm{outlier}} \right),$$

with a computationally tractable method, when R_{lasso} and R''_{outlier} are sufficiently small.

The result above closely resembles the previous one. While the assumption on the covariates has been relaxed, the error bound is less sharp compared to the previous result, with $\log(d/s)$ becoming $\log d$ and $\log(1/\delta)$ turning into $s^2\log(1/\delta)$. On the other hand, there is a slight improvement in the dependence on κ_{u}.

As a byproduct of the previous results, we have the fourth result, which deals with the case where only the outputs are contaminated by outliers. The fourth result is as follows (for a precise statement, see Corollary 18 in Section 5.2). The error bound in Theorem 13 is much sharper than the previous results because there is no dependence on $\Sigma_{\max}, \kappa_{\mathrm{l}}$ and κ_{u}. Additionally, n is not required to be larger than s^2.

Theorem 10 *Suppose that Assumption 3 and conditions (ii) and (iii) of Assumption 2 hold, and Σ is unknown to the algorithm. Assume that $\{y_i\}_{i=1}^n$ is generated by*

$$y_i = \mathbf{x}_i^\top \boldsymbol{\beta}^* + \xi_i + \sqrt{n}\theta_i, \quad i = 1, \cdots, n. \tag{7}$$

Then, with probability at least $1-3\delta$, we can construct an estimator $\hat{\boldsymbol{\beta}}$ such that

$$\|\Sigma^{\frac{1}{2}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_2 \lesssim_{c_{\mathrm{RE}}} \mathfrak{L}^3\sigma \left(R_{\mathrm{lasso}} + \frac{o}{n}\sqrt{\log\frac{n}{o}} \right), \tag{8}$$

with a computationally tractable method, when $R_{\mathrm{lasso}} + \frac{o}{n}\sqrt{\log\frac{n}{o}}$ is sufficiently small.

In Theorems 13 and 15 and Corollaries 17 and 18, we explicitly describe the relationship between the tuning parameters and the error bounds in Theorems 6, 7, 9 and 10 respectively. Additionally, in Theorems 13 and 15 and Corollaries 17 and 18, we derive error bounds not only for $\|\Sigma^{\frac{1}{2}}(\cdot)\|_2$ but also for $\|\cdot\|_2$ and $\|\cdot\|_1$.

2.2 Relationship to Previous Studies

As we mentioned in Section 1, Chen et al. (2013); Balakrishnan et al. (2017); Liu et al. (2020) dealt with the estimation problem of $\boldsymbol{\beta}^*$ from (3) under Assumption 1 with computationally tractable estimators.

We note that Lecué and Lerasle (2020); Chinot et al. (2020); Gao (2020) dealt with the estimation problem of $\boldsymbol{\beta}^*$ from (3) under a stronger assumption on outlier than Assumption 1, with computationally intractable estimators that derive sharp error bounds. Diakonikolas et al. (2024) derived a sharp error bound with tractable estimator when restrictive conditions: covariates and noises are Gaussian, $\|\boldsymbol{\beta}^*\|_2$ is small and outliers follow Huber's contamination, which is a stronger assumption on outliers than Assumption 1. Additionally, we note that Sasai (2022) dealt with a situation where $\{\mathbf{x}_i\}_{i=1}^n$ is a sequence of i.i.d. random vectors sampled from a heavy-tailed distribution, and Merad and Gaïffas (2023) dealt with more challenging situation weakening the assumptions for covariates than that of Sasai (2022). Their error bounds are looser than the results of the present paper because the weak assumptions restrict the techniques available. Therefore, these papers (Lecué and Lerasle, 2020; Chinot et al., 2020; Gao, 2020; Sasai, 2022; Merad and Gaïffas, 2023) treat computationally intractable estimators or suppose weaker assumptions than our method, and hence we do not mention such papers further because the interests of such papers are different from those of our paper. Therefore, we mainly discuss the results of Chen et al. (2013); Balakrishnan et al. (2017); Liu et al. (2020) in the remainder of Sections 2.2.1 – 2.2.3. About the case where only the output is contaminated, in other words, we estimate $\boldsymbol{\beta}^*$ from $\{\mathbf{x}_i, y_i\}_{i=1}^n$, where $\{y_i\}_{i=1}^n$ is generated by (7), we discuss in Section 2.2.4.

2.2.1 Case where Σ is allowed to be used in estimation

The results of Balakrishnan et al. (2017) and part of the results of Liu et al. (2020) use Σ in their estimation. Balakrishnan et al. (2017) and Liu et al. (2020) considered situations where the covariate vectors are sampled from the standard multivariate Gaussian distribution and the noises are sampled from a Gaussian distribution with mean 0 and variance σ^2. In contrast, our method works well for the case where the covariate vectors are sampled from an affine-transformed coordinate-wise independent $\mathfrak{L}$-subGaussian random vector with covariance which satisfies the restricted eigenvalue condition and the noises are sampled from a heavy-tailed distribution. Balakrishnan et al. (2017) and Liu et al. (2020) only considered the case where o/n is a sufficiently small constant. Let $o/n = \mathfrak{e}$. The ℓ_2-norm error bound of Balakrishnan et al. (2017) is $\lesssim_\sigma (\sqrt{1 + \|\boldsymbol{\beta}^*\|_2^2} \frac{o}{n} \log^2 \frac{n}{o})$ when $\frac{s^2}{\mathfrak{e}^2} \log d + \frac{s^2}{\mathfrak{e}^2} \log(1/\delta) \lesssim n$, where $\lesssim_\sigma$ is the inequality up to an absolute constant factor and the standard deviation of the random noise σ. The ℓ_2-norm error bound of Liu et al. (2020) is $\lesssim \sigma(\frac{o}{n} \log \frac{n}{o})$ when $\left(\frac{s^2}{\mathfrak{e}^2} \log(dT) + \frac{s^2}{\mathfrak{e}^2} \log(1/\delta)\right) \times T \lesssim n$, with $\log\left(\frac{\|\boldsymbol{\beta}^*\|_2}{\mathfrak{e}\sigma}\right) \lesssim T$. Under the same situation, the ℓ_2-norm error bound of our result becomes $\lesssim \sigma \frac{o}{n} \sqrt{\log \frac{n}{o}}$ when $\frac{s^2}{\mathfrak{e}^2} \log(d/s) + \frac{1}{\mathfrak{e}^2} \log(1/\delta) \lesssim n$. We see that our result does not depend on $\boldsymbol{\beta}^*$ and the error bound is sharper than the ones of Balakrishnan et al. (2017) and Liu et al. (2020). Additionally, our sample complexity is smaller than that of Balakrishnan et al. (2017) and Liu et al. (2020) because, in our sample complexity, the term such that $\frac{s^2}{\mathfrak{e}} \times \log(1/\delta)$ does not appear.

We consider the optimality of the error bound in (5). From Theorem D.3 of Cherapanamjeri et al. (2020), we see that the error bound cannot avoid a term such that *constant* $\times \sigma \frac{o}{n} \sqrt{\log \frac{n}{o}}$ even when $d = 1$. Detailed investigations of the influence of Σ in high dimensions on information-theoretic limits are a task for future research.

When there are no outliers ($o = 0$), our error bound coincides with that of the normal lasso, up to numerical and $\mathfrak{L}$ factors. The results in Balakrishnan et al. (2017) and Liu et al. (2020) do not have this property.

2.2.2 Case where Σ is not allowed to be used in estimation

Chen et al. (2013), and a part of the results of Liu et al. (2020) give error bounds with tractable methods and do not require Σ for estimation. However, the method in Liu et al. (2020) assumes a sparse structure of Σ, and the sample complexity depends on not only s^2 but also the sparsity of Σ. Chen et al. (2013) proposed some methods, however, the term in their error bounds containing o depends on $\log d$ and s even when $s^2 \log(d/s) \lesssim n$.

When there are no outliers ($o = 0$), similarly to the case where Σ is allowed to be used in estimation, our error bound coincides with that of the normal lasso, up to a numerical and $\mathfrak{L}$ factors. The results in Liu et al. (2020) and Chen et al. (2013) do not have this property.

2.2.3 Remaining problems

Not only our estimator but also the estimators proposed by Balakrishnan et al. (2017) and Liu et al. (2020) require that n to be proportional to s^2, which is not needed to derive (2) from (1). Similar phenomena can be observed in Wang et al. (2016); Fan et al. (2021); Liu et al. (2020); Balakrishnan et al. (2017); Diakonikolas et al. (2019c). Some relationships between computational tractability and similar quadratic dependencies are unraveled (Wang et al., 2016; Diakonikolas et al., 2019d, 2017b; Brennan and Bresler, 2020; Georgiev and Hopkins, 2022). We leave the analysis in our situation for future work.

Theorems 6 and 7 do not hold when the covariates are $\mathfrak{L}$-subGaussian random vectors. Instead, the assumption needs to be strengthened to an affine-transformed coordinate-wise independent $\mathfrak{L}$-subGaussian random vector. This is because the Hanson–Wright inequality (Hanson and Wright, 1971; Wright, 1973; Adamczak, 2015; Hsu et al., 2012; Rudelson and Vershynin, 2013), which is essential for analyzing our method, does not hold in its ideal form for $\mathfrak{L}$-subGaussian random vectors. For more advanced discussions on this point, see Spokoiny (2023); Dereziński (2023). Whether an error bound equivalent to ours can be derived under the assumption that the covariates are $\mathfrak{L}$-subGaussian random vectors is left for future work.

2.2.4 Case where only the output is contaminated

Nguyen and Tran (2012); Dalalyan and Thompson (2019); Chinot (2020); Thompson (2020); Minsker et al. (2024) considered the case where only the output is contaminated. Chinot (2020) dealt with 'weaker' outliers than ours, which maintain the independence of $\{\mathbf{x}_i\}_{i=1}^n$. Nguyen and Tran (2012); Dalalyan and Thompson (2019) explored the case where the covariates and noises are a multivariate Gaussian vectors and a Gaussian variables, respectively. Thompson (2020); Minsker et al. (2024) address the case where the covariates are $\mathfrak{L}$-subGaussian random vectors, and the the noises are drawn from an $\mathfrak{L}$-subGaussian distribution or a distribution with heavier tails than $\mathfrak{L}$-subGaussian distribution, deriving sharper error bounds than Nguyen and Tran (2012); Dalalyan and Thompson (2019). Our error bound (8) is sharper than those of Thompson (2020); Minsker et al. (2024) because

their error bounds depend on $\frac{o}{n}\log\frac{n}{o}$. However, the methods in Thompson (2020) and Minsker et al. (2024) do not require the normalization of the covariates. Additionally, the methods in Thompson (2020) do not require knowledge about s and o in constructing the estimator, by utilizing SLOPE techniques (Bellec et al., 2018). Minsker et al. (2024) used square root lasso techniques (Stucky and Van De Geer, 2017; Derumigny, 2018; Belloni et al., 2011) in addition to SLOPE, and their estimator does not require knowledge of s, o and σ. However, our estimator requires knowledge of the approximate values of s, o and σ. Whether we can achieve tuning-free estimation while maintaining the sharpness of the results remains a topic for future investigation.

2.3 Another Contribution

Up to this point, we have discussed the contributions of our paper from the perspective of the sharpness of the error bounds, but the technique we introduce to analyze the ℓ_1-penalized Huber loss may also be of independent interest. These points will be discussed in Section J.1.

3. Method and the First Result

Assume that $\mathbf{x}$ is a random vector drawn from the same distribution as $\{\mathbf{x}_i\}_{i=1}^n$. Hereafter, we often use the following simplified notations to express error orders:

$$r_{d,s} = \sqrt{\frac{\log(d/s)}{n}}, \quad r_\delta = \sqrt{\frac{\log(1/\delta)}{n}}, \quad r_o = \frac{o}{n}\sqrt{\log\frac{n}{o}}, \quad r'_o = \frac{o}{n}\log\frac{n}{o}.$$

3.1 Some Properties of $\mathfrak{L}$-subGaussian Random Vector

We show some additional properties of an $\mathfrak{L}$-subGaussian random vector $\mathbf{x}$. We note that, from (4), we have

$$\|\langle \mathbf{x}, \mathbf{v}\rangle\|_{\psi_2} \le \mathfrak{L}\left(\mathbb{E}|\langle \mathbf{x}, \mathbf{v}\rangle|^2\right)^{\frac{1}{2}} \le \mathfrak{L}\|\Sigma^{\frac{1}{2}}\mathbf{v}\|_2, \tag{9}$$

and from (2.14) - (2.16) of Vershynin (2018), for any $\mathbf{v} \in \mathbb{R}^d$ and $t \ge 0$, we have

$$\|\mathbf{v}^\top\mathbf{x}\|_{L_p}\left[:= \left\{\mathbb{E}|\mathbf{v}^\top\mathbf{x}|^p\right\}^{\frac{1}{p}}\right] \le c_{\mathfrak{L}}\sqrt{p}\|\mathbf{v}^\top\mathbf{x}\|_{\psi_2} \le c_{\mathfrak{L}}\sqrt{p}\mathfrak{L}\|\Sigma^{\frac{1}{2}}\mathbf{v}\|_2, \tag{10}$$

$$\mathbb{E}\exp(\mathbf{v}^\top\mathbf{x}) \le \exp(c_{\mathfrak{L}}^2\mathfrak{L}^2\|\Sigma^{\frac{1}{2}}\mathbf{v}\|_2^2),$$

$$\mathbb{E}\exp\left(\frac{(\mathbf{v}^\top\mathbf{x})^2}{c_{\mathfrak{L}}^2\mathfrak{L}^2\|\Sigma^{\frac{1}{2}}\mathbf{v}\|_2^2}\right) \le 2, \tag{11}$$

$$\mathbb{P}\left(|\mathbf{v}^\top\mathbf{x}| > t\right) \le 2\exp\left(-\frac{t^2}{c_{\mathfrak{L}}^2\mathfrak{L}^2\|\Sigma^{\frac{1}{2}}\mathbf{v}\|_2^2}\right), \tag{12}$$

where $c_{\mathfrak{L}}$ is a numerical constant. Define $L = \mathfrak{L} \times \max\{1, c_{\mathfrak{L}}\}$. We note that these properties hold for not only $\mathfrak{L}$-subGaussian random vectors but also affine-transformed coordinate-wise independent $\mathfrak{L}$-subGaussian random vectors.

3.2 Method

To estimate $\boldsymbol{\beta}^*$ in (3), we propose OUTLIER-ROBUST-AND-SPARSE-ESTIMATION (Algorithm 1). Algorithm 1 is similar to those used in Pensia et al. (2020) and Sasai (2022). However, Pensia et al. (2020) considered the non-sparse case, and Sasai (2022) considered heavy-tailed covariates. Therefore, to consider the sparsity of $\boldsymbol{\beta}^*$ or to derive a sharper error bound than that in Sasai (2022) taking advantage of the $\mathfrak{L}$-subGaussian assumption, our analysis is more involved than that of Pensia et al. (2020) and Sasai (2022). Concretely, unlike the previous studies, extensions of the Hanson–Wright inequality, which appear later in Proposition 20 and Corollary 28 proved via generic chaining, play important roles. We will analyze the ℓ_1-penalized Huber loss in step 3 of Algorithm 1. Our analysis of the ℓ_1-penalized Huber loss is similar to those in Alquier et al. (2019) and Lecué and Mendelson (2018), however, the analyses of Alquier et al. (2019) and Lecué and Mendelson (2018) are mainly interested in the case $\mathbb{E}\mathbf{x}\mathbf{x}^\top = I$, and limited effective for a more general covariance. We modify their analysis and our analysis is effective for a more general covariance. In particular, Proposition 21 is important and the modified analysis method is described in Appendices C and D.

Algorithm 1 OUTLIER-ROBUST-AND-SPARSE-ESTIMATION

Input: $\{y_i, \mathbf{X}_i\}_{i=1}^n$, $\Sigma(= \mathbb{E}\mathbf{x}\mathbf{x}^\top)$ and tuning parameters $\tau_{\mathrm{cut}}, \varepsilon, r_1, r_2, \lambda_o, \lambda_s$
Output: $\hat{\boldsymbol{\beta}}$
1: $\{\hat{w}_i\}_{i=1}^n \leftarrow \mathrm{WEIGHT}(\{\mathbf{X}_i\}_{i=1}^n, \tau_{\mathrm{cut}}, \varepsilon, r_1, r_2, \Sigma)$
2: $\{\hat{w}'_i\}_{i=1}^n \leftarrow \mathrm{TRUNCATION}(\{\hat{w}_i\}_{i=1}^n)$
3: $\hat{\boldsymbol{\beta}} \leftarrow$ WEIGHTED-PENALIZED-HUBER-REGRESSION $(\{y_i, \mathbf{X}_i\}_{i=1}^n, \{\hat{w}'_i\}_{i=1}^n, \lambda_o, \lambda_s)$

Here, we give simple explanations of the output steps. The details are provided in Sections 3.2.1, 3.2.2, and 3.2.3. The first step produces the weights $\{\hat{w}_i\}_{i=1}^n$ reducing adverse effects of covariate outliers. The second step is the truncation of the computed weights to zero or $1/n$, say $\{\hat{w}'_i\}_{i=1}^n$. The third step performs the ℓ_1-penalized Huber regression based on the weighted errors using the truncated weights $\{\hat{w}'_i\}_{i=1}^n$. The ℓ_1-penalization addresses the high-dimensional setting, and the Huber regression weakens the adverse effects of the response outliers.

3.2.1 WEIGHT

For a matrix $M = (m_{ij})_{1\le i\le d_1, 1\le j\le d_2} \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$, we define

$$\|M\|_1 = \sum_{i=1}^{d_1}\sum_{j=1}^{d_2} |m_{ij}|.$$

For a symmetric matrix M, we write $M \succeq 0$ if M is positive semi-definite. Define $\mathrm{Tr}(M)$ for a square matrix M as the trace of M. Define the following convex set:

$$\mathfrak{M}_{r_1,r_2,d}^{\ell_1,\mathrm{Tr}} = \left\{M \in \mathcal{S}(d) \mid \|M\|_1 \le r_1^2,\, \mathrm{Tr}(M) \le r_2^2,\, M \succeq 0\right\},$$

where $\mathcal{S}(d)$ is a set of symmetric matrices in $\mathbb{R}^d \times \mathbb{R}^d$. For a vector $\mathbf{v}$, we define the ℓ_∞ norm of $\mathbf{v}$ as $\|\mathbf{v}\|_\infty$ and define the probability simplex $\Delta^{n-1}(\varepsilon)$ with $0 < \varepsilon < 1$ as follows:

$$\Delta^{n-1}(\varepsilon) = \left\{ \mathbf{w} \in [0,1]^n \mid \sum_{i=1}^n w_i = 1, \quad \|\mathbf{w}\|_\infty \leq \frac{1}{n(1-\varepsilon)} \right\}.$$

The first step of Algorithm 1, WEIGHT, is stated as follows.

Algorithm 2 WEIGHT

Input: data $\{\mathbf{X}_i\}_{i=1}^n$, tuning parameters $\tau_{\mathrm{cut}}, \varepsilon, r_1, r_2$
Output: weight estimate $\hat{\mathbf{w}} = \{\hat{w}_1, \cdots, \hat{w}_n\}$

Let $\hat{\mathbf{w}}$ be the solution to

$$\min_{\mathbf{w} \in \Delta^{n-1}(\varepsilon)} \max_{M \in \mathfrak{M}_{r_1,r_2,d}^{\ell_1,\mathrm{Tr}}} \sum_{i=1}^n w_i \langle \mathbf{X}_i \mathbf{X}_i^\top - \Sigma, M \rangle \tag{13}$$

if the optimal value of (13) $\leq \tau_{\mathrm{cut}}$
  return $\hat{\mathbf{w}}$
else
  return $fail$

Algorithm 2 is a special case of Algorithm 3 of Balakrishnan et al. (2017). Therefore, as in Algorithm 3 of Balakrishnan et al. (2017), Algorithm 2 can also be computed efficiently. An intuitive meaning of (13) is given in Section 3.2.4. For details of the value of τ_{cut} and its validity, see Theorem 13 and Proposition 20, respectively.

3.2.2 TRUNCATION

The second step in Algorithm 1 is the discretized truncation of $\{\hat{w}_i\}_{i=1}^n$, say $\{\hat{w}_i'\}_{i=1}^n$, as in Algorithm 3. The discretized truncation makes it easy to analyze the estimator.

Algorithm 3 TRUNCATION

Input: weight vector $\hat{\mathbf{w}} = \{\hat{w}_i\}_{i=1}^n$
Output: truncated weight vector $\hat{\mathbf{w}}' = \{\hat{w}_i'\}_{i=1}^n$

For $i = 1 : n$
  if $\hat{w}_i \geq \frac{1}{2n}$
    $\hat{w}_i' = \frac{1}{n}$
  else
    $\hat{w}_i' = 0$
return $\hat{\mathbf{w}}'$

3.2.3 WEIGHTED-PENALIZED-HUBER-REGRESSION

The Huber loss function $H(t)$ is defined as follows:

$$H(t) = \begin{cases} |t| - 1/2 & (|t| > 1) \\ t^2/2 & (|t| \leq 1) \end{cases},$$

and let

$$h(t) = \frac{d}{dt}H(t) = \begin{cases} \mathrm{sgn}(t) & (|t| > 1) \\ t & (|t| \leq 1) \end{cases}.$$

We consider the ℓ_1-penalized Huber regression with the weighted samples $\{\hat{w}_i' y_i, \hat{w}_i' X_i\}_{i=1}^n$ in Algorithm 4. This is the third step in Algorithm 1.

Algorithm 4 WEIGHTED-PENALIZED-HUBER-REGRESSION

Input: data $\{y_i, \mathbf{X}_i\}_{i=1}^n$, truncated weight vector $\hat{\mathbf{w}}' = \{\hat{w}_i'\}_{i=1}^n$ and tuning parameters λ_o, λ_s

Output: estimator $\hat{\boldsymbol{\beta}}$

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^d}{\mathrm{argmin}} \sum_{i=1}^n \lambda_o^2 H\left(n\hat{w}_i' \frac{y_i - \mathbf{X}_i^\top \boldsymbol{\beta}}{\lambda_o \sqrt{n}}\right) + \lambda_s \|\boldsymbol{\beta}\|_1$$

return $\hat{\boldsymbol{\beta}}$

Several studies, such as Nguyen and Tran (2012); She and Owen (2011); Dalalyan and Thompson (2019); Sun et al. (2020); Chen and Zhou (2020); Chinot (2020); Pensia et al. (2020); Sasai (2022), have suggested that the Huber loss is effective for linear regression under heavy-tailed noise or the existence of outliers.

Lastly, we introduce the assumption on $\{\xi_i\}_{i=1}^n$:

Assumption 4 *(i)* $\{\xi_i\}_{i=1}^n$ *is a sequence of i.i.d. random variables such that*

$$\mathbb{P}\left(\frac{\xi_i}{\lambda_o \sqrt{n}} \geq \frac{1}{2}\right) \leq \frac{1}{144L^4}; \tag{14}$$

(ii) $\mathbb{E}h\left(\frac{\xi_i}{\lambda_o \sqrt{n}}\right) \times \mathbf{x}_i = 0.$

Remark 11 *For example, when* $\mathbb{E}\xi_i^2 \leq \sigma^2$*, from Markov's inequality, we have*

$$\mathbb{P}\left(\frac{\xi_i}{\lambda_o \sqrt{n}} \geq \frac{1}{2}\right) \leq \frac{4}{\lambda_o^2 n}\mathbb{E}\xi_i^2 \leq \frac{4\sigma^2}{\lambda_o^2 n}.$$

Therefore, to satisfy (14)*, it is sufficient to set*

$$24L^2\sigma \leq \lambda_o \sqrt{n}. \tag{15}$$

In this case, Condition (i) in Assumption 4 is weaker than $\mathbb{E}\xi_i^2 \leq \sigma^2$*.*

Remark 12 *Condition (ii) in Assumption 4 is weaker than the independence between* $\{\xi_i\}_{i=1}^n$ *and* $\{\mathbf{x}_i\}_{i=1}^n$*.*

3.2.4 An intuitive meaning of Algorithm 2

We explain an intuitive meaning of WEIGHT. For the sake of explanation, we introduce several definitions: For $l = 0, 1, 2$, we define d-dimensional ℓ_l-ball of radius a as $a\mathbb{B}_l^d = \{\mathbf{v} \in \mathbb{R}^d \mid \|\mathbf{v}\|_l \leq a\}$, and we also define $a\mathbb{B}_\Sigma^d = \{\mathbf{v} \in \mathbb{R}^d \mid \|\Sigma^{\frac{1}{2}}\mathbf{v}\|_2 \leq a\}$. In our analysis, it is necessary to derive a high-probability bound for the following quantity: for appropriate values of r_1, r_2, r_Σ,

$$\sup_{\mathbf{v} \in r_2\mathbb{B}_2^d \cap r_1\mathbb{B}_1^d \cap r_\Sigma\mathbb{B}_\Sigma^d} \frac{\lambda_o\sqrt{n}}{n} \sum_{i=1}^n \langle \mathbf{X}_i, \mathbf{v}\rangle \times h\left(\frac{\xi_i + \boldsymbol{\theta}_i}{\lambda_o\sqrt{n}}\right). \tag{16}$$

To give an upper bound of (16), from $r_1\mathbb{B}_1^d \cap r_2\mathbb{B}_2^d \cap r_\Sigma\mathbb{B}_\Sigma^d \subset r_1\mathbb{B}_1^d \cap r_2\mathbb{B}_2^d$, we consider

$$\sup_{\mathbf{v} \in r_2\mathbb{B}_2^d \cap r_1\mathbb{B}_1^d} \frac{\lambda_o\sqrt{n}}{n} \sum_{i=1}^n \langle \mathbf{X}_i, \mathbf{v}\rangle \times h\left(\frac{\xi_i + \boldsymbol{\theta}_i}{\lambda_o\sqrt{n}}\right) \tag{17}$$

to make the optimization easy. We can express (17) as follows:

$$\begin{aligned}
&\sup_{\mathbf{v} \in r_2\mathbb{B}_2^d \cap r_1\mathbb{B}_1^d a} \frac{1}{n} \sum_{i=1}^n h\left(\frac{\xi_i + \boldsymbol{\theta}_i}{\lambda_o\sqrt{n}}\right) \langle \mathbf{X}_i, \mathbf{v}\rangle \\
&= \sup_{\mathbf{v} \in r_2\mathbb{B}_2^d \cap r_1\mathbb{B}_1^d} \frac{1}{n} \left(\sum_{i=1}^n h\left(\frac{\xi_i}{\lambda_o\sqrt{n}}\right) \langle \mathbf{x}_i, \mathbf{v}\rangle + \sum_{i \in \mathcal{O}} h\left(\frac{\xi_i + \boldsymbol{\theta}_i}{\lambda_o\sqrt{n}}\right) \langle \mathbf{X}_i, \mathbf{v}\rangle + \sum_{i \in \mathcal{O}} h\left(\frac{\xi_i}{\lambda_o\sqrt{n}}\right) \langle \mathbf{x}_i, \mathbf{v}\rangle \right).
\end{aligned} \tag{18}$$

The second term on the right-hand side of the above equation is the most difficult to evaluate because it includes outliers. Therefore, we aim to find a weight vector $\mathbf{w} \in \Delta^{n-1}(\varepsilon)$ that allows for an appropriate evaluation of the second term. To do this, we make the following observation: For any $\mathbf{v} \in r_1\mathbb{B}_1^d \cap r_2\mathbb{B}_2^d$,

$$\begin{aligned}
\left(\sum_{i \in \mathcal{O}} w_i h\left(\frac{\xi_i + \boldsymbol{\theta}_i}{\lambda_o\sqrt{n}}\right) \langle \mathbf{X}_i, \mathbf{v}\rangle \right)^2 &\overset{(a)}{\leq} \sum_{i \in \mathcal{O}} w_i h\left(\frac{\xi_i + \boldsymbol{\theta}_i}{\lambda_o\sqrt{n}}\right)^2 \sum_{i \in \mathcal{O}} w_i (\mathbf{X}_i^\top \mathbf{v})^2 \\
&\leq \frac{1}{1-\varepsilon} \frac{o}{n} \sum_{i \in \mathcal{O}} w_i \langle \mathbf{X}_i \mathbf{X}_i^\top, \mathbf{v}\mathbf{v}^\top \rangle \\
&\leq \frac{1}{1-\varepsilon} \left(\frac{o}{n} \sum_{i \in \mathcal{O}} w_i \langle \mathbf{X}_i \mathbf{X}_i^\top - \Sigma, \mathbf{v}\mathbf{v}^\top \rangle + \frac{o}{n} \sum_{i \in \mathcal{O}} w_i \langle \Sigma, \mathbf{v}\mathbf{v}^\top \rangle \right) \\
&\overset{(b)}{\leq} \frac{1}{1-\varepsilon} \left(\frac{o}{n} \sum_{i \in \mathcal{O}} w_i \langle \mathbf{X}_i \mathbf{X}_i^\top - \Sigma, \mathbf{v}\mathbf{v}^\top \rangle + \frac{o^2}{n^2(1-\varepsilon)} \langle \Sigma, \mathbf{v}\mathbf{v}^\top \rangle \right),
\end{aligned}$$

where (a) follows from Hölder's inequality, and (b) follows from $w_i \leq 1/(n(1-\varepsilon))$. To evaluate $\sum_{i \in \mathcal{O}} w_i \langle \mathbf{X}_i \mathbf{X}_i^\top - \Sigma, \mathbf{v}\mathbf{v}^\top \rangle$, we see that it is sufficient to evaluate

$$\sum_{i=1}^n w_i \langle \mathbf{X}_i \mathbf{X}_i^\top - \Sigma, \mathbf{v}\mathbf{v}^\top \rangle. \tag{19}$$

Therefore, to make $\sum_{i\in\mathcal{O}} w_i h\left(\frac{\xi_i+\boldsymbol{\theta}_i}{\lambda_o\sqrt{n}}\right)\langle \mathbf{X}_i, \mathbf{v}\rangle$ sufficiently small, we want to minimize (19) in $\mathbf{w}\in\Delta^{n-1}(\varepsilon)$ for any $\mathbf{v}\in r_1\mathbb{B}_1^d\cap r_2\mathbb{B}_2^d$, in other words, we want to consider

$$\min_{\mathbf{w}\in\Delta^{n-1}(\varepsilon)}\max_{\mathbf{v}\in r_1\mathbb{B}_1^d\cap r_2\mathbb{B}_2^d}\sum_{i=1}^n w_i\langle \mathbf{X}_i\mathbf{X}_i^\top-\Sigma, \mathbf{v}\mathbf{v}^\top\rangle. \tag{20}$$

A convex relaxation of (20) is (13), which is an essential part in Algorithm 2. Lastly, we note that the weight vector $\mathbf{w}$ computed by Algorithm 2 is used to weight the left-hand side of (18). As a result, it affects not only the second term on the right-hand side of (18) but also the first and third terms. However, Lemma 27 shows that the impact on the first and third terms is small.

3.2.5 Approaches of the the previous studies

In this section, we explain the approaches of the previous studies Balakrishnan et al. (2017) and Liu et al. (2020). To estimate $\boldsymbol{\beta}^*$ in (3), Balakrishnan et al. (2017) used sparse and outlier-robust mean estimation on $y_i\mathbf{X}_i$, directly. Balakrishnan et al. (2017) assumes that $\{\mathbf{x}_i\}_{i=1}^n$ and $\{\xi_i\}_{i=1}^n$ are sequences of i.i.d. random vectors sampled from the standard multivariate Gaussian distribution and random variables sampled from the standard Gaussian, respectively, and $\{\mathbf{x}_i\}_{i=1}^n$ and $\{\xi_i\}_{i=1}^n$ are independent. Then, we have

$$\mathbb{E}y_i\mathbf{x}_i=\boldsymbol{\beta}^*,\quad \mathbb{V}(y_i\mathbf{x}_i)(y_i\mathbf{x}_i)^\top=(\|\boldsymbol{\beta}^*\|_2^2+1)I+\boldsymbol{\beta}^*\boldsymbol{\beta}^{*\top},$$

where we use Isserlis' theorem, which is a formula for the multivariate Gaussian distribution. From Remark 2.2 of Chen et al. (2018), we see that the error bound of any outlier-robust estimator for the mean vector cannot avoid the effect of the operator norm of the covariance $\mathbb{V}(y_i\mathbf{x}_i)(y_i\mathbf{x}_i)^\top$. Consequently, the approach of Balakrishnan et al. (2017) cannot remove $\|\boldsymbol{\beta}^*\|_2$ from their error bound. Additionally, we note that Liu et al. (2020) proposed a gradient based-method that repeatedly updates the estimator of $\boldsymbol{\beta}^*$. Define the t-step of the update of the estimator of $\boldsymbol{\beta}^*$ as $\boldsymbol{\beta}^t$. The gradient for the next update is based on the result of an outlier-robust estimation of $\mathbf{X}_i(\mathbf{X}_i^\top\boldsymbol{\beta}^t-y_i)$. Therefore, for reason similar to that of Balakrishnan et al. (2017), the sample complexity of Liu et al. (2020) is affected by $\|\boldsymbol{\beta}^*\|_2$. On the other hand, our estimator is based on ℓ_1-penalized Huber regression, and we can avoid this problem.

3.3 Result

We state our main theorem. Define

$$R_{d,n,o}=\rho c_{r_1}\sqrt{s}r_{d,s}+r_\delta+c_{r_2}\kappa_{\mathrm{u}}\left(\sqrt{\frac{o}{n}(sr_{d,s}+r_\delta)}+r_o\right),$$

and remember that we define $L=\mathfrak{L}\times\max\{1,c_{\mathfrak{L}}\}$, where $\mathfrak{L}$ is defined in (4) and $c_{\mathfrak{L}}$ is a numerical constant.

Theorem 13 *Suppose that (i) and (iii) of Assumption 2 and Assumption 4 hold. Suppose that the parameters* $\lambda_o, \lambda_s, \varepsilon, \tau_{\mathrm{cut}}, r_1, r_2, r_\Sigma$ *satisfy*

$$1 \geq 7c_o\left(4+c_s\right)c_{\max}^2\sqrt{1+\log L}L^2R_{d,n,o}, \tag{21}$$

$$\lambda_s = c_s c_{\max}^2 L\lambda_o\sqrt{n}\frac{1}{c_{r_1}\sqrt{s}}R_{d,n,o}, \quad \varepsilon = c_\varepsilon\frac{o}{n}, \quad \tau_{\mathrm{cut}} = c_{\mathrm{cut}}(L\kappa_{\mathrm{u}})^2\left(sr_{d,s}+r_\delta+r_o'\right)r_2^2,$$

$$r_1 = c_{r_1}\sqrt{s}r_\Sigma, \quad r_2 = c_{r_2}r_\Sigma, \quad r_\Sigma = 7\left(4+c_s\right)c_{\max}^2 L\lambda_o\sqrt{n}R_{d,n,o}, \tag{22}$$

where $c_o, c_s, c_\varepsilon, c_{\mathrm{cut}}, c_{r_1}, c_{r_2}$, *and* $c_{\max}$ *are sufficiently large numerical constants such that* $c_o \geq 4$, $c_s \geq 3(c_{\mathrm{RE}}+1)/(c_{\mathrm{RE}}-1)$, $2 > c_\varepsilon \geq 1$, $c_{\mathrm{cut}} \geq c_2$, $c_{r_1} = c_{r_1}^{\mathrm{num}}(1+c_{\mathrm{RE}})/\mathfrak{r}$, $c_{r_2} = c_{r_2}^{\mathrm{num}}(3+c_{\mathrm{RE}})/\kappa_{\mathrm{l}}$, $\min\{c_{r_1}^{\mathrm{num}}, c_{r_2}^{\mathrm{num}}\} \geq 2$ *and* $c_{r_1}^{\mathrm{num}}/c_{r_2}^{\mathrm{num}} \leq 1$. *In Proposition 20 and Definition 26,* c_2 *and* $c_{\max}$ *are defined, respectively. Suppose that* $sr_{d,s} \leq 1$ *and* $0 < o/n \leq 1/(5e)$ *hold. Then, the optimal solution* $\hat{\boldsymbol{\beta}}$ *satisfies the following:*

$$\|\Sigma^{\frac{1}{2}}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)\|_2 \leq r_\Sigma, \quad \|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\|_2 \leq r_2 \text{ and } \|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\|_1 \leq r_1, \tag{23}$$

with probability at least $1-4\delta$.

We note that the conditions (21) and (22) in Theorem 13 imply

$$\lambda_o\sqrt{n} \geq c_o L r_\Sigma\sqrt{1+\log L}.$$

Remark 14 *We consider the results of* (23) *in detail. Assume that* $\mathbb{E}\xi_i^2 \leq \sigma^2$ *and that the equality in* (15) *holds. Define* $C_{c_{\mathrm{RE}},1}$, $C_{c_{\mathrm{RE}},2}$ *and* $C_{c_{\mathrm{RE}},3}$ *as constants depending on* c_{RE}. *Then, we have*

$$\begin{aligned}\|\Sigma^{\frac{1}{2}}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)\|_2 &\leq C_{c_{\mathrm{RE}},1}\mathfrak{L}^3\sigma(R_{\mathrm{lasso}}+R_{\mathrm{outlier}}),\\ \|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\|_2 &\leq C_{c_{\mathrm{RE}},2}\mathfrak{L}^3\sigma\frac{1}{\kappa_{\mathrm{l}}}(R_{\mathrm{lasso}}+R_{\mathrm{outlier}}),\\ \|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\|_1 &\leq C_{c_{\mathrm{RE}},3}\mathfrak{L}^3\sigma\frac{1}{\mathfrak{r}}\sqrt{s}(R_{\mathrm{lasso}}+R_{\mathrm{outlier}}),\end{aligned} \tag{24}$$

and we see that (24) *recovers* (5).

4. The Second Result: Estimator Without the Covariance

In Theorem 13, we use the covariance of the covariates when we construct an estimator. On the other hand, especially in practical terms, the use of the covariance would be unfavorable. When we do not use the covariance in estimation, we need to modify algorithm WEIGHT as follows:

Algorithm 5 WEIGHT-WITHOUT-COVARIANCE

Input: data $\{\mathbf{X}_i\}_{i=1}^n$, tuning parameters $\tau_{\mathrm{cut}}, \varepsilon, r_1, r_2$
Output: weight estimate $\hat{\mathbf{w}} = \{\hat{w}_1, \cdots, \hat{w}_n\}$

Let $\hat{\mathbf{w}}$ be the solution to

$$\min_{\mathbf{w}\in\Delta^{n-1}(\varepsilon)} \max_{M\in\mathfrak{M}_{r_1,r_2,d}^{\ell_1,\mathrm{Tr}}} \sum_{i=1}^n w_i \langle \mathbf{X}_i\mathbf{X}_i^\top, M\rangle \tag{25}$$

if the optimal value of (13) $\leq \tau_{\mathrm{cut}}$
  return $\hat{\mathbf{w}}$
else
  return *fail*

In WEIGHT-WITHOUT-COVARIANCE, it is necessary to set the value of τ_{cut} larger than that in WEIGHT. For details, see Corollary 28. Then, Theorem 13 is changed as follows. Define

$$R'_{d,n,o} = \rho c_{r_1}\sqrt{s} r_{d,s} + r_\delta + c_{r_2}\sqrt{\frac{o}{n}}\sqrt{\kappa_{\mathrm{u}}^2\left(s r_{d,s} + r_\delta\right) + \Sigma_{\max}^2} r_2.$$

Theorem 15 *Suppose that (i) and (iii) of Assumption 2 and Assumption 4 hold. Suppose that the parameters $\lambda_o, \lambda_s, \varepsilon, \tau_{\mathrm{cut}}, r_1, r_2, r_\Sigma$ satisfy*

$$1 \geq 7c_o\left(4 + c_s\right) c_{\max}'^2 \sqrt{1+\log L} L^2 R'_{d,n,o}, \tag{26}$$
$$\lambda_s = c_s c_{\max}^2 L\lambda_o\sqrt{n}\frac{1}{c_{r_1}\sqrt{s}} R'_{d,n,o},\ \varepsilon = c_\varepsilon\frac{o}{n},\ \tau_{\mathrm{cut}} = c_{\mathrm{cut}}\left((L\kappa_{\mathrm{u}})^2\left(s r_{d,s} + r_\delta\right) + \Sigma_{\max}^2\right) r_2^2,$$
$$r_1 = c_{r_1}\sqrt{s} r_\Sigma, \quad r_2 = c_{r_2} r_\Sigma, \quad r_\Sigma = 7\left(4 + c_s\right) c_{\max}'^2 L\lambda_o\sqrt{n} R'_{d,n,o},$$

where $c_o, c_s, c_\varepsilon, c_{\mathrm{cut}}, c_{r_1}, c_{r_2}$, and $c'_{\max}$ are sufficiently large numerical constants such that $c_o \geq 4$, $c_s \geq 3(c_{\mathrm{RE}}+1)/(c_{\mathrm{RE}}-1)$, $2 > c_\varepsilon \geq 1$, $c_{\mathrm{cut}} \geq c_7$, $c_{r_1} = c_{r_1}^{\mathrm{num}}(1+c_{\mathrm{RE}})/\mathfrak{r}$, $c_{r_2} = c_{r_2}^{\mathrm{num}}(3+c_{\mathrm{RE}})/\kappa_{\mathrm{l}}$, $\min\{c_{r_1}^{\mathrm{num}}, c_{r_2}^{\mathrm{num}}\} \geq 2$ and $c_{r_1}^{\mathrm{num}}/c_{r_2}^{\mathrm{num}} \leq 1$. In Corollary 28 and Definition 31, c_7 and $c'_{\max}$ are defined, respectively. Suppose that $s r_{d,s} \leq 1$, $0 < o/n \leq 1/2$ hold. Then, the optimal solution $\hat{\boldsymbol{\beta}}$ satisfies the following:

$$\|\Sigma^{\frac{1}{2}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_2 \leq r_\Sigma,\ \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \leq r_2 \text{ and } \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \leq r_1, \tag{27}$$

with probability at least $1 - 3\delta$.

Remark 16 *We consider the results of (27) in detail. Assume that $\mathbb{E}\xi_i^2 \leq \sigma^2$ and that the equality in (15) holds. Define $C_{c_{\mathrm{RE}},4}$, $C_{c_{\mathrm{RE}},5}$ and $C_{c_{\mathrm{RE}},6}$ as constants depending on c_{RE}. Then, we have*

$$\|\Sigma^{\frac{1}{2}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_2 \leq C_{c_{\mathrm{RE}},4}\mathfrak{L}^3\sigma(R_{\mathrm{lasso}} + R'_{\mathrm{outlier}}), \tag{28}$$
$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \leq C_{c_{\mathrm{RE}},5}\mathfrak{L}^3\sigma\frac{1}{\kappa_{\mathrm{l}}}(R_{\mathrm{lasso}} + R'_{\mathrm{outlier}}),$$
$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \leq C_{c_{\mathrm{RE}},6}\mathfrak{L}^3\sigma\frac{1}{\mathfrak{r}}\sqrt{s}(R_{\mathrm{lasso}} + R'_{\mathrm{outlier}}),$$

and we see that (28) *recovers* (6). *Investigation of whether it is possible to achieve similar error bounds using an estimation method without covariance as in the case of with covariance, and the exploration of the trade-offs in such a scenario, are left as future research tasks.*

5. The Third and Fourth Results: $\mathfrak{L}$-subGaussian Case

In Section 5, we relax the assumption on the covariates and consider the case where the covariates are $\mathfrak{L}$-subGaussian.

5.1 The Third Result: Estimator Without Covariance

Even for the same situation as in Section 4 except that covariates are sampled from $\mathfrak{L}$-subGaussian random vector, Algorithm 1 substituting WEIGHT-WITHOUT-COVARIANCE (Algorithm 5) is valid with different parameter values. Define

$$r_d = \sqrt{\frac{\log d}{n}}, \quad R''_{d,n,o} = L\rho c_{r_1}\sqrt{s}r_{d,s} + Lr_\delta + c_{r_2}(L+\rho)\sqrt{\frac{o}{n}}\sqrt{s\left(r_d + r_\delta\right) + \Sigma^2_{\max}}r_2.$$

Then, we have the following result (Corollary 17). Corollary 17 is similar to Theorem 15, with the key differences being that the assumptions regarding the covariates have been weakened, and that the values of $R'_{d,n,o}$ and $c'_{\max}$ in Theorem 15 have been changed.

Corollary 17 *Suppose Assumption 3 and Assumption 4 hold. Suppose that the parameters* $\lambda_o, \lambda_s, \varepsilon, \tau_{\mathrm{cut}}, r_1, r_2, r_\Sigma$ *satisfy*

$$\begin{aligned} &1 \geq 7c_o\left(4+c_s\right)c''^2_{\max}\sqrt{1+\log L}LR''_{d,n,o}, \qquad (29)\\ &\lambda_s = c_s c''^2_{\max}\lambda_o\sqrt{n}\frac{1}{c_{r_1}\sqrt{s}}R''_{d,n,o},\ \varepsilon = c_\varepsilon\frac{o}{n},\ \tau_{\mathrm{cut}} = c_{\mathrm{cut}}\left(L^2 s\left(r_d+r_\delta\right)+\Sigma^2_{\max}\right)r_2^2,\\ &r_1 = c_{r_1}\sqrt{s}r_\Sigma, \quad r_2 = c_{r_2}r_\Sigma, \quad r_\Sigma = 7\left(4+c_s\right)c''^2_{\max}\lambda_o\sqrt{n}R''_{d,n,o}, \end{aligned}$$

where $c_o, c_s, c_\varepsilon, c_{\mathrm{cut}}, c_{r_1}, c_{r_2}$, *and* $c''_{\max}$ *are sufficiently large numerical constants such that* $c_o \geq 4$, $c_s \geq 3(c_{\mathrm{RE}}+1)/(c_{\mathrm{RE}}-1)$, $2 > c_\varepsilon \geq 1$, $c_{\mathrm{cut}} \geq c_{11}$, $c_{r_1} = c^{\mathrm{num}}_{r_1}(1+c_{\mathrm{RE}})/\mathfrak{r}$, $c_{r_2} = c^{\mathrm{num}}_{r_2}(3+c_{\mathrm{RE}})/\kappa_{\mathrm{l}}$, $\min\{c^{\mathrm{num}}_{r_1}, c^{\mathrm{num}}_{r_2}\} \geq 2$ *and* $c^{\mathrm{num}}_{r_1}/c^{\mathrm{num}}_{r_2} \leq 1$. *In Corollary 47 and Definition 50 in the appendix,* c_{11} *and* $c''_{\max}$ *are defined, respectively. Suppose that* $sr_{d,s} \leq 1$, $0 < o/n \leq 1/2$ *hold. Then, the output of Algorithm 1,* $\hat{\boldsymbol{\beta}}$, *substituting WEIGHT-WITHOUT-COVARIANCE (Algorithm 5) for WEIGHT, satisfies the following:*

$$\|\Sigma^{\frac{1}{2}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_2 \leq r_\Sigma,\ \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \leq r_2 \text{ and } \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \leq r_1,$$

with probability at least $1 - 3\delta$.

We note that the proof of Corollary 17 is in Section H.

5.2 The Fourth Result: Case Where Only the Outputs are Contaminated

In this section, we consider the case where only the output is contaminated by outliers and the covariates are sampled from an $\mathfrak{L}$-subGaussian random vector. Namely, we estimate

$\boldsymbol{\beta}^*$ from $\{y_i, \mathbf{x}_i\}_{i=1}^n$, where $\{\mathbf{x}_i\}_{i=1}^n$ are sampled from an $\mathfrak{L}$-subGaussian random vector and $\{y_i\}_{i=1}^n$ is generated by

$$y_i = \mathbf{x}_i^\top \boldsymbol{\beta}^* + \xi_i + \sqrt{n}\theta_i, \quad i = 1, \cdots, n,$$

where $\{\theta_i\}_{i=1}^n$ indicates outliers. In this situation, we use the following algorithm, which has no pre-processing step to reduce the effect of outlier for the covariates:

Algorithm 6 OUTLIER-ROBUST-AND-SPARSE-ESTIMATION-WHEN-ONLY-THE-OUTPUT-IS-CONTAMINATED-BY-OUTLIER

Input: data $\{y_i, \mathbf{x}_i\}_{i=1}^n$, and tuning parameters λ_o, λ_s
Output: estimator $\hat{\boldsymbol{\beta}}$

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^d}{\operatorname{argmin}} \sum_{i=1}^n \lambda_o^2 H\left(\frac{y_i - \mathbf{x}_i^\top \boldsymbol{\beta}}{\lambda_o \sqrt{n}}\right) + \lambda_s \|\boldsymbol{\beta}\|_1$$

return $\hat{\boldsymbol{\beta}}$

Define

$$R'''_{d,n,o} = \rho c_{r_1} \sqrt{s} r_{d,s} + r_\delta + r_o.$$

For the output of Algorithm 6, we have the following result. The assumptions of the next result is much simpler than those in Theorems 13 and 15 and Corollary 17, since no pre-processing is required to reduce outlier contamination for covariates, leading to a reduced sample complexity.

Corollary 18 *Suppose that (iii) of Assumption 2 and Assumption 4 hold. Suppose that the parameters $\lambda_o, \lambda_s, r_1, r_2, r_\Sigma$ satisfy*

$$\begin{aligned} 1 &\geq 7c_o\,(4 + c_s)\, c'''^2_{\max} \sqrt{1 + \log L} L^2 R'''_{d,n,o}, \\ \lambda_s &= c_s c'''^2_{\max} L \lambda_o \sqrt{n} \frac{1}{c_{r_1}\sqrt{s}} R'''_{d,n,o}, \\ r_1 &= c_{r_1}\sqrt{s} r_\Sigma, \quad r_2 = c_{r_2} r_\Sigma, \quad r_\Sigma = 15\,(2 + c_s)\, c'''^2_{\max} L \lambda_o \sqrt{n} R'''_{d,n,o}, \end{aligned} \tag{30}$$

where $c_o, c_s, c_{r_1}, c_{r_2}$, and $c'''_{\max}$ are sufficiently large numerical constants such that $c_o \geq 4$, $c_s \geq 2(c_{\mathrm{RE}}+1)/(c_{\mathrm{RE}}-1)$, $c_{r_1} = c_{r_1}^{\mathrm{num}}(1+c_{\mathrm{RE}})/\mathfrak{r}$, $c_{r_2} = c_{r_2}^{\mathrm{num}}(3+c_{\mathrm{RE}})/\kappa_{\mathrm{l}}$, $\min\{c_{r_1}^{\mathrm{num}}, c_{r_2}^{\mathrm{num}}\} \geq 2$ and $c_{r_1}^{\mathrm{num}}/c_{r_2}^{\mathrm{num}} \leq 1$. In Definition 52 in the appendix, $c'''_{\max}$ is defined. Suppose that $0 < o/n \leq 1/e$ holds. Then, the optimal solution $\hat{\boldsymbol{\beta}}$ satisfies the following:

$$\|\Sigma^{\frac{1}{2}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_2 \leq r_\Sigma, \quad \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \leq r_2 \text{ and } \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \leq r_1,$$

with probability at least $1 - 3\delta$.

The proof of Corollary 18 is in Section I.

Remark 19 *Our results involve a considerable number of tuning parameters, and these tuning parameters require knowledge of the upper bounds of certain unknown quantities. We explain in Section J.2 that adaptive tuning of these multiple parameters is generally difficult.*

6. Key Techniques

In Section 6, we introduce the key propositions and a lemma for Theorem 13, as well as key propositions and a corollary for Theorem 15.

6.1 Key Propositions and Lemma for Theorem 13

First, we introduce Proposition 20, which gives the condition on τ_{cut} when the covariance matrix Σ is known. The proof is given in Section 7.

Proposition 20 *Suppose (i) of Assumption 2 holds, and $r_1/r_2 \leq \sqrt{s}, sr_{d,s}$ and $r_\delta \leq 1$ hold. Define c_1 and c_2 as numerical constants. Then, with probability at least $1-\delta$, we have*

$$\max_{M \in \mathfrak{M}^{\ell_1,\mathrm{Tr}}_{r_1,r_2,d}} \sum_{i=1}^{n} \frac{\left\langle \mathbf{x}_i \mathbf{x}_i^\top - \Sigma, M \right\rangle}{n} \leq c_1 (L\kappa_{\mathrm{u}})^2 \left(sr_{d,s} + r_\delta\right) r_2^2. \tag{31}$$

Additionally, assume that $\varepsilon = c_\varepsilon \times (o/n)$ and $o/n \leq 1/(5e)$ with $1 \leq c_\varepsilon < 2$ hold. Then, we have with probability at least $1-2\delta$, we have

$$\max_{M \in \mathfrak{M}^{\ell_1,\mathrm{Tr}}_{r_1,r_2,d}} \sum_{i=1}^{n} \hat{w}_i \left\langle \mathbf{X}_i \mathbf{X}_i^\top - \Sigma, M \right\rangle \leq c_2 (L\kappa_{\mathrm{u}})^2 \left(sr_{d,s} + r_\delta + r'_o\right) r_2^2, \tag{32}$$

where $\{\hat{w}_i\}_{i=1}^n$ is a solution of (13).

Therefore, we see that, when $c_2(L\kappa_{\mathrm{u}})^2 \left(sr_{d,s} r_\delta + r'_o\right) r_2^2 \leq \tau_{\mathrm{cut}}$, Algorithm 2 succeeds in returning $\hat{\mathbf{w}}$ under (32). The key techniques of the proof of the proposition above are Theorem 1.1 of Rudelson and Vershynin (2013), which is one of the variants of the Hanson–Wright inequality (Hanson and Wright, 1971; Wright, 1973; Adamczak, 2015; Hsu et al., 2012), and generic chaining for a subexponential random variable (Corollary 5.2 of Dirksen (2015)). We note that a subexponential random variable is a random variable with a finite ψ_1 norm.

Next, we introduce a deterministic proposition related to Theorem 13. Let

$$r_{\mathbf{v},i} = \hat{w}'_i n \frac{y_i - \mathbf{X}_i^\top \mathbf{v}}{\lambda_o \sqrt{n}}, \quad X_{\mathbf{v},i} = \frac{\mathbf{X}_i^\top \mathbf{v}}{\lambda_o \sqrt{n}}, \quad x_{\mathbf{v},i} = \frac{\mathbf{x}_i^\top \mathbf{v}}{\lambda_o \sqrt{n}}, \quad \xi_{\lambda_o,i} = \frac{\xi_i}{\lambda_o \sqrt{n}},$$

and for $\eta \in (0,1)$,

$$\boldsymbol{\theta} = \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*, \quad \boldsymbol{\theta}_\eta = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\eta.$$

The following proposition is proved in a manner similar to the proof of Proposition 9.1 of Alquier et al. (2019), and the proof is given in Appendices C and D.

Proposition 21 *Suppose that, for any $\boldsymbol{\theta}_\eta \in r_1 \mathbb{B}_1^d \cap r_2 \mathbb{B}_2^d \cap r_\Sigma \mathbb{B}_\Sigma^d$,*

$$\left| \lambda_o \sqrt{n} \sum_{i=1}^{n} \hat{w}'_i h(r_{\boldsymbol{\beta}^*,i}) \mathbf{X}_i^\top \boldsymbol{\theta}_\eta \right| \leq r_{a,1} r_1 + r_{a,2} r_2 + r_{a,\Sigma} r_\Sigma, \tag{33}$$

$$b \|\Sigma^{\frac{1}{2}} \boldsymbol{\theta}_\eta\|_2^2 - r_{b,2} r_2 - r_{b,\Sigma} r_\Sigma - r_{b,1} r_1 \leq \sum_{i=1}^{n} \lambda_o \sqrt{n} \hat{w}'_i \left(-h(r_{\boldsymbol{\beta}^* + \boldsymbol{\theta}_\eta, i}) + h(r_{\boldsymbol{\beta}^*,i}) \right) \mathbf{X}_i^\top \boldsymbol{\theta}_\eta, \tag{34}$$

where $r_{a,1}, r_{a,2}, r_{a,\Sigma}, r_{b,1}, r_{b,2}, r_{b,\Sigma} \geq 0$, $b > 0$ *are some numbers. Suppose that* $\mathbb{E}\mathbf{x}_i\mathbf{x}_i^\top = \Sigma$ *satisfies* $\mathrm{RE}(s, c_{\mathrm{RE}}, \mathfrak{r})$, $\kappa_{\mathrm{l}} > 0$, *and*

$$\lambda_s - \left(r_{a,1} + \frac{c_{r_2} r_{a,2} + r_{a,\Sigma}}{c_{r_1}\sqrt{s}}\right) > 0, \quad \frac{\lambda_s + \left(r_{a,1} + \frac{c_{r_2} r_{a,2} + r_{a,\Sigma}}{c_{r_1}\sqrt{s}}\right)}{\lambda_s - \left(r_{a,1} + \frac{c_{r_2} r_{a,2} + r_{a,\Sigma}}{c_{r_1}\sqrt{s}}\right)} \leq c_{\mathrm{RE}}, \tag{35}$$

$$r_\Sigma \geq \frac{2}{b}\left(c_{r_1}\sqrt{s}(r_{a,1} + r_{b,1}) + c_{r_2}(r_{a,2} + r_{b,2}) + r_{a,\Sigma} + r_{b,\Sigma} + c_{r_1}\sqrt{s}\lambda_s\right), \quad r_1 = c_{r_1}\sqrt{s} r_\Sigma, \quad r_2 = c_{r_2} r_\Sigma \tag{36}$$

hold, where $c_{r_1} = c_{r_1}^{\mathrm{num}}(1 + c_{\mathrm{RE}})/\mathfrak{r}$, $c_{r_2} = c_{r_2}^{\mathrm{num}}(3 + c_{\mathrm{RE}})/\kappa_{\mathrm{l}}$, $\min\{c_{r_1}^{\mathrm{num}}, c_{r_2}^{\mathrm{num}}\} \geq 2$ *and* $c_{r_1}/c_{r_2} \leq 1$. *Then, we have the following:*

$$\|\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}\|_1 \leq r_1, \quad \|\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}\|_2 \leq r_2, \quad \|\Sigma^{\frac{1}{2}}(\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}})\|_2 \leq r_\Sigma. \tag{37}$$

In the remainder of Section 6, we introduce Propositions 22–25 and one lemma. In Section A, we prove Theorem 13 using the propositions. In the proof of Theorem 13, we prove that (33) – (36) are satisfied with high probability for appropriate values of $r_{a,1}, r_{a,2}, r_{a,\Sigma}, r_{b,1}, r_{b,2}, r_{b,\Sigma}$ and b under the assumptions in Theorem 13, and we see that (37) is also satisfied. Then, we have the result (23) in Theorem 13. We note that, for Proposition 25, similar statements are found, for example, in Sun et al. (2020); Chen and Zhou (2020), and the proof of Proposition 25 basically follows the same line as those in Sun et al. (2020); Chen and Zhou (2020). Propositions 23 and 24 are proved by relatively simple calculations based on the result of Proposition 20. Therefore, the proofs of Propositions 23–25 are given in Appendix F.

Note that we present the statements of Propositions 22 and 25 in a flexible manner, which can be used not only for the proof of Theorem 13, but also for the proofs of the other three main theorems.

Proposition 22 *Suppose that (i) of Assumption 2 or Assumption 3, and (ii) of Assumption 4 hold. Then, for any* $\mathbf{v} \in r_1\mathbb{B}_1^d \cap r_\Sigma\mathbb{B}_\Sigma^d$, *where* r_1 *and* r_Σ *are positive constants such that* $r_\Sigma/r_1 \leq 1/\sqrt{s}$, *with probability at least* $1 - \delta$, *we have*

$$\left|\sum_{i=1}^n \frac{1}{n} h(\xi_{\lambda_o,i})\mathbf{x}_i^\top \mathbf{v}\right| \leq c_3 L\left(\rho r_{d,s} r_1 + r_\delta r_\Sigma\right),$$

where c_3 *denotes a numerical constant.*

Proposition 23 *Suppose that the assumptions of Proposition 20 hold. Furthermore, suppose that* (31) *and* (32) *hold and that Algorithm 2 returns* $\hat{\mathbf{w}}$ *with* $\tau_{\mathrm{cut}} = c_{\mathrm{cut}}(L\kappa_{\mathrm{u}})^2\left(s r_{d,s} + r_\delta + r_o'\right) r_2^2$, *where* c_{cut} *is a constant such that* $c_{\mathrm{cut}} \geq c_2$. *For any* $\mathbf{u} \in \mathbb{R}^n$ *such that* $\|\mathbf{u}\|_\infty \leq 2$ *and for any* $\mathbf{v} \in r_1\mathbb{B}_1^d \cap r_2\mathbb{B}_2^d$, *we have*

$$\left|\sum_{i\in\mathcal{O}} \hat{w}_i' u_i \mathbf{X}_i^\top \mathbf{v}\right| \leq c_4 L\sqrt{1 + c_{\mathrm{cut}}}\left(\kappa_{\mathrm{u}}\sqrt{\frac{o}{n}}\left(\sqrt{s r_{d,s}} + \sqrt{r_\delta}\right) + \kappa_{\mathrm{u}} r_o\right) r_2,$$

where c_4 *is a numerical constant that depends on* c_1 *and* c_2.

Let I_m be an index set such that $I_m \subset \{1, \cdots, n\}$ and $|I_m| = m$.

Proposition 24 *Suppose that the assumptions of Proposition 20 hold, and suppose that* (31) *holds. Then, for any $m \in \mathbb{N}$ such that $m \leq (2c_\varepsilon + 1)o$, for any $\mathbf{u} \in \mathbb{R}^n$ such that $\|\mathbf{u}\|_\infty \leq 2$ and for any $\mathbf{v} \in r_1 \mathbb{B}_1^d \cap r_2 \mathbb{B}_2^d$, we have the following:*

$$\left| \sum_{i \in I_m} \frac{1}{n} u_i \mathbf{x}_i^\top \mathbf{v} \right| \leq c_5 L \left(\kappa_{\mathrm{u}} \sqrt{\frac{o}{n}} \left(\sqrt{s r_{d,s}} + \sqrt{r_\delta} \right) + \kappa_{\mathrm{u}} r_o \right) r_2,$$

where c_5 is a numerical constant that depends on c_1 and c_ε.

Proposition 25 *Suppose that (i) of Assumption 2 or Assumption 3, and Assumption 4 hold. Assume that n is taken to be large so that* (21) *holds when this proposition is applied to Theorem 13,* (26) *holds when applied to Theorem 15, and* (29) *holds when applied to Corollary 17,* (30) *holds when applied to Corollary 18. Then, for any $\mathbf{v} \in r_1 \mathbb{B}_1^d \cap r_\Sigma \mathbb{B}_\Sigma^d$, where r_1 and r_Σ are positive constants, with probability at least $1 - \delta$, we have*

$$\sum_{i=1}^n \frac{\lambda_o}{\sqrt{n}} \left(-h(\xi_{\lambda_o, i} - x_{\mathbf{v},i}) + h(\xi_{\lambda_o,i}) \right) \mathbf{x}_i^\top \mathbf{v} \geq \frac{\|\Sigma^{\frac{1}{2}} \mathbf{v}\|_2^2}{3} - c_6 L \lambda_o \sqrt{n} \left(\rho r_{d,s} r_1 + \sqrt{s} r_{d,s} r_\Sigma + r_\delta r_\Sigma \right), \tag{38}$$

where c_6 is a numerical constant.

We define $c_{\max}$, which is a numerical constant used in Theorem 13.

Definition 26 *Define*

$$c_{\max} = \max \left(1, c_3, c_4 \sqrt{1 + c_{\mathrm{cut}}}, c_5, c_6 \right).$$

Let $I_<$ and $I_\geq$ be the sets of indices such that $w_i < 1/(2n)$ and $w_i \geq 1/(2n)$, respectively. The following lemma makes the analysis of the ℓ_1-penalized Huber loss easy as we noted at the bottom of Section 3.2.4.

Lemma 27 *Suppose that $0 < \varepsilon < 1$. Then, for any $\mathbf{w} \in \Delta^{n-1}(\varepsilon)$, we have $|I_<| \leq 2n\varepsilon$.*

6.2 Key Propositions and Corollary for Theorem 15

First, we introduce Corollary 28, that gives the condition on τ_{cut} when we do not use Σ in the estimator. The proof is given in Appendix F.

Corollary 28 *Suppose that (i) of Assumption 2, $r_1/r_2 \leq \sqrt{s}$, where $r_1, r_2 > 0$, and $s r_{d,s}, r_\delta \leq 1$ hold. Define c_1' and c_7 as numerical constants. Then, with probability at least $1 - \delta$, we have*

$$\max_{M \in \mathfrak{M}_{r_1, r_2, d}^{\ell_1, \mathrm{Tr}}} \sum_{i=1}^n \frac{\left\langle \mathbf{x}_i \mathbf{x}_i^\top, M \right\rangle}{n} \leq c_1' L^2 \left(\kappa_{\mathrm{u}}^2 \left(s r_{d,s} + r_\delta \right) + \Sigma_{\max}^2 \right) r_2^2. \tag{39}$$

Additionally, assume that $\varepsilon = c_\varepsilon \times (o/n)$ *and* $o/n \leq 1/2$ *with* $1 \leq c_\varepsilon < 2$ *hold. Then, with probability at least* $1-\delta$*, we have*

$$\max_{M \in \mathfrak{M}^{\ell_1,\mathrm{Tr}}_{r_1,r_2,d}} \sum_{i=1}^{n} \hat{w}_i \left\langle \mathbf{X}_i \mathbf{X}_i^\top, M \right\rangle \leq c_7 L^2 \left(\kappa_{\mathrm{u}}^2 \left(s r_{d,s} + r_\delta \right) + \Sigma_{\max}^2 \right) r_2^2, \tag{40}$$

where $\{\hat{w}_i\}_{i=1}^n$ *is a solution of* (25)*.*

Therefore, under (40), when $c_7 \left((L\kappa_{\mathrm{u}})^2 \left(s r_{d,s} + r_\delta \right) + \Sigma_{\max}^2 \right) r_2^2 \leq \tau_{\mathrm{cut}}$, Algorithm 5 succeeds in returning $\hat{\mathbf{w}}$. We see that when the covariance is not used in the estimator, it is necessary to set the value of τ_{cut} to a higher magnitude compared to the case when covariance is used. Using the propositions, in Appendix B, we can prove that (33) – (36) are satisfied with a high probability for appropriate values of $r_{a,1}, r_{a,2}, r_{a,\Sigma}, r_{b,1}, r_{b,2}, r_{b,\Sigma}$ and b under the assumptions in Theorem 15, and we see that (37) is also satisfied. Then, we can have the result (27) in Theorem 15. When Σ is not used in the estimator, Propositions 21 and 25 and Lemma 27 are commonly used, and instead of Propositions 23 and 24, Propositions 29 and 30 are used, respectively. In Definition 31, $c'_{\max}$ is defined. The proofs of Propositions 29 and 30 are given in Appendix F because Propositions 29 and 30 can be proved by simple calculations based on the result of Corollary 28.

Proposition 29 *Suppose that assumptions of Corollary 28 hold. Furthermore, suppose that* (39) *and* (40) *hold and that Algorithm 25 returns* $\hat{\mathbf{w}}$ *with* $\tau_{\mathrm{cut}} = c_{\mathrm{cut}} \left((L\kappa_{\mathrm{u}})^2 \left(s r_{d,s} + r_\delta \right) + \Sigma_{\max}^2 \right) r_2^2$*, where* c_{cut} *is a constant such that* $c_{\mathrm{cut}} \geq c_7$*. Furthermore, suppose that* (39) *holds and that Algorithm 5 returns* $\hat{\mathbf{w}}$*. For any* $\mathbf{u} \in \mathbb{R}^n$ *such that* $\|\mathbf{u}\|_\infty \leq 2$ *and for any* $\mathbf{v} \in r_1 \mathbb{B}_1^d \cap r_2 \mathbb{B}_2^d$*, we have the following:*

$$\left| \sum_{i \in \mathcal{O}} \hat{w}'_i u_i \mathbf{X}_i^\top \mathbf{v} \right| \leq c_8 \sqrt{c_{\mathrm{cut}}} L \sqrt{\frac{o}{n}} \sqrt{\kappa_{\mathrm{u}}^2 \left(s r_{d,s} + r_\delta \right) + \Sigma_{\max}^2} r_2,$$

where c_8 *is a numerical constant.*

Proposition 30 *Suppose that the assumptions of Corollary 28 hold, and suppose that* (40) *holds. Then, for any* $m \in \mathbb{N}$ *such that* $m \leq (2c_\varepsilon + 1)o$*, for any* $\mathbf{u} \in \mathbb{R}^n$ *such that* $\|\mathbf{u}\|_\infty \leq 2$ *and for any* $\mathbf{v} \in r_1 \mathbb{B}_1^d \cap r_2 \mathbb{B}_2^d$*, we have the following :*

$$\left| \sum_{i \in I_m} \frac{1}{n} u_i \mathbf{x}_i^\top \mathbf{v} \right| \leq c_9 L \sqrt{\frac{o}{n}} \sqrt{\kappa_{\mathrm{u}}^2 \left(s r_{d,s} + r_\delta \right) + \Sigma_{\max}^2} r_2,$$

where c_9 *is a numerical constant that depends on* c'_1 *and* c_ε*.*

For Theorem 15, the following definition of $c'_{\max}$ is used.

Definition 31 *Define*

$$c'_{\max} = \max \left(1, c_3, c_6, c_8 \sqrt{c_{\mathrm{cut}}}, c_9 \right).$$

7. Proofs of Propositions 20 and 22

The value of the numerical constant C be allowed to change from line to line. Before the proofs of Propositions 20 and 22, we introduce Definition 32, Theorem 33, and Lemma 34. They play important roles in the proofs of Propositions 20 and 22.

First, we define $\gamma_\alpha(T,d)$-functional, which is used to state Theorem 33:

Definition 32 ($\gamma_\alpha(T,d)$-functional, Section 2 of Dirksen (2015)) *Let (T,d) be a semi-metric space with $d(x,z) \le d(x,y)+d(y,z)$ and $d(x,y)=d(y,x)$ for $x,y,z \in T$. A sequence $\mathcal{T}=\{T_m\}_{m\ge 0}$ with subsets of T is said to be admissible if $|T_0|=1$ and $|T_m| \le 2^{2^m}$ for all $m \ge 1$. For any $\alpha \in (0,\infty)$, the γ_α-functional of (T,d) is defined by*

$$\gamma_\alpha(T,d) = \inf_{\mathcal{T}} \sup_{t\in T} \sum_{m=0}^{\infty} 2^{\frac{m}{\alpha}} \inf_{s\in T_m} d(t,s),$$

where infimum is taken over all admissible sequences $\mathcal{T}=\{T_m\}_{m\ge 0}$.

Second, we introduce Theorem 33, which is used in the proof of Proposition 20.

Theorem 33 (Corollary 5.2 of Dirksen (2015)) *Assume that T consists of n-tuples $t=(t_1,\cdots,t_n)$. Let $\{Z_{t_i} \in \mathbb{R}\}_{i=1}^n$ as a sequence of independent subexponential random variable indexed by $t_i \in T$, and*

$$E_t = \frac{1}{n}\sum_{i=1}^n Z_{t_i} - \mathbb{E}Z_{t_i}.$$

Define

$$d_1(s,t) = \max_{1\le i\le m} \|Z_{t_i} - Z_{s_i}\|_{\psi_1}, \quad d_2(s,t) = \left(\frac{1}{m}\sum_{i=1}^n \|Z_{t_i} - Z_{s_i}\|_{\psi_1}^2\right)^{\frac{1}{2}}.$$

Assume that, for any $q=2,3,\cdots$,

$$\sup_{t\in T} \frac{1}{n}\sum_{i=1}^n \mathbb{E}|Z_{t_1} - \mathbb{E}Z_{t_i}|^q \le \frac{q!}{2}a^2 b^{q-2},$$

where a and b are positive numbers. Then, for any $u \ge 1$, we have

$$\mathbb{P}\left(\sup_{t\in T}|E_t| \gtrsim \frac{\gamma_1(T,d_1)}{n} + \frac{\gamma_2(T,d_1)}{\sqrt{n}} + a\frac{\sqrt{u}}{\sqrt{n}} + b\frac{u}{n}\right) \le e^{-u}.$$

Lastly, we introduce Lemma 34, which is used in the proof of Proposition 22. This lemma is used to evaluate $\gamma_\alpha(T,d)$-functional sharply especially when d is $\|\Sigma^{\frac{1}{2}}(\cdot)\|$ via the majorizing measure theorem (Theorem 2.4.1 of Talagrand (2014)). The proof of this lemma is given in Appendix E.2. Let $\mathbf{g}=(g_1,\cdots,g_d)^\top$ be the d-dimensional standard normal Gaussian random vector.

Lemma 34 *Suppose that the (i) of Assumption 2 or Assumption 3 holds. Then,for positive constants r_1 and r_Σ such that $r_\Sigma/r_1 \le 1/\sqrt{s}$, we have*

$$\mathbb{E}\sup_{\mathbf{v}\in r_1\mathbb{B}_1^d \cap r_\Sigma \mathbb{B}_\Sigma^d} \langle \Sigma^{\frac{1}{2}}\mathbf{g}, \mathbf{v}\rangle \le C\rho r_1\sqrt{\log(d/s)}r_\Sigma.$$

7.1 Preparation for the Proof of Proposition 20

For a matrix M, we define the operator norm and Frobenius norm of M as $\|M\|_{\mathrm{op}}$ and $\|M\|_{\mathrm{F}}$, respectively, and we define the number of nonzero elements of M as $\|M\|_0$. Additionally, we define $a\mathbb{B}_1^{d\times d} = \{M \in \mathbb{R}^{d\times d} \mid \|M\|_1 \leq a\}$, $a\mathbb{B}_{\mathrm{F}}^{d\times d} = \{M \in \mathbb{R}^{d\times d} \mid \|M\|_{\mathrm{F}} \leq a\}$, and $a\mathbb{B}_0^{d\times d} = \{M \in \mathbb{R}^{d\times d} \mid \|M\|_0 \leq a\}$.

Lemma 35 *Suppose that (i) of Assumptions 2 holds and r_2 is a positive constant. For any fixed $M, M' \in s^2\mathbb{B}_0^{d\times d} \cap r_2^2\mathbb{B}_{\mathrm{F}}^{d\times d}$, we have*

$$\mathbb{E}\left|\langle \mathbf{x}\mathbf{x}^\top - \Sigma, M\rangle\right|^p \leq p!(C(L\kappa_{\mathrm{u}})^2 r_2^2)^p, \tag{41}$$

$$\|\langle \mathbf{x}\mathbf{x}^\top, M - M'\rangle\|_{\psi_1} \leq C(L\kappa_{\mathrm{u}})^2\|M - M'\|_{\mathrm{F}}. \tag{42}$$

Proof In this proof, we define c as some positive numerical constant. From (11), for any index set J such that $|J| \leq 2s^2$ and $\mathbf{v} \in \mathbb{S}^{2s^2-1}$ we have

$$\mathbb{E}\exp\left(\frac{\langle \mathbf{v}, \mathbf{x}|_J\rangle^2}{c_{\mathfrak{L}}^2\mathfrak{L}^2\kappa_{\mathrm{u}}^2}\right) \leq 2,$$

and we have

$$\|\mathbf{x}|_J\|_{\psi_2} \leq c_{\mathfrak{L}}\mathfrak{L}\kappa_{\mathrm{u}} \leq L\kappa_{\mathrm{u}}.$$

For any matrix A and any index set $J \subset \{1, \cdots, d\}$, define $A|_{J,J}$ as the matrix such that all the elements of the i-th rows and columns are zero for $i \in J^c$. Fix $M \in s^2\mathbb{B}_0^{d\times d} \cap r_2^2\mathbb{B}_{\mathrm{F}}^{d\times d}$. For M, let $K \subset \{1, \cdots, d\}$ be an index set such that $M_{ij} = 0$ for $i \in K^c$ or $j \in K^c$. We note that $|K| \leq s^2$. From Theorem 1.1 of Rudelson and Vershynin (2013) and Exercise 6.3.3 of Vershynin (2018), for any $t > 0$, we have

$$\begin{aligned}\mathbb{P}\left(\left|\langle \mathbf{x}\mathbf{x}^\top - \Sigma, M\rangle\right| > t\right) &= \mathbb{P}\left(\left|\langle \mathbf{x}|_K\mathbf{x}|_K^\top - \Sigma|_{K,K}, M\rangle\right| > t\right) \\ &\leq 2\exp\left\{-c\min\left(\frac{t^2}{(L\kappa_{\mathrm{u}})^4\|M\|_{\mathrm{F}}^2}, \frac{t}{(L\kappa_{\mathrm{u}})^2\|M\|_{\mathrm{F}}}\right)\right\}. \end{aligned}\tag{43}$$

From (43), for $t \leq (L\kappa_{\mathrm{u}})^2\|M\|_{\mathrm{F}}$, we have

$$\mathbb{P}\left(\left|\langle \mathbf{x}\mathbf{x}^\top - \Sigma, M\rangle\right| > t\right) \leq 2\exp\left(-c\frac{t^2}{(L\kappa_{\mathrm{u}})^4\|M\|_{\mathrm{F}}^2}\right),$$

and for $t \geq (L\kappa_{\mathrm{u}})^2\|M\|_{\mathrm{F}}$, we have

$$\mathbb{P}\left(\left|\langle \mathbf{x}\mathbf{x}^\top - \Sigma, M\rangle\right| > t\right) \leq 2\exp\left(-c\frac{t}{(L\kappa_{\mathrm{u}})^2\|M\|_{\mathrm{F}}}\right).$$

We follow almost the same argument of the proof of Proposition 2.5.2 of Vershynin (2018). For any $1 \leq p < \infty$, we have

$$\begin{aligned}
\mathbb{E}\left|\langle \mathbf{x}\mathbf{x}^\top - \Sigma, M\rangle\right|^p &= \int_0^\infty \mathbb{P}\left(\left|\langle \mathbf{x}\mathbf{x}^\top - \Sigma, M\rangle\right|^p \geq u\right) du \\
&= \int_0^\infty \mathbb{P}\left(\left|\langle \mathbf{x}\mathbf{x}^\top - \Sigma, M\rangle\right| \geq t\right) pt^{p-1} dt \\
&\leq \int_0^\infty 2\exp\left(-c\frac{t^2}{(L\kappa_{\mathrm{u}})^4\|M\|_{\mathrm{F}}^2}\right) pt^{p-1}dt + \int_0^\infty 2\exp\left(-c\frac{t}{(L\kappa_{\mathrm{u}})^2\|M\|_{\mathrm{F}}}\right) pt^{p-1}dt \\
&\leq p\frac{(L\kappa_{\mathrm{u}})^{2p}\|M\|_{\mathrm{F}}^p}{\sqrt{c^p}}\Gamma(p/2) + 2p\frac{(L\kappa_{\mathrm{u}})^{2p}\|M\|_{\mathrm{F}}^p}{c^p}\Gamma(p) \\
&\leq C\left(L\kappa_{\mathrm{u}}\left(\frac{1}{c^{\frac{1}{4}}} + \frac{1}{\sqrt{c}}\right)\right)^{2p}\|M\|_{\mathrm{F}}^p p\left((p/2)^{p/2} + p^p\right),
\end{aligned}$$

and we have

$$\begin{aligned}
\left(\mathbb{E}\langle \mathbf{x}_i\mathbf{x}_i^\top - \Sigma, M\rangle^p\right)^{\frac{1}{p}} &\leq C\left(L\kappa_{\mathrm{u}}\left(\frac{1}{c^{\frac{1}{4}}} + \frac{1}{\sqrt{c}}\right)\right)^2\|M\|_{\mathrm{F}}\left(p(p/2)^{p/2} + p^{p+1}\right)^{\frac{1}{p}} \\
&\leq C\left(L\kappa_{\mathrm{u}}\left(\frac{1}{c^{\frac{1}{4}}} + \frac{1}{\sqrt{c}}\right)\right)^2\|M\|_{\mathrm{F}}\left(2p^{p+1}\right)^{\frac{1}{p}} \\
&\leq C\left(L\kappa_{\mathrm{u}}\left(\frac{1}{c^{\frac{1}{4}}} + \frac{1}{\sqrt{c}}\right)\right)^2\|M\|_{\mathrm{F}}p \leq C\left(L\kappa_{\mathrm{u}}\right)^2\|M\|_{\mathrm{F}}p. \qquad (44)
\end{aligned}$$

From Stirling's formula $p^p \leq p!e^p$, we have

$$\max_{M\in\mathfrak{M}_{r_1,r_2,d}^{\ell_1,\mathrm{Tr}}}\frac{1}{n}\sum_{i=1}^n\mathbb{E}\left|\langle \mathbf{x}_i\mathbf{x}_i^\top - \Sigma, M\rangle\right|^p \leq Cp!e^p(L\kappa_{\mathrm{u}})^{2p}r_2^{2p},$$

and the proof of (41) is complete.

For any fixed $M, M' \in \mathfrak{M}_{r_1,r_2,d}^{\ell_1,\mathrm{Tr}}$, let $K' \subset \{1,\cdots,d\}$ is the index set such that $(M - M')_{ij} = 0$ for $i \in K'^c$ or $j \in K'^c$. We note that $|K'| \leq 2 \times s^2$. From (44) exchanging M to $M - M'$, and the equivalence between (b) and (c) in Proposition 2.7.1 of Vershynin (2018) and Definition 2.7.5, we have

$$\begin{aligned}
\left\|\langle \mathbf{x}\mathbf{x}^\top, M - M'\rangle\right\|_{\psi_1} = \left\|\langle \mathbf{x}|_{K'}\mathbf{x}|_{K'}^\top, (M - M')|_{K',K'}\rangle\right\|_{\psi_1} &\leq C(L\kappa_{\mathrm{u}})^2\|(M - M')|_{K',K'}\|_{\mathrm{F}} \\
&\leq C(L\kappa_{\mathrm{u}})^2\|M - M'\|_{\mathrm{F}}, \qquad (45)
\end{aligned}$$

and the proof of (42) is complete. ■

Using the lemma above, we have the key lemma (Lemma 37) to prove Proposition 20. For a set K, we define $\mathrm{conv}(K)$ as its convex hull. Before the statement and the proof of Lemma 37, we introduce the following lemma, which slightly generalizes Lemma 3.1 of Plan and Vershynin (2013) and the proof is in Appendix E.

Lemma 36 *Suppose that r_1 and r_2 are positive constants. We have*

$$r_1^2\mathbb{B}_1^{d\times d}\cap r_2^2\mathbb{B}_{\mathrm{F}}^{d\times d}\subset 2\mathrm{conv}\left(\left(\frac{r_1}{r_2}\right)^4\mathbb{B}_0^{d\times d}\cap r_2^2\mathbb{B}_{\mathrm{F}}^{d\times d}\right), \tag{46}$$

$$r_1\mathbb{B}_1^{d}\cap r_2\mathbb{B}_2^{d}\subset 2\mathrm{conv}\left(\left(\frac{r_1}{r_2}\right)^2\mathbb{B}_0^{d}\cap r_2\mathbb{B}_2^{d}\right). \tag{47}$$

Then, we introduce Lemma 37.

Lemma 37 *Suppose that (i) of Assumptions 2 holds and r_1 and r_2 are positive constants such that $r_1/r_2\leq\sqrt{s}$. Then, with probability at least $1-\delta$, we have*

$$\max_{M\in\mathfrak{M}_{r_1,r_2,d}^{\ell_1,\mathrm{Tr}}}\left|\frac{1}{n}\sum_{i=1}^n\langle\mathbf{x}_i\mathbf{x}_i^\top-\Sigma,M\rangle\right|$$
$$\leq C(L\kappa_{\mathrm{u}})^2\left(\frac{\gamma_1(s^2\mathbb{B}_0^{d\times d}\cap r_2^2\mathbb{B}_{\mathrm{F}}^{d\times d},\|\cdot\|_{\mathrm{F}})}{n}+\frac{\gamma_2(s^2\mathbb{B}_0^{d\times d}\cap r_2^2\mathbb{B}_{\mathrm{F}}^{d\times d},\|\cdot\|_{\mathrm{F}})}{\sqrt{n}}+r_\delta^2r_2^2+r_\delta r_2^2\right).$$

Proof First, we have

$$\mathfrak{M}_{r_1,r_2,d}^{\ell_1,\mathrm{Tr}}\subset r_1^2\mathbb{B}_1^{d\times d}\cap r_2^2\mathbb{B}_{\mathrm{F}}^{d\times d}\overset{(a)}{\subset}2\mathrm{conv}(s^2\mathbb{B}_0^{d\times d}\cap r_2^2\mathbb{B}_{\mathrm{F}}^{d\times d}),$$

where (a) follows from $r_1/r_2\leq\sqrt{s}$ and Lemma 36, identifying vectors with matrices. Then, we have

$$\max_{M\in\mathfrak{M}_{r_1,r_2,d}^{\ell_1,\mathrm{Tr}}}\left|\frac{1}{n}\sum_{i=1}^n\langle\mathbf{x}_i\mathbf{x}_i^\top-\Sigma,M\rangle\right|$$
$$\leq\max_{M\in r_1^2\mathbb{B}_1^{d\times d}\cap r_2^2\mathbb{B}_{\mathrm{F}}^{d\times d}}\left|\frac{1}{n}\sum_{i=1}^n\langle\mathbf{x}_i\mathbf{x}_i^\top-\Sigma,M\rangle\right|\leq\max_{M\in2\mathrm{conv}(s^2\mathbb{B}_0^{d\times d}\cap r_2^2\mathbb{B}_{\mathrm{F}}^{d\times d})}\left|\frac{1}{n}\sum_{i=1}^n\langle\mathbf{x}_i\mathbf{x}_i^\top-\Sigma,M\rangle\right|.$$

Identifying vectors with matrices, from the fact that the extreme points of a linear program are the vertices of the feasible region (or from Lemma D.8 of Oymak (2018)), we have

$$\max_{M\in2\mathrm{conv}(s^2\mathbb{B}_0^{d\times d}\cap r_2^2\mathbb{B}_{\mathrm{F}}^{d\times d})}\left|\frac{1}{n}\sum_{i=1}^n\langle\mathbf{x}_i\mathbf{x}_i^\top-\Sigma,M\rangle\right|\leq\max_{M\in s^2\mathbb{B}_0^{d\times d}\cap r_2^2\mathbb{B}_{\mathrm{F}}^{d\times d}}\left|\frac{2}{n}\sum_{i=1}^n\langle\mathbf{x}_i\mathbf{x}_i^\top-\Sigma,M\rangle\right|.$$

Next, note that the upper bound of the first inequality in Lemma 35 is independent of i and M. Hence, from Theorem 33, with probability at least $1-\delta$, from the same argument used to have (45), we have

$$\max_{M\in\mathfrak{M}_{r_1,r_2,d}^{\ell_1,\mathrm{Tr}}}\left|\frac{1}{n}\sum_{i=1}^n\langle\mathbf{x}_i\mathbf{x}_i^\top-\Sigma,M\rangle\right| \tag{48}$$
$$\leq C\left(\frac{\gamma_1(s^2\mathbb{B}_0^{d\times d}\cap r_2^2\mathbb{B}_{\mathrm{F}}^{d\times d},d_1)}{n}+\frac{\gamma_2(s^2\mathbb{B}_0^{d\times d}\cap r_2^2\mathbb{B}_{\mathrm{F}}^{d\times d},d_2)}{\sqrt{n}}+(L\kappa_{\mathrm{u}})^2r_\delta r_2^2+(L\kappa_{\mathrm{u}})^2r_\delta^2r_2^2\right).$$

Since $\{\mathbf{x}_i\}_{i=1}^n$ is an i.i.d. sequence, two semi-metrics are the same and then, for any $i \in \{1, \cdots, n\}$, we have

$$d_1(M, M') = d_2(M, M') = \|\langle \mathbf{x}_i \mathbf{x}_i^\top, M - M' \rangle\|_{\psi_1} \leq C(L\kappa_{\mathrm{u}})^2 \|M - M'\|_{\mathrm{F}},$$

where the last inequality follows from (42). We know that $\gamma_1(s^2\mathbb{B}_0^{d\times d} \cap r_2^2\mathbb{B}_{\mathrm{F}}^{d\times d}, d_1)$ and $\gamma_2(s^2\mathbb{B}_0^{d\times d} \cap r_2^2\mathbb{B}_{\mathrm{F}}^{d\times d}, d_2)$ is monotone increasing with respect to d_1 and d_2, and for some constants $\mathfrak{c}_1, \mathfrak{c}_2$, we have

$$\begin{aligned}
\gamma_1(s^2\mathbb{B}_0^{d\times d} \cap r_2^2\mathbb{B}_{\mathrm{F}}^{d\times d}, \mathfrak{c}_1 d_1) &= \mathfrak{c}_1 \gamma_1(s^2\mathbb{B}_0^{d\times d} \cap r_2^2\mathbb{B}_{\mathrm{F}}^{d\times d}, d_1), \\
\gamma_2(s^2\mathbb{B}_0^{d\times d} \cap r_2^2\mathbb{B}_{\mathrm{F}}^{d\times d}, \mathfrak{c}_2 d_2) &= \mathfrak{c}_2 \gamma_2(s^2\mathbb{B}_0^{d\times d} \cap r_2^2\mathbb{B}_{\mathrm{F}}^{d\times d}, d_2).
\end{aligned}$$

Combining (48) with the above properties of the γ_α-functional, we have

$$\begin{aligned}
&\max_{M \in \mathfrak{M}_{r_1,r_2,d}^{\ell_1,\mathrm{Tr}}} \left| \frac{1}{n} \sum_{i=1}^n \langle \mathbf{x}_i \mathbf{x}_i^\top - \Sigma, M \rangle \right| \\
&\quad \leq C(L\kappa_{\mathrm{u}})^2 \left(\frac{\gamma_1(s^2\mathbb{B}_0^{d\times d} \cap r_2^2\mathbb{B}_{\mathrm{F}}^{d\times d}, \|\cdot\|_{\mathrm{F}})}{n} + \frac{\gamma_2(s^2\mathbb{B}_0^{d\times d} \cap r_2^2\mathbb{B}_{\mathrm{F}}^{d\times d}, \|\cdot\|_{\mathrm{F}})}{\sqrt{n}} + r_\delta^2 r_2^2 + r_\delta r_2^2 \right),
\end{aligned}$$

and the proof is complete. ■

7.2 Proof of Proposition 20

First, we prove (31) in Proposition 20 via Lemma 38, which is necessary to calculate $\gamma_1(s^2\mathbb{B}_0^{d\times d} \cap r_2^2\mathbb{B}_{\mathrm{F}}^{d\times d}, \|\cdot\|_{\mathrm{F}})$ and $\gamma_2(s^2\mathbb{B}_0^{d\times d} \cap r_2^2\mathbb{B}_{\mathrm{F}}^{d\times d}, \|\cdot\|_{\mathrm{F}})$, and which is proved in Appendix E.

Lemma 38 *Suppose that r_2 is a positive constant. Then, we have*

$$\begin{aligned}
\gamma_1(s^2\mathbb{B}_0^{d\times d} \cap r_2^2\mathbb{B}_{\mathrm{F}}^{d\times d}, \|\cdot\|_{\mathrm{F}}) &\leq C s^2 r_2^2 \log(d/s) \\
\gamma_2(s^2\mathbb{B}_0^{d\times d} \cap r_2^2\mathbb{B}_{\mathrm{F}}^{d\times d}, \|\cdot\|_{\mathrm{F}}) &\leq C s r_2^2 \sqrt{\log(d/s)},
\end{aligned}$$

First, we prove (31). From Lemmas 37 and 38, we have, with probability at least $1 - \delta$,

$$\max_{M \in \mathfrak{M}_{r_1,r_2,d}^{\ell_1,\mathrm{Tr}}} \left| \frac{1}{n} \sum_{i=1}^n \langle \mathbf{x}_i \mathbf{x}_i^\top - \Sigma, M \rangle \right| \leq C(L\kappa_{\mathrm{u}})^2 \left(\frac{s^2 r_2^2 \log(d/s)}{n} + \frac{s r_2^2 \sqrt{\log(d/s)}}{\sqrt{n}} + r_\delta^2 r_2^2 + r_\delta r_2^2 \right) \tag{49}$$

$$\overset{(a)}{\leq} C(L\kappa_{\mathrm{u}})^2 \left(s r_{d,s} + r_\delta \right) r_2^2, \tag{50}$$

where (a) follows from $s r_{d,s}, r_\delta \leq 1$.

Next, we prove (32). Define

$$w_i^\circ = \begin{cases} \frac{1}{n(1-o/n)} & i \in \mathcal{I} \\ 0 & i \in \mathcal{O} \end{cases}. \tag{51}$$

From the optimality of $\{\hat{w}_i\}_{i=1}^n$ and the fact that $\{w_i^\circ\}_{i=1}^n \in \Delta^{n-1}(\varepsilon)$ and the definition of $\mathbf{X}_i$, we have

$$\begin{aligned}\max_{M\in\mathfrak{M}^{\ell_1,\mathrm{Tr}}_{r_1,r_2,d}} \sum_{i=1}^n \hat{w}_i\langle \mathbf{X}_i\mathbf{X}_i^\top - \Sigma, M\rangle &\le \max_{M\in\mathfrak{M}^{\ell_1,\mathrm{Tr}}_{r_1,r_2,d}} \sum_{i=1}^n w_i^\circ\langle \mathbf{X}_i\mathbf{X}_i^\top - \Sigma, M\rangle \\ &= \max_{M\in\mathfrak{M}^{\ell_1,\mathrm{Tr}}_{r_1,r_2,d}} \sum_{i=1}^n w_i^\circ\langle \mathbf{x}_i\mathbf{x}_i^\top - \Sigma, M\rangle \\ &\le \max_{M\in\mathfrak{M}^{\ell_1,\mathrm{Tr}}_{r_1,r_2,d}} \left|\sum_{i=1}^n w_i^\circ\langle \mathbf{x}_i\mathbf{x}_i^\top - \Sigma, M\rangle\right|. \end{aligned} \tag{52}$$

From triangular inequality and the fact that $o/n \le 1/2$, we have

$$\begin{aligned}\left|\sum_{i=1}^n w_i^\circ\langle \mathbf{x}_i\mathbf{x}_i^\top - \Sigma, M\rangle\right| &= \frac{1}{1-o/n}\left|\sum_{i=1}^n \frac{\langle \mathbf{x}_i\mathbf{x}_i^\top - \Sigma, M\rangle}{n} - \sum_{i\in\mathcal{O}} \frac{\langle \mathbf{x}_i\mathbf{x}_i^\top - \Sigma, M\rangle}{n}\right| \\ &\le 2\left|\sum_{i=1}^n \frac{\langle \mathbf{x}_i\mathbf{x}_i^\top - \Sigma, M\rangle}{n} - \sum_{i\in\mathcal{O}} \frac{\langle \mathbf{x}_i\mathbf{x}_i^\top - \Sigma, M\rangle}{n}\right| \\ &\le 2\left|\sum_{i=1}^n \frac{\langle \mathbf{x}_i\mathbf{x}_i^\top - \Sigma, M\rangle}{n}\right| + 2\left|\sum_{i\in\mathcal{O}} \frac{\langle \mathbf{x}_i\mathbf{x}_i^\top - \Sigma, M\rangle}{n}\right|. \end{aligned} \tag{53}$$

From (52) and (53), we have

$$\begin{aligned}&\max_{M\in\mathfrak{M}^{\ell_1,\mathrm{Tr}}_{r_1,r_2,d}} \sum_{i=1}^n \hat{w}_i\langle \mathbf{X}_i\mathbf{X}_i^\top - \Sigma, M\rangle \\ &\le 2\max_{M\in\mathfrak{M}^{\ell_1,\mathrm{Tr}}_{r_1,r_2,d}} \left(\left|\sum_{i=1}^n \frac{\langle \mathbf{x}_i\mathbf{x}_i^\top - \Sigma, M\rangle}{n}\right| + \left|\sum_{i\in\mathcal{O}} \frac{\langle \mathbf{x}_i\mathbf{x}_i^\top - \Sigma, M\rangle}{n}\right|\right) \\ &\le 2\left(\max_{M\in\mathfrak{M}^{\ell_1,\mathrm{Tr}}_{r_1,r_2,d}} \left|\sum_{i=1}^n \frac{\langle \mathbf{x}_i\mathbf{x}_i^\top - \Sigma, M\rangle}{n}\right| + \max_{M\in\mathfrak{M}^{\ell_1,\mathrm{Tr}}_{r_1,r_2,d}} \left|\sum_{i\in\mathcal{O}} \frac{\langle \mathbf{x}_i\mathbf{x}_i^\top - \Sigma, M\rangle}{n}\right|\right) \\ &\le 2\left(\max_{M\in\mathfrak{M}^{\ell_1,\mathrm{Tr}}_{r_1,r_2,d}} \left|\sum_{i=1}^n \frac{\langle \mathbf{x}_i\mathbf{x}_i^\top - \Sigma, M\rangle}{n}\right| + \max_{|\mathcal{J}|=o} \max_{M\in\mathfrak{M}^{\ell_1,\mathrm{Tr}}_{r_1,r_2,d}} \left|\sum_{i\in\mathcal{J}} \frac{\langle \mathbf{x}_i\mathbf{x}_i^\top - \Sigma, M\rangle}{n}\right|\right), \end{aligned} \tag{54}$$

We evaluate the last term of (54) in a manner similar to the proof of Lemma 5 of Dalalyan and Minasyan (2022). From (49), we have

$$\max_{|\mathcal{J}|=o} \max_{M\in\mathfrak{M}^{\ell_1,\mathrm{Tr}}_{r_1,r_2,d}} \left|\sum_{i\in\mathcal{J}} \frac{\langle \mathbf{x}_i\mathbf{x}_i^\top - \Sigma, M\rangle}{o}\right| \ge C(L\kappa_{\mathrm{u}})^2\left(s^2\frac{\log(d/s)}{o} + s\sqrt{\frac{\log(d/s)}{o}} + \sqrt{\frac{t}{o}} + \frac{t}{o}\right) r_2^2, \tag{55}$$

with probability at most

$$\leq \binom{n}{o} \times$$
$$\mathbb{P}\left[\max_{M\in\mathfrak{M}^{\ell_1,\mathrm{Tr}}_{r_1,r_2,d}} \left|\sum_{i=1}^{o} \frac{\langle \mathbf{z}_i\mathbf{z}_i^\top - \Sigma, M\rangle}{o}\right| \geq C(L\kappa_{\mathrm{u}})^2 \left(s^2\frac{\log(d/s)}{o} + s\sqrt{\frac{\log(d/s)}{o}} + \sqrt{\frac{t}{o}} + \frac{t}{o}\right) r_2^2\right]$$
$$\leq \binom{n}{o} e^{-t},$$

where $\{\mathbf{z}_i\}_{i=1}^o$ is a sequence of i.i.d. random vectors sampled from the same distribution as $\{\mathbf{x}_i\}_{i=1}^n$. Let $t = o\log(ne/o) + \log(1/\delta)$. We have

$$\binom{n}{o} e^{-t} \leq \frac{\prod_{k=0}^{o-1}(n-k)}{o!}\left(\frac{o}{ne}\right)^o \delta = \prod_{k=0}^{o-1}\frac{n-k}{n}\frac{o^o}{o!e^o}\delta \leq \delta,$$

where the last inequality follows from Stirling's formula $o^o \leq o!e^o$. From (55), with probability at least $1-\delta$, we have

$$\max_{|\mathcal{J}|=o}\max_{M\in\mathfrak{M}^{\ell_1,\mathrm{Tr}}_{r_1,r_2,d}} \left|\sum_{i\in\mathcal{J}}\frac{1}{n}\langle \mathbf{x}_i\mathbf{x}_i^\top - \Sigma, M\rangle\right|$$
$$\leq C\frac{o}{n}(L\kappa_{\mathrm{u}})^2\left(s^2\frac{\log(d/s)}{o} + s\sqrt{\frac{\log(d/s)}{o}} + \sqrt{\frac{o\log(ne/o)+\log(1/\delta)}{o}} + \frac{o\log(ne/o)+\log(1/\delta)}{o}\right) r_2^2. \tag{56}$$

From $e \leq n/o$ and $r_\delta \leq 1$, we have

$$\frac{o\log(ne/o)+\log(1/\delta)}{n} \leq 2\frac{o}{n}\log(n/o) + \frac{1}{n}\log(1/\delta) = 2r_o'^2 + r_\delta^2 \leq 2r_o' + r_\delta,$$
$$\sqrt{\frac{o}{n}}\sqrt{\frac{o\log(ne/o)+\log(1/\delta)}{n}} \leq \sqrt{\frac{o}{n}}\sqrt{2\frac{o}{n}\log(n/o)} + \sqrt{\frac{o}{n}}\sqrt{\frac{1}{n}\log(1/\delta)} \leq 2r_o' + r_\delta.$$

Combining (56) with the above two inequalities, with probability at least $1-\delta$, we have

$$\max_{|\mathcal{J}|=o}\max_{M\in\mathfrak{M}^{\ell_1,\mathrm{Tr}}_{r_1,r_2,d}} \left|\sum_{i\in\mathcal{J}}\frac{1}{n}\langle \mathbf{x}_i\mathbf{x}_i^\top - \Sigma, M\rangle\right| \leq C(L\kappa_{\mathrm{u}})^2\left(s^2\frac{\log(d/s)}{n} + s\sqrt{\frac{\log(d/s)}{n}}\sqrt{\frac{o}{n}} + 4r_o'^2 + 2r_\delta\right) r_2^2$$
$$\overset{(a)}{\leq} C(L\kappa_{\mathrm{u}})^2\left(2s^2\frac{\log(d/s)}{n} + \frac{o}{n} + 4r_o'^2 + 2r_\delta\right) r_2^2$$
$$\overset{(b)}{\leq} C(L\kappa_{\mathrm{u}})^2\left(s^2\frac{\log(d/s)}{n} + 5r_o'^2 + 2r_\delta\right) r_2^2, \tag{57}$$

where (a) follows from Young's inequality, and (b) follows from $0 < o/n \leq 1/(5e)$ and $o/n \leq r_o'^2$. Combining (49), (54) and (57), with probability at least $1-2\delta$, we have

$$\max_{M\in\mathfrak{M}^{\ell_1,\mathrm{Tr}}_{r_1,r_2,d}} \sum_{i=1}^n \hat{w}_i\langle \mathbf{X}_i\mathbf{X}_i^\top - \Sigma, M\rangle \leq C(L\kappa_{\mathrm{u}})^2\left(sr_{d,s} + r_\delta + s^2\frac{\log(d/s)}{n} + 5r_o'^2 + 2r_\delta\right) r_2^2$$
$$\overset{(a)}{\leq} C(L\kappa_{\mathrm{u}})^2\left(sr_{d,s} + r_o'^2 + r_\delta\right) r_2^2,$$

where (a) follows from $sr_{d,s} \leq 1$, and the proof is complete.

7.3 Proof of Proposition 22

We will apply Theorem 3.2 of Dirksen (2015). To do this, we need to confirm $\{h(\xi_{\lambda_o,i})\langle \mathbf{x}_i, \mathbf{v}\rangle\}_{i=1}^n$ satisfies *the* ψ_α *condition* with $\alpha = 2$ in Dirksen (2015), which is for any $u > 0$ and for any fixed $\mathbf{v}_1, \mathbf{v}_2 \in \mathbb{R}^d$,

$$\mathbb{P}\left(\left|\sum_{i=1}^n h(\xi_{\lambda_o,i})\langle \mathbf{x}_i, \mathbf{v}_1\rangle - h(\xi_{\lambda_o,i})\langle \mathbf{x}_i, \mathbf{v}_2\rangle\right| \geq u d(\mathbf{v}_1, \mathbf{v}_2)\right) \leq 2\exp\left(-u^2\right), \tag{58}$$

where $d(\mathbf{v}_1, \mathbf{v}_2)$ is a distance between $\mathbf{v}_1$ and $\mathbf{v}_2$. Note that from (ii) of Assumption 4, we have $\mathbb{E}h(\xi_{\lambda_o,i})\langle \mathbf{x}_i, \mathbf{v}\rangle = 0$. We see that (58) is satisfied with $d(\mathbf{v}_1, \mathbf{v}_2) = \sqrt{n}\|\Sigma^{\frac{1}{2}}(\mathbf{v}_1 - \mathbf{v}_2)\|_2$ because

$$\begin{aligned}\left\|h(\xi_{\lambda_o,i})\langle \mathbf{x}_i, \mathbf{v}_1\rangle - h(\xi_{\lambda_o,i})\langle \mathbf{x}_i, \mathbf{v}_2\rangle\right\|_{L_p} &= \left\|h(\xi_{\lambda_o,i})\langle \mathbf{x}_i, \mathbf{v}_1 - \mathbf{v}_2\rangle\right\|_{L_p} \\ &\overset{(a)}{\leq} \left\|\langle \mathbf{x}_i, \mathbf{v}_1 - \mathbf{v}_2\rangle\right\|_{L_p} \\ &\overset{(b)}{\leq} L\sqrt{p}\|\Sigma^{\frac{1}{2}}(\mathbf{v}_1 - \mathbf{v}_2)\|_2,\end{aligned}$$

where (a) follows from $-1 \leq h(\cdot) \leq 1$, and (b) follows from (10), and then, from (12), for any $t > 0$ and the i.i.d. assumption on $\{\mathbf{x}_i, \xi_i\}_{i=1}^n$, we have

$$\mathbb{P}\left(\left|\sum_{i=1}^n h(\xi_{\lambda_o,i})\langle \mathbf{x}_i, \mathbf{v}_1 - \mathbf{v}_2\rangle\right| > t\right) \lesssim 2\exp\left(-\frac{t^2}{L^2 n\|\Sigma^{\frac{1}{2}}(\mathbf{v}_1 - \mathbf{v}_2)\|_2^2}\right).$$

Set $t^2/(L^2 n\|\Sigma^{\frac{1}{2}}(\mathbf{v}_1 - \mathbf{v}_2)\|_2^2) = u^2$, we have (58) with $d(\mathbf{v}_1, \mathbf{v}_2) = L\sqrt{n}\|\Sigma^{\frac{1}{2}}(\mathbf{v}_1 - \mathbf{v}_2)\|_2$. Then, from Theorem 3.2 of Dirksen (2015) with $t_0 = \mathbf{0}$, we have

$$\mathbb{P}\left(\sup_{\mathbf{v}\in r_1\mathbb{B}_1^d \cap r_\Sigma \mathbb{B}_\Sigma^d}\left|\frac{1}{n}\sum_{i=1}^n h(\xi_{\lambda_o,i})\langle \mathbf{x}_i, \mathbf{v}\rangle\right| \gtrsim L\frac{\gamma_2(r_1\mathbb{B}_1^d \cap r_\Sigma \mathbb{B}_\Sigma^d, \|\Sigma^{\frac{1}{2}}(\cdot)\|_2)}{\sqrt{n}} + L\frac{u}{\sqrt{n}}r_\Sigma\right) \leq \exp(-u^2).$$

Lastly from the majorizing measure theorem (Theorem 2.4.1 of Talagrand (2014)) and Lemma 34, we have

$$\mathbb{P}\left(\sup_{\mathbf{v}\in r_1\mathbb{B}_1^d \cap r_\Sigma \mathbb{B}_\Sigma^d}\left|\frac{1}{n}\sum_{i=1}^n h(\xi_{\lambda_o,i})\langle \mathbf{x}_i, \mathbf{v}\rangle\right| \gtrsim L\rho\sqrt{\frac{\log(d/s)}{n}}r_1 + L\sqrt{\frac{u}{n}}r_\Sigma\right) \leq \exp(-u),$$

and the proof is complete.

8. Numerical Experiments

In this section, we present numerical experiments about our methods. In all of the experiments, we used CVXPY (Diamond and Boyd, 2016). The data were generated from

$$y_i = \langle \boldsymbol{\beta}^*, \mathbf{x}_i + \boldsymbol{\varrho}_i\rangle + \xi_i + \theta_i, \quad i = 1, \cdots, 80,$$

where $\{\xi_i\}_{i=1}^{80}$ was drawn from Student's t-distribution with two degrees of freedom, $\{\mathbf{x}_i\}_{i=1}^{80}$ was drawn from 120-dimensional i.i.d. Gaussian, $\boldsymbol{\beta}^* = (20, -10, 0, \cdots, 0)$ and $\{\xi_i, \boldsymbol{\varrho}_i\}_{i=1}^{80}$ is a sequence of outliers. We note that due to the high computational cost of WEIGHT (Algorithm 2) in Algorithm 1, the sample size and dimensionality in the experiments were restricted. We increased the number of outliers o from 0 to 20 and we conducted 10 experiments for each value. For outliers, we randomly chose a set of indices $\mathcal{O} \subset [1, \cdots, 80]$ such that $|\mathcal{O}| = o$, and we tried two patterns of outlier values: $(\theta_i, \boldsymbol{\varrho}_i)_{i\in\mathcal{O}} = (10000, \mathbf{1})$ and $(1, \mathbf{10000})$. For both patterns, we set $(\theta_i, \boldsymbol{\varrho}_i)_{i\in[1,\cdots,80]\setminus\mathcal{O}} = (0, \mathbf{0})$. We compared the estimator from Algorithm 1, the estimator from Algorithm 6, and the estimator from the standard Lasso. The estimator of the standard Lasso is defined as follows:

$$\hat{\boldsymbol{\beta}} = \min_{\boldsymbol{\beta}\in\mathbb{R}^d} \left(\sum_{i=1}^{n} (y_i - \langle \mathbf{x}_i + \boldsymbol{\varrho}_i, \boldsymbol{\beta} \rangle)^2 + \lambda_s \|\boldsymbol{\beta}\|_1 \right).$$

The results are shown in Figure 1 and Figure 2.

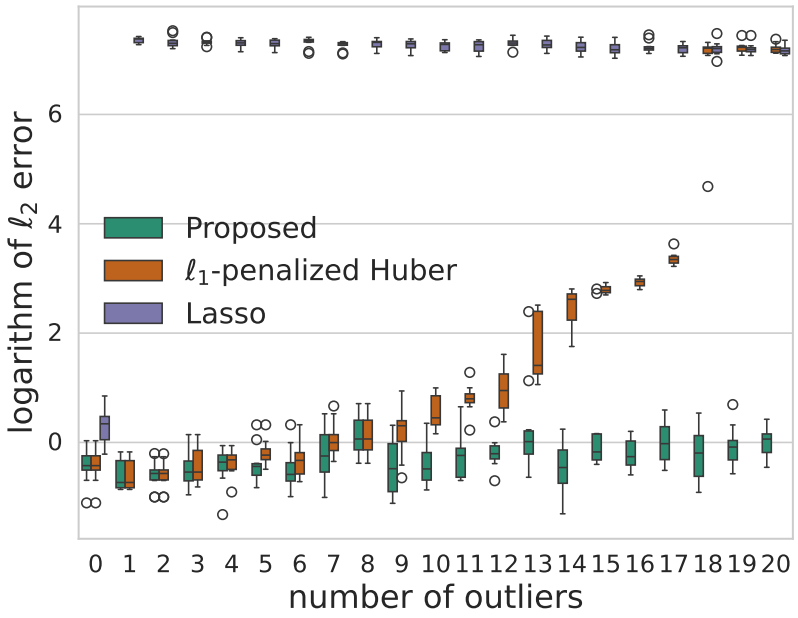

Figure 1: $(\theta_i, \boldsymbol{\varrho}_i)_{i\in\mathcal{O}} = (10000, \mathbf{1})$

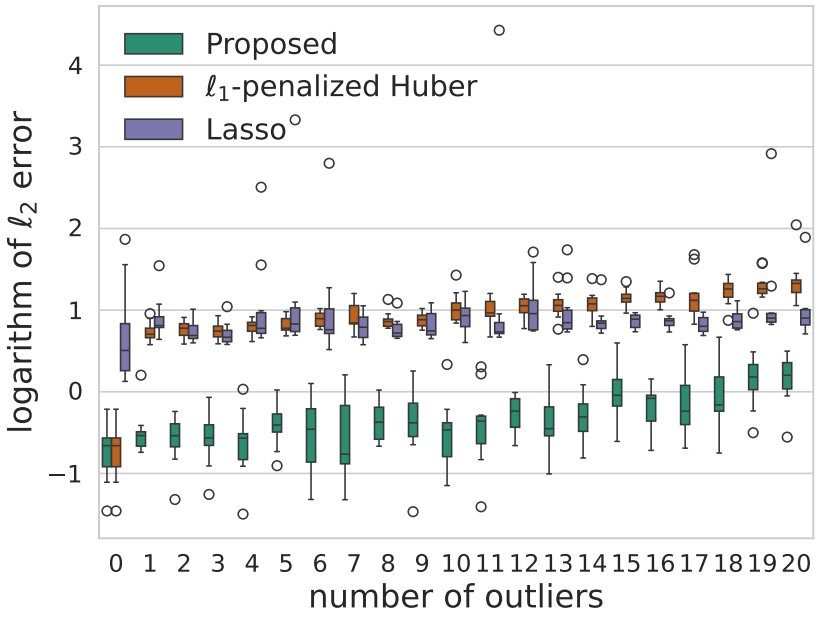

Figure 2: $(\theta_i, \boldsymbol{\varrho}_i)_{i\in\mathcal{O}} = (1, \mathbf{10000})$

The horizontal axis represents the number of outliers, and the vertical axis represents $\log \|\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}\|_2$, where $\hat{\boldsymbol{\beta}}$ is the estimator of each method. For Algorithm 1, we set

$$(\lambda_o, \lambda_s, r_1, r_2, \varepsilon) = \left(\frac{1}{\sqrt{n}}, \sqrt{\frac{\log(d/s)}{n}} + \frac{o}{n\sqrt{s}}\sqrt{\log\frac{n}{o}}, s\lambda_s, \sqrt{s}\lambda_s, \sqrt{\frac{\log(d/s)}{n}}, \frac{o}{n}\right),$$

and for simplicity we omitted the 'if' part of WEIGHT. For Algorithm 6 and the standard Lasso, we set

$$(\lambda_o, \lambda_s) = \left(\frac{1}{\sqrt{n}}, \sqrt{\frac{\log(d/s)}{n}} + \frac{o}{n\sqrt{s}}\sqrt{\log\frac{n}{o}}\right).$$

According to Figures 1 and Figure 2, when outliers are present, the standard Lasso does not seem to perform well. According to Figure 1, the ℓ_1-penalized Huber seems to perform relatively well when the magnitude of outliers in the covariates is small and the number of outliers is limited. This observation is somewhat in line with the results suggested by Corollary 18. On the other hand, Figure 2 shows that when the magnitude of outliers in the covariates is large, Algorithm 1 consistently demonstrates superior performance, regardless of the number of outliers.

Acknowledgments

The authors would like to thank the Action Editor and anonymous reviewers for their careful attention and insightful suggestions, which have significantly improved the quality of this paper. In particular, the authors are deeply grateful to the reviewer who pointed out a

critical mistake in the initial manuscript. This work was supported by JSPS KAKENHI Grant Number 17K00065 and 24K14870.

Appendix A. Proof of Theorem 13

Suppose that the assumptions in Theorem 13 hold. To prove Theorem 13, it is sufficient that we confirm (33) – (36) in Proposition 21 hold with high probability, as described already, and then we obtain the result (23) with concrete values of $r_{a,1}, r_{a,2}, r_{a,\Sigma}, r_{b,1}, r_{b,2}, r_{b,\Sigma}, b$. Obviously, we have $r_1 = c_{r_1}\sqrt{s}r_\Sigma$ and $r_2 = c_{r_2}r_\Sigma$ in (36) from (22). In the following, we show the remaining. We see that, under the assumptions in Theorem 13, Propositions 20, 22 and 25 hold immediately with probability at least $1-4\delta$, and then, Propositions 23 and 24 also hold because (31) and (32) in Proposition 20 hold and Algorithm 2 returns $\hat{w}$ from the definition of τ_{cut}.

A.1 Confirmation of (33)

From the following lemma, we can confirm (33). The proofs are given in Appendix G.

Lemma 39 *Assume that Propositions 20, 22-25 hold. For any $\boldsymbol{\theta}_\eta \in r_1\mathbb{B}_1^d \cap r_2\mathbb{B}_2^d \cap r_\Sigma\mathbb{B}_\Sigma^d$, we have*

$$\left|\sum_{i=1}^n \hat{w}_i' h(r_{\boldsymbol{\beta}^*,i})\mathbf{X}_i^\top \boldsymbol{\theta}_\eta\right| \leq 3c_{\max}^2 L\left(\rho r_{d,s}r_1 + r_\delta r_\Sigma + \kappa_{\mathrm{u}}\sqrt{\frac{o}{n}(sr_{d,s}+r_\delta)}r_2 + \kappa_{\mathrm{u}}r_o r_2\right).$$

From Lemma 39, we see that (33) holds with

$$\begin{aligned} &r_{a,1} = 3c_{\max}^2 L\lambda_o\sqrt{n}\rho r_{d,s}, \quad r_{a,2} = 3c_{\max}^2 L\lambda_o\sqrt{n}\left(\kappa_{\mathrm{u}}\sqrt{\frac{o}{n}(sr_{d,s}+r_\delta)} + \kappa_{\mathrm{u}}r_o\right), \\ &r_{a,\Sigma} = 3c_{\max}^2 L\lambda_o\sqrt{n}r_\delta. \end{aligned} \tag{59}$$

A.2 Confirmation of (35)

From (59),

$$(C_s :=)r_{a,1} + \frac{c_{r_2}r_{a,2} + r_{a,\Sigma}}{c_{r_1}\sqrt{s}} = 3c_{\max}^2 L\lambda_o\sqrt{n} \times \frac{1}{c_{r_1}\sqrt{s}}R_{d,n,o}.$$

From the definition of λ_s,

$$\frac{\lambda_s}{C_s} \geq \frac{c_s c_{\max}^2 L\lambda_o\sqrt{n}\frac{1}{c_{r_1}\sqrt{s}}R_{d,n,o}}{3c_{\max}^2 L\lambda_o\sqrt{n}\frac{1}{c_{r_1}\sqrt{s}}R_{d,n,o}} = \frac{c_s}{3} \geq \frac{c_{\mathrm{RE}}+1}{c_{\mathrm{RE}}-1} > 0.$$

Hence, we have $\lambda_s - C_s > 0$ and

$$\frac{\lambda_s + C_s}{\lambda_s - C_s} = \frac{1+\frac{C_s}{\lambda_s}}{1-\frac{C_s}{\lambda_s}} \leq \frac{1+\frac{c_{\mathrm{RE}}-1}{c_{\mathrm{RE}}+1}}{1-\frac{c_{\mathrm{RE}}-1}{c_{\mathrm{RE}}+1}} = c_{\mathrm{RE}}.$$

Therefore, we see that (35) holds.

A.3 Confirmation of (34)

From the following lemma, we can confirm (34). The proofs are given in the Appendix G.

Lemma 40 *Assume that Propositions 20, 22–25 hold. For any* $\boldsymbol{\theta}_\eta \in r_1\mathbb{B}_1^d \cap r_2\mathbb{B}_2^d \cap r_\Sigma\mathbb{B}_\Sigma^d$, *we have*

$$\begin{aligned}
&\sum_{i=1}^n \lambda_o\sqrt{n}\hat{w}_i' \left(-h(r_{\boldsymbol{\beta}^*+\boldsymbol{\theta}_\eta,i}) + h(r_{\boldsymbol{\beta}^*,i})\right) \mathbf{X}_i^\top \boldsymbol{\theta}_\eta \\
&\geq \frac{\|\Sigma^{\frac{1}{2}}\mathbf{v}\|_2^2}{3} - c_{\max}^2 L\lambda_o\sqrt{n}\left(\rho r_{d,s}r_1 + r_\delta r_\Sigma + \kappa_{\mathrm{u}}\sqrt{\frac{o}{n}(sr_{d,s}+r_\delta)}r_2 + \kappa_{\mathrm{u}} r_o r_2\right).
\end{aligned}$$

From Lemma 40, we see that (34) holds with

$$b = \frac{1}{3}, \quad r_{b,1} = \frac{r_{a,1}}{3}, \quad r_{b,2} = \frac{r_{a,2}}{3}, \quad r_{b,\Sigma} = \frac{r_{a,\Sigma}}{3}.$$

A.4 Confirmation of (36)

As we mentioned at the beginning of Appendix A, $r_1 = c_{r_1}\sqrt{s}r_\Sigma$ and $r_2 = c_{r_2}r_\Sigma$ is clear from their definitions. We confirm r_Σ. We see that

$$\begin{aligned}
&\frac{2}{b}\left(c_{r_1}\sqrt{s}(r_{a,1}+r_{b,1}) + c_{r_2}(r_{a,2}+r_{b,2}) + r_{a,\Sigma} + r_{b,\Sigma} + c_{r_1}\sqrt{s}\lambda_s\right) \\
&\overset{(a)}{\leq} 6c_{\max}^2 L\lambda_o\sqrt{n}\,(4+c_s)\,R_{d,n,o} \\
&\leq r_\Sigma,
\end{aligned}$$

and the proof is complete.

Appendix B. Proof of Theorem 15

The strategy of this proof is the same as that of Theorem 13. Suppose that the assumptions in Theorem 15 hold. To prove Theorem 15, it is sufficient that we confirm (33) – (36) in Proposition 21 hold with high probability, as described already, and then we obtain the result (27) with concrete values of $r_{a,1}, r_{a,2}, r_{a,\Sigma}, r_{b,1}, r_{b,2}, r_{b,\Sigma}, b$. Obviously, we have $r_1 = c_{r_1}\sqrt{s}r_\Sigma$ and $r_2 = c_{r_2}r_\Sigma$ in (36) from (22). In the following, we show the remaining. We see that, under the assumptions in Theorem 15, Corollary 28, Propositions 22 and 25 hold immediately with probability at least $1-3\delta$, and then, Propositions 29 and 30 also hold because (39) and (40) in Corollary 28 hold, and Algorithm 5 returns $\hat{w}$ from the definition of τ_{cut}.

B.1 Confirmation of (33)

From the following lemma, we can confirm (33). The proofs are given by Appendix G.

Lemma 41 *Assume that Corollary 28, Propositions 22, 25, 29 and 30 hold. For any* $\boldsymbol{\theta}_\eta \in r_1\mathbb{B}_1^d \cap r_2\mathbb{B}_2^d \cap r_\Sigma\mathbb{B}_\Sigma^d$, *we have*

$$\left|\sum_{i=1}^n \hat{w}_i' h(r_{\boldsymbol{\beta}^*,i})\mathbf{X}_i^\top\boldsymbol{\theta}_\eta\right| \leq 3c_{\max}'^2 L\left(\rho r_{d,s}r_1 + r_\delta r_\Sigma + \sqrt{\frac{o}{n}}\sqrt{\kappa_{\mathrm{u}}^2\left(sr_{d,s}+r_\delta\right) + \Sigma_{\max}^2}r_2\right).$$

From Lemma 39, we see that (33) holds with

$$r_{a,1} = 3c_{\max}'^2 L\lambda_o\sqrt{n}\rho r_{d,s}, \quad r_{a,2} = 3c_{\max}'^2 L\lambda_o\sqrt{n}\sqrt{\frac{o}{n}}\sqrt{\kappa_{\mathrm{u}}^2\left(sr_{d,s}+r_\delta\right)+\Sigma_{\max}^2},$$
$$r_{a,\Sigma} = 3c_{\max}'^2 L\lambda_o\sqrt{n}r_\delta. \tag{60}$$

B.2 Confirmation of (35)

From (60),

$$(C_s :=) r_{a,1} + \frac{c_{r_2}r_{a,2}+r_{a,\Sigma}}{c_{r_1}\sqrt{s}} = 3c_{\max}'^2 L\lambda_o\sqrt{n} \times \frac{1}{c_{r_1}\sqrt{s}}R'_{d,n,o}.$$

From the definition of λ_s,

$$\frac{\lambda_s}{C_s} \geq \frac{c_s c_{\max}'^2 L\lambda_o\sqrt{n}\frac{1}{c_{r_1}\sqrt{s}}R'_{d,n,o}}{3c_{\max}'^2 L\lambda_o\sqrt{n}\frac{1}{c_{r_1}\sqrt{s}}R'_{d,n,o}} = \frac{c_s}{3} \geq \frac{c_{\mathrm{RE}}+1}{c_{\mathrm{RE}}-1} > 0.$$

Hence, we have $\lambda_s - C_s > 0$ and

$$\frac{\lambda_s + C_s}{\lambda_s - C_s} = \frac{1+\frac{C_s}{\lambda_s}}{1-\frac{C_s}{\lambda_s}} \leq \frac{1+\frac{c_{\mathrm{RE}}-1}{c_{\mathrm{RE}}+1}}{1-\frac{c_{\mathrm{RE}}-1}{c_{\mathrm{RE}}+1}} = c_{\mathrm{RE}}.$$

Therefore, we see that (35) holds.

B.3 Confirmation of (34)

From the following lemma, we can confirm (34). The proofs are given by Appendix G.

Lemma 42 *Assume that Corollary 28, Propositions 22, 25, 29 and 30 hold. For any* $\boldsymbol{\theta}_\eta \in r_1\mathbb{B}_1^d \cap r_2\mathbb{B}_2^d \cap r_\Sigma\mathbb{B}_\Sigma^d$*, we have*

$$\sum_{i=1}^n \lambda_o\sqrt{n}\hat{w}_i'\left(-h(r_{\boldsymbol{\beta}^*+\boldsymbol{\theta}_\eta,i}) + h(r_{\boldsymbol{\beta}^*,i})\right)\mathbf{X}_i^\top\boldsymbol{\theta}_\eta$$
$$\geq \frac{\|\Sigma^{\frac{1}{2}}\mathbf{v}\|_2^2}{3} - c_{\max}'^2 L\lambda_o\sqrt{n}\left(\rho r_{d,s}r_1 + r_\delta r_\Sigma + \sqrt{\frac{o}{n}}\sqrt{\kappa_{\mathrm{u}}^2\left(sr_{d,s}+r_\delta\right)+\Sigma_{\max}^2}r_2\right).$$

From Lemma 42, we see that (34) holds with

$$b = \frac{1}{3}, \quad r_{b,1} = \frac{r_{a,1}}{3}, \quad r_{b,2} = \frac{r_{a,2}}{3}, \quad r_{b,\Sigma} = \frac{r_{a,\Sigma}}{3}.$$

B.4 Confirmation of (36)

As we mentioned at the beginning of Appendix B, $r_1 = c_{r_1}\sqrt{s}r_\Sigma$ and $r_2 = c_{r_2}r_\Sigma$ is clear from their definitions. We confirm r_Σ. We see that

$$\frac{2}{b}\left(c_{r_1}\sqrt{s}(r_{a,1}+r_{b,1}) + c_{r_2}(r_{a,2}+r_{b,2}) + r_{a,\Sigma} + r_{b,\Sigma} + c_{r_1}\sqrt{s}\lambda_s\right)$$
$$\leq 6c_{\max}'^2 L\lambda_o\sqrt{n}\left(4+c_s\right)\left(\rho c_{r_1}\sqrt{s}r_{d,s} + r_\delta + c_{r_2}\sqrt{\frac{o}{n}}\sqrt{\kappa_{\mathrm{u}}^2\left(sr_{d,s}+r_\delta\right)+\Sigma_{\max}^2}\right)$$
$$\leq r_\Sigma,$$

and the proof is complete.

Appendix C. Preparation of Proof of Proposition 21

We introduce the following four lemmas, which are used in the proof of Proposition 21.

Lemma 43 *Suppose that* (33), (35), $\|\boldsymbol{\theta}_\eta\|_1 \le r_1$, $\|\boldsymbol{\theta}_\eta\|_2 = r_2$ *and* $\|\Sigma^{\frac{1}{2}}\boldsymbol{\theta}_\eta\|_2 \le r_\Sigma$, *where* $r_1 = c_{r_1}\sqrt{s}r_\Sigma$ *and* $r_2 = c_{r_2}r_\Sigma$ *hold. Then, for any fixed* $\eta \in (0,1)$, *we have*

$$\|\boldsymbol{\theta}_\eta\|_2 \le \frac{3 + c_{\mathrm{RE}}}{\kappa_{\mathrm{l}}}\|\Sigma^{\frac{1}{2}}\boldsymbol{\theta}_\eta\|_2.$$

Lemma 44 *Suppose that* (33), (35), $\|\boldsymbol{\theta}_\eta\|_1 = r_1$, $\|\boldsymbol{\theta}_\eta\|_2 \le r_2$ *and* $\|\Sigma^{\frac{1}{2}}\boldsymbol{\theta}_\eta\|_2 \le r_\Sigma$, *where* $r_1 = c_{r_1}\sqrt{s}r_\Sigma$ *and* $r_2 = c_{r_2}r_\Sigma$ *hold. Then, for any fixed* $\eta \in (0,1)$, *we have*

$$\|\boldsymbol{\theta}_\eta\|_1 \le \frac{1 + c_{\mathrm{RE}}}{\mathfrak{r}}\sqrt{s}\|\Sigma^{\frac{1}{2}}\boldsymbol{\theta}_\eta\|_2.$$

Lemma 45 *Suppose that* (33) – (34), $\|\boldsymbol{\theta}_\eta\|_1 \le r_1$, $\|\boldsymbol{\theta}_\eta\|_2 \le r_2$ *and* $\|\Sigma^{\frac{1}{2}}\boldsymbol{\theta}_\eta\|_2 = r_2$, *where* $r_1 = c_{r_1}\sqrt{s}r_\Sigma$ *and* $r_2 = c_{r_2}r_\Sigma$ *hold. Then, for any* $\eta \in (0,1)$, *we have*

$$\|\Sigma^{\frac{1}{2}}\boldsymbol{\theta}_\eta\|_2 \le \frac{1}{b}\left(c_{r_1}\sqrt{s}(r_{a,1} + r_{b,1}) + c_{r_2}(r_{a,2} + r_{b,2}) + r_{a,\Sigma} + r_{b,\Sigma} + c_{r_1}\sqrt{s}\lambda_s\right).$$

C.1 Proof of Lemma 43

For a vector $\mathbf{v} = (v_1, \cdots, v_d)$, define $\{v_1^\sharp, \cdots, v_d^\sharp\}$ as a non-increasing rearrangement of $\{|v_1|, \cdots, |v_d|\}$, and $\mathbf{v}^\sharp \in \mathbb{R}^d$ as the vector such that $\mathbf{v}^\sharp|_i = v_i^\sharp$. For the sets $S_1 = \{1, ..., s\}$ and $S_2 = \{s+1, ..., d\}$, let $v^{\sharp 1} = v^\sharp_{S_1}$ and $v^{\sharp 2} = v^\sharp_{S_2}$.

In Section C.1.1, we have

$$\|\boldsymbol{\theta}_\eta\|_2 \le \frac{\sqrt{1 + c_{\mathrm{RE}}}}{\kappa_{\mathrm{l}}}\|\Sigma^{\frac{1}{2}}\boldsymbol{\theta}_\eta\|_2$$

assuming $\|\boldsymbol{\theta}_\eta\|_2 \le \|\boldsymbol{\theta}_\eta\|_1/\sqrt{s}$, and in Section C.1.2, we have

$$\|\boldsymbol{\theta}_\eta\|_2 \le 2\|\boldsymbol{\theta}_\eta^{\sharp 1}\|_2 \le \frac{2}{\kappa_{\mathrm{l}}}\|\Sigma^{\frac{1}{2}}\boldsymbol{\theta}_\eta\|_2$$

assuming $\|\boldsymbol{\theta}_\eta\|_2 \ge \|\boldsymbol{\theta}_\eta\|_1/\sqrt{s}$. From the above two inequalities, we have

$$\|\boldsymbol{\theta}_\eta\|_2 \le \frac{3 + c_{\mathrm{RE}}}{\kappa_{\mathrm{l}}}\|\Sigma^{\frac{1}{2}}\boldsymbol{\theta}_\eta\|_2.$$

C.1.1 Case I

In Section C.1.1, suppose that $\|\boldsymbol{\theta}_\eta\|_2 \le \|\boldsymbol{\theta}_\eta\|_1/\sqrt{s}$. Let

$$Q'(\eta) = \lambda_o\sqrt{n}\hat{w}_i'\sum_{i=1}^n\left(-h(r_{\boldsymbol{\beta}^* + \boldsymbol{\theta}_\eta, i}) + h(r_{\boldsymbol{\beta}^*, i})\right)\mathbf{X}_i^\top\boldsymbol{\theta}.$$

From the proof of Lemma F.2. of Fan et al. (2018), we have $\eta Q'(\eta) \leq \eta Q'(1)$ and this means

$$\sum_{i=1}^{n} \lambda_o \sqrt{n} \hat{w}_i' \left(-h(r_{\boldsymbol{\beta}^* + \boldsymbol{\theta}_\eta, i}) + h(r_{\boldsymbol{\beta}^*, i}) \right) \mathbf{X}_i^\top \boldsymbol{\theta}_\eta \leq \sum_{i=1}^{n} \lambda_o \sqrt{n} \hat{w}_i' \eta \left(-h(r_{\hat{\boldsymbol{\beta}}, i}) + h(r_{\boldsymbol{\beta}^*, i}) \right) \mathbf{X}_i^\top \boldsymbol{\theta}_\eta \tag{61}$$

Let $\partial \mathbf{v}$ be the sub-differential of $\|\mathbf{v}\|_1$. Adding $\eta \lambda_s (\|\hat{\boldsymbol{\beta}}\|_1 - \|\boldsymbol{\beta}^*\|_1)$ to both sides of (61), we have

$$\begin{aligned}
&\sum_{i=1}^{n} \lambda_o \sqrt{n} \hat{w}_i' \left(-h(r_{\boldsymbol{\beta}^* + \boldsymbol{\theta}_\eta, i}) + h(r_{\boldsymbol{\beta}^*, i}) \right) \mathbf{X}_i^\top \boldsymbol{\theta}_\eta + \eta \lambda_s (\|\hat{\boldsymbol{\beta}}\|_1 - \|\boldsymbol{\beta}^*\|_1) \\
&\overset{(a)}{\leq} \sum_{i=1}^{n} \lambda_o \sqrt{n} \hat{w}_i' \eta \left(-h(r_{\hat{\boldsymbol{\beta}}, i}) + h(r_{\boldsymbol{\beta}^*, i}) \right) \mathbf{X}_i^\top \hat{\boldsymbol{\theta}} + \eta \lambda_s \langle \partial \hat{\boldsymbol{\beta}}, \boldsymbol{\theta} \rangle \\
&\overset{(b)}{=} \sum_{i=1}^{n} \lambda_o \sqrt{n} \hat{w}_i' h(r_{\boldsymbol{\beta}^*, i}) \mathbf{X}_i^\top \boldsymbol{\theta}_\eta,
\end{aligned} \tag{62}$$

where (a) follows from $\|\hat{\boldsymbol{\beta}}\|_1 - \|\boldsymbol{\beta}^*\|_1 \leq \langle \partial \hat{\boldsymbol{\beta}}, \boldsymbol{\theta} \rangle$, which is the definition of the sub-differential, and (b) follows from the optimality of $\hat{\boldsymbol{\beta}}$.

From the convexity of Huber loss, the first term of the left-hand side of (62) is positive and we have

$$0 \leq \sum_{i=1}^{n} \lambda_o \sqrt{n} \hat{w}_i' h(r_{\boldsymbol{\beta}^*, i}) \mathbf{X}_i^\top \boldsymbol{\theta}_\eta + \eta \lambda_s (\|\boldsymbol{\beta}^*\|_1 - \|\hat{\boldsymbol{\beta}}\|_1). \tag{63}$$

From (33), the first term of the right-hand side of (63) is evaluated as

$$\begin{aligned}
\sum_{i=1}^{n} \lambda_o \sqrt{n} \hat{w}_i' h(r_{\boldsymbol{\beta}^*, i}) \mathbf{X}_i^\top \boldsymbol{\theta}_\eta &\leq r_{a,1} r_1 + r_{a,2} r_2 + r_{a,\Sigma} r_\Sigma \\
&\overset{(a)}{\leq} \left(\frac{c_{r_1}}{c_{r_2}} \sqrt{s} r_{a,1} + r_{a,2} + \frac{1}{c_{r_2}} r_{a,\Sigma} \right) \|\boldsymbol{\theta}_\eta\|_2 \\
&\overset{(b)}{\leq} \frac{1}{\sqrt{s}} \left(\frac{c_{r_1}}{c_{r_2}} \sqrt{s} r_{a,1} + r_{a,2} + \frac{1}{c_{r_2}} r_{a,\Sigma} \right) \|\boldsymbol{\theta}_\eta\|_1,
\end{aligned} \tag{64}$$

where (a) follows from $r_1 = c_{r_1} \sqrt{s} r_\Sigma$ and $r_2 = c_{r_2} r_\Sigma$ and (b) follows from the assumption of this section, $\|\boldsymbol{\theta}_\eta\|_2 \leq \|\boldsymbol{\theta}_\eta\|_1 / \sqrt{s}$. From (63), (64) and the assumption $\|\boldsymbol{\theta}_\eta\|_2 \leq \|\boldsymbol{\theta}_\eta\|_1 / \sqrt{s}$, we have

$$0 \leq \left(\frac{r_{a,2}}{\sqrt{s}} + \frac{c_{r_1} \sqrt{s} r_{a,1} + r_{a,\Sigma}}{c_{r_2} \sqrt{s}} \right) \|\boldsymbol{\theta}_\eta\|_1 + \eta \lambda_s (\|\boldsymbol{\beta}^*\|_1 - \|\hat{\boldsymbol{\beta}}\|_1).$$

Define $\mathcal{J}_{\mathbf{a}}$ as the index set of non-zero entries of $\mathbf{a}$, and $\boldsymbol{\theta}_{\eta,\mathcal{J}_{\mathbf{a}}}$ as a vector such that $\boldsymbol{\theta}_{\eta,\mathcal{J}_{\mathbf{a}}}|_i = \boldsymbol{\theta}_\eta|_i$ for $i \in \mathcal{J}_{\mathbf{a}}$ and $\boldsymbol{\theta}_{\eta,\mathcal{J}_{\mathbf{a}}}|_i = 0$ for $i \notin \mathcal{J}_{\mathbf{a}}$. Furthermore, we see

$$
\begin{aligned}
0 &\leq \left(\frac{r_{a,2}}{\sqrt{s}} + \frac{c_{r_1}\sqrt{s}r_{a,1} + r_{a,\Sigma}}{c_{r_2}\sqrt{s}}\right)\|\boldsymbol{\theta}_\eta\|_1 + \eta\lambda_s(\|\boldsymbol{\beta}^*\|_1 - \|\hat{\boldsymbol{\beta}}\|_1) \\
&\leq \left(\frac{r_{a,2}}{\sqrt{s}} + \frac{c_{r_1}\sqrt{s}r_{a,1} + r_{a,\Sigma}}{c_{r_2}\sqrt{s}}\right)(\|\boldsymbol{\theta}_{\eta,\mathcal{J}_{\boldsymbol{\beta}^*}}\|_1 + \|\boldsymbol{\theta}_{\eta,\mathcal{J}^c_{\boldsymbol{\beta}^*}}\|_1) + \eta\lambda_s(\|\boldsymbol{\beta}^*_{\mathcal{J}_{\boldsymbol{\beta}^*}} - \hat{\boldsymbol{\beta}}_{\mathcal{J}_{\boldsymbol{\beta}^*}}\|_1 - \|\hat{\boldsymbol{\beta}}_{\mathcal{J}^c_{\boldsymbol{\beta}^*}}\|_1) \\
&= \left(\lambda_s + \left(\frac{r_{a,2}}{\sqrt{s}} + \frac{c_{r_1}\sqrt{s}r_{a,1} + r_{a,\Sigma}}{c_{r_2}\sqrt{s}}\right)\right)\|\boldsymbol{\theta}_{\eta,\mathcal{J}_{\boldsymbol{\beta}^*}}\|_1 + \left(-\lambda_s + \left(\frac{r_{a,2}}{\sqrt{s}} + \frac{c_{r_1}\sqrt{s}r_{a,1} + r_{a,\Sigma}}{c_{r_2}\sqrt{s}}\right)\right)\|\boldsymbol{\theta}_{\eta,\mathcal{J}^c_{\boldsymbol{\beta}^*}}\|_1.
\end{aligned}
$$

Then, we have

$$
\|\boldsymbol{\theta}_{\eta,\mathcal{J}^c_{\boldsymbol{\beta}^*}}\|_1 \leq \frac{\lambda_s + \left(\frac{r_{a,2}}{\sqrt{s}} + \frac{c_{r_1}\sqrt{s}r_{a,1}+r_{a,\Sigma}}{c_{r_2}\sqrt{s}}\right)}{\lambda_s - \left(\frac{r_{a,2}}{\sqrt{s}} + \frac{c_{r_1}\sqrt{s}r_{a,1}+r_{a,\Sigma}}{c_{r_2}\sqrt{s}}\right)}\|\boldsymbol{\theta}_{\eta,\mathcal{J}_{\boldsymbol{\beta}^*}}\|_1 \overset{(a)}{\leq} \frac{\lambda_s + \left(r_{a,1} + \frac{c_{r_2}r_{a,2}+r_{a,\Sigma}}{c_{r_1}\sqrt{s}}\right)}{\lambda_s - \left(r_{a,1} + \frac{c_{r_2}r_{a,2}+r_{a,\Sigma}}{c_{r_1}\sqrt{s}}\right)}\|\boldsymbol{\theta}_{\eta,\mathcal{J}_{\boldsymbol{\beta}^*}}\|_1 \overset{(b)}{\leq} c_{\mathrm{RE}}\|\boldsymbol{\theta}_{\eta,\mathcal{J}_{\boldsymbol{\beta}^*}}\|_1,
$$

where (a) follows from the fact that $c_{r_2} \geq c_{r_1}$ and (b) follows from (35), and from the definition of $\|\boldsymbol{\theta}^{\sharp 2}_\eta\|_1$ and $\|\boldsymbol{\theta}^{\sharp 1}_\eta\|_1$, we have

$$
\|\boldsymbol{\theta}^{\sharp 2}_\eta\|_1 \leq c_{\mathrm{RE}}\|\boldsymbol{\theta}^{\sharp 1}_\eta\|_1.
$$

Then, from the standard shelling argument, we have

$$
\|\boldsymbol{\theta}^{\sharp 2}_\eta\|_2^2 = \sum_{i=s+1}^{d}(\boldsymbol{\theta}^\sharp_\eta|_i)^2 \leq \sum_{i=s+1}^{d}\left|\boldsymbol{\theta}^\sharp_\eta|_i\right|\left(\frac{1}{s}\sum_{j=1}^{s}\left|\boldsymbol{\theta}^\sharp_\eta|_j\right|\right) \leq \frac{1}{s}\|\boldsymbol{\theta}^{\sharp 1}_\eta\|_1\|\boldsymbol{\theta}^{\sharp 2}_\eta\|_1 \leq \frac{c_{\mathrm{RE}}\|\boldsymbol{\theta}^{\sharp 1}_\eta\|_1^2}{s} \leq c_{\mathrm{RE}}\|\boldsymbol{\theta}^{\sharp 1}_\eta\|_2^2.
$$

and from the definition of κ_{l}, we have

$$
\kappa_{\mathrm{l}}^2\|\boldsymbol{\theta}_\eta\|_2^2 \leq \kappa_{\mathrm{l}}^2\left(\|\boldsymbol{\theta}^{\sharp 1}_\eta\|_2^2 + \|\boldsymbol{\theta}^{\sharp 2}_\eta\|_2^2\right) \leq \kappa_{\mathrm{l}}^2(1 + c_{\mathrm{RE}})\|\boldsymbol{\theta}^{\sharp 1}_\eta\|_2^2 \leq (1 + c_{\mathrm{RE}})\|\Sigma^{\frac{1}{2}}\boldsymbol{\theta}_\eta\|_2^2
$$

C.1.2 Case II

In Section C.1.2, suppose that $\|\boldsymbol{\theta}_\eta\|_2 \geq \|\boldsymbol{\theta}_\eta\|_1/\sqrt{s}$.

$$
\|\boldsymbol{\theta}^{\sharp 2}_\eta\|_2^2 = \sum_{i=s+1}^{d}(\boldsymbol{\theta}^\sharp_\eta|_i)^2 \leq \sum_{i=s+1}^{d}\left|\boldsymbol{\theta}^\sharp_\eta|_i\right|\left(\frac{1}{s}\sum_{j=1}^{s}\left|\boldsymbol{\theta}^\sharp_\eta|_j\right|\right) \leq \frac{1}{s}\|\boldsymbol{\theta}^{\sharp 1}_\eta\|_1\|\boldsymbol{\theta}^{\sharp 2}_\eta\|_1 \leq \|\boldsymbol{\theta}^{\sharp 1}_\eta\|_2\|\boldsymbol{\theta}_\eta\|_2.
$$

Then, we have

$$
\|\boldsymbol{\theta}_\eta\|_2^2 \leq \|\boldsymbol{\theta}^{\sharp 1}_\eta\|_2^2 + \|\boldsymbol{\theta}^{\sharp 2}_\eta\|_2^2 \leq \|\boldsymbol{\theta}^{\sharp 1}_\eta\|_2\|\boldsymbol{\theta}_\eta\|_2 + \|\boldsymbol{\theta}^{\sharp 1}_\eta\|_2\|\boldsymbol{\theta}_\eta\|_2 \Rightarrow \|\boldsymbol{\theta}_\eta\|_2 \leq 2\|\boldsymbol{\theta}^{\sharp 1}_\eta\|_2,
$$

and we have

$$
\|\boldsymbol{\theta}_\eta\|_2 \leq 2\|\boldsymbol{\theta}^{\sharp 1}_\eta\|_2 \leq \frac{2}{\kappa_{\mathrm{l}}}\|\Sigma^{\frac{1}{2}}\boldsymbol{\theta}^{\sharp 1}_\eta\|_2 \leq \frac{2}{\kappa_{\mathrm{l}}}\|\Sigma^{\frac{1}{2}}\boldsymbol{\theta}\|_2.
$$

C.2 Proof of Lemma 44

From the same argument of the proof of Lemma 43, we have (63). From (33), the first term of the right-hand side of (63) is evaluated as

$$\sum_{i=1}^{n} \lambda_o \sqrt{n} \hat{w}_i' h(r_{\boldsymbol{\beta}^*,i}) \mathbf{X}_i^\top \boldsymbol{\theta}_\eta \leq r_{a,1} r_1 + r_{a,2} r_2 + r_{a,\Sigma} r_\Sigma \\ \overset{(a)}{\leq} \left(r_{a,1} + \frac{c_{r_2}}{c_{r_1}\sqrt{s}} r_{a,2} + \frac{1}{c_{r_1}\sqrt{s}} r_{a,\Sigma} \right) \|\boldsymbol{\theta}_\eta\|_1. \tag{65}$$

where (a) follows from $r_1 = c_{r_1}\sqrt{s} r_\Sigma$ and $r_2 = c_{r_2} r_\Sigma$. From (63) and (65), we have

$$0 \leq \left(r_{a,1} + \frac{c_{r_2} r_{a,2} + r_{a,\Sigma}}{c_{r_1}\sqrt{s}} \right) \|\boldsymbol{\theta}_\eta\|_1 + \eta \lambda_s (\|\boldsymbol{\beta}^*\|_1 - \|\hat{\boldsymbol{\beta}}\|_1).$$

Furthermore, we see

$$\begin{aligned} 0 &\leq \left(r_{a,1} + \frac{c_{r_2} r_{a,2} + r_{a,\Sigma}}{c_{r_1}\sqrt{s}} \right) \|\boldsymbol{\theta}_\eta\|_1 + \eta \lambda_s (\|\boldsymbol{\beta}^*\|_1 - \|\hat{\boldsymbol{\beta}}\|_1) \\ &\leq \left(r_{a,1} + \frac{c_{r_2} r_{a,2} + r_{a,\Sigma}}{c_{r_1}\sqrt{s}} \right) (\|\boldsymbol{\theta}_{\eta,\mathcal{J}_{\boldsymbol{\beta}^*}}\|_1 + \|\boldsymbol{\theta}_{\eta,\mathcal{J}^c_{\boldsymbol{\beta}^*}}\|_1) + \eta \lambda_s (\|\boldsymbol{\beta}^*_{\mathcal{J}_{\boldsymbol{\beta}^*}} - \hat{\boldsymbol{\beta}}_{\mathcal{J}_{\boldsymbol{\beta}^*}}\|_1 - \|\hat{\boldsymbol{\beta}}_{\mathcal{J}^c_{\boldsymbol{\beta}^*}}\|_1) \\ &= \left(\lambda_s + \left(r_{a,1} + \frac{c_{r_2} r_{a,2} + r_{a,\Sigma}}{c_{r_1}\sqrt{s}} \right) \right) \|\boldsymbol{\theta}_{\eta,\mathcal{J}_{\boldsymbol{\beta}^*}}\|_1 + \left(-\lambda_s + \left(r_{a,1} + \frac{c_{r_2} r_{a,2} + r_{a,\Sigma}}{c_{r_1}\sqrt{s}} \right) \right) \|\boldsymbol{\theta}_{\eta,\mathcal{J}^c_{\boldsymbol{\beta}^*}}\|_1. \end{aligned}$$

Then, we have

$$\|\boldsymbol{\theta}_{\eta,\mathcal{J}^c_{\boldsymbol{\beta}^*}}\|_1 \leq \frac{\lambda_s + \left(r_{a,1} + \frac{c_{r_2} r_{a,2} + r_{a,\Sigma}}{c_{r_1}\sqrt{s}} \right)}{\lambda_s - \left(r_{a,1} + \frac{c_{r_2} r_{a,2} + r_{a,\Sigma}}{c_{r_1}\sqrt{s}} \right)} \|\boldsymbol{\theta}_{\eta,\mathcal{J}_{\boldsymbol{\beta}^*}}\|_1 \overset{(a)}{\leq} c_{\mathrm{RE}} \|\boldsymbol{\theta}_{\eta,\mathcal{J}_{\boldsymbol{\beta}^*}}\|_1, \tag{66}$$

where (a) follows from (35), and we have

$$\|\boldsymbol{\theta}_\eta\|_1 = \|\boldsymbol{\theta}_{\eta,\mathcal{J}_{\boldsymbol{\beta}^*}}\|_1 + \|\boldsymbol{\theta}_{\eta,\mathcal{J}^c_{\boldsymbol{\beta}^*}}\|_1 \leq (1 + c_{\mathrm{RE}}) \|\boldsymbol{\theta}_{\eta,\mathcal{J}_{\boldsymbol{\beta}^*}}\|_1 \leq (1 + c_{\mathrm{RE}}) \sqrt{s} \|\boldsymbol{\theta}_{\eta,\mathcal{J}_{\boldsymbol{\beta}^*}}\|_2.$$

From (66) and the restricted eigenvalue condition, we have

$$\|\boldsymbol{\theta}_\eta\|_1 \leq (1 + c_{\mathrm{RE}}) \sqrt{s} \|\boldsymbol{\theta}_{\eta,\mathcal{J}_{\boldsymbol{\beta}^*}}\|_2 \leq \frac{1 + c_{\mathrm{RE}}}{\mathfrak{r}} \sqrt{s} \|\Sigma^{\frac{1}{2}} \boldsymbol{\theta}_\eta\|_2.$$

C.3 Proof of Lemma 45

From the same argument of the proof of Lemma 43, we have (63). From (63), we have

$$\sum_{i=1}^{n} \lambda_o \sqrt{n} \hat{w}_i' \left(-h(r_{\boldsymbol{\beta}^* + \boldsymbol{\theta}_\eta, i}) + h(r_{\boldsymbol{\beta}^*,i}) \right) \mathbf{X}_i^\top \boldsymbol{\theta}_\eta \leq \sum_{i=1}^{n} \lambda_o \sqrt{n} \hat{w}_i' h(r_{\boldsymbol{\beta}^*,i}) \mathbf{X}_i^\top \boldsymbol{\theta}_\eta + \eta \lambda_s (\|\boldsymbol{\beta}^*\|_1 - \|\hat{\boldsymbol{\beta}}\|_1). \tag{67}$$

We evaluate each term of (67). From (34), the left-hand side of (67) is evaluated as

$$\sum_{i=1}^{n} \lambda_o \sqrt{n} \hat{w}_i' \left(-h(r_{\boldsymbol{\beta}^*+\boldsymbol{\theta}_\eta,i}) + h(r_{\boldsymbol{\beta}^*,i})\right) \mathbf{X}_i^\top \boldsymbol{\theta}_\eta \geq b\|\Sigma^{\frac{1}{2}}\boldsymbol{\theta}_\eta\|_2^2 - r_{b,1}r_1 - r_{b,2}r_2 - r_{b,\Sigma}r_\Sigma$$
$$\overset{(a)}{\geq} b\|\Sigma^{\frac{1}{2}}\boldsymbol{\theta}_\eta\|_2^2 - \left(c_{r_1}\sqrt{s}r_{b,1} - c_{r_2}r_{b,2} - r_{b,\Sigma}\right)\|\Sigma^{\frac{1}{2}}\boldsymbol{\theta}_\eta\|_2,$$

where (a) follows from $r_1 = c_{r_1}\sqrt{s}r_\Sigma$ and $r_2 = c_{r_2}r_\Sigma$. From (33), the first term of the right-hand side of (67) is evaluated as

$$\sum_{i=1}^{n} \lambda_o \sqrt{n} \hat{w}_i' h(r_{\boldsymbol{\beta}^*,i}) \mathbf{X}_i^\top \boldsymbol{\theta}_\eta \leq r_{a,1}r_1 + r_{a,2}r_2 + r_{a,\Sigma}r_\Sigma$$
$$\overset{(a)}{\leq} \left(c_{r_1}\sqrt{s}r_{a,1} + c_{r_2}r_{a,2} + r_{a,\Sigma}\right)\|\Sigma^{\frac{1}{2}}\boldsymbol{\theta}_\eta\|_2,$$

where (a) follows from $r_1 = c_{r_1}\sqrt{s}r_\Sigma$ and $r_2 = c_{r_2}r_\Sigma$. The second term of the right-hand side of (67) is evaluated as

$$\eta\lambda_s(\|\boldsymbol{\beta}^*\|_1 - \|\hat{\boldsymbol{\beta}}\|_1) \leq \lambda_s\|\boldsymbol{\theta}_\eta\|_1 \overset{(a)}{\leq} c_{r_1}\sqrt{s}\lambda_s\|\Sigma^{\frac{1}{2}}\boldsymbol{\theta}_\eta\|_2,$$

where (a) follows from $r_1 = c_{r_1}\sqrt{s}r_\Sigma$.

Combining the inequalities above, we have

$$b\|\Sigma^{\frac{1}{2}}\boldsymbol{\theta}_\eta\|_2^2 \leq \left(c_{r_1}\sqrt{s}(r_{a,1} + r_{b,1}) + c_{r_2}(r_{a,2} + r_{b,2}) + r_{a,\Sigma} + r_{b,\Sigma} + c_{r_1}\sqrt{s}\lambda_s\right)\|\Sigma^{\frac{1}{2}}\boldsymbol{\theta}_\eta\|_2,$$

and from $\|\Sigma^{\frac{1}{2}}\boldsymbol{\theta}_\eta\|_2 \geq 0$, we have

$$\|\Sigma^{\frac{1}{2}}\boldsymbol{\theta}_\eta\|_2 \leq \frac{1}{b}\left(c_{r_1}\sqrt{s}(r_{a,1} + r_{b,1}) + c_{r_2}(r_{a,2} + r_{b,2}) + r_{a,\Sigma} + r_{b,\Sigma} + c_{r_1}\sqrt{s}\lambda_s\right),$$

and the proof is complete.

Appendix D. Proof of Proposition 21

In Appendix D, we prove Proposition 21.

D.1 Step1

We derive a contradiction if $\|\boldsymbol{\theta}\|_1 > r_1$, $\|\boldsymbol{\theta}\|_2 > r_2$ and $\|\Sigma^{\frac{1}{2}}\boldsymbol{\theta}\|_2 > r_\Sigma$ hold. Assume that $\|\boldsymbol{\theta}\|_1 > r_1$, $\|\boldsymbol{\theta}\|_2 > r_2$ and $\|\Sigma^{\frac{1}{2}}\boldsymbol{\theta}\|_2 > r_\Sigma$. Then we can find $\eta_1, \eta_2, \eta_2' \in (0,1)$ such that $\|\boldsymbol{\theta}_{\eta_1}\|_1 = r_1$, $\|\boldsymbol{\theta}_{\eta_2}\|_2 = r_2$ and $\|\Sigma^{\frac{1}{2}}\boldsymbol{\theta}_{\eta_2'}\|_2 = r_\Sigma$ hold. Define $\eta_3 = \min\{\eta_1, \eta_2, \eta_2'\}$. We consider the case $\eta_3 = \eta_2'$ in Section D.1.1, the case $\eta_3 = \eta_2$ in Section D.1.2, and the case $\eta_3 = \eta_1$ in Section D.1.3.

D.1.1 Step 1(a)

Assume that $\eta_3 = \eta_2'$. We see that $\|\Sigma^{\frac{1}{2}}\boldsymbol{\theta}_{\eta_3}\|_2 = r_\Sigma$, $\|\boldsymbol{\theta}_{\eta_3}\|_1 \leq r_1$ and $\|\boldsymbol{\theta}_{\eta_3}\|_2 \leq r_2$ hold. Then, from Lemma 45, we have

$$\|\Sigma^{\frac{1}{2}}\boldsymbol{\theta}_{\eta_3}\|_2 \leq \frac{1}{b}\left(c_{r_1}\sqrt{s}(r_{a,1}+r_{b,1}) + c_{r_2}(r_{a,2}+r_{b,2}) + r_{a,\Sigma} + r_{b,\Sigma} + c_{r_1}\sqrt{s}\lambda_s\right).$$

The case $\eta_3 = \eta_2'$ is a contradiction from $\|\Sigma^{\frac{1}{2}}\boldsymbol{\theta}_{\eta_3}\|_2 = r_\Sigma$ and (36).

D.1.2 Step 1(b)

Assume that $\eta_3 = \eta_2$. We see that $\|\Sigma^{\frac{1}{2}}\boldsymbol{\theta}_{\eta_3}\|_2 \leq r_\Sigma$, $\|\boldsymbol{\theta}_{\eta_3}\|_1 \leq r_1$ and $\|\boldsymbol{\theta}_{\eta_3}\|_2 = r_2$ hold. Then, from Lemma 43, we have

$$\|\boldsymbol{\theta}_{\eta_3}\|_2 \leq \frac{3+c_{\mathrm{RE}}}{\kappa_{\mathrm{l}}}\|\Sigma^{\frac{1}{2}}\boldsymbol{\theta}_{\eta_3}\|_2 \leq \frac{3+c_{\mathrm{RE}}}{\kappa_{\mathrm{l}}}r_\Sigma.$$

The case $\eta_3 = \eta_2$ is a contradiction from $\|\boldsymbol{\theta}_{\eta_3}\|_2 = r_2$ and (36).

D.1.3 Step 1(c)

Assume that $\eta_3 = \eta_1$. We see that $\|\Sigma^{\frac{1}{2}}\boldsymbol{\theta}_{\eta_3}\|_2 \leq r_\Sigma$, $\|\boldsymbol{\theta}_{\eta_3}\|_1 = r_1$ and $\|\boldsymbol{\theta}_{\eta_3}\|_2 \leq r_2$ hold. Then, from Lemma 44, for $\eta = \eta_3$, we have

$$\|\boldsymbol{\theta}_{\eta_3}\|_1 \leq \frac{1+c_{\mathrm{RE}}}{\mathfrak{r}}\sqrt{s}\|\Sigma^{\frac{1}{2}}\boldsymbol{\theta}_{\eta_3}\|_2 \leq \frac{1+c_{\mathrm{RE}}}{\mathfrak{r}}\sqrt{s}r_\Sigma.$$

The case $\eta_3 = \eta_1$ is a contradiction from $\|\boldsymbol{\theta}_{\eta_3}\|_1 = r_1$ and (36).

D.2 Step 2

From the arguments in Section D.1, we have that $\|\Sigma^{\frac{1}{2}}\boldsymbol{\theta}\|_2 \leq r_\Sigma$ or $\|\boldsymbol{\theta}\|_1 \leq r_1$ or $\|\boldsymbol{\theta}\|_2 \leq r_2$ holds.

(a) In Section D.2.1, assume that $\|\Sigma^{\frac{1}{2}}\boldsymbol{\theta}\|_2 \leq r_\Sigma$ and $\|\boldsymbol{\theta}\|_1 > r_1$ and $\|\boldsymbol{\theta}\|_2 > r_2$ hold and then derive a contradiction.

(b) In Section D.2.2, assume that $\|\Sigma^{\frac{1}{2}}\boldsymbol{\theta}\|_2 > r_\Sigma$ and $\|\boldsymbol{\theta}\|_1 \leq r_1$ and $\|\boldsymbol{\theta}\|_2 > r_2$ hold and then derive a contradiction.

(c) In Section D.2.3, assume that $\|\Sigma^{\frac{1}{2}}\boldsymbol{\theta}\|_2 > r_\Sigma$ and $\|\boldsymbol{\theta}\|_1 > r_1$ and $\|\boldsymbol{\theta}\|_2 \leq r_2$ hold and then derive a contradiction.

(d) In Section D.2.4, assume that $\|\Sigma^{\frac{1}{2}}\boldsymbol{\theta}\|_2 > r_\Sigma$ and $\|\boldsymbol{\theta}\|_1 \leq r_1$ and $\|\boldsymbol{\theta}\|_2 \leq r_2$ hold and then derive a contradiction.

(e) In Section D.2.5, assume that $\|\Sigma^{\frac{1}{2}}\boldsymbol{\theta}\|_2 \leq r_\Sigma$ and $\|\boldsymbol{\theta}\|_1 > r_1$ and $\|\boldsymbol{\theta}\|_2 \leq r_2$ hold and then derive a contradiction.

(f) In Section D.2.6, assume that $\|\Sigma^{\frac{1}{2}}\boldsymbol{\theta}\|_2 \leq r_\Sigma$ and $\|\boldsymbol{\theta}\|_1 \leq r_1$ and $\|\boldsymbol{\theta}\|_2 > r_2$ hold and then derive a contradiction.

Finally, we have

$$\|\Sigma^{\frac{1}{2}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_2 \leq r_\Sigma,\ \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \leq r_2, \text{ and } \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \leq r_1,$$

and the proof is complete.

D.2.1 Step 2(a)

Assume that $\|\Sigma^{\frac{1}{2}}\boldsymbol{\theta}\|_2 \leq r_\Sigma$ and $\|\boldsymbol{\theta}\|_1 > r_1$ and $\|\boldsymbol{\theta}\|_2 > r_2$ hold, and then we can find $\eta_4, \eta_4' \in (0,1)$ such that $\|\boldsymbol{\theta}_{\eta_4}\|_1 = r_1$ and $\|\boldsymbol{\theta}_{\eta_4'}\|_2 = r_2$ hold. We note that $\|\Sigma^{\frac{1}{2}}\boldsymbol{\theta}_{\eta_4}\|_2 \leq r_\Sigma$ and $\|\Sigma^{\frac{1}{2}}\boldsymbol{\theta}_{\eta_4'}\|_2 \leq r_\Sigma$ also hold. Then, from the same arguments as in Sections D.1.2 and D.1.3, we have a contradiction.

D.2.2 Step 2(b)

Assume that $\|\Sigma^{\frac{1}{2}}\boldsymbol{\theta}\|_2 > r_\Sigma$ and $\|\boldsymbol{\theta}\|_1 \leq r_1$ and $\|\boldsymbol{\theta}\|_2 > r_2$ hold, and then we can find $\eta_5, \eta_5' \in (0,1)$ such that $\|\Sigma^{\frac{1}{2}}\boldsymbol{\theta}_{\eta_5}\|_2 = r_\Sigma$ and $\|\boldsymbol{\theta}_{\eta_5'}\|_2 = r_2$ hold. We note that $\|\boldsymbol{\theta}_{\eta_5}\|_1 \leq r_1$ and $\|\boldsymbol{\theta}_{\eta_5'}\|_1 \leq r_1$ also hold. Then, from the same arguments as in Sections D.1.1 and D.1.2, we have a contradiction.

D.2.3 Step 2(c)

Assume that $\|\Sigma^{\frac{1}{2}}\boldsymbol{\theta}\|_2 > r_\Sigma$ and $\|\boldsymbol{\theta}\|_1 > r_1$ and $\|\boldsymbol{\theta}\|_2 \leq r_2$ hold, and then we can find $\eta_6, \eta_6' \in (0,1)$ such that $\|\boldsymbol{\theta}_{\eta_6}\|_1 = r_1$ and $\|\Sigma^{\frac{1}{2}}\boldsymbol{\theta}_{\eta_6'}\|_2 = r_\Sigma$ hold. We note that $\|\boldsymbol{\theta}_{\eta_6}\|_2 \leq r_2$ and $\|\boldsymbol{\theta}_{\eta_6'}\|_2 \leq r_2$ also hold. Then, from the same arguments as in Sections D.1.1 and D.1.3, we have a contradiction.

D.2.4 Step 2(d)

Assume that $\|\Sigma^{\frac{1}{2}}\boldsymbol{\theta}\|_2 > r_\Sigma$ and $\|\boldsymbol{\theta}\|_1 \leq r_1$ and $\|\boldsymbol{\theta}\|_2 \leq r_2$ hold, and then we can find $\eta_7 \in (0,1)$ such that $\|\Sigma^{\frac{1}{2}}\boldsymbol{\theta}_{\eta_7}\|_2 = r_\Sigma$ holds. We note that $\|\boldsymbol{\theta}_{\eta_7}\|_1 \leq r_1$ and $\|\boldsymbol{\theta}_{\eta_7}\|_2 \leq r_2$ also hold. Then, from the same arguments as in Section D.1.1, we have a contradiction.

D.2.5 Step 2(e)

Assume that $\|\Sigma^{\frac{1}{2}}\boldsymbol{\theta}\|_2 \leq r_\Sigma$ and $\|\boldsymbol{\theta}\|_1 > r_1$ and $\|\boldsymbol{\theta}\|_2 \leq r_2$ hold, and then we can find $\eta_8 \in (0,1)$ such that $\|\boldsymbol{\theta}_{\eta_8}\|_1 = r_1$ holds. We note that $\|\Sigma^{\frac{1}{2}}\boldsymbol{\theta}_{\eta_8}\|_2 \leq r_\Sigma$ and $\|\boldsymbol{\theta}_{\eta_8}\|_2 \leq r_2$ also hold. Then, from the same arguments as in Section D.1.3, we have a contradiction.

D.2.6 Step 2(f)

Assume that $\|\Sigma^{\frac{1}{2}}\boldsymbol{\theta}\|_2 \leq r_\Sigma$ and $\|\boldsymbol{\theta}\|_1 \leq r_1$ and $\|\boldsymbol{\theta}\|_2 > r_2$ hold, and then we can find $\eta_9 \in (0,1)$ such that $\|\boldsymbol{\theta}_{\eta_9}\|_2 = r_2$ holds. We note that $\|\Sigma^{\frac{1}{2}}\boldsymbol{\theta}_{\eta_9}\|_2 \leq r_\Sigma$ and $\|\boldsymbol{\theta}_{\eta_9}\|_1 \leq r_1$ also hold. Then, from the same arguments as in Section D.1.2, we have a contradiction.

Appendix E. Proofs of Lemmas 27, 34, 36 and 38

In Section E, we prove Lemmas 27, 34, 36 and 38.

E.1 Proof of Lemma 27

We assume $|I_<| > 2\varepsilon n$, and then we derive a contradiction. From the constraint about w_i, we have $0 \leq w_i \leq \frac{1}{(1-\varepsilon)n}$ for any $i \in \{1, \cdots, n\}$ and we have

$$\begin{aligned}
\sum_{i=1}^n w_i &= \sum_{i \in I_<} w_i + \sum_{i \in I_\geq} w_i \\
&\leq |I_<| \times \frac{1}{2n} + (n - |I_<|) \times \frac{1}{(1-\varepsilon)\, n} \\
&= 2\varepsilon n \times \frac{1}{2n} + (|I_<| - 2\varepsilon n) \times \frac{1+\varepsilon}{2n} + (n - 2\varepsilon n) \times \frac{1}{(1-\varepsilon)\, n} + (2\varepsilon n - |I_<|) \times \frac{1}{(1-\varepsilon)\, n} \\
&= \varepsilon + (n - 2\varepsilon n) \times \frac{1}{(1-\varepsilon)\, n} + (|I_<| - 2\varepsilon n) \times \left(\frac{1}{2n} - \frac{1}{(1-\varepsilon)\, n} \right) \\
&< \varepsilon + \frac{n - 2\varepsilon n}{(1-\varepsilon)\, n} \\
&= \varepsilon + \frac{1 - 2\varepsilon}{1 - \varepsilon} \\
&\leq \frac{1 - \varepsilon - \varepsilon^2}{1 - \varepsilon} \\
&< 1.
\end{aligned}$$

This is a contradiction because $\sum_{i=1}^n w_i = 1$. Then, we have $|I_<| \leq 2\varepsilon n$.

E.2 Proofs of Lemmas 34, 36 and 38

In Section E.2, we prove Lemmas 34, 36 and 38.

E.2.1 Proof of Lemma 34

The following argument is essentially identical to a part of the proof of Proposition 9 of Bellec (2019). We note that

$$r_1 \mathbb{B}_1^d \cap r_\Sigma \mathbb{B}_\Sigma^d \subset 2r_1 \mathbb{B}_1^d \cap r_\Sigma \mathbb{B}_\Sigma^d = 2r_1 \left(1 \times \mathbb{B}_1^d \cap \frac{r_\Sigma}{2r_1} \mathbb{B}_\Sigma^d \right),$$

and we have

$$\mathbb{E} \sup_{\mathbf{v} \in r_1 \mathbb{B}_1^d \cap r_\Sigma \mathbb{B}_\Sigma^d} \langle \Sigma^{\frac{1}{2}} \mathbf{g}, \mathbf{v} \rangle \leq \mathbb{E} \sup_{\mathbf{v} \in 2r_1 \left(1 \times \mathbb{B}_1^d \cap \frac{r_\Sigma}{2r_1} \mathbb{B}_\Sigma^d \right)} \langle \Sigma^{\frac{1}{2}} \mathbf{g}, \mathbf{v} \rangle.$$

Let $\{\mathbf{u}_1, \cdots, \mathbf{u}_d\}$ be the columns of $\Sigma^{\frac{1}{2}\top}$. Then, we have

$$\begin{aligned}
\mathbb{E} \sup_{\mathbf{v} \in 2r_1 \left(1 \times \mathbb{B}_1^d \cap \frac{r_\Sigma}{2r_1} \mathbb{B}_\Sigma^d \right)} \langle \Sigma^{\frac{1}{2}} \mathbf{g}, \mathbf{v} \rangle &\leq \mathbb{E} \sup_{\mathbf{v} \in 2r_1 \left(\mathrm{conv}(\{\pm \mathbf{u}_1, \cdots, \pm \mathbf{u}_d\}) \cap \frac{r_\Sigma}{2r_1} \mathbb{B}_2^d \right)} \langle \mathbf{g}, \mathbf{v} \rangle \\
&\leq \mathbb{E} \sup_{\mathbf{v} \in 2r_1 \rho \left(\mathrm{conv}\left(\frac{1}{\rho} \{\pm \mathbf{u}_1, \cdots, \pm \mathbf{u}_d\} \right) \cap \frac{r_\Sigma}{2r_1} \mathbb{B}_2^d \right)} \langle \mathbf{g}, \mathbf{v} \rangle.
\end{aligned}$$

From Proposition 1 of Bellec (2019), we have

$$\begin{aligned}\mathbb{E}\sup_{\mathbf{v}\in 2r_1\rho\left(\mathrm{conv}\left(\frac{1}{\rho}\{\pm\mathbf{u}_1,\cdots,\pm\mathbf{u}_d\}\right)\cap\frac{r_\Sigma}{2r_1}\mathbb{B}_2^d\right)}\langle\mathbf{g},\mathbf{v}\rangle &\le 4\rho r_1\sqrt{\log\left(4ed\left(\frac{r_\Sigma}{2r_1}\right)^2\right)}\\ &\overset{(a)}{\le} 4\rho r_1\sqrt{\log\left(\frac{ed}{s}\right)}\\ &\overset{(b)}{\le} C\rho r_1\sqrt{\log(d/s)},\end{aligned}$$

where (a) follows from $r_\Sigma/r_1 \le 1/\sqrt{s}$, and (b) follows from $d/s \ge 3$.

E.2.2 Proof of Lemma 36

First, we prove (47). For any $\mathbf{v} \in r_1\mathbb{B}_1^d \cap r_2\mathbb{B}_2^d$, we see that

$$\frac{\mathbf{v}}{r_2} \in \left\{\mathbf{v}\in\mathbb{R}^d \mid \|\mathbf{v}\|_2 \le 1,\ \|\mathbf{v}\|_1 \le \frac{r_1}{r_2}\right\} = \frac{r_1}{r_2}\mathbb{B}_1^d \cap \mathbb{B}_2^d.$$

From Lemma 3.1 of Plan and Vershynin (2013), we have

$$\frac{r_1}{r_2}\mathbb{B}_1^d \cap \mathbb{B}_2^d \subset 2\mathrm{conv}\left(\left(\frac{r_1}{r_2}\right)^2\mathbb{B}_0^d \cap \mathbb{B}_2^d\right),$$

and

$$\frac{\mathbf{v}}{r_2} \in 2\mathrm{conv}\left(\left(\frac{r_1}{r_2}\right)^2\mathbb{B}_0^d \cap \mathbb{B}_2^d\right). \tag{68}$$

From the definition of convex hull, this means

$$\frac{\mathbf{v}}{r_2} = 2\sum_{i=1}^a \lambda_i\mathbf{a}_i.$$

for some $\{\lambda_i\}_{i=1}^a$ and $\{\mathbf{a}_i\}_{i=1}^a$ such that $\sum_{i=1}^a\lambda_i = 1$, $\lambda_i \ge 0$ and $\mathbf{a}_i \in \left(\frac{r_1}{r_2}\right)^2\mathbb{B}_0^d \cap \mathbb{B}_2^d$ for any $i \in \{1,\cdots,a\}$. We note that

$$\|r_2\mathbf{a}_i\|_0 \le \left(\frac{r_1}{r_2}\right)^2, \quad \|r_2\mathbf{a}_i\|_2 \le r_2. \tag{69}$$

From (68) and (69), we see

$$\mathbf{v} \in 2\mathrm{conv}\left(\left(\frac{r_1}{r_2}\right)^2\mathbb{B}_0^d \cap r_2\mathbb{B}_2^d\right).$$

Then the proof of (47) is complete. Identifying the vectors and the matrices, the proof of (46) is also complete.

E.2.3 Proof of Lemma 38

We note that

$$\int_0^\infty \frac{x}{e^x}dx = 1,\ \int_0^\infty \frac{\sqrt{x}}{e^x}dx \leq \int_0^1 \frac{1}{e^x}dx + \int_0^\infty \frac{x}{e^x}dx \leq \left[\frac{-1}{e^x}\right]_0^\infty + 1 = 2. \tag{70}$$

First, we evaluate $\gamma_1(s^2\mathbb{B}_0^{d\times d}\cap r_2^2\mathbb{B}_{\mathrm{F}}^{d\times d}, \|\cdot\|_{\mathrm{F}})$. From a standard entropy bound from chaining theory Lemma D.17 of Oymak (2018), we have

$$\gamma_1(s^2\mathbb{B}_0^{d\times d}\cap r_2^2\mathbb{B}_{\mathrm{F}}^{d\times d}, \|\cdot\|_{\mathrm{F}}) \leq C\int_0^{r_2^2} \log N(s^2\mathbb{B}_0^{d\times d}\cap r_2^2\mathbb{B}_{\mathrm{F}}^{d\times d}, \epsilon)d\epsilon,$$

where $N(s^2\mathbb{B}_0^{d\times d}\cap r_2^2\mathbb{B}_{\mathrm{F}}^{d\times d}, \epsilon)$ is the covering number of $s^2\mathbb{B}_0^{d\times d}\cap r_2^2\mathbb{B}_{\mathrm{F}}^{d\times d}$, that is the minimal cardinality of an ϵ-net of $s^2\mathbb{B}_0^{d\times d}\cap r_2^2\mathbb{B}_{\mathrm{F}}^{d\times d}$. From Lemma 36 and $d/s \geq 3$, we have

$$\begin{aligned}
\int_0^{r_2^2} \log N(s^2\mathbb{B}_0^{d\times d}\cap r_2^2\mathbb{B}_{\mathrm{F}}^{d\times d}, \epsilon)d\epsilon &\leq C\int_0^{r_2^2} s^2\log r_2^2\frac{9d^2}{\epsilon s^2}d\epsilon \\
&= Cs^2\int_0^{r_2^2}\left(\log(9d^2/s^2) + \log(r_2^2/\epsilon)\right)d\epsilon \\
&= Cs^2\left(r_2^2\log(9d^2/s^2) + r_2^2\int_1^\infty \frac{x}{e^x}dx\right) \\
&\leq Cs^2\left(r_2^2\log(9d^2/s^2) + r_2^2\int_0^\infty \frac{x}{e^x}dx\right) \\
&\overset{(a)}{=} Cs^2\left(r_2^2\log(9d^2/s^2) + r_2^2\right) \\
&\overset{(b)}{\leq} Cs^2r_2^2\log(d/s),
\end{aligned}$$

where (a) follows from (70), and (b) follows from $3 \leq d/s$. Consequently, we have

$$\gamma_1(s^2\mathbb{B}_0^{d\times d}\cap r_2^2\mathbb{B}_{\mathrm{F}}^{d\times d}, \|\cdot\|_{\mathrm{F}}) \leq Cs^2r_2^2\log(d/s).$$

Second, we evaluate $\gamma_2(s^2\mathbb{B}_0^{d\times d}\cap r_2^2\mathbb{B}_{\mathrm{F}}^{d\times d}, \|\cdot\|_{\mathrm{F}})$. From similar argument of the case $\gamma_1(s^2\mathbb{B}_0^{d\times d}\cap r_2^2\mathbb{B}_{\mathrm{F}}^{d\times d}, \|\cdot\|_{\mathrm{F}})$, we have

$$\begin{aligned}
\int_0^{r_2^2} \sqrt{\log N(s^2\mathbb{B}_0^{d\times d}\cap r_2^2\mathbb{B}_{\mathrm{F}}^{d\times d}, \epsilon)}d\epsilon &\leq C\int_0^{r_2^2} s\sqrt{\log r_2^2\frac{9d^2}{\epsilon s^2}}d\epsilon \\
&\overset{(a)}{\leq} C\left(s\int_0^{r_2^2}\sqrt{\log(d/s)} + \int_1^{r_2^2}\sqrt{\log\frac{r_2^2}{\epsilon}}d\epsilon\right) \\
&\leq C\left(sr_2^2\sqrt{\log(d/s)} + \int_1^{r_2^2}\sqrt{\log\frac{r_2^2}{\epsilon}}d\epsilon\right) \\
&\leq C\left(sr_2^2\sqrt{\log(d/s)} + \int_0^{r_2^2}\sqrt{\log\frac{r_2^2}{\epsilon}}d\epsilon\right) \\
&= C\left(sr_2^2\sqrt{\log(d/s)} + \int_0^\infty \frac{\sqrt{x}}{e^x}dx\right) \\
&\overset{(b)}{\leq} Csr_2^2\sqrt{\log(d/s)},
\end{aligned}$$

where (a) follows from triangular inequality, and (b) follows from (70) and $d/s \geq 3$.

Appendix F. Proofs of Corollary 28 and Propositions 23, 24, 25, 29 and 30

Define

$$\mathfrak{M}^{\ell_1,\ell_2}_{r_1,r_2,d,\mathbf{v}} = \{M \in \mathcal{S}(d) : M = \mathbf{v}\mathbf{v}^\top, \mathbf{v} \in \mathbb{R}^d,\ \|\mathbf{v}\|_1 \leq r_1, \|\mathbf{v}\|_2 \leq r_2\}.$$

F.1 Proof of Corollary 28

First, we prove (39). We see that, for any $M \in \mathfrak{M}^{\ell_1,\mathrm{Tr}}_{r_1,r_2,d}$,

$$\begin{aligned}\left|\frac{1}{n}\sum_{i=1}^n \langle \mathbf{x}_i\mathbf{x}_i^\top, M\rangle\right| &\leq \left|\frac{1}{n}\sum_{i=1}^n \langle \mathbf{x}_i\mathbf{x}_i^\top - \Sigma, M\rangle\right| + |\langle \Sigma, M\rangle| \\ &\leq \left|\frac{1}{n}\sum_{i=1}^n \langle \mathbf{x}_i\mathbf{x}_i^\top - \Sigma, M\rangle\right| + \Sigma^2_{\max} r_2^2, \end{aligned} \tag{71}$$

where we use Hölder's inequality. From (50) and (71), with probability at least $1-\delta$, we have

$$\max_{M \in \mathfrak{M}^{\ell_1,\mathrm{Tr}}_{r_1,r_2,d}} \left|\frac{1}{n}\sum_{i=1}^n \langle \mathbf{x}_i\mathbf{x}_i^\top, M\rangle\right| \leq C(L\kappa_{\mathrm{u}})^2 \left(sr_{d,s} + r_\delta\right) r_2^2 + \Sigma^2_{\max} r_2^2.$$

Next, we prove (40). Remember (51), we have

$$\begin{aligned}\max_{M \in \mathfrak{M}^{\ell_1,\mathrm{Tr}}_{r_1,r_2,d}} \sum_{i=1}^n \hat{w}_i \langle \mathbf{X}_i\mathbf{X}_i^\top, M\rangle &\overset{(a)}{\leq} \max_{M \in \mathfrak{M}^{\ell_1,\mathrm{Tr}}_{r_1,r_2,d}} \sum_{i=1}^n w_i^\circ \langle \mathbf{X}_i\mathbf{X}_i^\top, M\rangle \\ &= \frac{1}{1-o/n} \max_{M \in \mathfrak{M}^{\ell_1,\mathrm{Tr}}_{r_1,r_2,d}} \sum_{i=1}^n \frac{1}{n}\langle \mathbf{x}_i\mathbf{x}_i^\top, M\rangle, \end{aligned} \tag{72}$$

where (a) follows from the optimality of $\{\hat{w}_i\}_{i=1}^n$ and $\{w_i^\circ\}_{i=1}^n \in \Delta^{n-1}(\varepsilon)$. From (39), (72) and $1/(1-o/n) \leq 2$, with probability at least $1-\delta$, we have

$$\max_{M \in \mathfrak{M}^{\ell_1,\mathrm{Tr}}_{r_1,r_2,d}} \sum_{i=1}^n \hat{w}_i \langle \mathbf{X}_i\mathbf{X}_i^\top, M\rangle \leq C\left((L\kappa_{\mathrm{u}})^2 \left(sr_{d,s} + r_\delta\right) + \Sigma^2_{\max}\right) r_2^2,$$

and from $L \geq 1$, the proof is complete.

F.2 Proof of Proposition 23

We have

$$\begin{aligned}\max_{\mathbf{v}\in r_1\mathbb{B}_1^d\cap r_2\mathbb{B}_2^d}\left|\sum_{i\in\mathcal{O}}\hat{w}_i' u_i\mathbf{X}_i^\top\mathbf{v}\right| &\overset{(a)}{\leq} \max_{\mathbf{v}\in 2\mathrm{conv}\left(\left(\frac{r_1}{r_2}\right)^2\mathbb{B}_0^d\cap r_2\mathbb{B}_2^d\right)}\left|\sum_{i\in\mathcal{O}}\hat{w}_i' u_i\mathbf{X}_i^\top\mathbf{v}\right| \\ &\leq \max_{\mathbf{v}\in \mathrm{conv}\left(\left(\frac{r_1}{r_2}\right)^2\mathbb{B}_0^d\cap r_2\mathbb{B}_2^d\right)} 2\left|\sum_{i\in\mathcal{O}}\hat{w}_i' u_i\mathbf{X}_i^\top\mathbf{v}\right| \\ &\overset{(b)}{\leq} 2\max_{\mathbf{v}\in \left(\frac{r_1}{r_2}\right)^2\mathbb{B}_0^d\cap r_2\mathbb{B}_2^d}\left|\sum_{i\in\mathcal{O}}\hat{w}_i' u_i\mathbf{X}_i^\top\mathbf{v}\right|,\end{aligned}$$

where (a) follows from Lemma 36, and (b) follows from Lemma D.8 of Oymak (2018). We note that, for any $\mathbf{v}\in\mathbb{R}^d$,

$$\left|\sum_{i\in\mathcal{O}}\hat{w}_i' u_i\mathbf{X}_i^\top\mathbf{v}\right|^2 \overset{(a)}{\leq} 4\frac{o}{n}\sum_{i\in\mathcal{O}}\hat{w}_i'|\mathbf{X}_i^\top\mathbf{v}|^2 \overset{(b)}{\leq} 8\frac{o}{n}\sum_{i\in\mathcal{O}}\hat{w}_i|\mathbf{X}_i^\top\mathbf{v}|^2, \tag{73}$$

where (a) follows from Hölder's inequality and $\sum_{i\in\mathcal{O}}u_i^2\leq 4o$, and (b) follows from the fact that $\hat{w}_i'\leq 2\hat{w}_i$ for any $i\in(1,\cdots,n)$. We focus on $\sum_{i\in\mathcal{O}}\hat{w}_i|\mathbf{X}_i^\top\mathbf{v}|^2$.

First, we have

$$
\begin{aligned}
&\max_{\mathbf{v}\in\left(\frac{r_1}{r_2}\right)^2\mathbb{B}_0^d\cap r_2\mathbb{B}_2^d}\sum_{i\in\mathcal{O}}\hat{w}_i|\mathbf{X}_i^\top\mathbf{v}|^2\\
&=\max_{\mathbf{v}\in\left(\frac{r_1}{r_2}\right)^2\mathbb{B}_0^d\cap r_2\mathbb{B}_2^d}\sum_{i\in\mathcal{O}}\hat{w}_i\langle\mathbf{X}_i\mathbf{X}_i^\top,\mathbf{v}\mathbf{v}^\top\rangle\\
&\le\max_{\mathbf{v}\in\left(\frac{r_1}{r_2}\right)^2\mathbb{B}_0^d\cap r_2\mathbb{B}_2^d}\sum_{i=1}^n\hat{w}_i\langle\mathbf{X}_i\mathbf{X}_i^\top,\mathbf{v}\mathbf{v}^\top\rangle-\max_{\mathbf{v}\in\left(\frac{r_1}{r_2}\right)^2\mathbb{B}_0^d\cap r_2\mathbb{B}_2^d}\sum_{i\in\mathcal{I}}\hat{w}_i\langle\mathbf{X}_i\mathbf{X}_i^\top,\mathbf{v}\mathbf{v}^\top\rangle\\
&=\max_{\mathbf{v}\in\left(\frac{r_1}{r_2}\right)^2\mathbb{B}_0^d\cap r_2\mathbb{B}_2^d}\sum_{i=1}^n\hat{w}_i\langle\mathbf{X}_i\mathbf{X}_i^\top,\mathbf{v}\mathbf{v}^\top\rangle-\max_{\mathbf{v}\in\left(\frac{r_1}{r_2}\right)^2\mathbb{B}_0^d\cap r_2\mathbb{B}_2^d}\sum_{i\in\mathcal{I}}\hat{w}_i\langle\mathbf{x}_i\mathbf{x}_i^\top,\mathbf{v}\mathbf{v}^\top\rangle\\
&=\max_{\mathbf{v}\in\left(\frac{r_1}{r_2}\right)^2\mathbb{B}_0^d\cap r_2\mathbb{B}_2^d}\sum_{i=1}^n\hat{w}_i\langle\mathbf{X}_i\mathbf{X}_i^\top-\Sigma,\mathbf{v}\mathbf{v}^\top\rangle-\max_{\mathbf{v}\in\left(\frac{r_1}{r_2}\right)^2\mathbb{B}_0^d\cap r_2\mathbb{B}_2^d}\sum_{i\in\mathcal{I}}\hat{w}_i\langle\mathbf{x}_i\mathbf{x}_i^\top-\Sigma,\mathbf{v}\mathbf{v}^\top\rangle\\
&\quad+\max_{\mathbf{v}\in\left(\frac{r_1}{r_2}\right)^2\mathbb{B}_0^d\cap r_2\mathbb{B}_2^d}\sum_{i\in\mathcal{O}}\hat{w}_i\langle\Sigma,\mathbf{v}\mathbf{v}^\top\rangle\\
&\le\max_{\mathbf{v}\in\mathrm{conv}\left(\left(\frac{r_1}{r_2}\right)^2\mathbb{B}_0^d\cap r_2\mathbb{B}_2^d\right)}\sum_{i=1}^n\hat{w}_i\langle\mathbf{X}_i\mathbf{X}_i^\top-\Sigma,\mathbf{v}\mathbf{v}^\top\rangle-\max_{\mathbf{v}\in\mathrm{conv}\left(\left(\frac{r_1}{r_2}\right)^2\mathbb{B}_0^d\cap r_2\mathbb{B}_2^d\right)}\sum_{i\in\mathcal{I}}\hat{w}_i\langle\mathbf{x}_i\mathbf{x}_i^\top-\Sigma,\mathbf{v}\mathbf{v}^\top\rangle\\
&\quad+\max_{\mathbf{v}\in\left(\frac{r_1}{r_2}\right)^2\mathbb{B}_0^d\cap r_2\mathbb{B}_2^d}\sum_{i\in\mathcal{O}}\hat{w}_i\langle\Sigma,\mathbf{v}\mathbf{v}^\top\rangle\\
&\overset{(a)}{\le}\max_{\mathbf{v}\in r_1\mathbb{B}^d\cap r_2\mathbb{B}_2^d}\sum_{i=1}^n\hat{w}_i\langle\mathbf{X}_i\mathbf{X}_i^\top-\Sigma,\mathbf{v}\mathbf{v}^\top\rangle-\max_{\mathbf{v}\in r_1\mathbb{B}^d\cap r_2\mathbb{B}_2^d}\sum_{i\in\mathcal{I}}\hat{w}_i\langle\mathbf{x}_i\mathbf{x}_i^\top-\Sigma,\mathbf{v}\mathbf{v}^\top\rangle+\max_{\mathbf{v}\in s\mathbb{B}_0^d\cap r_2\mathbb{B}_2^d}\sum_{i\in\mathcal{O}}\hat{w}_i\langle\Sigma,\mathbf{v}\mathbf{v}^\top\rangle\\
&\overset{(b)}{\le}\max_{M\in\mathfrak{M}_{r_1,r_2,d}^{\ell_1,\mathrm{Tr}}}\sum_{i=1}^n\hat{w}_i\langle\mathbf{X}_i\mathbf{X}_i^\top-\Sigma,M\rangle-\max_{M\in\mathfrak{M}_{r_1,r_2,d}^{\ell_1,\mathrm{Tr}}}\sum_{i\in\mathcal{I}}\hat{w}_i\langle\mathbf{x}_i\mathbf{x}_i^\top-\Sigma,M\rangle+\max_{\mathbf{v}\in s\mathbb{B}_0^d\cap r_2\mathbb{B}_2^d}\sum_{i\in\mathcal{O}}\hat{w}_i\langle\Sigma,\mathbf{v}\mathbf{v}^\top\rangle,
\end{aligned}
\tag{74}
$$

where (a) follows from Lemma 36 and $r_1/r_2\le\sqrt{s}$, and (b) follows from $\mathfrak{M}_{r_1,r_2,d,\mathbf{v}}^{\ell_1,\ell_2}\subset\mathfrak{M}_{r_1,r_2,d}^{\ell_1,\mathrm{Tr}}$ because $\mathrm{Tr}(M)=\mathrm{Tr}(\mathbf{v}\mathbf{v}^\top)=\|\mathbf{v}\|_2^2$ for $M\in\mathfrak{M}_{r_1,r_2,d,\mathbf{v}}^{\ell_1,\ell_2}$. We note that, from $1/(1-\varepsilon)\le 2$ and matrix Hölder's inequality,

$$
\max_{\mathbf{v}\in s\mathbb{B}_0^d\cap r_2\mathbb{B}_2^d}\sum_{i\in\mathcal{O}}\hat{w}_i\langle\Sigma,\mathbf{v}\mathbf{v}^\top\rangle\overset{(a)}{\le}2\frac{o}{n}\max_{\mathbf{v}\in s\mathbb{B}_0^d\cap r_2\mathbb{B}_2^d}\sum_{i\in\mathcal{O}}\langle\Sigma,\mathbf{v}\mathbf{v}^\top\rangle\overset{(b)}{\le}2\kappa_{\mathrm{u}}^2r_o'r_2^2.
\tag{75}
$$

where (a) follows from $\hat{w}_i'\le 1/(1-\varepsilon)\le 2$, and (b) follows from $o/n\le 1/(5e)$ and the definition of κ_{u}. From (74) and (75), we have

$$
\begin{aligned}
\sum_{i\in\mathcal{O}}\hat{w}_i|\mathbf{X}_i^\top\mathbf{v}|^2&\le\max_{M\in\mathfrak{M}_{r_1,r_2,d}^{\ell_1,\mathrm{Tr}}}\sum_{i=1}^n\hat{w}_i\langle\mathbf{X}_i\mathbf{X}_i^\top-\Sigma,M\rangle+\max_{M\in\mathfrak{M}_{r_1,r_2,d}^{\ell_1,\mathrm{Tr}}}\left|\sum_{i\in\mathcal{I}}\hat{w}_i\langle\mathbf{x}_i\mathbf{x}_i^\top-\Sigma,M\rangle\right|+2\kappa_{\mathrm{u}}^2r_o'r_2^2\\
&\le\tau_{\mathrm{cut}}+\max_{M\in\mathfrak{M}_{r_1,r_2,d}^{\ell_1,\mathrm{Tr}}}\left|\sum_{i\in\mathcal{I}}\hat{w}_i\langle\mathbf{x}_i\mathbf{x}_i^\top-\Sigma,M\rangle\right|+2\kappa_{\mathrm{u}}^2r_o'r_2^2,
\end{aligned}
\tag{76}
$$

where the last line follows from the assumption of success of Algorithm 2.

Next, we consider the second term of right-hand side of (76). We have $\frac{1-o/n}{1-\varepsilon} \leq 2$ because $\varepsilon = c_\varepsilon \times \frac{o}{n}$ with $1 \leq c_\varepsilon < 2$ and $o/n \leq 1/(5e)(\leq 1/3)$, and we have

$$\begin{aligned}
\max_{M \in \mathfrak{M}^{\ell_1,\mathrm{Tr}}_{r_1,r_2,d}} \left| \sum_{i \in \mathcal{I}} \hat{w}_i \langle \mathbf{x}_i \mathbf{x}_i^\top - \Sigma, M \rangle \right| &= \sum_{j \in \mathcal{I}} \hat{w}_j \max_{M \in \mathfrak{M}^{\ell_1,\mathrm{Tr}}_{r_1,r_2,d}} \left| \sum_{i \in \mathcal{I}} \frac{\hat{w}_i}{\sum_{j \in \mathcal{I}} \hat{w}_j} \langle \mathbf{x}_i \mathbf{x}_i^\top - \Sigma, M \rangle \right| \\
&\leq \frac{1 - o/n}{1 - \varepsilon} \max_{M \in \mathfrak{M}^{\ell_1,\mathrm{Tr}}_{r_1,r_2,d}} \left| \sum_{i \in \mathcal{I}} \frac{\hat{w}_i}{\sum_{j \in \mathcal{I}} \hat{w}_j} \langle \mathbf{x}_i \mathbf{x}_i^\top - \Sigma, M \rangle \right| \\
&\leq 2 \max_{M \in \mathfrak{M}^{\ell_1,\mathrm{Tr}}_{r_1,r_2,d}} \left| \sum_{i \in \mathcal{I}} \frac{\hat{w}_i}{\sum_{j \in \mathcal{I}} \hat{w}_j} \langle \mathbf{x}_i \mathbf{x}_i^\top - \Sigma, M \rangle \right|.
\end{aligned}$$

Define the following three sets:

$$\begin{aligned}
\Delta^{\mathcal{I}}(\varepsilon + o/n) &= \left\{ (w_1, \ldots, w_n) \mid 0 \leq w_i \leq \frac{1}{n\{1 - (\varepsilon + o/n)\}}, \sum_{i \in \mathcal{I}} w_i = \sum_{i=1}^n w_i = 1, \right\}, \\
\Delta^{\mathcal{I}}(3o/n) &= \left\{ (w_1, \ldots, w_n) \mid 0 \leq w_i \leq \frac{1}{n(1 - 3o/n)}, \sum_{i \in \mathcal{I}} w_i = \sum_{i=1}^n w_i = 1 \right\}, \\
\Delta^{n-1}(3o/n) &= \left\{ (w_1, \ldots, w_n) \mid 0 \leq w_i \leq \frac{1}{n(1 - 3o/n)}, \sum_{i=1}^n w_i = 1 \right\}.
\end{aligned}$$

From $\sum_{j \in \mathcal{I}} \hat{w}_j \geq 1 - \frac{o}{n(1-\varepsilon)} = \frac{1-\varepsilon-o/n}{1-\varepsilon}$, for any $i \in \mathcal{I}$, we have $0 \leq \frac{\hat{w}_i}{\sum_{j \in \mathcal{I}} \hat{w}_j} \leq \frac{1}{n(1-(\varepsilon+o/n))}$, and from $\varepsilon = c_\varepsilon \times o/n$ with $1 \leq c_\varepsilon < 2$, we have $\Delta^{\mathcal{I}}(\varepsilon + o/n) \subset \Delta^{\mathcal{I}}(3o/n) \subset \Delta^{n-1}(3o/n)$. Therefore, we have

$$\begin{aligned}
\max_{M \in \mathfrak{M}^{\ell_1,\mathrm{Tr}}_{r_1,r_2,d}} \left| \sum_{i \in \mathcal{I}} \frac{\hat{w}_i}{\sum_{j \in \mathcal{I}} \hat{w}_j} \langle \mathbf{x}_i \mathbf{x}_i^\top - \Sigma, M \rangle \right| &\leq \max_{\mathbf{w} \in \Delta^{\mathcal{I}}(\varepsilon + o/n)} \max_{M \in \mathfrak{M}^{\ell_1,\mathrm{Tr}}_{r_1,r_2,d}} \left| \sum_{i \in \mathcal{I}} w_i \langle \mathbf{x}_i \mathbf{x}_i^\top - \Sigma, M \rangle \right| \\
&\leq \max_{\mathbf{w} \in \Delta^{n-1}(3o/n)} \max_{M \in \mathfrak{M}^{\ell_1,\mathrm{Tr}}_{r_1,r_2,d}} \left| \sum_{i=1}^n w_i \langle \mathbf{x}_i \mathbf{x}_i^\top - \Sigma, M \rangle \right|.
\end{aligned}$$

From Lemma 1 of Dalalyan and Minasyan (2022) and $1/(1-3o/n) \le 1/(1-3/(5e)) \le 2$, we have

$$\begin{aligned}
&\max_{\mathbf{w}\in\Delta^{n-1}(3o/n)} \max_{M\in\mathfrak{M}^{\ell_1,\mathrm{Tr}}_{r_1,r_2,d}} \left|\sum_{i=1}^n w_i \langle \mathbf{x}_i\mathbf{x}_i^\top - \Sigma, M\rangle\right| \\
&\le \max_{|\mathcal{J}|=n-3o} \max_{M\in\mathfrak{M}^{\ell_1,\mathrm{Tr}}_{r_1,r_2,d}} \left|\sum_{i\in\mathcal{J}} \frac{1}{n(1-3o/n)} \langle \mathbf{x}_i\mathbf{x}_i^\top - \Sigma, M\rangle\right| \\
&\le 2 \max_{|\mathcal{J}|=n-3o} \max_{M\in\mathfrak{M}^{\ell_1,\mathrm{Tr}}_{r_1,r_2,d}} \left|\sum_{i\in\mathcal{J}} \frac{1}{n} \langle \mathbf{x}_i\mathbf{x}_i^\top - \Sigma, M\rangle\right| \\
&\le 2\left(\max_{M\in\mathfrak{M}^{\ell_1,\mathrm{Tr}}_{r_1,r_2,d}} \left|\sum_{i=1}^n \frac{\langle \mathbf{x}_i\mathbf{x}_i^\top - \Sigma, M\rangle}{n}\right| + \max_{M\in\mathfrak{M}^{\ell_1,\mathrm{Tr}}_{r_1,r_2,d}} \left|\sum_{i\in\mathcal{J}} \frac{\langle \mathbf{x}_i\mathbf{x}_i^\top - \Sigma, M\rangle}{n}\right|\right).
\end{aligned}$$

Consequently, from almost the same calculation for (54), we have

$$\max_{M\in\mathfrak{M}^{\ell_1,\mathrm{Tr}}_{r_1,r_2,d}} \left|\sum_{i\in\mathcal{I}} \hat{w}_i \langle \mathbf{x}_i\mathbf{x}_i^\top - \Sigma, M\rangle\right| \le CL\kappa_{\mathrm{u}}^2 \left(sr_{d,s} + r_\delta + r'_o\right) r_2^2. \tag{77}$$

Lastly, from (73), (76) and (77), we have

$$\begin{aligned}
\left|\sum_{i\in\mathcal{O}} \hat{w}_i u_i \mathbf{X}_i^\top \mathbf{v}\right| &\le C\sqrt{\frac{o}{n}} \sqrt{\tau_{\mathrm{cut}} + (L\kappa)_{\mathrm{u}}^2 \left(sr_{d,s} + r_\delta + r'_o\right) r_2^2 + \kappa_{\mathrm{u}}^2 r'_o r_2^2} \\
&\overset{(a)}{\le} CL\sqrt{1+c_{\mathrm{cut}}} \sqrt{\frac{o}{n}} \sqrt{\kappa_{\mathrm{u}}^2 sr_{d,s} r_2^2 + \kappa_{\mathrm{u}}^2 r_\delta r_2^2 + \kappa_{\mathrm{u}}^2 r'_o r_2^2} \\
&\overset{(b)}{\le} CL\sqrt{1+c_{\mathrm{cut}}}\kappa_{\mathrm{u}} \left(\sqrt{\frac{o}{n}}\left(\sqrt{sr_{d,s}} + \sqrt{r_\delta}\right) + r_o\right) r_2,
\end{aligned}$$

where (a) follows from the definition of τ_{cut}, and (b) follows from triangular inequality, and the proof is complete.

F.3 Proof of Proposition 24

From $\log \frac{n}{m} \ge 1$ and arguments similar to the proof of Proposition 23, we have

$$\begin{aligned}
\sum_{i\in I_m} u_i \mathbf{x}_i^\top \mathbf{v} &\le CL\sqrt{\frac{m}{n}} \sqrt{\kappa_{\mathrm{u}}^2 \left(sr_{d,s} r_2^2 + r_\delta r_2^2 + \frac{m}{n}\log\frac{n}{m} r_2^2\right) + \kappa_{\mathrm{u}}^2 \frac{m}{n} r_2^2} \\
&\overset{(a)}{\le} CL\sqrt{\frac{(1+2c_\varepsilon)o}{n}}\kappa_{\mathrm{u}} \sqrt{sr_{d,s} r_2^2 + r_\delta r_2^2 + \frac{(1+2c_\varepsilon)o}{n}\log\frac{n}{(1+2c_\varepsilon)o} r_2^2} \\
&\overset{(b)}{\le} CL\sqrt{1+2c_\varepsilon}\kappa_{\mathrm{u}} \sqrt{\frac{o}{n}} \sqrt{sr_{d,s} r_2^2 + r_\delta r_2^2 + \frac{o}{n}\log\frac{n}{o} r_2^2}.
\end{aligned}$$

where (a) follows from the fact that $0 < x < 1/e$, $x\log(1/x)$ is increasing, and (b) follows from $\log \frac{1}{1+2c_\varepsilon} \le 1 \le \log\frac{n}{o}$. From triangular inequality, the proof is complete.

F.4 Proof of Proposition 25

First, we introduce Lemma 46, which is used in the proof of Proposition 25.

Lemma 46 *Suppose that (i) of Assumption 2 or Assumption 3 holds. Define $\{a_i\}_{i=1}^n$ as a sequence of i.i.d. Rademacher random variables which are independent of $\{\mathbf{x}_i\}_{i=1}^n$. Then, we have*

$$\mathbb{E} \sup_{\mathbf{v} \in r_1 \mathbb{B}_1^d \cap r_\Sigma \mathbb{B}_\Sigma^d} \left| \frac{1}{n} \sum_{i=1}^n a_i \mathbf{x}_i^\top \mathbf{v} \right| \leq L \rho r_{d,s} r_1.$$

Proof Define $\mathbb{E}^*$ as the conditional expectation of $\{a_i\}_{i=1}^n$ given $\{\mathbf{x}_i\}_{i=1}^n$. From Exercise 2.2.2 of Talagrand (2014), for any $\mathbf{v}_0 \in r_1 \mathbb{B}_1^d \cap r_\Sigma \mathbb{B}_\Sigma^d$, we have

$$\mathbb{E}^* \sup_{\mathbf{v} \in r_1 \mathbb{B}_1^d \cap r_\Sigma \mathbb{B}_\Sigma^d} \left| \frac{1}{n} \sum_{i=1}^n a_i \mathbf{x}_i^\top \mathbf{v} \right| \leq 2\mathbb{E}^* \sup_{\mathbf{v} \in r_1 \mathbb{B}_1^d \cap r_\Sigma \mathbb{B}_\Sigma^d} \frac{1}{n} \sum_{i=1}^n a_i \mathbf{x}_i^\top \mathbf{v} + \mathbb{E} \left| \frac{1}{n} \sum_{i=1}^n a_i \mathbf{x}_i^\top \mathbf{v}_0 \right|,$$

and taking $\mathbf{v}_0 = 0$, we have

$$\mathbb{E}^* \sup_{\mathbf{v} \in r_1 \mathbb{B}_1^d \cap r_\Sigma \mathbb{B}_\Sigma^d} \left| \frac{1}{n} \sum_{i=1}^n a_i \mathbf{x}_i^\top \mathbf{v} \right| \leq 2\mathbb{E}^* \sup_{\mathbf{v} \in r_1 \mathbb{B}_1^d \cap r_\Sigma \mathbb{B}_\Sigma^d} \frac{1}{n} \sum_{i=1}^n a_i \mathbf{x}_i^\top \mathbf{v}. \tag{78}$$

Taking the expectation with respect to $\{\mathbf{x}_i\}_{i=1}^n$ on both sides of (78), we have

$$\mathbb{E} \sup_{\mathbf{v} \in r_1 \mathbb{B}_1^d \cap r_\Sigma \mathbb{B}_\Sigma^d} \left| \frac{1}{n} \sum_{i=1}^n a_i \mathbf{x}_i^\top \mathbf{v} \right| \leq 2\mathbb{E} \sup_{\mathbf{v} \in r_1 \mathbb{B}_1^d \cap r_\Sigma \mathbb{B}_\Sigma^d} \frac{1}{n} \sum_{i=1}^n a_i \mathbf{x}_i^\top \mathbf{v}. \tag{79}$$

For any i and any $\mathbf{v}_1, \mathbf{v}_2 \in r_1 \mathbb{B}_1^d \cap r_\Sigma \mathbb{B}_\Sigma^d$ we have

$$\|\langle a_i \mathbf{x}_i, \mathbf{v}_1 - \mathbf{v}_2 \rangle\|_{\psi_2} \overset{(a)}{\leq} \|\langle \mathbf{x}_i, \mathbf{v}_1 - \mathbf{v}_2 \rangle\|_{\psi_2} \overset{(b)}{\leq} \mathfrak{L} \|\langle \mathbf{x}_i, \mathbf{v}_1 - \mathbf{v}_2 \rangle\|_{L_2}, \tag{80}$$

where (a) follows from $|a_i| = 1$ and the definition of $\|\cdot\|_{\psi_2}$ and (b) follows from (9), and we see that $\langle a_i \mathbf{x}_i, \mathbf{v} \rangle$ is a subGaussian random variable. Then, from Proposition 2.6.1 of Vershynin (2018), we have

$$\left\| \left\langle \frac{1}{n} \sum_{i=1}^n a_i \mathbf{x}_i, \mathbf{v}_1 - \mathbf{v}_2 \right\rangle \right\|_{\psi_2}^2 \leq C \sum_{i=1}^n \left\| \left\langle \frac{a_i \mathbf{x}_i}{n}, \mathbf{v}_1 - \mathbf{v}_2 \right\rangle \right\|_{\psi_2}^2 \overset{(a)}{\leq} C \frac{\mathfrak{L}^2}{n} \| \langle \mathbf{x}, \mathbf{v}_1 - \mathbf{v}_2 \rangle \|_{L_2}^2, \tag{81}$$

where (a) follows from (80). From the assumption on $\mathbf{x}$, we have

$$\| \langle \mathbf{x}, \mathbf{v}_1 - \mathbf{v}_2 \rangle \|_{L_2}^2 = \|\langle \Sigma^{\frac{1}{2}}, \mathbf{v}_1 - \mathbf{v}_2 \rangle\|_2^2 = \|\langle \Sigma^{\frac{1}{2}} \mathbf{g}, \mathbf{v}_1 - \mathbf{v}_2 \rangle\|_{L_2}^2. \tag{82}$$

From (81) and (82), we have

$$\left\| \left\langle \frac{1}{n} \sum_{i=1}^n a_i \mathbf{x}_i, \mathbf{v}_1 - \mathbf{v}_2 \right\rangle \right\|_{\psi_2} \leq C \frac{\mathfrak{L}}{\sqrt{n}} \|\langle \Sigma^{\frac{1}{2}} \mathbf{g}, \mathbf{v}_1 - \mathbf{v}_2 \rangle\|_{L_2}.$$

Then, from Corollary 8.6.2 of Vershynin (2018), we have

$$\mathbb{E} \sup_{\mathbf{v} \in r_1 \mathbb{B}_1^d \cap r_\Sigma \mathbb{B}_\Sigma^d} \left\langle \frac{1}{n} \sum_{i=1}^n a_i \mathbf{x}_i, \mathbf{v} \right\rangle \leq C \frac{\mathfrak{L}}{\sqrt{n}} \mathbb{E} \sup_{\mathbf{v} \in r_1 \mathbb{B}_1^d \cap r_\Sigma \mathbb{B}_\Sigma^d} \langle \Sigma^{\frac{1}{2}} \mathbf{g}, \mathbf{v} \rangle.$$

Then, from Lemma 34 and (79), the proof is complete. ■

Then, we proceed the proof of Proposition 25. The left-hand side of (38) divided by λ_o^2 can be expressed as

$$\sum_{i=1}^n \left(-h(\xi_{\lambda_o,i} - x_{\mathbf{v},i}) + h\left(\xi_{\lambda_o,i}\right)\right) x_{\mathbf{v},i}.$$

From the convexity of Huber loss, for any $a, b \in \mathbb{R}$, we have

$$H(a) - H(b) \geq h(b)(a-b) \text{ and } H(b) - H(a) \geq h(a)(b-a),$$

and we have

$$0 \leq \left(h(a) - h(b)\right)(a-b). \tag{83}$$

Let I_{E_i} be the indicator function of the event

$$E_i := (|\xi_{\lambda_o,i}| \leq 1/2) \cap (|x_{\mathbf{v},i}| \leq 1/2).$$

From (83), we have

$$\sum_{i=1}^n \left(-h(\xi_{\lambda_o,i} - x_{\mathbf{v},i}) + h\left(\xi_{\lambda_o,i}\right)\right) x_{\mathbf{v},i} \geq \sum_{i=1}^n \left(-h(\xi_{\lambda_o,i} - x_{\mathbf{v},i}) + h\left(\xi_{\lambda_o,i}\right)\right) x_{\mathbf{v},i} \mathrm{I}_{E_i} = \sum_{i=1}^n x_{\mathbf{v},i}^2 \mathrm{I}_{E_i}.$$

Define the functions

$$\varphi(x) = \begin{cases} x^2 & \text{if } |x| \leq 1/4 \\ (x-1/2)^2 & \text{if } 1/4 \leq x \leq 1/2 \\ (x+1/2)^2 & \text{if } -1/2 \leq x \leq -1/4 \\ 0 & \text{if } |x| > 1/2 \end{cases} \quad \text{and} \quad \psi(x) = \mathrm{I}_{(|x| \leq 1/2)},$$

where $\mathrm{I}_{(|x| \leq 1/2)}$ is the indicator function of the event $|x| \leq 1/2$. Let $f_i(\mathbf{v}) = \varphi(x_{\mathbf{v},i})\psi(\xi_{\lambda_o,i})$ and we have

$$\sum_{i=1}^n x_{\mathbf{v},i}^2 \mathrm{I}_{E_i} \overset{(a)}{\geq} \sum_{i=1}^n \varphi(x_{\mathbf{v},i})\psi(\xi_{\lambda_o,i}) = \sum_{i=1}^n f_i(\mathbf{v}), \tag{84}$$

where (a) follows from $\varphi(v) \leq v^2$ for $|v| \leq 1/2$. We note that

$$f_i(\mathbf{v}) \leq \varphi(x_{\mathbf{v},i}) \leq \min\left(x_{\mathbf{v},i}^2, 1\right). \tag{85}$$

To bound $\sum_{i=1}^n f_i(\mathbf{v})$ from below, we have

$$\inf_{\mathbf{v}\in r_1\mathbb{B}_1^d\cap r_\Sigma\mathbb{B}_\Sigma^d}\sum_{i=1}^n f_i(\mathbf{v}) \geq \inf_{\mathbf{v}\in r_1\mathbb{B}_1^d\cap r_\Sigma\mathbb{B}_\Sigma^d}\mathbb{E}\sum_{i=1}^n f(\mathbf{v}) - \sup_{\mathbf{v}\in r_1\mathbb{B}_1^d\cap r_\Sigma\mathbb{B}_\Sigma^d}\left|\sum_{i=1}^n f_i(\mathbf{v}) - \mathbb{E}\sum_{i=1}^n f_i(\mathbf{v})\right|$$

Define the supremum of a random process indexed by $r_1\mathbb{B}_1^d\cap r_\Sigma\mathbb{B}_\Sigma^d$:

$$\Delta := \sup_{\mathbf{v}\in r_1\mathbb{B}_1^d\cap r_\Sigma\mathbb{B}_\Sigma^d}\left|\sum_{i=1}^n f_i(\mathbf{v}) - \mathbb{E}\sum_{i=1}^n f_i(\mathbf{v})\right|. \tag{86}$$

Define

$$\mathrm{I}_{|x_{\mathbf{v},i}|\geq 1/2} \text{ and } \mathrm{I}_{|\xi_{\lambda_o,i}|\geq 1/2}$$

as the indicator functions of the events $|x_{\mathbf{v},i}|\geq 1/2$ and $|\xi_{\lambda_o,i}|\geq 1/2$, respectively. From $\mathrm{I}_{E_i} = 1 - \mathrm{I}_{|x_{\mathbf{v},i}|\geq 1/2} - \mathrm{I}_{|\xi_{\lambda_o,i}|>1/2}$ and (84), we have

$$\begin{aligned}\sum_{i=1}^n x_{\mathbf{v},i}^2\mathrm{I}_{E_i} \geq \mathbb{E}\sum_{i=1}^n f_i(\mathbf{v}) &\geq \sum_{i=1}^n \mathbb{E}x_{\mathbf{v},i}^2 - \sum_{i=1}^n \mathbb{E}x_{\mathbf{v},i}^2\mathrm{I}_{|x_{\mathbf{v},i}|\geq 1/2} - \sum_{i=1}^n \mathbb{E}x_{\mathbf{v},i}^2\mathrm{I}_{|\xi_{\lambda_o,i}|\geq 1/2} \\ &\geq \frac{\|\Sigma^{\frac{1}{2}}\mathbf{v}\|_2^2}{\lambda_o^2} - \sum_{i=1}^n \mathbb{E}x_{\mathbf{v},i}^2\mathrm{I}_{|x_{\mathbf{v},i}|\geq 1/2} - \sum_{i=1}^n \mathbb{E}x_{\mathbf{v},i}^2\mathrm{I}_{|\xi_{\lambda_o,i}|\geq 1/2}. \end{aligned}\tag{87}$$

We note that, from (10)

$$\mathbb{E}(\mathbf{x}_i^\top\mathbf{v})^4 \leq 16L^4\|\Sigma^{\frac{1}{2}}\mathbf{v}\|_2^4. \tag{88}$$

We evaluate the right-hand side of (87) at each term. First, for any $\mathbf{v}\in r_1\mathbb{B}_1^d\cap r_\Sigma\mathbb{B}_\Sigma^d$, we have

$$\begin{aligned}\sum_{i=1}^n \mathbb{E}x_{\mathbf{v},i}^2\mathrm{I}_{|x_{\mathbf{v},i}|\geq 1/2} &\overset{(a)}{\leq} \sum_{i=1}^n \sqrt{\mathbb{E}x_{\mathbf{v},i}^4}\sqrt{\mathbb{E}\mathrm{I}_{|x_{\mathbf{v},i}|\geq 1/2}} \\ &\overset{(b)}{=} \sum_{i=1}^n \sqrt{\mathbb{E}x_{\mathbf{v},i}^4}\sqrt{\mathbb{P}\left(|x_{\mathbf{v},i}|\geq 1/2\right)} \\ &\overset{(c)}{\leq} \sum_{i=1}^n \sqrt{\mathbb{E}x_{\mathbf{v},i}^4}\sqrt{2\exp\left(-\frac{\lambda_o^2 n}{c_{\mathfrak{L}}^2\mathfrak{L}^2\|\Sigma^{\frac{1}{2}}\mathbf{v}\|_2^2}\right)} \\ &\overset{(d)}{\leq} \frac{4}{\lambda_o^2}L^2\|\Sigma^{\frac{1}{2}}\mathbf{v}\|_2^2\sqrt{2\exp\left(-\frac{\lambda_o^2 n}{L^2\|\Sigma^{\frac{1}{2}}\mathbf{v}\|_2^2}\right)} \\ &\overset{(e)}{\leq} \frac{1}{3\lambda_o^2}\|\Sigma^{\frac{1}{2}}\mathbf{v}\|_2^2, \end{aligned}\tag{89}$$

where (a) follows from Hölder's inequality, (b) follows from the relation between indicator function and expectation, (c) follows from (12), (d) follows from (88) and the definition of

L, and (e) follows from the assumption on n. Second, for any $\mathbf{v} \in r_1\mathbb{B}_1^d \cap r_\Sigma \mathbb{B}_\Sigma^d$, we have

$$\begin{aligned}
\sum_{i=1}^n \mathbb{E}x_{\mathbf{v},i}^2 \mathrm{I}_{|\xi_{\lambda_o,i}|\geq 1/2} &\overset{(a)}{\leq} \sum_{i=1}^n \sqrt{\mathbb{E}x_{\mathbf{v},i}^4}\sqrt{\mathbb{E}\mathrm{I}_{|\xi_{\lambda_o,i}|\geq 1/2}} \\
&\overset{(b)}{=} \sum_{i=1}^n \sqrt{\mathbb{E}x_{\mathbf{v},i}^4}\sqrt{\mathbb{P}(|\xi_{\lambda_o,i}| \geq 1/2)} \\
&= \frac{1}{\lambda_o^2}\sqrt{\mathbb{E}(\mathbf{v}^\top \mathbf{x}_i)^4}\sqrt{\mathbb{P}(|\xi_{\lambda_o,i}| \geq 1/2)} \\
&\overset{(c)}{\leq} \frac{4L^2}{\lambda_o^2}\sqrt{\mathbb{P}(|\xi_{\lambda_o,i}| \geq 1/2)}\|\Sigma^{\frac{1}{2}}\mathbf{v}\|_2^2 \\
&\overset{(d)}{\leq} \frac{1}{3\lambda_o^2}\|\Sigma^{\frac{1}{2}}\mathbf{v}\|_2^2, \qquad (90)
\end{aligned}$$

where (a) follows from Hölder's inequality, (b) follows from relation between indicator function and expectation, and (c) follows from (88), and (d) follows from (14). Consequently, from (89) and (90) we have

$$\frac{\|\Sigma^{\frac{1}{2}}\mathbf{v}\|_2^2}{3\lambda_o^2} - \Delta \leq \sum_{i=1}^n \left(-h(\xi_{\lambda_o,i} - x_{\mathbf{v},i}) + h\left(\xi_{\lambda_o,i}\right)\right) x_{\mathbf{v},i}. \qquad (91)$$

Next, we evaluate the stochastic term Δ defined in (86). From (85) and Theorem 3 of Massart (2000), with probability at least $1-\delta$, we have

$$\Delta \leq 2\mathbb{E}\Delta + \sigma_f\sqrt{8\log(1/\delta)} + 18\log(1/\delta), \qquad (92)$$

where

$$\sigma_f^2 = \sup_{\mathbf{v}\in r_1\mathbb{B}_1^d \cap r_\Sigma \mathbb{B}_\Sigma^d} \sum_{i=1}^n \mathbb{E}(f_i(\mathbf{v}) - \mathbb{E}f_i(\mathbf{v}))^2.$$

From (85) and (88), we have

$$\mathbb{E}(f_i(\mathbf{v}) - \mathbb{E}f_i(\mathbf{v}))^2 \leq \mathbb{E}f_i^2(\mathbf{v}) \leq \mathbb{E}f_i(\mathbf{v}) \leq \mathbb{E}x_{\mathbf{v},i}^2 \leq \frac{L^2\|\Sigma^{\frac{1}{2}}\mathbf{v}\|_2^2}{\lambda_o^2 n}.$$

Combining this and (92), with probability at least $1-\delta$, we have

$$\Delta \leq 2\mathbb{E}\Delta + \frac{L}{\lambda_o}\sqrt{8\log(1/\delta)}\|\Sigma^{\frac{1}{2}}\mathbf{v}\|_2 + 18\log(1/\delta). \qquad (93)$$

From symmetrization inequality (Theorem 11.4 of Boucheron et al. (2013)), we have $\mathbb{E}\Delta \leq 2\,\mathbb{E}\sup_{\mathbf{v}\in r_1\mathbb{B}_1^d\cap r_\Sigma\mathbb{B}_\Sigma^d}|\mathbb{G}_\mathbf{v}| \leq 2\,\mathbb{E}\sup_{\mathbf{v}\in r_1\mathbb{B}_1^d\cap r_\Sigma\mathbb{B}_\Sigma^d}|\mathbb{G}_\mathbf{v}|$, where

$$\mathbb{G}_\mathbf{v} := \sum_{i=1}^n a_i\varphi(x_{\mathbf{v},i})\psi(\xi_{\lambda_o,i}),$$

and $\{a_i\}_{i=1}^n$ is a sequence of i.i.d. Rademacher random variables which are independent of $\{\mathbf{x}_i, \xi_i\}_{i=1}^n$. We denote $\mathbb{E}^*$ as a conditional expectation of $\{a_i\}_{i=1}^n$ given $\{\mathbf{x}_i, \xi_i\}_{i=1}^n$. Since φ is 1-Lipschitz and $\varphi(0) = 0$, from contraction principle (Theorem 11.5 of Boucheron et al. (2013)), we have

$$\mathbb{E}^* \sup_{\mathbf{v} \in r_1 \mathbb{B}_1^d \cap r_\Sigma \mathbb{B}_\Sigma^d} \left| \sum_{i=1}^n a_i \varphi(x_{\mathbf{v},i}) \psi(\xi_{\lambda_o,i}) \right| \leq \frac{1}{2\lambda_o \sqrt{n}} \mathbb{E}^* \sup_{\mathbf{v} \in r_1 \mathbb{B}_1^d \cap r_\Sigma \mathbb{B}_\Sigma^d} \left| \sum_{i=1}^n a_i \mathbf{x}_i^\top \mathbf{v} \right|.$$

and from the basic property of the expectation, we have

$$\mathbb{E} \sup_{\mathbf{v} \in r_1 \mathbb{B}_1^d \cap r_\Sigma \mathbb{B}_\Sigma^d} \left| \sum_{i=1}^n a_i \varphi(x_{\mathbf{v},i}) \psi(\xi_{\lambda_o,i}) \right| \leq \frac{1}{2\lambda_o \sqrt{n}} \mathbb{E} \sup_{\mathbf{v} \in r_1 \mathbb{B}_1^d \cap r_\Sigma \mathbb{B}_\Sigma^d} \left| \sum_{i=1}^n a_i \mathbf{x}_i^\top \mathbf{v} \right|.$$

From Lemma 46, we have

$$\lambda_o^2 \mathbb{E}\Delta \leq \lambda_o^2 \mathbb{E} \sup_{\mathbf{v} \in r_1 \mathbb{B}_1^d \cap r_\Sigma \mathbb{B}_\Sigma^d} \left| \sum_{i=1}^n a_i \varphi(x_{\mathbf{v},i}) \psi(\xi_{\lambda_o,i}) \right| \leq CL\lambda_o \sqrt{n} \rho r_1 r_{d,s}. \tag{94}$$

From (93) and (94), we have

$$\begin{aligned} \lambda_o^2 \Delta &\leq CL\lambda_o \sqrt{n} \rho r_{d,s} r_1 + CL\lambda_o \sqrt{\log(1/\delta)} r_\Sigma + C\lambda_o^2 n r_\delta^2 \\ &\overset{(a)}{\leq} CL\lambda_o \sqrt{n} (\rho r_{d,s} r_1 + r_\delta r_\Sigma), \end{aligned} \tag{95}$$

where (a) follows from $\lambda_o \sqrt{n} r_\delta \leq r_\Sigma$ and $L \geq 1$. From (91) and (95), we have

$$\inf_{\mathbf{v} \in r_1 \mathbb{B}_1^d \cap r_\Sigma \mathbb{B}_\Sigma^d} \lambda_o^2 \sum_{i=1}^n \left(-h(\xi_{\lambda_o,i} - x_{\mathbf{v},i}) + h\left(\xi_{\lambda_o,i}\right) \right) x_{\mathbf{v},i} \geq \frac{\|\Sigma^{\frac{1}{2}} \mathbf{v}\|_2^2}{3} - CL\lambda_o \sqrt{n} (\rho r_{d,s} r_1 + r_\delta r_\Sigma),$$

and the proof is complete.

F.5 Proof of Proposition 29

We note that

$$\left| \sum_{i \in \mathcal{O}} \hat{w}_i' u_i \mathbf{X}_i^\top \mathbf{v} \right|^2 \overset{(a)}{\leq} 4 \frac{o}{n} \sum_{i \in \mathcal{O}}^n \hat{w}_i' |\mathbf{X}_i^\top \mathbf{v}|^2 \overset{(b)}{\leq} 8 \frac{o}{n} \sum_{i=1}^n \hat{w}_i |\mathbf{X}_i^\top \mathbf{v}|^2,$$

where (a) follows from Hölder's inequality and $\sum_{i \in \mathcal{O}} u_i^2 \leq 4o$, and (b) follows from the fact that $\hat{w}_i' \leq 2\hat{w}_i$ for any $i \in (1, \cdots, n)$. We focus on $\sum_{i=1}^n \hat{w}_i |\mathbf{X}_i^\top \mathbf{v}|^2$. First, for any $\mathbf{v} \in r_1 \mathbb{B}_1^d \cap r_2 \mathbb{B}_2^d$, we have

$$\sum_{i=1}^n \hat{w}_i |\mathbf{X}_i^\top \mathbf{v}|^2 \overset{(a)}{\leq} \max_{M \in \mathfrak{M}_{r_1, r_2, d}^{\ell_1, \mathrm{Tr}}} \sum_{i=1}^n \hat{w}_i \langle \mathbf{X}_i \mathbf{X}_i^\top, M \rangle \overset{(b)}{\leq} \tau_{\mathrm{cut}},$$

where (a) follows from the fact that $\mathfrak{M}^{\ell_1,\ell_2}_{r_1,r_2,d,\mathbf{v}} \subset \mathfrak{M}^{\ell_1,\mathrm{Tr}}_{r_1,r_2,d}$ and (b) follows from the succeeding condition of Algorithm 5.

Combining the arguments above, we have

$$\sum_{i\in\mathcal{O}} \hat{w}_i \mathbf{X}_i^\top \mathbf{v} \leq CL\sqrt{c_{\mathrm{cut}}}\sqrt{\frac{o}{n}}\sqrt{\kappa_{\mathrm{u}}^2\left(sr_{d,s}+r_\delta\right)+\Sigma_{\max}^2}r_2,$$

where we use the definition of τ_{cut} and the proof is complete.

F.6 Proof of Proposition 30

We note that, from Hölder's inequality, we have

$$\left|\sum_{i\in I_m}\frac{u_i\mathbf{x}_i^\top\mathbf{v}}{n}\right|^2 \overset{(a)}{\leq} \sum_{i\in I_m}\frac{1}{n}u_i^2\sum_{i=1}^n\frac{1}{n}|\mathbf{x}_i^\top\mathbf{v}|^2 \overset{(b)}{\leq} 4\frac{m}{n}\sum_{i=1}^n\frac{1}{n}|\mathbf{x}_i^\top\mathbf{v}|^2,$$

where (a) follows from Hölder's inequality, and (b) follows from the fact that $\|\mathbf{u}\|_2^2 \leq 4m$. From the fact that $\mathfrak{M}^{\ell_1,\ell_2}_{r_1,r_2,d,\mathbf{v}} \subset \mathfrak{M}^{\ell_1,\mathrm{Tr}}_{r_1,r_2,d}$, we have

$$\sum_{i=1}^n\frac{(\mathbf{x}_i^\top\mathbf{v})^2}{n} \leq \sup_{M\in\mathfrak{M}^{\ell_1,\mathrm{Tr}}_{r_1,r_2,d}}\sum_{i=1}^n\frac{\langle\mathbf{x}_i\mathbf{x}_i^\top,M\rangle}{n}.$$

From Corollary 28 and triangular inequality, we have

$$\begin{aligned}\sum_{i\in I_m}u_i\mathbf{x}_i^\top\mathbf{v} &\leq \sqrt{c_1'\frac{m}{n}}\sqrt{(L\kappa_{\mathrm{u}})^2\left(sr_{d,s}+r_\delta\right)+\Sigma_{\max}^2}r_2\\ &\overset{(a)}{\leq} CL\sqrt{\frac{o}{n}}\sqrt{\kappa_{\mathrm{u}}^2\left(sr_{d,s}+r_\delta\right)+\Sigma_{\max}^2}r_2,\end{aligned}$$

where (a) follows from $m \leq (2c_\varepsilon+1)o$ and $L \geq 1$, and the proof is complete.

Appendix G. Proof of Lemmas 39, 40, 41 and 42

Proof [Proof of Lemma 39] From the triangular inequality, we have

$$\begin{aligned}\left|\sum_{i=1}^n\hat{w}_i'h(r_{\boldsymbol{\beta}^*,i})\mathbf{X}_i^\top\boldsymbol{\theta}_\eta\right| &\leq \left|\sum_{i\in\mathcal{I}}\hat{w}_i'h(r_{\boldsymbol{\beta}^*,i})\mathbf{X}_i^\top\boldsymbol{\theta}_\eta\right| + \left|\sum_{i\in\mathcal{O}}\hat{w}_i'h(r_{\boldsymbol{\beta}^*,i})\mathbf{X}_i^\top\boldsymbol{\theta}_\eta\right|\\ &= \left|\sum_{i\in\mathcal{I}}\hat{w}_i'h(\xi_{\lambda_o,i})\mathbf{x}_i^\top\boldsymbol{\theta}_\eta\right| + \left|\sum_{i\in\mathcal{O}}\hat{w}_i'h(r_{\boldsymbol{\beta}^*,i})\mathbf{X}_i^\top\boldsymbol{\theta}_\eta\right|\\ &= \left|\sum_{i=1}^n\frac{1}{n}h(\xi_{\lambda_o,i})\mathbf{x}_i^\top\boldsymbol{\theta}_\eta - \sum_{i\in\mathcal{O}\cup(\mathcal{I}\cap I_<)}\frac{1}{n}h(\xi_{\lambda_o,i})\mathbf{x}_i^\top\boldsymbol{\theta}_\eta\right| + \left|\sum_{i\in\mathcal{O}}\hat{w}_i'h(r_{\boldsymbol{\beta}^*,i})\mathbf{X}_i^\top\boldsymbol{\theta}_\eta\right|\\ &\leq \left|\sum_{i=1}^n\frac{1}{n}h(\xi_{\lambda_o,i})\mathbf{x}_i^\top\boldsymbol{\theta}_\eta\right| + \left|\sum_{i\in\mathcal{O}}\hat{w}_i'h(\xi_{\lambda_o,i})\mathbf{X}_i^\top\boldsymbol{\theta}_\eta\right| + \left|\sum_{i\in\mathcal{O}\cup(\mathcal{I}\cap I_<)}\frac{1}{n}h(r_{\boldsymbol{\beta}^*,i})\mathbf{x}_i^\top\boldsymbol{\theta}_\eta\right|.\end{aligned}$$

We note that $|h(\cdot)| \leq 1$ and from Lemma 27, $|\mathcal{O} \cup (\mathcal{I} \cap I_<)| \leq (1 + 2c_\varepsilon)o$. Therefore, from Propositions 22 - 24, we have

$$\left|\sum_{i=1}^{n} \frac{1}{n} h(\xi_{\lambda_o,i}) \mathbf{x}_i^\top \boldsymbol{\theta}_\eta\right| \leq c_3 L \left(\rho r_{d,s} r_1 + r_\delta r_\Sigma\right)$$
$$\left|\sum_{i\in\mathcal{O}} \frac{1}{n} h(\xi_{\lambda_o,i}) \mathbf{x}_i^\top \boldsymbol{\theta}_\eta\right| \leq c_4 L \sqrt{1 + c_{\mathrm{cut}}} \left(\kappa_{\mathrm{u}} \sqrt{\frac{o}{n}} \left(\sqrt{s r_{d,s}} + \sqrt{r_\delta}\right) r_2 + \kappa_{\mathrm{u}} r_o r_2\right)$$
$$\left|\sum_{i\in\mathcal{O}\cup(\mathcal{I}\cap I_<)} \frac{1}{n} h(r_{\boldsymbol{\beta}^*,i}) \mathbf{x}_i^\top \boldsymbol{\theta}_\eta\right| \leq c_5 L \left(\kappa_{\mathrm{u}} \sqrt{\frac{o}{n}} \left(\sqrt{s r_{d,s}} + \sqrt{r_\delta}\right) r_2 + \kappa_{\mathrm{u}} r_o r_2\right),$$

and we see that

$$\left|\sum_{i=1}^{n} \hat{w}_i' h(r_{\boldsymbol{\beta}^*,i}) \mathbf{X}_i^\top \boldsymbol{\theta}_\eta\right| \leq 3 c_{\max}^2 L \left(\rho r_{d,s} r_1 + r_\delta r_\Sigma + \kappa_{\mathrm{u}} \sqrt{\frac{o}{n}} \left(\sqrt{s r_{d,s}} + \sqrt{r_\delta}\right) r_2 + \kappa_{\mathrm{u}} r_o r_2\right).$$

■

Proof [Proof of Lemma 40] We have

$$\begin{aligned}
&\sum_{i=1}^{n} \lambda_o \sqrt{n} \hat{w}_i' \left(-h(r_{\boldsymbol{\beta}^*+\boldsymbol{\theta}_\eta,i}) + h(r_{\boldsymbol{\beta}^*,i})\right) \mathbf{X}_i^\top \boldsymbol{\theta}_\eta \\
&= \sum_{i\in\mathcal{I}} \lambda_o \sqrt{n} \hat{w}_i' \left(-h(r_{\boldsymbol{\beta}^*+\boldsymbol{\theta}_\eta,i}) + h(r_{\boldsymbol{\beta}^*,i})\right) \mathbf{X}_i^\top \boldsymbol{\theta}_\eta + \sum_{i\in\mathcal{O}} \lambda_o \sqrt{n} \hat{w}_i' \left(-h(r_{\boldsymbol{\beta}^*+\boldsymbol{\theta}_\eta,i}) + h(r_{\boldsymbol{\beta}^*,i})\right) \mathbf{X}_i^\top \boldsymbol{\theta}_\eta \\
&= \sum_{i\in\mathcal{I}} \lambda_o \sqrt{n} \hat{w}_i' \left(-h(\xi_{\lambda_o,i} - x_{\boldsymbol{\theta}_\eta,i}) + h(\xi_{\lambda_o,i})\right) \mathbf{x}_i^\top \boldsymbol{\theta}_\eta + \sum_{i\in\mathcal{O}} \lambda_o \sqrt{n} \hat{w}_i' \left(-h(r_{\boldsymbol{\beta}^*+\boldsymbol{\theta}_\eta,i}) + h(r_{\boldsymbol{\beta}^*,i})\right) \mathbf{X}_i^\top \boldsymbol{\theta}_\eta \\
&= \left(\sum_{i=1}^{n} - \sum_{i\in\mathcal{O}\cup(\mathcal{I}\cap I_<)}\right) \frac{\lambda_o}{\sqrt{n}} \left(-h(\xi_{\lambda_o,i} - x_{\boldsymbol{\theta}_\eta,i}) + h(\xi_{\lambda_o,i})\right) \mathbf{x}_i^\top \boldsymbol{\theta}_\eta \\
&\quad + \sum_{i\in\mathcal{O}} \lambda_o \sqrt{n} \hat{w}_i' \left(-h(r_{\boldsymbol{\beta}^*+\boldsymbol{\theta}_\eta,i}) + h(r_{\boldsymbol{\beta}^*,i})\right) \mathbf{X}_i^\top \boldsymbol{\theta}_\eta \\
&\geq \sum_{i=1}^{n} \frac{\lambda_o}{\sqrt{n}} \left(-h(\xi_{\lambda_o,i} - x_{\boldsymbol{\theta}_\eta,i}) + h(\xi_{\lambda_o,i})\right) \mathbf{x}_i^\top \boldsymbol{\theta}_\eta - \left|\sum_{i\in\mathcal{O}\cup(\mathcal{I}\cap I_<)} \frac{\lambda_o}{\sqrt{n}} \left(-h(\xi_{\lambda_o,i} - x_{\boldsymbol{\theta}_\eta,i}) + h(\xi_{\lambda_o,i})\right) \mathbf{x}_i^\top \boldsymbol{\theta}_\eta\right| \\
&\quad - \left|\sum_{i\in\mathcal{O}} \lambda_o \sqrt{n} \hat{w}_i' \left(-h(r_{\boldsymbol{\beta}^*+\boldsymbol{\theta}_\eta,i}) + h(r_{\boldsymbol{\beta}^*,i})\right) \mathbf{X}_i^\top \boldsymbol{\theta}_\eta\right|.
\end{aligned}$$

We note that $|h(\cdot)| \leq 1$ and from Lemma 27, $|\mathcal{O} \cup (\mathcal{I} \cap I_<)| \leq (1+2c_\varepsilon)o$. From Propositions 25, Propositions 23 and 24, we have

$$\sum_{i=1}^{n} \frac{\lambda_o}{\sqrt{n}} \left(-h(\xi_{\lambda_o,i} - x_{\boldsymbol{\theta}_\eta,i}) + h(\xi_{\lambda_o,i})\right) \mathbf{x}_i^\top \boldsymbol{\theta}_\eta \geq \frac{\|\Sigma^{\frac{1}{2}}\mathbf{v}\|_2^2}{3} r_2^2 - c_{\max} L \lambda_o \sqrt{n} \left(\rho r_{d,s} r_1 + r_\delta r_\Sigma\right)$$

$$\left|\sum_{i\in\mathcal{O}} \lambda_o \sqrt{n} \hat{w}_i' \left(-h(r_{\boldsymbol{\beta}^*+\boldsymbol{\theta}_\eta,i}) + h(r_{\boldsymbol{\beta}^*,i})\right) \mathbf{X}_i^\top \boldsymbol{\theta}_\eta\right| \leq c_{\max}^2 L \lambda_o \sqrt{n} \left(\kappa_{\mathrm{u}} \sqrt{\frac{o}{n}} \left(\sqrt{s r_{d,s}} + \sqrt{r_\delta}\right) r_2 + \kappa_{\mathrm{u}} r_o r_2\right)$$

$$\left|\sum_{i\in\mathcal{O}\cup(\mathcal{I}\cap I_<)} \frac{\lambda_o}{\sqrt{n}} \left(-h(\xi_{\lambda_o,i} - x_{\boldsymbol{\theta}_\eta,i}) + h(\xi_{\lambda_o,i})\right) \mathbf{x}_i^\top \boldsymbol{\theta}_\eta\right| \leq c_{\max}^2 L \lambda_o \sqrt{n} \left(\kappa_{\mathrm{u}} \sqrt{\frac{o}{n}} \left(\sqrt{s r_{d,s}} + \sqrt{r_\delta}\right) r_2 + \kappa_{\mathrm{u}} r_o r_2\right).$$

Combining the arguments above, we see that

$$\begin{aligned}
&\sum_{i=1}^{n} \lambda_o \sqrt{n} \hat{w}_i' \left(-h(r_{\boldsymbol{\beta}^*+\boldsymbol{\theta}_\eta,i}) + h(r_{\boldsymbol{\beta}^*,i})\right) \mathbf{X}_i^\top \boldsymbol{\theta}_\eta \\
&\geq \frac{\|\Sigma^{\frac{1}{2}}\mathbf{v}\|_2^2}{3} - c_{\max}^2 L \lambda_o \sqrt{n} \left(\rho r_{d,s} r_1 + r_\delta r_\Sigma + \kappa_{\mathrm{u}} \sqrt{\frac{o}{n}} \left(\sqrt{s r_{d,s}} + \sqrt{r_\delta}\right) r_2 + \kappa_{\mathrm{u}} r_o r_2\right),
\end{aligned}$$

and the proof is complete. ■

Proof [Proof of Lemma 41] From the triangular inequality, we have

$$\begin{aligned}
\left|\sum_{i=1}^{n} \hat{w}_i' h(r_{\boldsymbol{\beta}^*,i}) \mathbf{X}_i^\top \boldsymbol{\theta}_\eta\right| &\leq \left|\sum_{i\in\mathcal{I}} \hat{w}_i' h(r_{\boldsymbol{\beta}^*,i}) \mathbf{X}_i^\top \boldsymbol{\theta}_\eta\right| + \left|\sum_{i\in\mathcal{O}} \hat{w}_i' h(r_{\boldsymbol{\beta}^*,i}) \mathbf{X}_i^\top \boldsymbol{\theta}_\eta\right| \\
&= \left|\sum_{i\in\mathcal{I}} \hat{w}_i' h(\xi_{\lambda_o,i}) \mathbf{x}_i^\top \boldsymbol{\theta}_\eta\right| + \left|\sum_{i\in\mathcal{O}} \hat{w}_i' h(r_{\boldsymbol{\beta}^*,i}) \mathbf{X}_i^\top \boldsymbol{\theta}_\eta\right| \\
&= \left|\sum_{i=1}^{n} \frac{1}{n} h(\xi_{\lambda_o,i}) \mathbf{x}_i^\top \boldsymbol{\theta}_\eta - \sum_{i\in\mathcal{O}\cup(\mathcal{I}\cap I_<)} \frac{1}{n} h(\xi_{\lambda_o,i}) \mathbf{x}_i^\top \boldsymbol{\theta}_\eta\right| + \left|\sum_{i\in\mathcal{O}} \hat{w}_i' h(r_{\boldsymbol{\beta}^*,i}) \mathbf{X}_i^\top \boldsymbol{\theta}_\eta\right| \\
&\leq \left|\sum_{i=1}^{n} \frac{1}{n} h(\xi_{\lambda_o,i}) \mathbf{x}_i^\top \boldsymbol{\theta}_\eta\right| + \left|\sum_{i\in\mathcal{O}} \hat{w}_i' h(\xi_{\lambda_o,i}) \mathbf{X}_i^\top \boldsymbol{\theta}_\eta\right| + \left|\sum_{i\in\mathcal{O}\cup(\mathcal{I}\cap I_<)} \frac{1}{n} h(r_{\boldsymbol{\beta}^*,i}) \mathbf{x}_i^\top \boldsymbol{\theta}_\eta\right|.
\end{aligned}$$

We note that $|h(\cdot)| \leq 1$ and from Lemma 27, $|\mathcal{O} \cup (\mathcal{I} \cap I_<)| \leq (1 + 2c'_\varepsilon)o$. Therefore, from Propositions 22, 29 and 30, we have

$$\begin{aligned}
\left|\sum_{i=1}^{n} \frac{1}{n} h(\xi_{\lambda_o,i}) \mathbf{x}_i^\top \boldsymbol{\theta}_\eta\right| &\leq c_3 L \left(\rho r_{d,s} r_1 + r_\delta r_\Sigma\right), \\
\left|\sum_{i\in\mathcal{O}} \frac{1}{n} h(\xi_{\lambda_o,i}) \mathbf{x}_i^\top \boldsymbol{\theta}_\eta\right| &\leq c_8 \sqrt{c_{\mathrm{cut}}} L \sqrt{\frac{o}{n}} \sqrt{\kappa_{\mathrm{u}}^2 \left(s r_{d,s} + r_\delta\right) + \kappa_{\mathrm{u}}^2} r_2, \\
\left|\sum_{i\in\mathcal{O}\cup(\mathcal{I}\cap I_<)} \frac{1}{n} h(r_{\boldsymbol{\beta}^*,i}) \mathbf{x}_i^\top \boldsymbol{\theta}_\eta\right| &\leq c_9 L \sqrt{\frac{o}{n}} \sqrt{\kappa_{\mathrm{u}}^2 \left(s r_{d,s} + r_\delta\right) + \kappa_{\mathrm{u}}^2} r_2,
\end{aligned}$$

and we see that

$$\left|\sum_{i=1}^{n} \hat{w}'_i h(r_{\boldsymbol{\beta}^*,i}) \mathbf{X}_i^\top \boldsymbol{\theta}_\eta\right| \leq 3 c'^2_{\max} L \left(\rho r_{d,s} r_1 + r_\delta r_\Sigma + \sqrt{\frac{o}{n}} \sqrt{\kappa_{\mathrm{u}}^2 \left(s r_{d,s} + r_\delta\right) + \kappa_{\mathrm{u}}^2} r_2\right).$$

■

Proof [Proof of Lemma 42] We have

$$\begin{aligned}
&\sum_{i=1}^{n} \lambda_o \sqrt{n} \hat{w}'_i \left(-h(r_{\boldsymbol{\beta}^*+\boldsymbol{\theta}_\eta,i}) + h(r_{\boldsymbol{\beta}^*,i})\right) \mathbf{X}_i^\top \boldsymbol{\theta}_\eta \\
&= \sum_{i\in\mathcal{I}} \lambda_o \sqrt{n} \hat{w}'_i \left(-h(r_{\boldsymbol{\beta}^*+\boldsymbol{\theta}_\eta,i}) + h(r_{\boldsymbol{\beta}^*,i})\right) \mathbf{X}_i^\top \boldsymbol{\theta}_\eta + \sum_{i\in\mathcal{O}} \lambda_o \sqrt{n} \hat{w}'_i \left(-h(r_{\boldsymbol{\beta}^*+\boldsymbol{\theta}_\eta,i}) + h(r_{\boldsymbol{\beta}^*,i})\right) \mathbf{X}_i^\top \boldsymbol{\theta}_\eta \\
&= \sum_{i\in\mathcal{I}} \lambda_o \sqrt{n} \hat{w}'_i \left(-h(\xi_{\lambda_o,i} - x_{\boldsymbol{\theta}_\eta,i}) + h(\xi_{\lambda_o,i})\right) \mathbf{x}_i^\top \boldsymbol{\theta}_\eta + \sum_{i\in\mathcal{O}} \lambda_o \sqrt{n} \hat{w}'_i \left(-h(r_{\boldsymbol{\beta}^*+\boldsymbol{\theta}_\eta,i}) + h(r_{\boldsymbol{\beta}^*,i})\right) \mathbf{X}_i^\top \boldsymbol{\theta}_\eta \\
&= \left(\sum_{i=1}^{n} - \sum_{i\in\mathcal{O}\cup(\mathcal{I}\cap I_<)}\right) \frac{\lambda_o}{\sqrt{n}} \left(-h(\xi_{\lambda_o,i} - x_{\boldsymbol{\theta}_\eta,i}) + h(\xi_{\lambda_o,i})\right) \mathbf{x}_i^\top \boldsymbol{\theta}_\eta \\
&\quad + \sum_{i\in\mathcal{O}} \lambda_o \sqrt{n} \hat{w}'_i \left(-h(r_{\boldsymbol{\beta}^*+\boldsymbol{\theta}_\eta,i}) + h(r_{\boldsymbol{\beta}^*,i})\right) \mathbf{X}_i^\top \boldsymbol{\theta}_\eta \\
&\geq \sum_{i=1}^{n} \frac{\lambda_o}{\sqrt{n}} \left(-h(\xi_{\lambda_o,i} - x_{\boldsymbol{\theta}_\eta,i}) + h(\xi_{\lambda_o,i})\right) \mathbf{x}_i^\top \boldsymbol{\theta}_\eta - \left|\sum_{i\in\mathcal{O}\cup(\mathcal{I}\cap I_<)} \frac{\lambda_o}{\sqrt{n}} \left(-h(\xi_{\lambda_o,i} - x_{\boldsymbol{\theta}_\eta,i}) + h(\xi_{\lambda_o,i})\right) \mathbf{x}_i^\top \boldsymbol{\theta}_\eta\right| \\
&\quad - \left|\sum_{i\in\mathcal{O}} \lambda_o \sqrt{n} \hat{w}'_i \left(-h(r_{\boldsymbol{\beta}^*+\boldsymbol{\theta}_\eta,i}) + h(r_{\boldsymbol{\beta}^*,i})\right) \mathbf{X}_i^\top \boldsymbol{\theta}_\eta\right|.
\end{aligned}$$

We note that $|h(\cdot)| \leq 1$ and from Lemma 27, $|\mathcal{O} \cup (\mathcal{I} \cap I_<)| \leq (1+2c_\varepsilon)o$. Therefore, from Proposition 25, we have

$$\sum_{i=1}^{n} \frac{\lambda_o}{\sqrt{n}} \left(-h(\xi_{\lambda_o,i} - x_{\boldsymbol{\theta}_\eta,i}) + h(\xi_{\lambda_o,i})\right) \mathbf{x}_i^\top \boldsymbol{\theta}_\eta \geq \frac{\|\Sigma^{\frac{1}{2}}\mathbf{v}\|_2^2}{3} r_2^2 - c_{\max} L \lambda_o \sqrt{n} \left(\rho r_{d,s} r_1 + r_\delta r_\Sigma\right)$$

$$\left| \sum_{i \in \mathcal{O}} \lambda_o \sqrt{n} \hat{w}_i' \left(-h(r_{\boldsymbol{\beta}^* + \boldsymbol{\theta}_\eta, i}) + h(r_{\boldsymbol{\beta}^*,i})\right) \mathbf{X}_i^\top \boldsymbol{\theta}_\eta \right| \leq c_8 \sqrt{c_{\mathrm{cut}}} L \sqrt{\frac{o}{n}} \sqrt{\kappa_{\mathrm{u}}^2 \left(s r_{d,s} + r_\delta\right) + \kappa_{\mathrm{u}}^2} r_2,$$

$$\left| \sum_{i \in \mathcal{O} \cup (\mathcal{I} \cap I_<)} \frac{\lambda_o}{\sqrt{n}} \left(-h(\xi_{\lambda_o,i} - x_{\boldsymbol{\theta}_\eta,i}) + h(\xi_{\lambda_o,i})\right) \mathbf{x}_i^\top \boldsymbol{\theta}_\eta \right| \leq c_8 \sqrt{c_{\mathrm{cut}}} L \sqrt{\frac{o}{n}} \sqrt{\kappa_{\mathrm{u}}^2 \left(s r_{d,s} + r_\delta\right) + \kappa_{\mathrm{u}}^2} r_2, .$$

Combining the arguments above, we see that

$$\begin{aligned}
&\sum_{i=1}^{n} \lambda_o \sqrt{n} \hat{w}_i' \left(-h(r_{\boldsymbol{\beta}^* + \boldsymbol{\theta}_\eta, i}) + h(r_{\boldsymbol{\beta}^*,i})\right) \mathbf{X}_i^\top \boldsymbol{\theta}_\eta \\
&\geq \frac{\|\Sigma^{\frac{1}{2}}\mathbf{v}\|_2^2}{3} - c_{\max}'^2 L \lambda_o \sqrt{n} \left(\rho r_{d,s} r_1 + r_\delta r_\Sigma + \sqrt{\frac{o}{n}} \sqrt{\kappa_{\mathrm{u}}^2 \left(s r_{d,s} + r_\delta\right) + \kappa_{\mathrm{u}}^2} r_2\right),
\end{aligned}$$

and the proof is complete. ■

Appendix H. Proof of Corollary 17

The proof of Corollary 17 is almost the same as the one of Theorem 15. The difference lies in using Corollary 47 instead of Corollary 28. The proof of Corollary 47 is in Section H.1.

Corollary 47 *Suppose that Assumption 3, and $r_d, r_\delta \leq 1, 0 < r_1, r_2$, $r_1/r_2 \leq \sqrt{s}$ hold. Define c_{10} and c_{11} as numerical constants. Then, with probability at least $1-\delta$, we have*

$$\max_{M \in \mathfrak{M}_{r_1,r_2,d}^{\ell_{1,\mathrm{Tr}}}} \sum_{i=1}^{n} \frac{\left\langle \mathbf{x}_i \mathbf{x}_i^\top, M \right\rangle}{n} \leq c_{10}(L^2 + \rho^2) \left(s\left(r_{d,s} + r_\delta\right) + \Sigma_{\max}^2\right) r_2^2. \tag{96}$$

Additionally, assume that $\varepsilon = c_\varepsilon \times (o/n)$, where $1 \leq c_\varepsilon < 2$ and $(0 <) o/n \leq 1/2$ hold, then, with probability at least $1-\delta$, we have

$$\max_{M \in \mathfrak{M}_{r_1,r_2,d}^{\ell_{1,\mathrm{Tr}}}} \sum_{i=1}^{n} \hat{w}_i \left\langle \mathbf{X}_i \mathbf{X}_i^\top, M \right\rangle \leq c_{11}(L^2 + \rho^2) \left(s\left(r_{d,s} + r_\delta\right) + \Sigma_{\max}^2\right) r_2^2, \tag{97}$$

where $\{\hat{w}\}_{i=1}^n$ is a solution of (25).

As a result, Propositions 29 and 30 used in the proof of Theorem 15 will be replaced Corollaries 48 and 49, respectively.

Corollary 48 *Suppose that the assumptions in Corollary 47 hold. Set*

$$\tau_{\mathrm{cut}} = c_{\mathrm{cut}} \left(L^2 \left(s \left(r_{d,s} + r_\delta \right) + \Sigma_{\max}^2 \right) r_2^2 \right),$$

where $c_{\mathrm{cut}} \geq c_{11}$*, and suppose that* (97) *holds and that Algorithm 5 returns* $\hat{\mathbf{w}}$*. For any* $\mathbf{u} \in \mathbb{R}^n$ *such that* $\|\mathbf{u}\|_\infty \leq 2$ *and for any* $\mathbf{v} \in r_1 \mathbb{B}_1^d \cap r_2 \mathbb{B}_2^d$*, we have the following:*

$$\left| \sum_{i \in \mathcal{O}} \hat{w}_i' u_i \mathbf{X}_i^\top \mathbf{v} \right| \leq c_{12} \sqrt{c_{\mathrm{cut}}} (L + \rho) \sqrt{\frac{o}{n}} \sqrt{s \left(r_d + r_\delta \right) + \Sigma_{\max}^2} r_2,$$

where c_{12} *is a numerical constant.*

Corollary 49 *Suppose that the assumptions in Corollary 47 hold. Then, for any* $m \in \mathbb{N}$ *such that* $m \leq (2c_\varepsilon + 1)o$*, where* c_ε *is some numerical constant such that* $1 \leq c_\varepsilon < 2$*, for any* $\mathbf{u} \in \mathbb{R}^n$ *such that* $\|\mathbf{u}\|_\infty \leq 2$ *and for any* $\mathbf{v} \in r_1 \mathbb{B}_1^d \cap r_2 \mathbb{B}_2^d$*, we have the following:*

$$\left| \sum_{i \in I_m} \frac{1}{n} u_i \mathbf{x}_i^\top \mathbf{v} \right| \leq c_{13} (L + \rho) \sqrt{\frac{o}{n}} \sqrt{s \left(r_d + r_\delta \right) + \Sigma_{\max}^2} r_2,$$

where c_{13} *is a numerical constant that depends on* c_{10} *and* c_ε*.*

Lastly, we define $c''_{\max}$.

Definition 50 *Define*

$$c''_{\max} = \max \left(1, c_3, c_6, c_{12} \sqrt{c_{\mathrm{cut}}}, c_{13} \right).$$

The steps to derive Corollaries 48 and 49 from Corollary 47 are almost the same as those used to derive Propositions 29 and 30 from Corollary 28. Therefore, the proofs of Corollaries 48 and 49 are omitted. Additionally, since the proof of Theorem 17 is almost identical to that of Theorem 15, the proof of Theorem 17 is also omitted.

H.1 Proof of Proposition 47

First, we prove (96). This proof is based on a simple union bound. First, we have

$$\begin{aligned} \max_{M \in \mathfrak{M}_{r_1, r_2, d}^{\ell_1, \mathrm{Tr}}} \sum_{i=1}^n \frac{\left\langle \mathbf{x}_i \mathbf{x}_i^\top, M \right\rangle}{n} &\leq \max_{M \in \mathfrak{M}_{r_1, r_2, d}^{\ell_1, \mathrm{Tr}}} \left| \left\langle \sum_{i=1}^n \frac{\mathbf{x}_i \mathbf{x}_i^\top}{n} - \Sigma, M \right\rangle \right| + \max_{M \in \mathfrak{M}_{r_1, r_2, d}^{\ell_1, \mathrm{Tr}}} |\langle \Sigma, M \rangle| \\ &\overset{(a)}{\leq} \left\| \sum_{i=1}^n \frac{\mathbf{x}_i \mathbf{x}_i^\top}{n} - \Sigma \right\|_\infty r_1^2 + \Sigma_{\max}^2 r_2^2, \end{aligned}$$

where (a) follows from Hölder's inequality. We note that the (k, l)-element of $\mathbf{x}_i \mathbf{x}_i^\top$ is $\mathbf{x}_i|_k \times \mathbf{x}_i|_l$ and its mean is $\Sigma_{k,l}$. Then, we have

$$\|\mathbf{x}_i|_k \times \mathbf{x}_i|_l - \Sigma_{k,l}\|_{\psi_1} \lesssim \|\mathbf{x}_i|_k \times \mathbf{x}_i|_l\|_{\psi_1} + \rho^2 \overset{(a)}{\leq} C \|\mathbf{x}_i|_k\|_{\psi_2} \|\mathbf{x}_i|_l\|_{\psi_2} + \rho^2 \overset{(b)}{\leq} C(L^2 + \rho^2),$$

where (a) (b) follows from (9). Then, from Theorem 2.8.1 of Vershynin (2018), we have

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i|_k \times \mathbf{x}_i|_l - \Sigma_{k,l}\right| \leq C(L^2+\rho^2)\left(\sqrt{\frac{\log(1/t)}{n}} + \frac{\log(1/t)}{n}\right)\right) \geq 1-t.$$

Let $t = \delta^2/d^2$ and from union bound, we have

$$\mathbb{P}\left(\left\|\sum_{i=1}^{n}\frac{\mathbf{x}_i\mathbf{x}_i^\top}{n} - \Sigma\right\|_\infty \leq C(L^2+\rho^2)\left(\sqrt{\frac{\log(d/\delta)}{n}} + \frac{\log(d/\delta)}{n}\right)\right) \geq 1-\delta^2.$$

From the fact that $0 < \delta < 1$, $r_1^2/r_2^2 \leq s$ and $\log(d/\delta)/n \leq 1$, we complete the proof of (96). The proof of (97) is almost the same to the one of (40). Therefore, the proof of (97) is omitted.

Appendix I. Proof of Corollary 18

For the proof of Corollary 18, the following Proposition 51 plays an important role, whose proof is in Section 56:

Proposition 51 *Suppose that Assumption 3 holds. Furthermore, suppose that* (31) *holds. Then, for any* $\mathbf{u} \in \mathbb{R}^n$ *such that* $\|\mathbf{u}\|_\infty \leq 2$ *and for any* $\mathbf{v} \in r_1\mathbb{B}_1^d \cap r_\Sigma\mathbb{B}_\Sigma^d$*, we have the following:*

$$\left|\sum_{i\in\mathcal{O}}\frac{1}{n}u_i\mathbf{x}_i^\top\mathbf{v}\right| \leq c_{14}L\left(\rho r_{d,s}r_1 + (r_\delta + r_o)r_\Sigma\right),$$

where c_{14} *is a numerical constant.*

Additionally, we define $c'''_{\max}$.

Definition 52 *Define*

$$c'''_{\max} = \max\left(1, c_3, c_6, c_{14}\right).$$

Then we proceed to the proof of Corollary 18. The proof is similar to those of Theorems 13 and 15. To prove Corollary 18, it is sufficient to confirm Proposition 6.2 with $\{\mathbf{X}\}_{i=1}^n = \{\mathbf{x}\}_{i=1}^n$ and $\{\hat{w}_i'\}_{i=1}^n = \{1/n\}_{i=1}^n$, in other words, we will confirm (98), (99), (100) and (101) in the following Proposition 53 under the assumptions in Corollary 18.

Proposition 53 *Suppose that, for any* $\boldsymbol{\theta}_\eta \in r_1\mathbb{B}_1^d \cap r_2\mathbb{B}_2^d \cap r_\Sigma\mathbb{B}_\Sigma^d$*,*

$$\left|\lambda_o\sqrt{n}\frac{1}{n}\sum_{i=1}^{n}h(r_{\boldsymbol{\beta}^*,i})\mathbf{x}_i^\top\boldsymbol{\theta}_\eta\right| \leq r_{a,1}r_1 + r_{a,2}r_2 + r_{a,\Sigma}r_\Sigma, \tag{98}$$

$$b\|\Sigma^{\frac{1}{2}}\boldsymbol{\theta}_\eta\|_2^2 - r_{b,2}r_2 - r_{b,\Sigma}r_\Sigma - r_{b,1}r_1 \leq \frac{1}{n}\sum_{i=1}^{n}\lambda_o\sqrt{n}\left(-h(r_{\boldsymbol{\beta}^*+\boldsymbol{\theta}_\eta,i}) + h(r_{\boldsymbol{\beta}^*,i})\right)\mathbf{x}_i^\top\boldsymbol{\theta}_\eta, \tag{99}$$

where $r_{a,1}, r_{a,2}, r_{a,\Sigma}, r_{b,1}, r_{b,2}, r_{b,\Sigma} \geq 0$, $b > 0$ *are some numbers. Suppose that* $\mathbb{E}\mathbf{x}_i\mathbf{x}_i^\top = \Sigma$ *satisfies* $\mathrm{RE}(s, c_{\mathrm{RE}}, \mathfrak{r})$, $\kappa_{\mathrm{l}} > 0$, *and*

$$\lambda_s - \left(r_{a,1} + \frac{c_{r_2} r_{a,2} + r_{a,\Sigma}}{c_{r_1}\sqrt{s}}\right) > 0, \quad \frac{\lambda_s + \left(r_{a,1} + \frac{c_{r_2} r_{a,2} + r_{a,\Sigma}}{c_{r_1}\sqrt{s}}\right)}{\lambda_s - \left(r_{a,1} + \frac{c_{r_2} r_{a,2} + r_{a,\Sigma}}{c_{r_1}\sqrt{s}}\right)} \leq c_{\mathrm{RE}}, \tag{100}$$

$$r_\Sigma \geq \frac{2}{b}\left(c_{r_1}\sqrt{s}(r_{a,1} + r_{b,1}) + c_{r_2}(r_{a,2} + r_{b,2}) + r_{a,\Sigma} + r_{b,\Sigma} + c_{r_1}\sqrt{s}\lambda_s\right), \quad r_1 = c_{r_1}\sqrt{s} r_\Sigma, \quad r_2 = c_{r_2} r_\Sigma \tag{101}$$

hold, where $c_{r_1} = c_{r_1}^{\mathrm{num}}(1 + c_{\mathrm{RE}})/\mathfrak{r}$, $c_{r_2} = c_{r_2}^{\mathrm{num}}(3 + c_{\mathrm{RE}})/\kappa_{\mathrm{l}}$, $\min\{c_{r_1}^{\mathrm{num}}, c_{r_2}^{\mathrm{num}}\} \geq 2$ *and* $c_{r_1}/c_{r_2} \leq 1$. *Then, we have the following:*

$$\|\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}\|_1 \leq r_1, \quad \|\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}\|_2 \leq r_2, \quad \|\Sigma^{\frac{1}{2}}(\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}})\|_2 \leq r_\Sigma.$$

I.1 Confirmation of (33)

From the following lemma, we can confirm (33). The proof is given in Appendix G.

Lemma 54 *Assume that Propositions 20 and 51 hold. For any* $\boldsymbol{\theta}_\eta \in r_1\mathbb{B}_1^d \cap r_2\mathbb{B}_2^d \cap r_\Sigma\mathbb{B}_\Sigma^d$, *we have*

$$\left|\frac{1}{n}\sum_{i=1}^n h(r_{\boldsymbol{\beta}^*,i})\mathbf{x}_i^\top\boldsymbol{\theta}_\eta\right| \leq 2c_{\max}'''^2 L\left(\rho r_{d,s} r_1 + (r_\delta + r_o) r_\Sigma\right).$$

From Lemma 54, we see that (33) holds with

$$r_{a,1} = 2c_{\max}'''^2\lambda_o\sqrt{n}L\rho r_{d,s}, \quad r_{a,2} = 0, \quad r_{a,\Sigma} = 2c_{\max}'''^2\lambda_o\sqrt{n}L(r_\delta + r_o). \tag{102}$$

I.2 Confirmation of (35)

From (102),

$$(C_s :=) r_{a,1} + \frac{c_{r_2} r_{a,2} + r_{a,\Sigma}}{c_{r_1}\sqrt{s}} = 2c_{\max}'''^2 L\lambda_o\sqrt{n} \times \frac{1}{c_{r_1}\sqrt{s}} R_{d,n,o}'''.$$

From the definition of λ_s,

$$\frac{\lambda_s}{C_s} \geq \frac{c_s c_{\max}'''^2 L\lambda_o\sqrt{n}\frac{1}{c_{r_1}\sqrt{s}}R_{d,n,o}'''}{2c_{\max}'''^2 L\lambda_o\sqrt{n}\frac{1}{c_{r_1}\sqrt{s}}R_{d,n,o}'''} = \frac{c_s}{2} \geq \frac{c_{\mathrm{RE}} + 1}{c_{\mathrm{RE}} - 1} > 0.$$

Hence, we have $\lambda_s - C_s > 0$ and

$$\frac{\lambda_s + C_s}{\lambda_s - C_s} = \frac{1 + \frac{C_s}{\lambda_s}}{1 - \frac{C_s}{\lambda_s}} \leq \frac{1 + \frac{c_{\mathrm{RE}}-1}{c_{\mathrm{RE}}+1}}{1 - \frac{c_{\mathrm{RE}}-1}{c_{\mathrm{RE}}+1}} = c_{\mathrm{RE}}.$$

Therefore, we see that (35) holds.

I.3 Confirmation of (34)

From the following lemma, we can confirm (34). The proof is given in Appendix I.6.

Lemma 55 *Assume that Propositions 25 and 51 hold. For any $\boldsymbol{\theta}_\eta \in r_1\mathbb{B}_1^d \cap r_2\mathbb{B}_2^d \cap r_\Sigma\mathbb{B}_\Sigma^d$, we have*

$$\frac{1}{n}\sum_{i=1}^n \lambda_o\sqrt{n}\left(-h(r_{\boldsymbol{\beta}^*+\boldsymbol{\theta}_\eta,i}) + h(r_{\boldsymbol{\beta}^*,i})\right)\mathbf{x}_i^\top\boldsymbol{\theta}_\eta \geq \frac{\|\Sigma^{\frac{1}{2}}\mathbf{v}\|_2^2}{3} - 3c'''_{\max}\lambda_o\sqrt{n}L\left(\rho r_{d,s}r_1 + (r_\delta + r_o)r_\Sigma\right).$$

From Lemma 55, we see that (34) holds with

$$b = \frac{1}{3}, \quad r_{b,1} = \frac{3}{2}r_{a,1}, \quad r_{b,2} = 0, \quad r_{b,\Sigma} = \frac{3}{2}r_{a,\Sigma}.$$

I.4 Confirmation of (36)

As we mentioned at the beginning of Appendix A, $r_1 = c_{r_1}\sqrt{s}r_\Sigma$ and $r_2 = c_{r_2}r_\Sigma$ is clear from their definitions. We confirm r_Σ. We see that

$$\begin{aligned}&\frac{2}{b}\left(c_{r_1}\sqrt{s}(r_{a,1}+r_{b,1}) + c_{r_2}(r_{a,2}+r_{b,2}) + r_{a,\Sigma} + r_{b,\Sigma} + c_{r_1}\sqrt{s}\lambda_s\right)\\ &\leq 15c'''^2_{\max}L\lambda_o\sqrt{n}\left(2+c_s\right)R'''_{d,n,o} \leq r_\Sigma,\end{aligned}$$

and the proof is complete.

I.5 Proof of Proposition 51

Before proving Proposition 51, we prove the following proposition:

Proposition 56 *Suppose that Assumption 3 holds. Then, with probability at least $1-t$, we have*

$$\sup_{\mathbf{v}\in r_1\mathbb{B}_1^d\cap r_\Sigma\mathbb{B}_\Sigma^d}\left|\sum_{i=1}^n\frac{(\mathbf{x}_i^\top\mathbf{v})^2}{n} - \mathbb{E}(\mathbf{x}_i^\top\mathbf{v})^2\right| \leq CL^2\left(\rho^2\frac{\log(d/s)}{n}r_1^2 + \rho\sqrt{\frac{\log(d/s)}{n}}r_\Sigma r_1 + \sqrt{\frac{t}{n}}r_\Sigma^2 + \frac{t}{n}r_\Sigma^2\right). \tag{103}$$

Proof This proof heavily relies on Theorem 5.5 of Dirksen (2015). To apply the theorem, we note that, from (9), we have the following four inequalities:

$$\begin{aligned}&\max_{i=1,\cdots,n}\|\mathbf{x}_i^\top(\mathbf{v}-\mathbf{v}')\|_{\psi_2} \leq L\|\Sigma^{\frac{1}{2}}(\mathbf{v}-\mathbf{v}')\|_2,\\ &\sup_{\mathbf{v}\in r_1\mathbb{B}_1^d\cap r_\Sigma\mathbb{B}_\Sigma^d}\max_{i=1,\cdots,n}\|\mathbf{x}_i^\top\mathbf{v}\|_{\psi_2} \leq \sup_{\mathbf{v}\in r_1\mathbb{B}_1^d\cap r_\Sigma\mathbb{B}_\Sigma^d}L\|\Sigma^{\frac{1}{2}}\mathbf{v}\|_2 \leq Lr_\Sigma,\\ &\sup_{\mathbf{v}\in r_1\mathbb{B}_1^d\cap r_\Sigma\mathbb{B}_\Sigma^d}\left(\frac{1}{n}\sum_{i=1}^n\|\mathbf{x}_i^\top\mathbf{v}\|_{\psi_2}^4\right)^{1/2} \leq \sup_{\mathbf{v}\in r_1\mathbb{B}_1^d\cap r_\Sigma\mathbb{B}_\Sigma^d}\left(L^4\|\Sigma^{\frac{1}{2}}\mathbf{v}\|_2^4\right)^{1/2} \leq L^2r_\Sigma^2,\\ &\sup_{\mathbf{v}\in r_1\mathbb{B}_1^d\cap r_\Sigma\mathbb{B}_\Sigma^d}\max_{i=1,\cdots,n}\|\mathbf{x}_i^\top\mathbf{v}\|_{\psi_2}^2 \leq \sup_{\mathbf{v}\in r_1\mathbb{B}_1^d\cap r_\Sigma\mathbb{B}_\Sigma^d}\|\Sigma^{\frac{1}{2}}\mathbf{v}\|_2^2 \leq L^2r_\Sigma^2.\end{aligned}$$

Then, from the inequalities above, applying Theorem 5.5 of Dirksen (2015), with probability at least $1-t$, we have

$$\begin{aligned}
&\sup_{\mathbf{v}\in r_1\mathbb{B}_1^d\cap r_\Sigma\mathbb{B}_\Sigma^d}\left|\frac{1}{n}\sum_{i=1}^n(\mathbf{x}_i^\top\mathbf{v})^2-\mathbb{E}(\mathbf{x}_i^\top\mathbf{v})^2\right| \\
&\leq CL^2\left(\frac{\gamma_2^2(r_1\mathbb{B}_1^d\cap r_\Sigma\mathbb{B}_\Sigma^d,\|\Sigma^{\frac{1}{2}}(\cdot)\|_2)}{n}+\frac{\gamma_2(r_1\mathbb{B}_1^d\cap r_\Sigma\mathbb{B}_\Sigma^d,\|\Sigma^{\frac{1}{2}}(\cdot)\|_2)}{\sqrt{n}}r_\Sigma+\sqrt{\frac{t}{n}}r_\Sigma^2+\frac{t}{n}r_\Sigma^2\right) \\
&\overset{(a)}{\leq} CL^2\left(\frac{\left(\sup_{\mathbf{v}\in r_1\mathbb{B}_1^d\cap r_\Sigma\mathbb{B}_\Sigma^d}\langle\mathbf{v},\Sigma^{\frac{1}{2}}\mathbf{g}\rangle\right)^2}{n}+\frac{\sup_{\mathbf{v}\in r_1\mathbb{B}_1^d\cap r_\Sigma\mathbb{B}_\Sigma^d}\langle\mathbf{v},\Sigma^{\frac{1}{2}}\mathbf{g}\rangle}{\sqrt{n}}r_\Sigma+\sqrt{\frac{t}{n}}r_\Sigma^2+\frac{t}{n}r_\Sigma^2\right) \\
&\overset{(b)}{\leq} CL^2\left(\rho^2\frac{\log(d/s)}{n}r_1^2+\rho\sqrt{\frac{\log(d/s)}{n}}r_\Sigma r_1+\sqrt{\frac{t}{n}}r_\Sigma^2+\frac{t}{n}r_\Sigma^2\right),
\end{aligned}$$

where (a) from the majorizing measure theorem (Theorem 2.4.1. of Talagrand (2014)), and (b) follows from Lemma 34. ■

Then, we proceed to the proof of Proposition 51. We start from a simple algebra:

$$\begin{aligned}
\sup_{\mathbf{v}\in r_1\mathbb{B}_1^d\cap r_\Sigma\mathbb{B}_\Sigma^d}\left|\sum_{i\in\mathcal{O}}\frac{1}{n}u_i\mathbf{x}_i^\top\mathbf{v}\right| &\overset{(a)}{\leq} \sup_{\mathbf{v}\in r_1\mathbb{B}_1^d\cap r_\Sigma\mathbb{B}_\Sigma^d}\sqrt{\frac{4o}{n}}\sqrt{\frac{1}{n}\sum_{i\in\mathcal{O}}\left((\mathbf{x}_i^\top\mathbf{v})^2-\mathbb{E}(\mathbf{x}_i^\top\mathbf{v})^2\right)+\frac{4o}{n}\mathbb{E}(\mathbf{x}_i^\top\mathbf{v})^2} \\
&\overset{(b)}{\leq} \sup_{\mathbf{v}\in r_1\mathbb{B}_1^d\cap r_\Sigma\mathbb{B}_\Sigma^d}\sqrt{\frac{4o}{n}}\sqrt{\frac{1}{n}\sum_{i\in\mathcal{O}}\left((\mathbf{x}_i^\top\mathbf{v})^2-\mathbb{E}(\mathbf{x}_i^\top\mathbf{v})^2\right)}+2\frac{o}{n}r_\Sigma, \quad (104)
\end{aligned}$$

where (a) follows from Hölder's inequality and $\|\mathbf{u}\|_\infty\leq 2$, and (b) follows from triangular inequality. Then, we focus on evaluating $\sup_{\mathbf{v}\in r_1\mathbb{B}_1^d\cap r_\Sigma\mathbb{B}_\Sigma^d}\sum_{i\in\mathcal{O}}\frac{(\mathbf{x}_i^\top\mathbf{v})^2}{n}-\mathbb{E}(\mathbf{x}_i^\top\mathbf{v})^2$, in a manner similar to the proof of (32). From (103), we have

$$\max_{|\mathcal{J}|=o}\sup_{\mathbf{v}\in r_1\mathbb{B}_1^d\cap r_\Sigma\mathbb{B}_\Sigma^d}\left|\sum_{i\in\mathcal{J}}\frac{(\mathbf{x}_i^\top\mathbf{v})^2}{o}-\mathbb{E}(\mathbf{x}_i^\top\mathbf{v})^2\right|\geq CL^2\left(\rho^2\frac{\log(d/s)}{o}r_1^2+\rho\sqrt{\frac{\log(d/s)}{o}}r_\Sigma r_1+\sqrt{\frac{t}{o}}r_\Sigma^2+\frac{t}{o}r_\Sigma^2\right), \quad (105)$$

with probability at most

$$\begin{aligned}
&\leq\binom{n}{o}\times \\
&\mathbb{P}\left[\sup_{\mathbf{v}\in r_1\mathbb{B}_1^d\cap r_\Sigma\mathbb{B}_\Sigma^d}\left|\sum_{i\in\mathcal{J}}\frac{(\mathbf{z}_i^\top\mathbf{v})^2}{o}-\mathbb{E}(\mathbf{z}_i^\top\mathbf{v})^2\right|\geq CL^2\left(\rho^2\frac{\log(d/s)}{o}r_1^2+\rho\sqrt{\frac{\log(d/s)}{o}}r_\Sigma r_1+\sqrt{\frac{t}{o}}r_\Sigma^2+\frac{t}{o}r_\Sigma^2\right)\right] \\
&\leq\binom{n}{o}e^{-t},
\end{aligned}$$

where $\{\mathbf{z}_i\}_{i=1}^o$ is a sequence of i.i.d. random vectors sampled from the same distribution as $\{\mathbf{x}_i\}_{i=1}^n$. Let $t=o\log(ne/o)+\log(1/\delta)$. From Stirling's formula, we have $\binom{n}{o}e^{-t}\leq\delta$. From

(105), with probability at least $1-\delta$, we have

$$\begin{aligned}
&\max_{|\mathcal{J}|=o} \sup_{\mathbf{v}\in r_1\mathbb{B}_1^d \cap r_\Sigma \mathbb{B}_\Sigma^d} \left| \sum_{i\in\mathcal{J}} \frac{(\mathbf{x}_i^\top \mathbf{v})^2}{n} - \mathbb{E}(\mathbf{x}_i^\top \mathbf{v})^2 \right| \\
&\leq C\frac{o}{n}L^2 \left(\rho^2 \frac{\log(d/s)}{o} r_1^2 + \rho\sqrt{\frac{\log(d/s)}{o}} r_\Sigma r_1 + \sqrt{\frac{o\log\frac{ne}{o} + \log(1/\delta)}{o}} r_\Sigma + \frac{o\log\frac{ne}{o}+\log(1/\delta)}{o} r_\Sigma^2 \right) \\
&\overset{(a)}{\leq} CL^2 \left(\rho^2 \frac{\log(d/s)}{n} r_1^2 + \rho\sqrt{\frac{o}{n}}\sqrt{\frac{\log(d/s)}{n}} r_\Sigma r_1 + \left(\frac{\log(1/\delta)}{n} + \frac{o}{n}\log\frac{n}{o} \right) r_\Sigma^2 \right) \\
&\overset{(b)}{\leq} CL^2 \left(\rho^2 \frac{\log(d/s)}{n} r_1^2 + \left(\frac{\log(1/\delta)}{n} + \frac{o}{n}\log\frac{n}{o} \right) r_\Sigma^2 \right), \qquad (106)
\end{aligned}$$

where (a) follows from $e \leq n/o$ and $r_\delta \leq 1$, and (b) follows from Young's inequality and again, $e \leq n/o$. From (104) and (106), with probability at least $1-\delta$, we have

$$\begin{aligned}
\sup_{\mathbf{v}\in r_1\mathbb{B}_1^d \cap r_\Sigma \mathbb{B}_\Sigma^d} \left| \sum_{i\in\mathcal{J}} \frac{1}{n} u_i \mathbf{x}_i^\top \mathbf{v} \right| &\leq \sqrt{\frac{4o}{n}} \sqrt{CL^2 \left(\rho^2 \frac{\log(d/s)}{n} r_1^2 + \left(\frac{\log(1/\delta)}{n} + \frac{o}{n}\log\frac{n}{o} \right) r_\Sigma^2 \right)} + 2\frac{o}{n} r_\Sigma \\
&\overset{(a)}{\leq} CL \left(\rho\sqrt{\frac{\log(d/s)}{n}} r_1 + \sqrt{\frac{\log(1/\delta)}{n}} r_\Sigma + \frac{o}{n}\sqrt{\log\frac{n}{o}} r_\Sigma \right),
\end{aligned}$$

where (a) follows from Young's inequality and $e \leq n/o$ and $L \geq 1$.

I.6 Proof of Lemmas 54 and 55

Proof [Proof of Lemma 54] From the triangular inequality, we have

$$\begin{aligned}
\left| \frac{1}{n}\sum_{i=1}^n h(r_{\boldsymbol{\beta}^*,i}) \mathbf{x}_i^\top \boldsymbol{\theta}_\eta \right| &\leq \left| \frac{1}{n}\sum_{i\in\mathcal{I}} h(r_{\boldsymbol{\beta}^*,i}) \mathbf{x}_i^\top \boldsymbol{\theta}_\eta \right| + \left| \frac{1}{n}\sum_{i\in\mathcal{O}} h(r_{\boldsymbol{\beta}^*,i}) \mathbf{x}_i^\top \boldsymbol{\theta}_\eta \right| \\
&\leq \left| \sum_{i=1}^n \frac{1}{n} h(\xi_{\lambda_o,i}) \mathbf{x}_i^\top \boldsymbol{\theta}_\eta \right| + 2\left| \frac{1}{n}\sum_{i\in\mathcal{O}} h(\xi_{\lambda_o,i}) \mathbf{x}_i^\top \boldsymbol{\theta}_\eta \right|.
\end{aligned}$$

We note that $|h(\cdot)| \leq 1$. From Propositions 22 and 56, we have

$$\begin{aligned}
\left| \sum_{i=1}^n \frac{1}{n} h(\xi_{\lambda_o,i}) \mathbf{x}_i^\top \boldsymbol{\theta}_\eta \right| &\leq c_3 L \left(\rho r_{d,s} r_1 + r_\delta r_\Sigma \right), \\
\left| \sum_{i\in\mathcal{O}} \frac{1}{n} h(\xi_{\lambda_o,i}) \mathbf{x}_i^\top \boldsymbol{\theta}_\eta \right| &\leq c_{14} L \left(\rho r_{d,s} r_1 + (r_\delta + r_o) r_\Sigma \right),
\end{aligned}$$

and we see that

$$\left| \frac{1}{n}\sum_{i=1}^n h(r_{\boldsymbol{\beta}^*,i}) \mathbf{x}_i^\top \boldsymbol{\theta}_\eta \right| \leq 2c_{\max}'^2 L \left(\rho r_{d,s} r_1 + (r_\delta + r_o) r_\Sigma \right).$$

■

Proof [Proof of Lemma 55] We have

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\lambda_o\sqrt{n}\left(-h(r_{\boldsymbol{\beta}^*+\boldsymbol{\theta}_\eta,i})+h(r_{\boldsymbol{\beta}^*,i})\right)\mathbf{x}_i^\top\boldsymbol{\theta}_\eta \\
&=\frac{1}{n}\sum_{i\in\mathcal{I}}\lambda_o\sqrt{n}\left(-h(r_{\boldsymbol{\beta}^*+\boldsymbol{\theta}_\eta,i})+h(r_{\boldsymbol{\beta}^*,i})\right)\mathbf{x}_i^\top\boldsymbol{\theta}_\eta+\frac{1}{n}\sum_{i\in\mathcal{O}}\lambda_o\sqrt{n}\left(-h(r_{\boldsymbol{\beta}^*+\boldsymbol{\theta}_\eta,i})+h(r_{\boldsymbol{\beta}^*,i})\right)\mathbf{x}_i^\top\boldsymbol{\theta}_\eta \\
&\geq\sum_{i=1}^{n}\frac{\lambda_o}{\sqrt{n}}\left(-h(\xi_{\lambda_o,i}-x_{\boldsymbol{\theta}_\eta,i})+h(\xi_{\lambda_o,i})\right)\mathbf{x}_i^\top\boldsymbol{\theta}_\eta-2\left|\frac{1}{n}\sum_{i\in\mathcal{O}}\lambda_o\sqrt{n}\left(-h(r_{\boldsymbol{\beta}^*+\boldsymbol{\theta}_\eta,i})+h(r_{\boldsymbol{\beta}^*,i})\right)\mathbf{x}_i^\top\boldsymbol{\theta}_\eta\right|.
\end{aligned}$$

We note that $|h(\cdot)|\leq 1$ From Propositions 25 and 56, we have

$$\sum_{i=1}^{n}\frac{\lambda_o}{\sqrt{n}}\left(-h(\xi_{\lambda_o,i}-x_{\boldsymbol{\theta}_\eta,i})+h(\xi_{\lambda_o,i})\right)\mathbf{x}_i^\top\boldsymbol{\theta}_\eta\geq\frac{\|\Sigma^{\frac{1}{2}}\mathbf{v}\|_2^2}{3}r_2^2-c'''_{\max}L\lambda_o\sqrt{n}\left(\rho r_{d,s}r_1+r_\delta r_\Sigma\right)$$

$$\left|\frac{1}{n}\sum_{i\in\mathcal{O}}\lambda_o\sqrt{n}\hat{w}_i'\left(-h(r_{\boldsymbol{\beta}^*+\boldsymbol{\theta}_\eta,i})+h(r_{\boldsymbol{\beta}^*,i})\right)\mathbf{x}_i^\top\boldsymbol{\theta}_\eta\right|\leq c'''_{\max}L\left(\rho r_{d,s}r_1+(r_\delta+r_o)r_\Sigma\right).$$

Combining the arguments above, we see that

$$\frac{1}{n}\sum_{i=1}^{n}\lambda_o\sqrt{n}\left(-h(r_{\boldsymbol{\beta}^*+\boldsymbol{\theta}_\eta,i})+h(r_{\boldsymbol{\beta}^*,i})\right)\mathbf{x}_i^\top\boldsymbol{\theta}_\eta\geq\frac{\|\Sigma^{\frac{1}{2}}\mathbf{v}\|_2^2}{3}-3c'''_{\max}L\left(\rho r_{d,s}r_1+(r_\delta+r_o)r_\Sigma\right),$$

and the proof is complete. ■

Appendix J. Discussion

In this section, we discuss the technical contributions of our paper, as mentioned in Section 2.3, as well as the adaptive setting of the tuning parameters, as mentioned in Remark 19.

J.1 About the Analysis of the ℓ_1-Penalized Huber Loss

There are two primary approaches known for analyzing the ℓ_1-penalized Huber loss. The first approach is a direct analysis, where the estimator $\hat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}^*$ is defined as

$$\hat{\boldsymbol{\beta}}=\underset{\boldsymbol{\beta}\in\mathbb{R}^d}{\operatorname{argmin}}\sum_{i=1}^{n}\lambda_o^2H\left(\frac{y_i-\mathbf{X}_i^\top\boldsymbol{\beta}}{\lambda_o\sqrt{n}}\right)+\lambda_s\|\boldsymbol{\beta}\|_1.$$

The second approach involves analyzing the following optimization problem, which provides the same solution for $\hat{\boldsymbol{\beta}}$ as the one above

$$(\hat{\boldsymbol{\beta}},\hat{\mathbf{o}})=\underset{\boldsymbol{\beta}\in\mathbb{R}^d,\boldsymbol{\theta}\in\mathbb{R}^n}{\operatorname{argmin}}\frac{1}{2n}\sum_{i=1}^{n}(y_i-\langle\mathbf{X}_i,\boldsymbol{\beta}-\sqrt{n}\mathbf{o}\rangle)^2+\lambda_o\|\mathbf{o}\|_1+\lambda_s\|\boldsymbol{\beta}\|_1. \tag{107}$$

For the equivalence of these two optimization problems, refer to She and Owen (2011).

In prior work, several studies have applied the ℓ_1-penalized Huber loss in situations where outliers are present. These include Nguyen and Tran (2012); Dalalyan and Thompson (2019); Thompson (2020); Minsker et al. (2024), all of which adopt the latter approach. These papers address cases where only the output is contaminated by outliers. However, it is challenging to extend this approach to situations where outliers also affect the input because the approach requires evaluating the following quantity:

$$\sup_{\mathbf{v}\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n}\langle \mathbf{x}_i + \boldsymbol{\varrho}_i, \mathbf{v}\rangle \times (\xi_i + \boldsymbol{\theta}_i). \tag{108}$$

In the case where $\boldsymbol{\varrho}_i = 0$ for all $i = 1, \cdots, n$, that is, when only the output is contaminated by outliers, this can be effectively evaluated by leveraging the fact that $\{\mathbf{x}_i\}_{i=1}^n$ satisfies a variant of Chevet's inequality (Proposition 4 of Dalalyan and Thompson (2019)) and combining this with the penalization on $\boldsymbol{\theta}$ in the optimization problem (107). However, when the input is contaminated, the variant of Chevet's inequality is not applicable effectively, making the evaluation much more difficult.

Therefore, in this paper, we adopt the former approach. The former approach is also employed in papers such as Pensia et al. (2020). Although Pensia et al. (2020) deals with the case where outliers are present in both the input and output, it does not use an ℓ_1 penalty since $\boldsymbol{\beta}^*$ is not sparse. Notably, the analysis of the Huber loss using the former approach is also adopted in papers such as Sun et al. (2020) and Chen and Zhou (2020). These papers prove an error bound for $\|\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}\|_2$ by assuming that any large value of $\|\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}\|_2$ would lead to a contradiction. In this approach, the quantity corresponding to (108) in the latter approach changes as follows. With a appropriately set to a positive number:

$$\sup_{\mathbf{v}\in a\mathbb{B}_2^d} \frac{\lambda_o\sqrt{n}}{n}\sum_{i=1}^{n}\langle \mathbf{x}_i + \boldsymbol{\varrho}_i, \mathbf{v}\rangle \times h\left(\frac{\xi_i + \boldsymbol{\theta}_i}{\lambda_o\sqrt{n}}\right).$$

There are two advantages to this approach: the first is the restriction on the range of $\mathbf{v}$, and the second is that $\xi_i + \boldsymbol{\theta}_i$ is bounded by the derivative of the Huber loss $h(\cdot)$. In Pensia et al. (2020), when analyzing the above quantity, the fact that $-1 \le h(\cdot) \le 1$ is used to leverage existing algorithms for outlier-robust mean estimation.

However, this approach is insufficient for our purposes because, to fully exploit sparsity, $\mathbf{v}$ needs to be constrained not only in terms of its ℓ_2 norm, $\mathbf{v} \in a\mathbb{B}_2^d$ but also in terms of its ℓ_1 norm. The technique of proceeding under the assumption that both the ℓ_1 and ℓ_2 norms are small has been developed in works such as Lugosi and Mendelson (2019); Alquier et al. (2019); Lecué and Lerasle (2020). However, these papers consider only the case where the covariance matrix is the identity and do not address cases where the covariance matrix has a general form. If we forcibly apply this method to our setting, the final predictive error bound ($\|\Sigma^{\frac{1}{2}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_2$) would involve the inverse of the smallest eigenvalue of the covariance matrix. In contrast, for standard Lasso, the predictive error typically depends on the inverse of the restricted eigenvalue (Definition 5), which is generally a larger value than the smallest eigenvalue. Given the recent interest in predictive error in the context of overparameterization (for example, Bartlett et al. (2020); Tsigler and Bartlett (2023); Cheng and Montanari (2022)), we believe that it is important to produce predictive error

bounds. To this end, we have modified the methods from previous research and developed an approach that allows us to control $\mathbf{v}$ not only with respect to the ℓ_1 and ℓ_2 norms but also the predictive norm. As a result, we have obtained results consistent with those of standard Lasso without outliers. Specifically, we analyze the following expression, where a, b, c are appropriately chosen positive constants:

$$\sup_{\mathbf{v}\in a\mathbb{B}_2^d\cap b\mathbb{B}_1^d\cap c\mathbb{B}_\Sigma^d} \frac{\lambda_o\sqrt{n}}{n}\sum_{i=1}^n \langle \mathbf{x}_i+\boldsymbol{\varrho}_i, \mathbf{v}\rangle \times h\left(\frac{\xi_i+\boldsymbol{\theta}_i}{\lambda_o\sqrt{n}}\right).$$

This is discussed in Proposition 21. While Proposition 21 is specialized for the ℓ_1-penalized Huber loss, we believe that with techniques from works such as Pan and Zhou (2021); Chinot et al. (2020), this approach can be extended to other loss functions.

J.2 About Adaptive Tuning of the Tuning Parameter

In this section, we discuss the adaptive setting of the tuning parameter in Corollary 18. For simplicity, we assume that $\mathbb{E}\xi_i^2 \le \sigma^2$ holds and (15) is satisfied. Under these assumptions, the result of Theorem 18 is given as follows: When (30) holds and we set

$$24L^2\sigma \le \lambda_o\sqrt{n}, \quad c_s c_{\max}'''^2 L\lambda_o\sqrt{n}\frac{1}{c_{r_1}\sqrt{s}}\left(\rho c_{r_1}\sqrt{s}r_{d,s}+r_\delta+r_o\right) \le \lambda_s,$$

then, we have

$$\begin{aligned}\|\Sigma^{\frac{1}{2}}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)\|_2 &\le 15\,(2+c_s)\,c_{\max}'''^2 L\lambda_o\sqrt{n}R_{d,n,o}''' \\ &= \frac{15\,(2+c_s)}{c_s}c_{r_1}\sqrt{s}\times\lambda_s. \end{aligned} \tag{109}$$

The tuning parameters λ_o and λ_s contain many unknown quantities, first, let us assume that only o/n is unknown. We set $\lambda_o\sqrt{n} = 24L^2\sigma$, $c_s = 3(c_{\mathrm{RE}}+1)/(c_{\mathrm{RE}}-1)$, and $c_{r_1}^{\mathrm{num}} = c_{r_2}^{\mathrm{num}} = 2$. Here, the conditions for λ_s and the error bound of (109) are organized as follows:

$$\begin{aligned}(F(o/n) :=) c_s c_{\max}'''^2 L\lambda_o\sqrt{n}\frac{1}{c_{r_1}\sqrt{s}}\left(\rho c_{r_1}\sqrt{s}r_{d,s}+r_\delta+r_o\right) &\le \lambda_s, \\ \|\Sigma^{\frac{1}{2}}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)\|_2 \le \frac{15\,(2+c_s)}{c_s}c_{r_1}\sqrt{s}\times\lambda_s &:= G(o/n).\end{aligned}$$

Here, note that both $F(\cdot)$ and $G(\cdot)$ are monotonic increasing functions of o/n. In other words, if o/n can be overestimated, the setting for λ_s will be successful; however, if the estimate is too lenient, the error bound will become considerably loose. In this case, Lepskii's method Lepskii (1992); Lepski and Spokoiny (1997) can be used to adaptively tune λ_s. To do this, we consider applying Lemma 3 from Jain et al. (2022) here. Define $a > 1$ and

$$\varepsilon_{\max} = \max\left\{1/e, \operatorname*{argmax}_{o/n}(1 \ge 7c_o\,(4+c_s)\,c_{\max}'''^2\sqrt{1+\log L}L^2R_{d,n,o}''')\right\}.$$

Let $K \in \mathbb{N}$ be a positive integer and we construct an increasing sequence $\{a_1, \cdots, a_{K+1}\}$, where $a_i = \varepsilon_{\max}/a^{K+1-i}$ for $i = 1, \cdots, K+1$. For a_i, $i \in \{1, \cdots, K+1\}$, define $\hat{\boldsymbol{\beta}}_{a_i}$ as an estimator with tuning parameter

$$\lambda_s = F(a_i).$$

We then define an adaptive estimator, $\hat{\boldsymbol{\beta}}_{\mathrm{ada}}$, as $\hat{\boldsymbol{\beta}}_{a_{k_{\mathrm{ada}}}}$, where for all $k_{\mathrm{ada}} \leq k$, the following condition holds:

$$\|\Sigma^{\frac{1}{2}}(\hat{\boldsymbol{\beta}}_{a_{k_{\mathrm{ada}}}} - \hat{\boldsymbol{\beta}}_{a_k})\|_2 \leq G(a_{k_{\mathrm{ada}}}) + G(a_k). \tag{110}$$

We will explain that $\hat{\boldsymbol{\beta}}_{\mathrm{ada}}$ serves as a good estimator of $\boldsymbol{\beta}^*$, following the proof of Lemma 3 in Jain et al. (2022).

Consider an index $k_{o/n} \in 1, \ldots, l$ such that $a_{k_{o/n}} \leq o/n \leq a_{k_{o/n}+1}$. For any $k \geq k_{o/n}+1$, it follows with probability at least $1 - 3K\delta$ that

$$\|\Sigma^{1/2}(\hat{\boldsymbol{\beta}}_{a_k} - \boldsymbol{\beta}^*)\|_2 \leq G(a_k).$$

Applying the triangle inequality, we have

$$\|\Sigma^{1/2}(\hat{\boldsymbol{\beta}}_{a_k} - \hat{\boldsymbol{\beta}}_{a_{k_{o/n}+1}})\|_2 \leq \|\Sigma^{1/2}(\hat{\boldsymbol{\beta}}_{a_k} - \boldsymbol{\beta}^*)\|_2 + \|\Sigma^{1/2}(\hat{\boldsymbol{\beta}}_{a_{k_{o/n}+1}} - \boldsymbol{\beta}^*)\|_2 \leq G(a_k) + G(a_{k_{o/n}+1}),$$

where the last inequality holds with probability at least $1-3K\delta$. Therefore, by the definition of $\hat{\boldsymbol{\beta}}_{\mathrm{ada}}$, we conclude that $k_{\mathrm{ada}} \leq k_{o/n} + 1$ with probability at least $1 - 3K\delta$. Again, from the triangle inequality, we have

$$\begin{aligned}
\|\Sigma^{\frac{1}{2}}(\hat{\boldsymbol{\beta}}_{\mathrm{ada}} - \boldsymbol{\beta}^*)\|_2 &\leq \|\Sigma^{\frac{1}{2}}(\hat{\boldsymbol{\beta}}_{\mathrm{ada}} - \hat{\boldsymbol{\beta}}_{a_{k_{o/n}-1}})\|_2 + \|\Sigma^{\frac{1}{2}}(\hat{\boldsymbol{\beta}}_{a_{k_{o/n}-1}} - \boldsymbol{\beta}^*)\|_2 \\
&\overset{(a)}{\leq} G(a_{k_{\mathrm{ada}}}) + G(a_{k_{o/n}-1}) + G(a_{k_{o/n}-1}) \\
&\overset{(b)}{\leq} 3G(a_{k_{o/n}-1}),
\end{aligned} \tag{111}$$

where (a) follows from (110) and $k_{\mathrm{ada}} \leq k_{o/n} + 1$, and (b) follows from $k_{\mathrm{ada}} \leq k_{o/n} + 1$ and the fact that $G(\cdot)$ is an increasing function. We note that when K is sufficiently large and $a(> 1)$ is sufficiently small, the error bound in (111) converges to the one in the case where o/n is known, except for the need to take δ to be sufficiently small.

However, when there are multiple unknown parameters, this approach does not work well. Here, we consider the case where not only o/n but also σ is unknown. Let the true value of o/n be 0.2, and the sequence of candidates for exploration be $[0.1, 0.2, 0.3]$. Let the true value of σ be 3, and the sequence of candidates for exploration be $[1, 2, 3, 4]$. We set $\boldsymbol{\beta}^* = (1, 0 \cdots, 0)$. For simplicity, we assume that n is sufficiently large, such that $\rho c_{r_1}\sqrt{s} r_{d,s} + r_\delta + r_o \leq 1.1 r_o$. Re-define $F(\cdot,\cdot)$ and $G(\cdot,\cdot)$ as

$$\begin{aligned}
F(\sigma, o/n) &:= c_s c'''^2_{\max} L \lambda_o \sqrt{n} \frac{1.1}{c_{r_1}\sqrt{s}} r_o, \\
G(\sigma, o/n) &:= \frac{15\,(2 + c_s)}{c_s} c_{r_1}\sqrt{s} \times F(\sigma, o/n),
\end{aligned}$$

where $\lambda_o\sqrt{n} = 24L^2\sigma$. Additionally, we assume that the product of the constant in $G(\cdot,\cdot)$ and $\sqrt{\log(n/o)}(\leq \sqrt{\log(10/3)})$ is less than or equal to 1.5/1.1. In other words, we can define $G(a,b) = 1.5ab$ for $a \in \{1,2,3,4\}$ and $b \in \{0.1, 0.2, 0.3\}$. The estimator where $a = i$ and $b = j$ will be denoted as $\hat{\boldsymbol{\beta}}_{i,j}$.

Here, suppose that the estimate of $\boldsymbol{\beta}^*$ is $(1.9, 0, \cdots, 0)$ for all combinations of $(\sigma, o/n)$ among the candidates. In this case, we note that the estimation is successful for $(\sigma, o/n) = (3, 0.2), (3, 0.3), (4, 0.2), (4, 0.3)$, while it fails for all other combinations. However, for a failed combinations $(i,j) = (2, 0.1)$, since $|\hat{\boldsymbol{\beta}}_{2,0.1} - \hat{\boldsymbol{\beta}}_{3,0.2}| = 0 < G(2, 0.1) + G(3, 0.2)$, we must conclude that the estimate $(1.9, 0, \cdots, 0)$ for $(\sigma, o/n) = (1, 0.1)$ is valid, similar to the case where only o/n is unknown. While it is expected that the estimate satisfies $|\hat{\boldsymbol{\beta}}_{1,0.1} - \boldsymbol{\beta}^*| \leq 3 \times G(1, 0.1)$, it is clear that this inequality does not hold. In other words, in this case, depending on how we set the sequence of candidate parameters, this constant factor can be arbitrarily worse. This phenomenon contrasts with the case where only o/n is unknown, where the quantity corresponding to $G(1, 0.1)$ was bounded by three times that amount. For a more general discussion on this point, refer to Jain et al. (2022). Of course, even in more complex settings such as Theorems 13 and 15 and Corollary 17, Lepskii's method may fail when multiple parameters are unknown.

References

Radoslaw Adamczak. A note on the hanson-wright inequality for random vectors with dependencies. *Electronic Communications in Probability*, 20:1–13, 2015.

Pierre Alquier, Vincent Cottet, and Guillaume Lecué. Estimation bounds and sharp oracle inequalities of regularized procedures with lipschitz loss functions. *The Annals of Statistics*, 47(4):2117–2144, 2019.

Ainesh Bakshi and Adarsh Prasad. Robust linear regression: Optimal rates in polynomial time. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 102–115, 2021.

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.

Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

Pierre C Bellec. Localized gaussian width of m-convex hulls with applications to lasso and convex aggregation. *Bernoulli*, 25(4A):3016–3040, 2019.

Pierre C Bellec, Guillaume Lecué, and Alexandre B Tsybakov. Slope meets lasso: improved oracle bounds and optimality. *The Annals of Statistics*, 46(6B):3603–3642, 2018.

Alexandre Belloni, Victor Chernozhukov, and Lie Wang. Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika*, 98(4):791–806, 2011.

Peter J Bickel, Ya'acov Ritov, and Alexandre B Tsybakov. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

Matthew Brennan and Guy Bresler. Reducibility and statistical-computational gaps from secret leakage. In *Conference on Learning Theory*, pages 648–847. PMLR, 2020.

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

Emmanuel Candes and Terence Tao. The dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6):2313–2351, 2007.

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

Xi Chen and Wen-Xin Zhou. Robust inference via multiplier bootstrap. *The Annals of Statistics*, 48(3):1665–1691, 2020.

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

Chen Cheng and Andrea Montanari. Dimension free ridge regression. *arXiv preprint arXiv:2210.08571*, 2022.

Yu Cheng, Ilias Diakonikolas, and Rong Ge. High-dimensional robust mean estimation in nearly-linear time. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2755–2771. SIAM, 2019a.

Yu Cheng, Ilias Diakonikolas, Rong Ge, and David Woodruff. Faster algorithms for high-dimensional robust covariance estimation. *arXiv preprint arXiv:1906.04661*, 2019b.

Yu Cheng, Ilias Diakonikolas, Rong Ge, Shivam Gupta, Daniel Kane, and Mahdi Soltanolkotabi. Outlier-robust sparse estimation via non-convex optimization. *Advances in Neural Information Processing Systems*, 35:7318–7327, 2022.

Yeshwanth Cherapanamjeri, Efe Aras, Nilesh Tripuraneni, Michael I Jordan, Nicolas Flammarion, and Peter L Bartlett. Optimal robust linear regression in nearly linear time. *arXiv preprint arXiv:2007.08137*, 2020.

Geoffrey Chinot. Erm and rerm are optimal estimators for regression problems when malicious outliers corrupt the labels. *Electronic Journal of Statistics*, 14(2):3563–3605, 2020.

Geoffrey Chinot, Guillaume Lecué, and Matthieu Lerasle. Robust high dimensional learning for lipschitz and convex losses. *Journal of Machine Learning Research*, 21:233, 2020.

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 ℓ_1-penalized Huber's m-estimator. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 13188–13198. Curran Associates, Inc., 2019.

Arnak S Dalalyan and Arshak Minasyan. All-in-one robust estimator of the gaussian mean. *The Annals of Statistics*, 50(2):1193–1219, 2022.

Jules Depersin and Guillaume Lecué. Robust sub-gaussian estimation of a mean vector in nearly linear time. *The Annals of Statistics*, 50(1):511–536, 2022.

Michał Dereziński. Algorithmic gaussianization through sketching: Converting data into sub-gaussian random designs. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 3137–3172. PMLR, 2023.

Alexis Derumigny. Improved bounds for square-root lasso and square-root slope. *Electronic Journal of Statistics*, 12(1):741–766, 2018.

Ilias Diakonikolas and Daniel M Kane. Recent advances in algorithmic high-dimensional robust statistics. *arXiv preprint arXiv:1911.05911*, 2019.

Ilias Diakonikolas, Gautam Kamath, Daniel M Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Being robust (in high dimensions) can be practical. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 999–1008. JMLR. org, 2017a.

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, 2017b.

Ilias Diakonikolas, Gautam Kamath, Daniel M Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robustly learning a gaussian: Getting optimal error, efficiently. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2683–2702. Society for Industrial and Applied Mathematics, 2018.

Ilias Diakonikolas, Themis Gouleakis, and Christos Tzamos. Distribution-independent pac learning of halfspaces with massart noise. *Advances in Neural Information Processing Systems*, 32, 2019a.

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, 2019b.

Ilias Diakonikolas, Daniel Kane, Sushrut Karmalkar, Eric Price, and Alistair Stewart. Outlier-robust high-dimensional sparse estimation via iterative filtering. In *Advances in Neural Information Processing Systems*, pages 10689–10700, 2019c.

Ilias Diakonikolas, Weihao Kong, and Alistair Stewart. Efficient algorithms and lower bounds for robust linear regression. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2745–2754. SIAM, 2019d.

Ilias Diakonikolas, Daniel M Kane, and Pasin Manurangsi. The complexity of adversarially robust proper learning of halfspaces with agnostic noise. *Advances in Neural Information Processing Systems*, 33:20449–20461, 2020.

Ilias Diakonikolas, Daniel M Kane, Vasilis Kontonis, Christos Tzamos, and Nikos Zarifis. Efficiently learning halfspaces with tsybakov noise. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 88–101, 2021.

Ilias Diakonikolas, Daniel Kane, Sushrut Karmalkar, Ankit Pensia, and Thanasis Pittas. List-decodable sparse mean estimation via difference-of-pairs filtering. *Advances in Neural Information Processing Systems*, 35:13947–13960, 2022a.

Ilias Diakonikolas, Daniel M Kane, Sushrut Karmalkar, Ankit Pensia, and Thanasis Pittas. Robust sparse mean estimation via sum of squares. In *Conference on Learning Theory*, pages 4703–4763. PMLR, 2022b.

Ilias Diakonikolas, Daniel M Kane, Sushrut Karmalkar, Ankit Pensia, and Thanasis Pittas. Robust sparse estimation for gaussians with optimal error under huber contamination. *arXiv preprint arXiv:2403.10416*, 2024.

Steven Diamond and Stephen Boyd. Cvxpy: A python-embedded modeling language for convex optimization. *Journal of Machine Learning Research*, 17(83):1–5, 2016.

Sjoerd Dirksen. Tail bounds via generic chaining. *Electronic Journal of Probability*, 20: 1–29, 2015.

Yihe Dong, Samuel Hopkins, and Jerry Li. Quantum entropy scoring for fast robust mean estimation and improved outlier detection. In *Advances in Neural Information Processing Systems*, pages 6067–6077, 2019.

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

Jianqing Fan, Quefeng Li, and Yuyan Wang. Estimation of high dimensional mean regression in the absence of symmetry and light tail assumptions. *Journal of the Royal Statistical Society. Series B, Statistical methodology*, 79(1):247, 2017.

Jianqing Fan, Han Liu, Qiang Sun, and Tong Zhang. I-lamm for sparse learning: Simultaneous control of algorithmic complexity and statistical error. *Annals of statistics*, 46(2): 814, 2018.

Jianqing Fan, Weichen Wang, and Ziwei Zhu. A shrinkage principle for heavy-tailed data: High-dimensional robust low-rank matrix recovery. *Annals of statistics*, 49(3):1239, 2021.

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

Kristian Georgiev and Samuel Hopkins. Privacy induces robustness: Information-computation gaps and sparse mean estimation. *Advances in neural information processing systems*, 35:6829–6842, 2022.

David Lee Hanson and Farroll Tim Wright. A bound on tail probabilities for quadratic forms in independent random variables. *The Annals of Mathematical Statistics*, 42(3): 1079–1083, 1971.

Daniel Hsu, Sham Kakade, and Tong Zhang. A tail inequality for quadratic forms of subgaussian random vectors. *Electronic Communications in Probability*, 17:1–6, 2012.

Ayush Jain, Alon Orlitsky, and Vaishakh Ravindrakumar. Robust estimation algorithms don't need to know the corruption level. *arXiv preprint arXiv:2202.05453*, 2022.

Pravesh K Kothari, Jacob Steinhardt, and David Steurer. Robust moment estimation and improved clustering via sum of squares. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1035–1046. ACM, 2018.

Kevin A Lai, Anup B Rao, and Santosh Vempala. Agnostic estimation of mean and covariance. In *Foundations of Computer Science (FOCS), 2016 IEEE 57th Annual Symposium on*, pages 665–674. IEEE, 2016.

Guillaume Lecué and Matthieu Lerasle. Robust machine learning by median-of-means: theory and practice. *The Annals of Statistics*, 48(2):906–931, 2020.

Guillaume Lecué and Shahar Mendelson. Regularization and the small-ball method i: sparse recovery. *The Annals of Statistics*, 46(2):611–641, 2018.

Oleg V Lepski and Vladimir G Spokoiny. Optimal pointwise adaptive methods in nonparametric estimation. *The Annals of Statistics*, 25(6):2512–2546, 1997.

OV Lepskii. Asymptotically minimax adaptive estimation. i: Upper bounds. optimally adaptive estimates. *Theory of Probability & Its Applications*, 36(4):682–697, 1992.

Allen Liu and Ankur Moitra. Settling the robust learnability of mixtures of gaussians. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 518–531, 2021.

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.

Gabor Lugosi and Shahar Mendelson. Risk minimization by median-of-means tournaments. *Journal of the European Mathematical Society*, 22(3):925–965, 2019.

Gabor Lugosi and Shahar Mendelson. Robust multivariate mean estimation: the optimality of trimmed mean. *The Annals of Statistics*, 49(1):393–410, 2021.

Pascal Massart. About the constants in talagrand's concentration inequalities for empirical processes. *The Annals of Probability*, 28(2):863–884, 2000.

Ibrahim Merad and Stéphane Gaïffas. Robust methods for high-dimensional linear learning. *Journal of Machine Learning Research*, 24(165):1–44, 2023.

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.

Omar Montasser, Surbhi Goel, Ilias Diakonikolas, and Nathan Srebro. Efficiently learning adversarially robust halfspaces with noise. In *International Conference on Machine Learning*, pages 7010–7021. PMLR, 2020.

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

Samet Oymak. Learning compact neural networks with regularization. In *International Conference on Machine Learning*, pages 3966–3975. PMLR, 2018.

Xiaoou Pan and Wen-Xin Zhou. Multiplier bootstrap for quantile regression: non-asymptotic theory under random design. *Information and Inference: A Journal of the IMA*, 10(3):813–861, 2021.

Ankit Pensia, Varun Jog, and Po-Ling Loh. Robust regression with covariate filtering: Heavy tails and adversarial contamination. *arXiv preprint arXiv:2009.12976*, 2020.

Yaniv Plan and Roman Vershynin. One-bit compressed sensing by linear programming. *Communications on Pure and Applied Mathematics*, 66(8):1275–1297, 2013.

Adarsh Prasad, Sivaraman Balakrishnan, and Pradeep Ravikumar. A robust univariate mean estimator is all you need. In *International Conference on Artificial Intelligence and Statistics*, pages 4034–4044. PMLR, 2020.

Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Restricted eigenvalue properties for correlated gaussian designs. *The Journal of Machine Learning Research*, 11:2241–2259, 2010.

Mark Rudelson and Roman Vershynin. Hanson-wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, 18:1–9, 2013.

Takeyuki Sasai. Robust and sparse estimation of linear regression coefficients with heavy-tailed noises and covariates. *arXiv preprint arXiv:2206.07594*, 2022.

Yiyuan She and Art B Owen. Outlier detection using nonconvex penalized regression. *Journal of the American Statistical Association*, 106(494):626–639, 2011.

Vidyashankar Sivakumar, Arindam Banerjee, and Pradeep K Ravikumar. Beyond sub-gaussian measurements: High-dimensional structured estimation with sub-exponential designs. *Advances in neural information processing systems*, 28, 2015.

Vladimir Spokoiny. Concentration of a high dimensional sub-gaussian vector. *arXiv preprint arXiv:2305.07885*, 2023.

Benjamin Stucky and Sara Van De Geer. Sharp oracle inequalities for square root regularization. *Journal of Machine Learning Research*, 18(67):1–29, 2017.

Weijie Su and Emmanuel Candes. Slope is adaptive to unknown sparsity and asymptotically minimax. *The Annals of Statistics*, 44(3):1038–1068, 2016.

Qiang Sun, Wen-Xin Zhou, and Jianqing Fan. Adaptive Huber regression. *Journal of the American Statistical Association*, 115(529):254–265, 2020.

Michel Talagrand. *Upper and lower bounds for stochastic processes*, volume 60. Springer, 2014.

Philip Thompson. Outlier-robust sparse/low-rank least-squares regression and robust matrix completion. *arXiv preprint arXiv:2012.06750*, 2020.

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

Alexander Tsigler and Peter L Bartlett. Benign overfitting in ridge regression. *Journal of Machine Learning Research*, 24(123):1–76, 2023.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Tengyao Wang, Quentin Berthet, and Richard J Samworth. Statistical and computational trade-offs in estimation of sparse principal components. *The Annals of Statistics*, 44(5): 1896–1930, 2016.

Farrol Tim Wright. A bound on tail probabilities for quadratic forms in independent random variables whose distributions are not necessarily symmetric. *The Annals of Probability*, 1(6):1068–1070, 1973.

Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B*, 68(1):49–67, 2006.

Shiwei Zeng and Jie Shen. List-decodable sparse mean estimation. *Advances in Neural Information Processing Systems*, 35:24031–24045, 2022.

Cun-Hui Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894–942, 2010.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B*, 67(2):301–320, 2005.