DISTRIBUTED AND ROBUST STATISTICAL LEARNING

Ziwei Zhu

A Dissertation
Presented to the Faculty
of Princeton University
in Candidacy for the Degree
of Doctor of Philosophy

RECOMMENDED FOR ACCEPTANCE
by the Department of
Operations Research and Financial Engineering
Adviser: Professor Jianqing Fan

June 2018
© Copyright by Ziwei Zhu, 2018.
All rights reserved.
Abstract

Decentralized and corrupted data are nowadays ubiquitous, which impose fundamental challenges for modern statistical analysis. Illustrative examples are massive and decentralized data produced by distributed data collection systems of giant IT companies, corrupted measurement in genetic micro-array analysis, heavy-tailed returns of stocks and etc. These notorious features of modern data often contradict conventional theoretical assumptions in statistics research and invalidate standard statistical procedures. My dissertation addresses these problems by proposing new methodologies with strong statistical guarantees. When data are distributed over different places with limited communication budget, we propose to do local statistical analysis first and aggregate the local results rather than the data themselves to generate a final result. We applied this approach to low-dimensional regression, high-dimensional sparse regression and principal component analysis. When data are not over-scattered, our distributed approach is proved to achieve the same statistical performance as the full sample oracle, i.e., the standard procedure based on all the data. To handle heavy-tailed corruption, we propose a generic principle of data shrinkage for robust estimation and inference. To illustrate this principle, we apply it to estimate regression coefficients in the trace regression model and generalized linear model with heavy-tailed noise and design. The proposed method achieves nearly the same statistical error rate as the standard procedure while requiring only bounded moment conditions on data. This widens the scope of high-dimensional techniques, reducing the moment conditions from sub-exponential or sub-Gaussian distributions to merely bounded second or fourth moment.
Acknowledgements

I would like to express the deepest appreciation to my advisor Professor Jianqing Fan, who has the attitude and the substance of a genius: he continually and convincingly conveyed a spirit of adventure in regard to research and scholarship, and an excitement in regard to teaching. Without his guidance and persistent help this dissertation would not have been possible.

I would like to thank my committee members, Professor Mengdi Wang and Professor Yuxin Chen, whose works always gave me new insights and perspectives in my research. I would also like to thank Prof. Samory Kpotufe for being the reader of my thesis and providing helpful comments and suggestions.

Finally I would like to thank my collaborators: Professor Heather Battey, Professor Han Liu, Professor Qiang Sun, Professor Wenxin Zhou, Dr. Weichen Wang, Wenyan Gong, Junwei Lu, Dong Wang, Kaizheng Wang for their significant contributions and support.
To my parents, grandparents, aunts and uncles.
# Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>iii</td>
</tr>
<tr>
<td>Acknowledgements</td>
<td>iv</td>
</tr>
<tr>
<td>List of Tables</td>
<td>xi</td>
</tr>
<tr>
<td>List of Figures</td>
<td>xii</td>
</tr>
</tbody>
</table>

1 Introduction

1.1 Features of modern data                     | 1    |
1.2 Distributed statistical inference           | 2    |
1.3 Robust estimation                           | 8    |
1.4 Nuclear-norm regularization                 | 12   |
1.5 Notations                                  | 14   |

2 Distributed inference under sparse high dimensional models

2.1 Problem setup                              | 17   |
2.1.1 General likelihood based framework       | 17   |
2.2 Divide and conquer hypothesis tests        | 20   |
2.2.1 Two divide and conquer Wald type constructions | 20   |
2.2.2 Divide and conquer score test            | 29   |
2.3 Accuracy of distributed estimation         | 32   |
2.3.1 The high-dimensional linear model        | 34   |
2.3.2 The high-dimensional generalized linear model | 37   |
List of Tables

2.1  Computation time for the divide and conquer testing and estimation, where $k = 1$ corresponds to the non-splitting case and $k > 1$ corresponds to the distributed case. ........................................... 42

5.1  Testing classification error using original deep features and shrunk deep features .................................................. 121
# List of Figures

1.1 Illustration of distributed statistical inference .......................... 3

1.2 Distributions of kurtosis of macroeconomic variables and gene expressions. Red dashline marks variables with empirical kurtosis equals to that of $t_5$-distribution. Left panel: For 131 macroeconomics variables in [111]. Right panel: For logarithm of expression profiles of 383 genes based on RNA-seq for autism data [52], whose kurtosis is bigger than that of $t_5$ among 19122 genes. ......................................................... 8

2.1 QQ plots of the $p$-values of the Wald (A) and score (B) divide and conquer test statistics against the theoretical quantiles of the uniform $[0,1]$ distribution under the null hypothesis. ........................................ 40

2.2 (A) Estimated probabilities of type I error for the Wald and score tests as a function of $k$. (B) Estimated power with signal strength 0.05 and 0.06 for the Wald, and score tests as a function of $k$. .......................... 42

2.3 (A): Statistical error of the DC estimator, split Lasso and the full sample Lasso for $k \in \{1, 2, 5, 10, 20, 25, 40, 50, 100, 200\}$ when $n = 5000$, $d = 5000$. (B): Euclidean norm difference between the DC thresholded debiased estimator and its full sample analogue. ......................... 44
3.1 Statistical error rate with respect to: (a) the dimension $d$ when $\lambda = 50$ and $n = 2000$; (b) the number of servers $m$ when $\lambda = 50$ and $n = 2000$; (c) the subsample size $n$ when $\lambda = 50$ and $m = 50$; (d) the eigengap $\delta$ when $d = 800$ and $n = 2000$.

3.2 Observed and fitted values of $\log(\rho(\tilde{V}_K, V_K))$.

3.3 Statistical error with respect to the number of machines when the total sample size $N = 6,000$ is fixed.

3.4 Comparison between DP, FP and DP5: (a) $m = 20$, $n = 2000$ and $\lambda = 50$; (b) $d = 1600$, $n = 1000$ and $\lambda = 30$; (c) $d = 800$, $m = 5$ and $\lambda = 30$; (d) $d = 1600$, $m = 10$ and $n = 500$.

4.1 Statistical errors of $\ln\|\hat{\Theta} - \Theta^*\|_F$ v.s. logarithmic sample size $\ln N$ for different dimensions $d$ in matrix compressed sensing.

4.2 Statistical errors of $\ln\|\hat{\Theta} - \Theta^*\|_F$ v.s. logarithmic sample size $\ln N$ for different dimensions $d$ in matrix completion.

4.3 Statistical errors of $\ln\|\hat{\Theta} - \Theta^*\|_F$ v.s. logarithmic sample size $\ln N$ for different dimensions $d$ in multi-task regression.

4.4 Statistical errors of $\|\hat{\Sigma} - \Sigma\|_{op}$ v.s. sample size $n$ for different dimensions $d$ in covariance estimation.

5.1 High dimensional sparse linear regression with light-tailed features (left) and heavy-tailed features (right).

5.2 Statistical Error of MLE based on minimizing $\ell_w(\beta)$ with 10% mislabeled data.

6.1 Illustration of M-estimation over data stream.
A.1 (A) The ratio between the loss of the divide and conquer procedure
and the statistical error of the estimator based on the whole sample
with \( d = \sqrt{n}/2 \) and different growth rates of \( k \). (B) Statistical error of
the DC estimator against \( \log k \). . . . . . . . . . . . . . . . . . . . . . . . . . . . 130

A.2 (A) The ratio between the loss of the divide and conquer procedure
and the statistical error of the estimator based on the whole sample
when \( d = 20 \). (B) Statistical error of the DC estimator. . . . . . . . . 132
Chapter 1

Introduction

1.1 Features of modern data

Modern data sets exhibit the following three features that challenge classical statistical inference.

- **Decentralized**: Data are generated from multiple sources across which the communication is constrained by bandwidth or privacy.

- **Contaminated**: Data quality suffers from heavy-tailed corruption.

- **High-Dimensional**: Parameters have high ambient dimensions but low intrinsic dimensions.

One illustrating example of these features is the large-scale recommendation system or collaborative filtering. Worldwide customers continually generate massive data through the Internet; it will be painfully slow to centralize all the raw data before statistical analysis. Besides, corruption is significant in customer data such as ratings, votings, etc. Finally, customer behavior usually has clear patterns that imply a low-dimensional latent structure of the data. Understanding and exploiting this structure is crucial for overcoming the curse of high dimensions.
My research focuses on new methodologies that adapt to this complex data setting with strong statistical guarantees. The goal is to characterize how splitting, contamination and high dimensions affect statistical efficiency, thus inspiring more economic, robust and accurate statistical procedures. The following sections will revolve around the statistical methodologies that cope with the three features above respectively. We will introduce the problem setups, review the related past works and summarize the contributions we have made.

1.2 Distributed statistical inference

Massive datasets are usually scattered across distant places such that to fuse or aggregate them is extremely difficult due to communication cost, privacy, data security and ownerships, among others. Consider giant IT companies that collect data simultaneously from places all around the world. Constraints on communication budget and network bandwidth make it nearly impossible to aggregate and maintain global data in a single data center. Another example is that health records are scattered across many hospitals or countries. It is hard to process the data in a central location due to privacy and ownership concerns.

To resolve these issues, the statistics community has witnessed a surge in activities exploiting distributed computing architectures and developing distributed estimators or testing statistics based on data scattered around different locations. A typical distributed statistical method first calculates local statistics based on each sub-dataset and then combines all the subsample-based statistics to produce an aggregated statistic. This method transforms a single large-scale estimation or testing problem into multiple smaller problem on each local machine. It fully adapts to the parallel data collection procedures and thus significantly reduces the communication cost. Figure 1.1 describes this distributed algorithm.
Figure 1.1: Illustration of distributed statistical inference

Following this framework of distributed statistical inference, we develop communication-efficient distributed algorithms for the Wald and Rao’s score tests for the sparse high dimensional scheme, as well as for the estimation of regression coefficients in the sparse high dimensional and low dimensional linear and generalized linear models. With the total sample size $N$ fixed, we give the upper bound on the number of data splits $m$ for preserving the statistical error of the analogous full sample procedure. While hypothesis testing in a low dimensional context is straightforward, in the sparse high dimensional setting, nuisance parameters introduce a non-negligible bias, causing classical low dimensional theory to break down. In our high dimensional Wald construction, the phenomenon is remedied through a debiasing of the estimator, which gives rise to a test statistic with tractable limiting distribution, as documented in the $m = 1$ setting in [134] and [120]. For the high dimensional analogue of Rao’s score statistic, the incorporation of a correction factor increases the convergence rate of higher order terms, thereby vanquishing the effect of the nuisance parameters. The approach is introduced in the $m = 1$ setting in [96], where the test statistic is shown to possess a tractable limit distribution. However, the computation complexity for the debiased estimators increases by an order of
magnitude, due to solving $d$ high-dimensional regularization problems, where $d$ is the dimension of regression coefficients. This motivates us to appeal to the distributed strategy, or the divide and conquer (DC) strategy.

We develop the theory and methodology for DC versions of these tests. In the case of $m = 1$, each of the above test statistics can be decomposed into a dominant term with tractable limit distribution and a negligible remainder term. The DC extension requires delicate control of these remainder terms to ensure the error accumulation remains sufficiently small so as not to materially contaminate the leading term. We obtain an upper bound on the number of permitted subsamples, $m$, subject to a statistical guarantee. More specifically, we find that the theoretical upper bound on the number of subsamples guaranteeing the same inferential or estimation efficiency as the whole-sample procedure is $m = o((s \log d)^{-1} \sqrt{N})$ in the linear model, where $s$ is the sparsity of the parameter vector. In the generalized linear model the scaling is $m = o(((s \vee s_1) \log d)^{-1} \sqrt{N})$, where $s_1$ is the sparsity of the inverse information matrix.

For sparse high dimensional estimation problems, we use the same debiasing technique introduced in the high dimensional testing problems to obtain a thresholded divide and conquer estimator that achieves the full sample minimax rate. The appropriate scaling is found to be $m = O(\sqrt{N/(s^2 \log d)})$ for the estimation of the sparse parameter vector in the high dimensional linear model and $m = O(\sqrt{N/((s \vee s_1)^2 \log d)})$ for the high dimensional generalized linear model. Moreover, we find that the loss incurred by the divide and conquer strategy, as quantified by the distance between the DC estimator and the full sample estimator, is negligible in comparison to the statistical error of the full sample estimator provided that $k$ is not too large. In the context of estimation, the optimal scaling of $m$ with $N$ and $d$ is also developed for the low dimensional linear and generalized linear model. This theory is of independent
interest. It also allows us to study a refitted estimation procedure under a minimal
signal strength assumption.

A partial list of references covering DC algorithms from a statistical perspective
is [29, 137, 105, 74]. Specifically, [137] consider the distributed estimator for kernel
ridge regression. In the context of \( d < N \), [137] propose the distributed estimator by
averaging the kernel ridge regression estimators for each data split. They obtain an
explicit upper bound on the number of splits yielding the minimax optimal rates for
the mean squared error. However, it is not straightforward to generalize their estima-
tor to the high dimensional setting. In an independent work, [74] propose the same
debiasing approach of [120] to allow aggregation of local estimates on distributed data
splits in the context of sparse high dimensional linear and generalized linear models.
Though using different techniques of proofs, the conclusions of [74] in terms of the
optimal choice of tuning parameter scaling and the upper bound on the permissible
number of sample splits is of the same order. Our work differs from theirs in two
aspects: (1) our work also contributes to the distributed testing in sparse high di-
mensional models and (2) we propose a refitted distributed estimator which has the
oracle rate. Our results on hypothesis testing reveal a different phenomenon to that
found in estimation, as we observe through the different requirements on the scaling
of \( m \). On the estimation side, our results also differ from those of [74] in that our
additional refitting step allows us to achieve the oracle rate. [105] consider the dis-
tributed empirical risk minimization for \( M \)-estimators. They require the dimension of
the interest parameter to satisfy the scaling condition \( d/N \to \kappa \in (0, 1) \), which rules
out the \( d \gg N \) case. They quantify the accuracy loss over the full sample estimator
in terms of the number of splits.

Besides regression problems, there has also been rapid advancement on distributed
principal component analysis (PCA) over the past two decades. Unlike the traditional
PCA where we have the complete data matrix \( X \in \mathbb{R}^{N \times d} \) with \( d \) features of \( N \),
samples at one place, the distributed PCA needs to handle data that are partitioned and stored across multiple servers. There are two data partition regimes: “horizontal” and “vertical”. In the horizontal partition regime, each server contains all the features of a subset of subjects, while in the vertical partition regime, each server has a subset of features of all the subjects. To conduct distributed PCA in the horizontal regime, [99] proposes that each server computes several top eigenvalues and eigenvectors on its local data and then sends them to the central server that aggregates the information together. Yet there is no theoretical guarantee on the approximation error of the proposed algorithm. [77], [63] and [12] aim to find a good rank-$K$ approximation $\hat{X}$ of $X$. To assess the approximation quality, they compare $\|\hat{X} - X\|_F$ against $\min_{\text{rank}(B) \leq K} \|B - X\|_F$ and study the excess risk. For the distributed PCA in the vertical data partition regime, there is also a great amount of literature, for example, [64], [76], [6], [107], etc. This line of research is often motivated from sensor networks and signal processing where the vertically partitioned data are common. Our work focuses on the horizontal partition regime, i.e., we have partitions over the samples rather than the features.

Despite these achievements, very few papers establish rigorous statistical error analysis of the proposed distributed PCA methods. To our best knowledge, the only works that provide statistical analysis so far are [34] and [27]. To estimate the leading singular vectors of a large target matrix, both papers propose to aggregate singular vectors of multiple random approximations of the original matrix. [34] adopts sparse approximation of the matrix by sampling the entries, while [27] uses Gaussian random sketches. The works are related to ours, since we can perceive sub-datasets in the distributed PCA problem as random approximations. To see why, suppose $\Sigma$ is the true $d$-by-$d$ covariance matrix. For any sub-sample matrix $X^{(\ell)} = (x_1^{(\ell)}, ..., x_n^{(\ell)})^\top \in$
Given \( \{Z^{(t)}\}_{\ell=1}^{m}\) are independent, \( \{X^{(t)\top}\}_{\ell=1}^{m}\) can thus be regarded as multiple independent random sketches of \( \Sigma_{1/2} \). Therefore by applying the distributed algorithm proposed in [27] to \( \{X^{(t)\top}\}_{\ell=1}^{m}\), we can obtain the eigenvectors of \( \Sigma \). However, our analysis is more general, since it does not rely on any matrix incoherence assumption as required by [34] and it explicitly characterizes how the probability distribution affects the final statistical error in finite sample error bounds. Besides, our aggregation algorithm is much simpler than the one in [27]. The manuscript [49] came out after we submitted the first draft of our work. The authors focused on estimation of the first principal component rather than the multi-dimensional eigenspaces, based on very different approaches.

We propose a distributed algorithm with only one-shot communication to solve for the top \( K \) eigenvectors of the population covariance matrix \( \Sigma \) when samples are scattered across \( m \) servers. We first calculate for each subset of data \( \ell \) its top \( K \) eigenvectors \( \{\hat{V}_{K}^{(t)} = (\hat{v}_{1}^{(t)}, \cdots, \hat{v}_{K}^{(t)})\}_{\ell=1}^{m} \) of the sample covariance matrix there, then compute the average of projection matrices of the eigenspaces \( \bar{\Sigma} = (1/m) \sum_{\ell=1}^{m} \hat{V}_{K}^{(t)} \hat{V}_{K}^{(t)\top} \), and finally take the top \( K \) eigenvectors of \( \bar{\Sigma} \) as the final estimator \( \bar{V}_{m} = (\bar{v}_{1}^{(t)}, \cdots, \bar{v}_{K}^{(t)}) \).

The communication cost of this method is of order \( O(mKd) \). We establish rigorous non-asymptotic analysis of the statistical error \( \|\bar{V}_{K} \bar{V}_{K}^{\top} - V_{K} V_{K}^{\top}\|_{F} \), and show that as long as we have a sufficiently large number of samples in each server, \( \bar{V}_{K} \) enjoys the same statistical error rate as the standard PCA over the full sample. The eigenvalues of \( \Sigma \) are easily estimated once we get good estimators of the eigenvectors, using another round of communication.
1.3 Robust estimation

Heavy-tailed distributions are ubiquitous in modern statistical analysis and machine learning problems. They are stylized features of high-dimensional data. By chance alone, some of observable variables in high-dimensional datasets can have heavy or moderately heavy tails (see right panel of Figure 1.2). It has been widely known that financial returns and macroeconomic variables exhibit heavy tails, and large-scale imaging datasets in biological studies are often corrupted by heavy-tailed noises due to limited measurement precisions. Figure 1.2 provides some empirical evidence on this which is pandemic to high-dimensional data. These stylized features and phenomena contradict the popular assumption of sub-Gaussian or sub-exponential noises in the theoretical analysis of standard statistical procedures. They also have adverse impacts on the methods that are popularly used. Simple and effective principles are needed for dealing with moderately heavy or heavy tailed data.

Figure 1.2: Distributions of kurtosis of macroeconomic variables and gene expressions. Red dashline marks variables with empirical kurtosis equals to that of $t_5$-distribution. Left panel: For 131 macroeconomics variables in [111]. Right panel: For logarithm of expression profiles of 383 genes based on RNA-seq for autism data [52], whose kurtosis is bigger than that of $t_5$ among 19122 genes.

Recent years have witnessed increasing literature on the robust mean estimation when the population distribution is heavy-tailed. [24] proposed a novel approach that is through minimizing a robust empirical loss. Unlike the traditional $\ell_2$ loss, the robust
loss function therein penalizes large deviations, thereby making the correspondent M-estimator insensitive to extreme values. It turns out that when the population has only finite second moment, the estimator has exponential concentration around the true mean and enjoys the same rate of statistical consistency as the sample average for sub-Gaussian distributions. [13] pursued the Catoni’s mean estimator further by applying it to empirical risk minimization. [36] utilized the Huber loss with diverging threshold, called robust approximation to quadratic (RA-quadratic), in a sparse regression problem and showed that the derived M-estimator can also achieve the minimax statistical error rate. [79] studied the statistical consistency and asymptotic normality of a general robust $M$-estimator and provided a set of sufficient conditions to achieve the minimax rate in the high-dimensional regression problem.

Another effective approach to handle heavy-tailed distribution is the so-called “median of means” approach, which can be traced back to [94]. The main idea is to first divide the whole samples into several parts and take the median of the means from all pieces of sub-samples as the final estimator. This “median of means” estimator also enjoys exponential large deviation bound around the true mean. [57] and [88] generalized this idea to multivariate cases and applied it to robust PCA, high-dimensional sparse regression and matrix regression, achieving minimax optimal rates up to logarithmic factors.

We propose a simple and effective principle: truncation of univariate data and more generally shrinkage of multivariate data to achieve the robustness. We will illustrate our ideas through a general model called the trace regression

\[ Y = \text{Tr}(\Theta^* \mathbf{X}) + \epsilon, \]

which embraces linear regression, matrix or vector compressed sensing, matrix completion and multi-tasking regression as specific examples. The goal is to estimate the
coefficient matrix $\Theta^* \in \mathbb{R}^{d_1 \times d_2}$, which is assumed to have a nearly low-rank structure in the sense that its Schatten norm is constrained: 

$$\min_{d_1, d_2} \sum_{i=1}^{\min(d_1, d_2)} \sigma_i(\Theta^*)^q \leq \rho \quad \text{for} \quad 0 \leq q < 1,$$

where $\sigma_i(\Theta^*)$ is the $i^{th}$ singular value of $\Theta^*$, i.e., the square-root of the $i^{th}$ eigenvalue of $\Theta^* \Sigma \Theta^*$. In other words, the singular values of $\Theta^*$ decay fast enough so that $\Theta^*$ can be well approximated by a low-rank matrix. We always consider the high-dimensional setting where the sample size $n \ll d_1 d_2$. As we shall see, appropriate data shrinkage allows us to recover $\Theta^*$ with only bounded moment conditions on noise and design.

Our work aims to handle the presence of heavy-tailed, asymmetrical and heteroscedastic noises in the general trace regression. Based on the shrinkage of data, we developed a new loss function called the robust quadratic loss, which is constructed by plugging robust covariance estimators in the $\ell_2$ risk function. Then we obtain the estimator $\hat{\Theta}$ by minimizing this new robust quadratic loss plus nuclear-norm penalty. By tailoring the analysis of [93] to this new loss, we can establish statistical rates in estimating the matrix $\Theta^*$ that are the same as those in [93] for the sub-Gaussian distributions, while allowing the noise and design to have much heavier tails. This result is very generic and applicable to all four specific aforementioned examples.

Our robust approach is particularly simple: it truncates or shrinks appropriately the response variables, depending on whether the responses are univariate or multivariate. Under the setting of sub-Gaussian design, unusually large responses are very likely to be due to the outliers of noises. This explains why we need to truncate the responses when we have light-tailed covariates. Under the setting of heavy-tailed covariates, we need to truncate the designs as well. It turns out that appropriate truncation does not induce significant bias or hurt the restricted strong convexity of the loss function. With these data robustfications, we can then apply penalized least-squares method to recover sparse vectors or low-rank matrices. Under only bounded moment conditions for either noise or covariates, our robust estimator achieves the same statistical error rate as that under the case of the sub-Gaussian design and noise.
The crucial component in our analysis is the sharp spectral-norm convergence rate of robust covariance matrices based on data shrinkage. Of course, other robustifications of estimated covariance matrices, such as the RA-covariance estimation in [36], are also possible to enjoy similar statistical error rates, but we will only focus on the shrinkage method, as it is easier to analyze and always semi-positive definite.

It is worth emphasis that the successful application of the shrinkage sample covariance in multi-tasking regression inspires us to also study its statistical error in covariance estimation. It turns out that as long as the random samples \( \{x_i \in \mathbb{R}^d\}_{i=1}^N \) have bounded fourth moment in the sense that \( \sup_{v \in S^{d-1}} E(v^\top x_i)^4 \leq R < \infty \), where \( S^{d-1} \) is the \( d \)-dimensional unit sphere, our \( \ell_4 \)-norm shrinkage sample covariance \( \tilde{\Sigma}_N \) achieves the statistical error rate of order \( O_P(\sqrt{\log d/N}) \) in terms of the spectral norm. This rate is the same, up to a logarithmic term, as that of the standard sample covariance matrix \( \Sigma_N \) with sub-Gaussian samples under the low-dimensional regime. Under the high-dimensional regime, \( \tilde{\Sigma} \) even outperforms \( \Sigma_N \) for sub-Gaussian random samples, since now the error rate of \( \Sigma_N \) deteriorates to \( O_P(d/N) \) while the error rate of \( \tilde{\Sigma} \) is still \( O_P(\sqrt{\log d/N}) \). This means even with light-tailed data, standard sample covariance can be inadmissible in terms of convergence rate when dimension is high. Therefore, shrinkage not only overcomes heavy-tailed corruption, but also mitigates curse of dimensionality. In terms of the elementwise max-norm, it is not hard to show that appropriate elementwise truncation of the data delivers a sample covariance with statistical error rate of order \( O_P(\sqrt{\log d/N}) \). This estimator can further be regularized if the true covariance has sparsity and other structure. See, for example, [86], [8], [72], [19], [18], [39], among others.
1.4 Nuclear-norm regularization

In modern data analytics, the parameters of interest often exhibit high ambient dimensions but low intrinsic dimensions that can be exploited to circumvent the curse of dimensionality. One of the most illustrating examples is the sparse signal recovery through incorporating sparsity regularization into empirical risk minimization ([113, 26, 37]). As shown in the profound works ([20, 42, 43, 141, 133], among others), the statistical rate of the appropriately regularized M-estimator has mere logarithmic dependence on the ambient dimension $d$. This implies that consistent signal recovery is feasible even when $d$ grows exponentially with respect to the sample size $n$. In econometrics, sparse models and methods have also been intensively studied and are proven to be powerful. For example, [4] studied estimation of optimal instruments under sparse high-dimensional models and showed that the instrumental variable (IV) estimator based on Lasso and post-Lasso methods enjoys root-n consistency and asymptotic normality. [54] and [23] investigated instrument selection using high-dimensional regularization methods. [68] established oracle inequalities for high dimensional vector autoregressions and [25] applied group Lasso in threshold autoregressive models and established near-optimal rates in the estimation of threshold parameters. [5] employed high-dimensional techniques for program evaluation and causal inference.

When the parameter of interest arises in the matrix form, elementwise sparsity is not the sole way of constraining model complexity; another structure that is exclusive to matrices comes into play: the rank. Low-rank matrices have much fewer degrees of freedom than its ambient dimensions $d_1 \cdot d_2$. To determine a rank-$r$ matrix $\Theta \in \mathbb{R}^{d_1 \times d_2}$, we only need $r$ left and right singular vectors and $r$ singular values, which correspond to $r(d_1 + d_2 - 1)$ degrees of freedom, without accounting the orthogonality. As a novel regularization approach, low-rankness motivates matrix representations of the parameters of interest in various statistical and econometric models. If we rearrange
the coefficient in the traditional linear model as a matrix, we obtain the so-called trace regression model:

$$Y = \text{Tr}(\Theta^* \mathbf{X}) + \epsilon,$$

where $\text{Tr}(\cdot)$ denotes the trace, $\mathbf{X}$ is a matrix of explanatory variables, $\Theta^* \in \mathbb{R}^{d_1 \times d_2}$ is the matrix of regression coefficients, $Y$ is the response and $\epsilon$ is the noise. In predictive econometric applications, $\mathbf{X}$ can be a large panel of time series data such as stock returns or macroeconomic variables \cite{110, 83}, whereas in statistical machine learning $\mathbf{X}$ can be images. The rank of a matrix is controlled by the $\ell_q$-norm for $q \in [0, 1)$ of its singular values:

$$B_q(\Theta^*) := \sum_{j=1}^{d_1 \wedge d_2} \sigma_j(\Theta^*)^q \leq \rho,$$  

where $\sigma_j(\Theta^*)$ is the $j$th largest singular value of $\Theta^*$, and $\rho$ is a positive constant that can grow to infinity. Note that when $q = 0$, it controls the rank of $\Theta^*$ at $\rho$. Trace regression is a natural model for matrix-type covariates, such as the panel data, images, genomics microarrays, etc. In addition, particular forms of $\mathbf{X}$ can reduce trace regression to several well-known problem setups. For example, when $\mathbf{X}$ contains only a column and the response $Y$ is multivariate, (1.4.1) becomes reduced-rank regression model (\cite{1}, \cite{59}). When $\mathbf{X} \in \mathbb{R}^{d_1 \times d_2}$ is a singleton in the sense that all entries of $\mathbf{X}$ are zeros except for one entry that equals one, (1.4.1) characterizes the matrix completion problem in item response problems and online recommendation systems.

To explore the low rank structure of $\Theta^*$ in (1.4.1), a natural approach is the penalized least-squares with the nuclear norm penalty. Specifically, consider the following optimization problem:

$$\hat{\Theta} = \arg\min \left\{ \frac{1}{n} \sum_{i=1}^{n} (\langle \Theta, \mathbf{X}_i \rangle - Y_i)^2 + \lambda \| \Theta \|_* \right\},$$  

(1.4.3)
where $\|\Theta\|_* = \sum_{j=1}^{d_1 \wedge d_2} \sigma_j(\Theta)$ is the nuclear norm of $\Theta$. As $\ell_1$-norm regularization yields sparse estimators, nuclear norm regularization enforces the solution to have sparse singular values, in other words, to be low-rank. Recent literatures have rigorously studied the statistical properties of $\hat{\Theta}$. [90] analyzed the nuclear norm penalization in estimating nearly low-rank matrices under the trace regression model. Specifically, they derived non-asymptotic estimation error bounds in terms of the Frobenius norm when the noise is sub-Gaussian. [101] proposed to use a Schatten-$p$ quasi-norm penalty where $p \leq 1$, and they derived non-asymptotic bounds on the prediction risk and Schatten-$q$ risk of the estimator, where $q \in [p, 2]$. Another effective method is through nuclear norm minimization under affine fitting constraint. Other important contributions include [101], [22], [16], [17], etc. When the true low-rank matrix $\Theta^*$ satisfies certain restricted isometry property (RIP) or similar properties, this approach can exactly recover $\Theta^*$ under the noiseless setting and enjoy sharp statistical error rate with sub-Gaussian and sub-exponential noise.

1.5 Notations

Here we collect the general notations before presenting the main results. More specialized notation is introduced in context.

We adopt the common convention of using boldface letters for vectors and matrices, while regular font is used for scalars. $|\cdot|$ denotes both absolute value and cardinality of a set, with the context ensuring no ambiguity. We denote the set $\{1, 2, 3, ..., d\}$ by $[d]$ for convenience. Given $a, b \in \mathbb{R}$, let $a \lor b$ and $a \land b$ denote the maximum and minimum of $a$ and $b$. For two scalar sequences $\{a_n\}_{n \geq 1}$ and $\{b_n\}_{n \geq 1}$, we say $a_n \gtrsim b_n$ ($a_n \lesssim b_n$) if there exists a universal constant $C > 0$ such that $a_n \geq C b_n$ ($a_n \leq C b_n$), and $a_n \asymp b_n$ if both $a_n \gtrsim b_n$ and $a_n \lesssim b_n$ hold. For $x = (x_1, \ldots, x_d)^\top \in \mathbb{R}^d$, and $1 \leq q \leq \infty$, we define $\|x\|_q = (\sum_{j=1}^{d} |x_j|^q)^{1/q}$, $\|x\|_0 = |\text{supp}(x)|$, where
\[ \text{supp}(x) = \{ j : x_j \neq 0 \}. \] Specifically, Write \( \| x \|_\infty = \max_{1 \leq j \leq d} |x_j| \), while for a matrix \( M = [M_{jk}] \), let \( \| M \|_{\text{max}} = \max_{j,k} |M_{jk}| \), \( \| M \|_1 = \sum_{j,k} |M_{jk}| \). For any matrix \( M \) we use \( M_\ell \) to index the transposed \( \ell \)th row of \( M \) and \( [M]_\ell \) to index the \( \ell \)th column. For \( q \geq r \), \( O_{q \times r} \) denotes the space of \( q \times r \) matrices with orthonormal columns. For a matrix \( M \in \mathbb{R}^{n \times d} \), we use \( \| M \|_F, \| M \|_* \) and \( \| M \|_{\text{op}} \) to denote the Frobenius norm, nuclear norm and spectral norm of \( M \), respectively. \( \text{Col}(M) \) represents the linear space spanned by column vectors of \( M \). We denote the Moore-Penrose pseudo inverse of a matrix \( M \in \mathbb{R}^{d \times d} \) by \( M^\dagger \). For a symmetric matrix \( M \), we use \( \lambda_j(M) \) and \( \lambda_{\text{min}}(A) \) to refer to its \( j \)th largest eigenvalue and minimum eigenvalue respectively. Denote the Euclidean and \( \ell_1 \)-norm ball with the center \( \beta^* \) and radius \( r \) by \( B_2(\beta^*, r) \) and \( B_1(\beta^*, r) \) respectively. For any \( \beta^* \in \mathbb{R}^d \) and any differential map \( f : \mathbb{R}^d \rightarrow \mathbb{R} \), define the first-order Taylor remainder of \( f(\beta) \) at \( \beta = \beta^* \) to be
\[ \delta f(\beta; \beta^*) := f(\beta) - f(\beta^*) - \nabla f(\beta^*)^\top (\beta - \beta^*). \]

For a random variable \( X \in \mathbb{R} \), we define \( \| X \|_{\psi_2} = \sup_{p \geq 1} (\mathbb{E}|X|^p)^{\frac{1}{p}} / \sqrt{p} \) and define \( \| X \|_{\psi_1} = \sup_{p \geq 1} (\mathbb{E}|X|^p)^{\frac{1}{p}} / p \). Please refer to \[123\] for equivalent definitions of \( \psi_2 \)-norm and \( \psi_1 \)-norm. For a random vector \( X \in \mathbb{R}^d \), its \( \psi_2 \) is defined as \( \| X \|_{\psi_2} = \sup_{x \in S^{d-1}} \| \langle X, x \rangle \|_{\psi_2} \), where \( S^{d-1} \) denotes the unit sphere in \( \mathbb{R}^d \). For two random variables \( X \) and \( Y \), we use \( X \overset{d}{=} Y \) to denote that \( X \) and \( Y \) have identical distributions. Let \( I_d \) denote the \( d \times d \) identity matrix; when the dimension is clear from the context, we omit the subscript. We also denote the Hadamard product of two matrices \( A \) and \( B \) as \( A \odot B \) and \( (A \odot B)_{jk} = A_{jk}B_{jk} \) for any \( j, k \). \( \{e_1, \ldots, e_d\} \) denotes the canonical basis for \( \mathbb{R}^d \). For a vector \( v \in \mathbb{R}^d \) and a set of indices \( S \subseteq \{1, \ldots, d\} \), \( v_S \) is the vector of length \( |S| \) whose components are \( \{v_j : j \in S\} \). Additionally, for a vector \( v \) with \( j \)th element \( v_j \), we use the notation \( v_{\setminus j} \) to denote the remaining vector when the \( j \)th element is removed. With slight abuse of notation, we write \( v = (v_j, v_{\setminus j}) \) when we
wish to emphasize the dependence of $v$ on $v_j$ and $v_{-j}$ individually. The gradient of a function $f(x)$ is denoted by $\nabla f(x)$, while $\nabla_x f(x, y)$ denotes the gradient of $f(x, y)$ with respect to $x$, and $\nabla^2_{xy} f(x, y)$ denotes the matrix of cross partial derivatives with respect to the elements of $x$ and $y$. For a scalar $\eta$, we simply write $f'(\eta) := \nabla_\eta f(\eta)$ and $f''(\eta) := \nabla^2_{\eta\eta} f(\eta)$. For a random variable $X$ and a sequence of random variables, \{$X_n$\}, we write $X_n \weak X$ when \{$X_n$\} converges weakly to $X$. If $X$ is a random variable with standard distribution, say $F_X$, we simply write $X_n \weak F_X$. 

Chapter 2

Distributed inference under sparse high dimensional models

2.1 Problem setup

2.1.1 General likelihood based framework

Let \((X_1^\top, Y_1^\top, \ldots, X_n^\top, Y_n^\top)\) be \(n\) i.i.d. copies of the random vector \((X^\top, Y)^\top\), whose realizations take values in \(\mathbb{R}^d \times \mathcal{Y}\). Write the collection of these \(n\) i.i.d. random couples as \(\mathcal{D} = \{(X_1^\top, Y_1^\top), \ldots, (X_n^\top, Y_n^\top)\}\) with \(Y = (Y_1, \ldots, Y_n)^\top\) and \(X = (X_1, \ldots, X_n)^\top \in \mathbb{R}^{n \times d}\). Conditional on \(X_i\), we assume \(Y_i\) is distributed as \(F_{\beta^*}\) for all \(i \in \{1, \ldots, n\}\), where \(F_{\beta^*}\) is a known distribution parameterized by a sparse \(d\)-dimensional vector \(\beta^*\) and has a density or mass function \(f_{\beta^*}\). We thus define the negative log-likelihood function, \(\ell_n(\beta)\), as

\[
\ell_n(\beta) = \frac{1}{n} \sum_{i=1}^{n} \ell_i(\beta) = -\frac{1}{n} \sum_{i=1}^{n} \log f_{\beta}(Y_i|X_i). \tag{2.1.1}
\]

We use \(J^* = J(\beta^*)\) to denote the information matrix and \(\Theta^*\) to denote \((J^*)^{-1}\), where \(J(\beta) = \mathbb{E}[\nabla_{\beta\beta}^2 \ell_n(\beta)]\).
For testing problems, our goal is to test $H_0 : \beta^*_v = \beta^*_v$ for a specific fixed index $v \in \{1, \ldots, d\}$. We partition $\beta^*$ as $\beta^* = (\beta^*_v, \beta^*_{-v}^\top)^\top \in \mathbb{R}^d$, where $\beta^*_{-v} \in \mathbb{R}^{d-1}$ is a vector of nuisance parameters and $\beta^*_v$ is the parameter of interest. To handle the curse of dimensionality, we exploit a penalized M-estimator defined as,

$$\hat{\beta}^\lambda = \arg\min_{\beta} \{\ell_n(\beta) + \mathcal{P}_\lambda(\beta)\},$$

(2.1.2)

with $\mathcal{P}_\lambda(\beta)$ a sparsity inducing penalty function with a regularization parameter $\lambda$. Examples of $\mathcal{P}_\lambda(\beta)$ include the convex $\ell_1$ penalty, $\mathcal{P}_\lambda(\beta) = \lambda \|\beta\|_1 = \lambda \sum_{j=1}^d |\beta_j|$ which, in the context of the linear model, gives rise to the Lasso estimator [114],

$$\hat{\beta}^\lambda_{\text{Lasso}} = \arg\min_{\beta} \left\{ \frac{1}{2n} \|Y - X\beta\|_2^2 + \lambda \|\beta\|_1 \right\}.$$  

(2.1.3)

Other penalties include folded concave penalties such as the smoothly clipped absolute deviation (SCAD) penalty [38] and minimax concave MCP penalty [132], which eliminate the estimation bias and attain the oracle rates of convergence [81, 128]. The SCAD penalty is defined as

$$\mathcal{P}_\lambda(\beta) = \sum_{j=1}^d p_\lambda(\beta_j), \quad \text{where} \quad p_\lambda(t) = \int_0^{|t|} \left\{ \lambda 1_{(z \leq \lambda)} + \frac{a_\lambda - z}{a-1} 1_{(z > \lambda)} \right\} dz,$$

(2.1.4)

for a given parameter $a > 0$ and MCP penalty is given by

$$\mathcal{P}_\lambda(\beta) = \sum_{j=1}^d p_\lambda(\beta_j), \quad \text{where} \quad p_\lambda(t) = \lambda \int_0^{|t|} \left( 1 - \frac{z}{\lambda b} \right)_+ dz$$

(2.1.5)

where $b > 0$ is a fixed parameter. The only requirement we have on $\mathcal{P}_\lambda(\beta)$ is that it induces an estimator satisfying the following condition.
Condition 2.1.1. For any $\delta \in (0, 1)$, if $\lambda \approx \sqrt{\log(d/\delta)/n}$,

$$\mathbb{P}\left( \|\hat{\beta}^\lambda - \beta^*\|_1 > C s n^{-1/2} \sqrt{\log(d/\delta)} \right) \leq \delta,$$  \hspace{1cm} (2.1.6)

where $s$ is the sparsity of $\beta^*$, i.e., $s = \|\beta^*\|_0$.

Condition 2.1.1 is crucial for the theory developed in Sections 2.2 and 2.3. Under suitable conditions on the design matrix $X$, it holds for the Lasso, SCAD and MCP. See [14, 38, 132] respectively and [136].

The DC algorithm randomly and evenly partitions $D$ into $k$ disjoint subsets $D_1, \ldots, D_k$, so that $D = \bigcup_{j=1}^k D_j$, $D_j \cap D_\ell = \emptyset$ for all $j, \ell \in \{1, \ldots, k\}$, and $|D_1| = |D_2| = \cdots = |D_k| = n_k = n/k$, where it is implicitly assumed that $n$ can be divided evenly. Let $I_j \subset \{1, \ldots, n\}$ be the index set corresponding to the elements of $D_j$.

Then for an arbitrary $n \times d$ matrix $A$, $A^{(j)} = [A_{i\ell}]_{i \in I_j, 1 \leq \ell \leq d}$. For an arbitrary estimator $\hat{\tau}$, we write $\hat{\tau}(D_j)$ when the estimator is constructed based only on $D_j$. Finally, we write $\ell_{n_k}^{(j)}(\beta) = \sum_{i \in I_j} \ell_i(\beta)$ to denote the negative log-likelihood function based on $D_j$.

While the results of this paper hold in a general likelihood based framework, for simplicity we state conditions at the population level for the generalized linear model (GLM) with canonical link. A much more general set of statements appear in the auxiliary lemmas upon which our main results are based. Under GLM with the canonical link, the response follows the distribution,

$$f_n(Y; X, \beta^*) = \prod_{i=1}^n f(Y_i; \eta_i^*) = \prod_{i=1}^n \left\{ c(Y_i) \exp \left[ \frac{Y_i \eta_i^* - b(\eta_i^*)}{\phi} \right] \right\}, \hspace{1cm} (2.1.7)$$
where \( \eta^*_i = X_i^\top \beta^* \). The negative log-likelihood corresponding to (2.1.7) is given, up to an affine transformation, by

\[
\ell_n(\beta) = \frac{1}{n} \sum_{i=1}^{n} -Y_i X_i^\top \beta + b(X_i^\top \beta) = \frac{1}{n} \sum_{i=1}^{n} -Y_i \eta_i + b(\eta_i) = \frac{1}{n} \sum_{i=1}^{n} \ell_i(\beta), \quad (2.1.8)
\]

and the gradient and Hessian of \( \ell_n(\beta) \) are respectively

\[
\nabla \ell_n(\beta) = -\frac{1}{n} X^\top (Y - \mu(X\beta)) \quad \text{and} \quad \nabla^2 \ell_n(\beta) = \frac{1}{n} X^\top D(X\beta) X,
\]

where \( \mu(\beta) = (b'(\eta_1), \ldots, b'(\eta_n))^\top \) and \( D(\beta) = \text{diag}\{b''(\eta_1), \ldots, b''(\eta_n)\} \). In this setting, \( J(\beta) = \mathbb{E}[b''(X_i^\top \beta)X_i X_i^\top] \) and \( J^* = \mathbb{E}[b''(X_i^\top \beta^*)X_i X_i^\top] \).

## 2.2 Divide and conquer hypothesis tests

In the context of the two classical testing frameworks, the Wald and Rao’s score tests, our objective is to construct a test statistic \( S_n \) with low communication cost and a tractable limiting distribution \( F \). From this statistic we define a test of size \( \alpha \) of the null hypothesis, \( H_0 : \beta^*_v = \beta^*_v^H \), against the alternative, \( H_1 : \beta^*_v \neq \beta^*_v^H \), as a partition of the sample space described by

\[
T_n^\alpha = \begin{cases} 
0 & \text{if } |S_n| \leq F^{-1}(1 - \alpha/2) \\
1 & \text{if } |S_n| > F^{-1}(1 - \alpha/2) 
\end{cases} \quad (2.2.1)
\]

for a two sided test.

### 2.2.1 Two divide and conquer Wald type constructions

For the high dimensional linear model, \([135]\), \([121]\) and \([60]\) propose methods for debiasing the Lasso estimator with a view to constructing high dimensional analogues of Wald statistics and confidence intervals for low-dimensional coordinates. As pointed
out by [135], the debiased estimator does not impose the minimum signal condition used in establishing oracle properties of regularized estimators [38, 44, 82, 129, 136] and hence has wider applicability than those inferences based on the oracle properties. The method of [121] is appealing in that it accommodates a general penalized likelihood based framework, while the [60] approach is appealing in that it optimizes asymptotic variance and requires a weaker condition than [121] in the specific case of the linear model. We consider the DC analogues of [60] and [121] in the following respectively.

**Lasso based Wald Test for the linear model**

The linear model assumes

\[ Y_i = X_i^\top \beta^* + \varepsilon_i, \quad (2.2.2) \]

where \( \{\varepsilon_i\}_{i=1}^n \) are i.i.d. with \( \mathbb{E}(\varepsilon_i) = 0 \) and variance \( \sigma^2 \). For concreteness, we focus on a Lasso based method, but our procedure is also valid when other pilot estimators are used. We describe a modification of the bias correction method introduced in [60] as a means to testing hypotheses on low dimensional coordinates of \( \beta^* \) via pivotal test statistics.

On each subset \( D_j \), we compute the debiased estimator of \( \beta^* \) as in [60] as

\[
\hat{\beta}^d(D_j) = \hat{\beta}_{Lasso}^\lambda(D_j) + \frac{1}{n_k} M^{(j)} (X^{(j)})^\top (Y^{(j)} - X^{(j)} \hat{\beta}_{Lasso}^\lambda(D_j)), \quad (2.2.3)
\]

where the superscript \( d \) is used to indicate the debiased version of the estimator, \( M^{(j)} = (m_1^{(j)}, \ldots, m_d^{(j)})^\top \) and \( m_v \) is the solution of

\[
m_v^{(j)} = \arg\min_m m^\top \hat{\Sigma}^{(j)} m \quad \text{s.t.} \quad \|\hat{\Sigma}^{(j)} m - e_v\|_\infty \leq \theta_1, \|X^{(j)} m\|_\infty \leq \theta_2. \quad (2.2.4)
\]
The choice of tuning parameters $\vartheta_1$ and $\vartheta_2$ is discussed in [60] and [139] and they suggest to choose $\vartheta_1 \approx \sqrt{\log d/n}$, $\vartheta_2 n^{-1/2} = o(1)$. In the context of our DC procedure, $\vartheta_1$ and $\vartheta_2$ rely on $k$ and should be chosen as $\vartheta_1 \approx \sqrt{k \log d/n}$, $\vartheta_2 n^{-1/2} = o(1)$, as quantified in Theorem 2.2.1. Above, $\hat{\Sigma}^{(j)} = n_k^{-1} \sum_{i \in I_j} X_i^{(j)} X_i^{(j)\top}$ is the sample covariance based on $D_j$, whose population counterpart is $\Sigma = \mathbb{E}(X_1 X_1\top)$ and $M^{(j)}$ is its regularized inverse. The second term in (2.2.3) is a bias correction term, while $\sigma^2 m_v^{(j)\top} \hat{\Sigma}^{(j)} m_v^{(j)}/n_k$ is shown in [60] to be the variance of the $v^{th}$ component of $\hat{\beta}^d (D_j)$. The parameter $\vartheta_1$, which tends to zero, controls the bias of the debiased estimator (2.2.3) and the optimization in (2.2.4) minimizes the variance of the resulting estimator.

Solving $d$ optimization problems in (2.2.4) increases an order of magnitude of computation complexity even for $k = 1$. Thus, it is necessary to appeal to the divide and conquer strategy to reduce the computation burden. This gives rise to the question how large $k$ can be in order to maintain the same statistical properties as the whole sample one ($k = 1$).

Because our DC procedure gives rise to smaller samples, $\hat{\Sigma}$ is singular. This singularity does not pose a statistical problem but it does make the optimization problem ill-posed. To overcome the singularity in $\hat{\Sigma}$ and the resulting instability of the algorithm, we propose a change of variables. More specifically, noting that $M^{(j)}$ is not required explicitly, but rather the product $M^{(j)}(X^{(j)})\top$, we propose

$$b_v^{(j)} = \arg\min_b \frac{b^{(j)\top} b^{(j)}}{n_k} \text{ s.t. } \left\| \frac{X^{(j)\top} b^{(j)}}{n_k} - e_v \right\|_\infty \leq \vartheta_1, \quad \|b^{(j)}\|_\infty \leq \vartheta_2,$$

from which we construct $M^{(j)}(X^{(j)})\top = B\top$, where $B = (b_1, \ldots, b_d)$. The algorithm in equation (2.2.5) is crucial to the success of our procedure in practice.
The following conditions on the data generating process and the tail behavior of the design vectors are imposed in [60]. Both conditions are used to derive the theoretical properties of the DC Wald test statistic based on the aggregated debiased estimator, \( \hat{\beta}^d = k^{-1} \sum_{j=1}^{k} \hat{\beta}^d(D_j) \).

**Condition 2.2.1.** \( \{(Y_i, X_i)\}_{i=1}^{n} \) are i.i.d. and \( \Sigma \) satisfies \( 0 < C_{\min} \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq C_{\max} \).

**Condition 2.2.2.** The rows of \( X \) are sub-Gaussian with \( \|X_i\|_{\psi_2} \leq \kappa, i = 1, \ldots, n. \)

Note that under the two conditions above, there exists a constant \( \kappa_1 > 0 \) such that \( \|X_1 \Sigma^{-\frac{1}{2}}\|_{\psi_2} \leq \kappa_1 \). Without loss of generality, we set \( \kappa_1 = \kappa \). Our first main theorem provides the relative scaling of the various tuning parameters involved in the construction of \( \hat{\beta}^d \).

**Theorem 2.2.1.** Suppose Conditions 2.1.1 [2.2.1] and 2.2.2 are fulfilled. Suppose \( \mathbb{E}[\varepsilon_1^4] < \infty \) and choose \( \vartheta_1, \vartheta_2 \) and \( k \) such that \( \vartheta_1 \asymp k \log d/n, \vartheta_2 n^{-1/2} = o(1) \) and \( k = o((s \log d)^{-1} \sqrt{n}) \). For any \( v \in \{1, \ldots, d\} \), we have

\[
\sqrt{n} k^{-1} \sum_{j=1}^{k} \frac{\hat{\beta}_v^d(D_j) - \beta_v^*}{\hat{Q}_v^{(j)}} \Rightarrow N(0, \sigma^2), \quad (2.2.6)
\]

where \( \hat{Q}_v^{(j)} = (m_v^{(j)} \Sigma^{(j)} m_v^{(j)})^{1/2} \).

Theorem 2.2.1 entertains the prospect of a divide and conquer Wald statistic of the form

\[
\overline{S}_n = \sqrt{n} k^{-1} \sum_{j=1}^{k} \frac{\hat{\beta}_v^d(D_j) - \beta_v^H}{\Sigma(m_v^{(j)} \Sigma^{(j)} m_v^{(j)})^{1/2}} \quad (2.2.7)
\]

for \( \beta_v^* \), where \( \Sigma \) is an estimator for \( \sigma \) based on the \( k \) subsamples. On the left hand side of equation (2.2.7) we suppress the dependence on \( v \) to simplify notation. As an estimator for \( \sigma \), a simple suggestion with the same computational complexity is \( \Sigma \)
\[ \hat{\sigma}^2 = \frac{1}{k} \sum_{j=1}^{k} \hat{\sigma}^2(D_j) \quad \text{and} \quad \hat{\sigma}^2(D_j) = \frac{1}{nk} \sum_{i \in I_j} (Y_i^{(j)} - X_i^{(j)} \lambda)^2. \] (2.2.8)

One can use the refitted cross-validation procedure of [35] to reduce the bias of the estimate. In Lemma 2.2.1 we show that with the scaling of \( k \) and \( \lambda \) required for the weak convergence results of Theorem 2.2.1, consistency of \( \hat{\sigma}^2 \) is also achieved.

**Lemma 2.2.1.** Suppose \( \mathbb{E}[\epsilon_i | X_i] = 0 \) for all \( i \in \{1, \ldots, n\} \). Then with \( \lambda \asymp \sqrt{k \log d/n} \) and \( k = o(\sqrt{n}(s \log d)^{-1}) \), \( |\hat{\sigma}^2 - \sigma^2| = o_P(1) \).

With Lemma 2.2.1 and Theorem 2.2.1 at hand, we establish in Corollary 2.2.1 the asymptotic distribution of \( S_n \) under the null hypothesis \( H_0 : \beta^*_v = \beta^H_v \). This holds for each component \( v \in \{1, \ldots, d\} \).

**Corollary 2.2.1.** Suppose Conditions 2.2.1 and 2.2.2 are fulfilled, \( \mathbb{E}[\epsilon_i^4] < \infty \), and \( \lambda, \vartheta_1 \) and \( \vartheta_2 \) are chosen as \( \lambda \asymp \sqrt{k \log d/n} \), \( \vartheta_1 \asymp \sqrt{k \log d/n} \) and \( \vartheta_2(n^{-1/2}) = o(1) \). Then provided \( k = o((s \log d)^{-1}\sqrt{n}) \), under \( H_0 : \beta^*_v = \beta^H_v \), we have

\[ \lim_{n \to \infty} \left| P(S_n \leq t) - \Phi(t) \right| = 0, \] (2.2.9)

where \( \Phi(\cdot) \) is the cdf of a standard normal distribution.

**Wald Test in the likelihood based framework**

An alternative route to debiasing the Lasso estimator of \( \beta^* \) is the one proposed in [121]. Their so called desparsified estimator of \( \beta^* \) is more general than the debiased estimator of [60] in that it accommodates generic estimators of the form (2.1.2) as pilot estimators, but the latter optimizes the variance of the resulting estimator. The
desparsified estimator for subsample $D_j$ is

$$\hat{\beta}^d(D_j) = \hat{\beta}^\lambda(D_j) - \hat{\Theta}^{(j)} \nabla \ell_n^{(j)}(\hat{\beta}^\lambda(D_j)),$$  \hfill (2.2.10)

where $\hat{\Theta}^{(j)}$ is a regularized inverse of the Hessian matrix of second order derivatives of $\ell^{(j)}_n(\beta)$ at $\hat{\beta}^\lambda(D_j)$, denoted by $\hat{\mathbf{J}}^{(j)} = \nabla^2 \ell_n^{(j)}(\hat{\beta}^\lambda(D_j))$. We will make this explicit in due course. The estimator resembles the classical one-step estimator [7], but now in the high-dimensional setting via regularized inverse of the Hessian matrix $\hat{\mathbf{J}}^{(j)}$, which reduces to the empirical covariance of the design matrix in the case of the linear model. From equation (2.2.10), the aggregated debiased estimator over the $k$ subsamples is defined as $\bar{\beta}^d = k^{-1} \sum_{j=1}^k \hat{\beta}^d(D_j)$.

We now use the nodewise Lasso [87] to approximately invert $\hat{\mathbf{J}}^{(j)}$ via $L_1$-regularization. The basic idea is to find the regularized invert row by row via a penalized $L_1$-regression, which is the same as regressing the variable $X_v$ on $X_{-v}$ but expressed in the sample covariance form. For each row $v \in 1, \ldots, d$, consider the optimization

$$\hat{\kappa}_v(D_j) = \arg\min_{\kappa \in \mathbb{R}^{d-1}} \left( \hat{\mathbf{J}}^{(j)}_{vv} - 2\hat{\mathbf{J}}^{(j)}_{v,v',v} \kappa + \kappa^\top \hat{\mathbf{J}}^{(j)}_{v,v',v} \kappa + 2\lambda_v \|\kappa\|_1 \right),$$  \hfill (2.2.11)

where $\hat{\mathbf{J}}^{(j)}_{v,v',v}$ denotes the $v$th row of $\hat{\mathbf{J}}^{(j)}$ without the $(v,v)$th diagonal element, and $\hat{\mathbf{J}}^{(j)}_{v,v',v}$ is the principal submatrix without the $v$th row and $v$th column. Introduce

$$\hat{\mathbf{C}}^{(j)} := \begin{pmatrix} 1 & -\hat{\kappa}_{1,2}(D_j) & \cdots & -\hat{\kappa}_{1,d}(D_j) \\ -\hat{\kappa}_{2,1}(D_j) & 1 & \cdots & -\hat{\kappa}_{2,d}(D_j) \\ \vdots & \vdots & \ddots & \vdots \\ -\hat{\kappa}_{d,1}(D_j) & -\hat{\kappa}_{d,2}(D_j) & \cdots & 1 \end{pmatrix}$$  \hfill (2.2.12)
and $\hat{\Xi}^{(j)} = \text{diag} (\hat{\tau}_1(D_j), \ldots, \hat{\tau}_d(D_j))$, where $\hat{\tau}_v(D_j)^2 = \hat{J}_{vv}^{(j)} - \hat{J}_{v,v}^{(j)} \hat{\kappa}_v(D_j)$. $\hat{\Theta}^{(j)}$ in equation (2.2.10) is given by

$$\hat{\Theta}^{(j)} = (\hat{\Xi}^{(j)})^{-2} \hat{C}^{(j)},$$

and we define $\hat{\Theta}_v^{(j)}$ as the transposed $v^{th}$ row of $\hat{\Theta}^{(j)}$.

Theorem 2.2.2 establishes the limit distribution of the term,

$$S_n = \sqrt{n} \sum_{j=1}^k \hat{\beta}^{d}_v(D_j) - \beta^H_v \sqrt{\Theta}_{vv}^{*}$$

for any $v \in \{1, \ldots, d\}$ under the null hypothesis $H_0 : \beta^*_v = \beta^H_v$. This provides the basis for the statistical testing based on divide-and-conquer. We need the following condition. Recall that $J^* = \mathbb{E}[\nabla \beta \ell_n(\beta^*)]$ and consider the generalized linear model (2.1.7).

**Condition 2.2.3.** (i) $\{(Y_i, X_i)\}_{i=1}^n$ are i.i.d., $0 < C_{\min} \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq C_{\max}$, $\lambda_{\min}(J^*) \geq L_{\min} > 0$, $\|J^*\|_{\max} < U_1 < \infty$. (ii) For some constant $M < \infty$, $\max_{1 \leq i \leq n} |X_i^\top \beta^*| \leq M$ and $\max_{1 \leq i \leq n} \|X_i\|_{\infty} \leq M$. (iii) There exist finite constants $U_2, U_3 > 0$ such that $b''(\eta) < U_2$ and $b'''(\eta) < U_3$ for all $\eta \in \mathbb{R}$.

The same assumptions appear in [121]. In the case of the Gaussian GLM, the condition on $\lambda_{\min}(J^*)$ reduces to the requirement that the covariance of the design has minimal eigenvalue bounded away from zero, which is a standard assumption. We require $\|J^*\|_{\max} < \infty$ to control the estimation error of different functionals of $J^*$. The restriction in (ii) on the covariates and the projection of the covariates are imposed for technical simplicity; it can be extended to the case of exponential tails [see 45]. Note that $\text{Var}(Y_i) = \phi b''(X_i^\top \beta^*)$ where $\phi$ is the dispersion parameter in (2.1.7), so $b''(\eta) < U_2$ essentially implies an upper bound on the variance of the response. In fact, Lemma A.5.2 shows that $b''(\eta) < U_2$ can guarantee that the response is sub-Gaussian.
$b''(\eta) < U_3$ is used to derive the Lipschitz property of $b''(X_i^\top \beta)$ with respect to $\beta$ as shown in Lemma A.5.5. We emphasize that no requirement in Condition 2.2.3 is specific to the divide and conquer framework.

The assumption of bounded design in (ii) can be relaxed to the sub-Gaussian design. However, the price to pay is that the allowable number of subsets $k$ is smaller than the bounded case, which means we need a larger sub-sample size. To be more precise, the order of maximum $k$ for the sub-Gaussian design has an extra factor, which is a polynomial of $\sqrt{\log d}$, compared to the order for the bounded design. This logarithmic factor comes from different Lipschitz properties of $b''(X_i^\top \beta)$ in the two designs, which is fully explained in Lemma A.5.5 in the Supplementary Material. In the following theorems, we only present results for the case of bounded design for technical simplicity.

In addition, recalling that $\Theta^* = (J^*)^{-1}$, where $J^* := J(\beta^*) = \mathbb{E}[\nabla^2_{\beta\beta} \ell_n(\beta^*)]$, we impose Condition 2.2.4 on $\Theta^*$ and its estimator $\hat{\Theta}$.

**Condition 2.2.4.** (i) $\min_{1 \leq v \leq d} \Theta_{vv}^* > \theta_{\min} > 0$. (ii) $\max_{1 \leq i \leq n} \|X_i^\top \Theta^*\|_\infty \leq M$. (iii) For $v = 1, \ldots, d$, whenever $\lambda_v \asymp \sqrt{k \log d/n}$ in (2.2.11), we have

$$\mathbb{P}\left(\|\hat{\Theta}_v - \Theta_v^*\|_1 \geq C s_1 \sqrt{\log d/n}\right) \leq d^{-1},$$

where $C$ is a constant and $s_1$ is such that $\|\Theta_v^*\|_0 \lesssim s_1$ for all $v \in \{1, \ldots, d\}$.

Part (i) of Corollary 2.2.4 ensures that the variances of each component of the debiased estimator exist, guaranteeing the existence of the Wald statistic. Parts (ii) and (iii) are imposed directly for technical simplicity. Results of this nature have been established under a similar set of assumptions in [121] and [92] for convex penalties and in [128] and [82] for folded concave penalties.
As a step towards deriving the limit distribution of the proposed divide and conquer Wald statistic in the GLM framework, we establish the asymptotic behavior of the aggregated debiased estimator $\beta_d^v$ for every given $v \in [d]$.

**Theorem 2.2.2.** Under Conditions 2.1.1, 2.2.3 and 2.2.4 with $\lambda \asymp \sqrt{k \log d/n}$, we have

$$\beta_d^v - \beta^*_v = -\frac{1}{k} \sum_{j=1}^{k} \Theta_v^{(j)^T} \nabla f_{n_k}^{(j)}(\beta^*) + o_P(n^{-1/2}) \quad (2.2.15)$$

for any $k \ll d$ satisfying $k = o\left(\left(\left(s \lor s_1\right) \log d\right)^{-1} \sqrt{n}\right)$, where $\Theta_v^{(j)}$ is the transposed $v^{th}$ row of $\Theta^{(j)}$.

The proof of Theorem 3.8 shows that for the Wald test procedure, the divide and conquer estimator $\beta_d^v$ is asymptotically as efficient as the full sample estimator $\hat{\beta}_v$, i.e.,

$$\lim_{n \to \infty} \frac{\text{Var}(\beta_d^v)}{\text{Var}(\beta_v^d)} - 1 = 0.$$

A corollary of Theorem 2.2.2 provides the asymptotic distribution of the Wald statistic in equation (2.2.14) under the null hypothesis.

**Corollary 2.2.2.** Let $S_n$ be as in equation (2.2.14), with $\Theta_*^{vv}$ replaced with an estimator $\hat{\Theta}_{vv}$. Then under the conditions of Theorem 2.2.2 and $H_0 : \beta^*_v = \beta_v^H$, provided $|\hat{\Theta}_{vv} - \Theta_{vv}| = o_P(1)$ under the scaling $k = o\left(\left(\left(s \lor s_1\right) \log d\right)^{-1} \sqrt{n}\right)$, we have

$$\lim_{n \to \infty} \sup_{t \in \mathbb{R}} |\mathbb{P}(S_n \leq t) - \Phi(t)| = 0.$$

**Remark 2.2.1.** Although Theorem 2.2.2 and Corollary 2.2.2 are stated only for the GLM, their proofs are in fact an application of two more general results. Further details are available in Lemmas A.5.6 and A.5.7 in the Supplementary Material.
We return to the issue of estimating $\Theta^*_v$ in Section 2.3, where we introduce a consistent estimator of $\Theta^*_v$ that preserves the scaling of Theorem 2.2.2 and Corollary 2.2.2.

### 2.2.2 Divide and conquer score test

In this section, we use $\nabla_v f(\beta)$ and $\nabla_{-v} f(\beta)$ to denote, respectively, the partial derivative of $f$ with respect to $\beta_v$ and the partial derivative vector of $f$ with respect to $\beta_{-v}$. $\nabla^2_v f(\beta), \nabla^2_{v,-v} f(\beta), \nabla^2_{-v,v} f(\beta)$ and $\nabla^2_{-v,-v} f(\beta)$ are analogously defined.

In the low dimensional setting (where $d$ is fixed), Rao’s score test of $H_0 : \beta^*_v = \beta^H_v$ against $H_1 : \beta^*_v \neq \beta^H_v$ is based on $\nabla_v \ell_n(\beta^H_v, \tilde{\beta}_{-v})$, where $\tilde{\beta}_{-v}$ is a constrained maximum likelihood estimator of $\beta^*_{-v}$, constructed as $\tilde{\beta}_{-v} = \arg\min_{\beta_{-v}} \ell_n(\beta^H_v, \beta_{-v}) = \arg\max_{\beta_{-v}} \{-\ell_n(\beta^H_v, \beta_{-v})\}$. If $H_0$ is false, imposing the constraint postulated by $H_0$ significantly violates the first order conditions from M-estimation with high probability; this is the principle underpinning the classical score test. Under regularity conditions, it can be shown [e.g. 31] that

$$\sqrt{n}(\nabla_v \ell_n(\beta^H_v, \tilde{\beta}_{-v})) J^*_{v|-v}^{1/2} \overset{d}{\sim} N(0, 1),$$

where $J^*_{v|-v}$ is given by $J^*_{v|-v} = J^*_{v,v} - J^*_{v,-v} J^*_{-v,v} J^*_{-v,-v}$, with $J^*_{v,v}, J^*_{v,-v}, J^*_{-v,v}$ and $J^*_{-v,-v}$ the partitions of the information matrix $J^* = J(\beta^*)$,

$$J(\beta) = \begin{pmatrix} J_{v,v} & J_{v,-v} \\ J_{-v,v} & J_{-v,-v} \end{pmatrix} = \begin{pmatrix} \mathbb{E} \nabla^2_{v,v} \ell_n(\beta) & \mathbb{E} \nabla^2_{v,-v} \ell_n(\beta) \\ \mathbb{E} \nabla^2_{-v,v} \ell_n(\beta) & \mathbb{E} \nabla^2_{-v,-v} \ell_n(\beta) \end{pmatrix}. \quad (2.2.16)$$

The problems associated with the use of the classical score statistic in the presence of a high dimensional nuisance parameter are brought to light by [95], who propose a remedy via the decorrelated score. The problem stems from the inversion of the
matrix $J^*_{v,-v}$ in high dimensions. The decorrelated score is defined as

$$S(\beta^*_v, \beta^*_u) = \nabla_v \ell_n(\beta^*_v, \beta^*_u) - w^* v \nabla_v \ell_n(\beta^*_v, \beta^*_u), \quad (2.2.17)$$

where $w^* = J^*_{v,-v} J^*_{v,-v}^{-1}$. For a regularized estimator $\hat{w}$ of $w^*$, to be defined below, we consider the score estimator

$$\hat{S}(\beta^*_v, \beta^*_u) = \nabla_v \ell_n(\beta^*_v, \beta^*_u) - \hat{w} \nabla_v \ell_n(\beta^*_v, \beta^*_u). \quad (2.2.18)$$

Hence, provided $w^*$ is sufficiently sparse to avoid excessive noise accumulation, we are able to achieve consistency of $\hat{w}$ under the high dimensional setting, ultimately giving rise to a tractable limit distribution of a suitable rescaling of $\hat{S}(\beta^*_v, \beta^*_u)$. Since $\beta^*_v$ is restricted under the null hypothesis, $H_0 : \beta^*_v = \beta^*_v$, the statistic in (2.2.18) is accessible once $H_0$ is imposed. As [95] point out, $w^*$ is the solution to

$$w^* = \arg\min_w \mathbb{E}[\nabla_v \ell_n(\beta^*_v, \beta^*_u) - w^* \nabla_v \ell_n(\beta^*_v, \beta^*_u)]^2$$

under $H_0 : \beta^*_v = \beta^*_v$.

Our divide and conquer score statistic under $H_0 : \beta^*_v = \beta^*_v$ is

$$\overline{S}(\beta^*_v) = \frac{1}{k} \sum_{j=1}^k \hat{S}^{(j)}(\beta^*_v, \beta^*_u(D_j)),$$  \quad (2.2.19)$$

$$\hat{S}^{(j)}(\beta_v, \beta^*_v(D_j)) = \nabla_v \ell_{n_k}(\beta_v, \beta^*_v(D_j)) - \hat{w}(D_j)^\top \nabla_v \ell_{n_k}(\beta_v, \beta^*_v(D_j))$$ and we estimate $w^*$ using the Dantzig selector of [21]

$$\hat{w}(D_j) = \arg\min_w \|w\|_1, \quad \text{s.t.} \quad \|\nabla^2 v, n_k \ell^{(j)}(\beta_v(D_j), \beta^*_v(D_j)) - w^\top \nabla^2 v, n_k \ell^{(j)}(\beta^*_v(D_j), \beta^*_v(D_j))\|_\infty \leq \mu.$$

30
Theorem 2.2.3. Let \( \hat{J}_{v|v} \) be a consistent estimator of \( J^*_{v|v} \) and

\[
S^{(j)}(\hat{\beta}_v^H, \hat{\beta}_v^*) = \nabla_v \ell^{(j)}_{n_k} (\hat{\beta}_v^H, \hat{\beta}_v^*) - \mathbf{w}^* \nabla_v \ell^{(j)}_{n_k} (\hat{\beta}_v^H, \hat{\beta}_v^*).
\]

Suppose \( \|\mathbf{w}^*\|_1 \lesssim s_1 \) and Conditions 2.1.1 and 2.2.3 are fulfilled. Then under \( H_0 : \beta_v^* = \beta_v^H \) with \( \lambda \asymp \mu \asymp \sqrt{k \log d/n} \),

\[
\sqrt{n} \mathbf{S}(\hat{\beta}_v^H) = \sqrt{n} \frac{1}{k} \sum_{j=1}^k S^{(j)}(\hat{\beta}_v^H, \hat{\beta}_v^*) + o_P(1)
\]

and

\[
\limsup_{n \to \infty} \mathbb{P} \left( \sqrt{n} \cdot \mathbf{S}(\hat{\beta}_v^H) \hat{J}_{v|v}^{-1/2} \leq t \right) - \Phi(t) = 0,
\]

for any \( k \ll d \) satisfying \( k = o\left(\left( (s \vee s_1) \log d \right)^{-1} \sqrt{n} \right) \), where \( \mathbf{S}(\hat{\beta}_v^H) \) is defined in equation (2.2.19).

Remark 2.2.2. By the definition of \( \mathbf{w}^* \) and the block matrix inversion formula for \( \Theta^* = (J^*)^{-1} \), sparsity of \( \mathbf{w}^* \) is implied by sparsity of \( \Theta^* \) as assumed in [121] and Condition 2.2.4 of Section 2.2.1. In turn, \( \|\mathbf{w}^*\|_0 \lesssim s_1 \) implies \( \|\mathbf{w}^*\|_1 \lesssim s_1 \) provided that the elements of \( \mathbf{w}^* \) are bounded.

Remark 2.2.3. Although Theorem 2.2.3 is stated in the penalized GLM setting, the result holds more generally; further details are available in Lemma A.5.11 in the Supplementary Material.

To maintain the same computational complexity, an estimator of the conditional information needs to be constructed using a DC procedure. For this, we propose to use

\[
\hat{J}_{v|v} = k^{-1} \sum_{j=1}^k \left( \nabla^2_{v,v} \ell^{(j)}_{n_k} (\hat{\beta}_v^d, \overline{\beta}_{v|-v}) - \hat{\mathbf{w}}^T \nabla^2_{v,v} \ell^{(j)}_{n_k} (\hat{\beta}_v^d, \overline{\beta}_{v|-v}) \right),
\]

where the divide and conquer estimator \( \overline{\beta}_v^d = k^{-1} \sum_{j=1}^k \hat{\beta}_v^d (\mathcal{D}_j) \), \( \overline{\beta}_{v|-v} = k^{-1} \sum_{j=1}^k \hat{\beta}_{v|-v} (\mathcal{D}_j) \) and \( \hat{\mathbf{w}} = k^{-1} \sum_{j=1}^k \hat{\mathbf{w}} (\mathcal{D}_j) \). Note that for certain \( v \), the communication cost for calculating \( \hat{J}_{v|v} \) is not high, since all the involved quantities \( \{ \nabla^2_{v,v} \ell^{(j)}_{n_k} (\hat{\beta}_v^d, \overline{\beta}_{v|-v}) \}_{j=1}^k \)
\[
\{\nabla^2_{-v,v} f_{\beta\cdot}(j)\} j=1^k \quad \text{and} \quad \{\hat{\mathbf{w}}(D_j)\} j=1^k
\]
are scalars, \(d\)-dimensional vectors and \(d\)-dimensional vectors respectively. The communication cost is thus of order \(O(kd)\). We do not communicate the entire huge Hessian matrix here.

**Lemma 2.2.2.** Suppose \(\|\mathbf{w}^*\|_1 = O(s_1)\) and Conditions 2.1.1 and 2.2.3 are fulfilled. Then for any \(k \ll d\) satisfying \(k = o\left((s \lor s_1) \log d\right)^{-1} \sqrt{n}\), \(|J_{\mathbf{v}|-\mathbf{v}} - J_{\mathbf{v}|-\mathbf{v}}^*| = o_p(1)\).

By Lemma 2.2.2 we show that \(J_{\mathbf{v}|-\mathbf{v}}\) is consistent and can be applied to Theorem 2.2.3.

### 2.3 Accuracy of distributed estimation

This section focuses on high-dimensional \((d \gg n)\) divide-and-conquer estimators for linear and generalized linear models. As explained below Theorem 2.2.2 in Section 3, the efficiency loss from the divide-and-conquer process is asymptotically zero. This motivates us to consider \(\|\hat{\beta}^d - \hat{\beta}^d\|\), the loss incurred by the divide and conquer strategy in comparison with the practically unavailable full sample debiased estimator \(\hat{\beta}^d\), where \(\|\cdot\|\) is certain norm. Indeed, it turns out that, for \(k\) not too large, \(\beta^d - \hat{\beta}^d\) appears only as a higher order term in the decomposition of \(\beta^d - \beta^*\) and thus \(\|\beta^d - \hat{\beta}^d\|\) is negligible compared to the statistical error, \(\|\hat{\beta}^d - \beta^*\|\). In other words, the divide-and-conquer errors are statistically negligible.

Compared with calculating the full sample debiased Lasso estimator, our proposed DC strategy enjoys computational advantages since it is highly parallel and each subsample problem has a much smaller scale than the full sample problem given a suitably large \(k\). However, relative to just the full sample penalized M-estimator (e.g., Lasso), distributed point estimation does not entail a computational gain like distributed testing, since our distributed algorithm requires debiasing each component of the Lasso estimator and hence brings high expense of computation. The bottleneck of computation of our DC procedure comes from the \(d\) extra debiasing steps. To
mitigate this problem, we can actually debias each component of $\hat{\beta}$ in a parallel fashion. According to the optimization procedures (3.4) and (3.11), debiasing one component of the Lasso estimator is entirely independent of the debiasing of another component. Therefore, as long as each branch computer in the cluster shares the sub-dataset $D_j$ and the Lasso estimator $\hat{\beta}^{(j)}$, they can work in parallel and collectively return to a central server all the components of the debiased Lasso estimator. This parallelization reduces the time complexity significantly.

When the minimum signal strength is sufficiently strong, thresholding $\bar{\beta}^d$ achieves exact support recovery, motivating a refitting procedure based on the low dimensional selected variables. As a means to understanding the theoretical properties of this refitting procedure, as well as for independent interest, we develop new theories and methodologies for the low dimensional ($d < n$) linear and generalized linear models in Appendixes A.1 and A.2 in the Supplementary Material respectively. We show that simple averaging of low dimensional OLS or GLM estimators (denoted uniformly as $\hat{\beta}^{(j)}$, without superscript $d$ as debiasing is not necessary) suffices to preserve the statistical error, i.e., achieving the same statistical accuracy as the estimator based on the full sample. This is because, in contrast to the high dimensional setting, parameters are not penalized in the low dimensional case. With $\bar{\beta}$ the average of $\hat{\beta}^{(j)}$ over the $k$ machines and $\hat{\beta}$ the full sample counterpart ($k = 1$), we derive the rate of convergence of $\|\beta - \hat{\beta}\|_2$. Refitted estimation using only the selected covariates allows us to eliminate the $\log d$ term in the statistical rate of convergence of the estimator under high-dimensional settings. We present theoretical results on the refitting estimation as corollaries to the low-dimensional regression results in Appendixes A.1 and A.2 in the Supplementary Material.
2.3.1 The high-dimensional linear model

Recall that the high dimensional DC estimator is \( \bar{\beta}^d = k^{-1} \sum_{j=1}^{k} \hat{\beta}^d(D_j) \), where \( \hat{\beta}^d(D_j) \) for \( 1 \leq j \leq k \) is the debiased estimator defined in (2.2.3). We also denote the debiased Lasso estimator using the entire dataset as \( \hat{\beta}^d = \beta^d(\cup_{j=1}^{k} D_j) \). The following lemma shows that not only is \( \bar{\beta}^d \) asymptotically normal, it approximates the full sample estimator \( \hat{\beta}^d \) so well that it has the same statistical error as \( \hat{\beta}^d \) provided the number of subsamples \( k \) is not too large.

Lemma 2.3.1. Consider the linear model (2.2.2). Under the Conditions 2.2.1 and 2.2.2 if \( \lambda, \vartheta_1 \) and \( \vartheta_2 \) are chosen as \( \lambda \asymp \sqrt{k \log d/n} \), \( \vartheta_1 \asymp \sqrt{k \log d/n} \) and \( \vartheta_2 n^{-1/2} = o(1) \), we have with probability \( 1 - c/d \),

\[
\|\bar{\beta}^d - \hat{\beta}^d\|_\infty \leq C \frac{sk \log d}{n} \quad \text{and} \quad \|\bar{\beta}^d - \beta^*\|_\infty \leq C \left( \sqrt{\frac{\log d}{n}} + \frac{sk \log d}{n} \right). \tag{2.3.1}
\]

Remark 2.3.1. The term \( \sqrt{\log d/n} \) in (2.3.1) is the estimation error of \( \|\hat{\beta}^d - \beta^*\|_\infty \), while the term \( (sk \log d)/n \) is the rate of the distance between the divide and conquer estimator and the full sample estimator. Lemma 2.3.1 does not rely on any specific choice of \( k \). However, in order for the aggregated estimator \( \bar{\beta}^d \) to attain the same \( \|\cdot\|_\infty \) norm estimation error as the full sample Lasso estimator, \( \hat{\beta}_{\text{Lasso}} \), the required scaling is \( k = O(\sqrt{n/(s^2 \log d)}) \). This is a weaker scaling requirement than that of Theorem 2.2.1 because the latter entails a guarantee of asymptotic normality, which is a stronger result. It is for the same reason that our estimation results only require \( O(\cdot) \) scaling whilst those for testing require \( o(\cdot) \) scaling.

[105] show that in the high-dimensional regime where \( d/n_k \to \kappa \in (0, 1) \), the divide and conquer procedure suffers from first-order accuracy loss. This seems a contradiction to our result, since our dimension is even higher than their context, but we have no first-order accuracy loss while averaging debiased estimators based on subsamples, as long as we have an appropriate number of data splits. In fact, in
the high-dimensional sparse linear regression, the intrinsic dimension is the sparsity \( s \) rather than \( d \), which is regarded instead as the ambient dimension. The sparsity assumption changes the original high-dimensional problem to be an intrinsically low-dimensional one and thus allows us to escape from any first-order accuracy loss of the divide and conquer procedure. Given \( s = o(n_k) \), we can treat high-dimensional sparse linear regression approximately as the classical linear regression setting where \( d = o(n_k) \). Hence we expect no first-order accuracy loss from the divide and conquer procedure here.

Although \( \beta^d \) achieves the same rate as the Lasso estimator under the infinity norm, it cannot achieve the minimax rate in \( \ell_2 \) norm since it is not a sparse estimator. To obtain an estimator with the \( \ell_2 \) minimax rate, we sparsify \( \beta^d \) by hard thresholding. For any \( \beta \in \mathbb{R}^d \), define the hard thresholding operator \( T_\nu \) such that the \( j \)-th entry of \( T_\nu(\beta^d) \) is
\[
[T_\nu(\beta)]_j = \beta_j 1\{|\beta_j| \geq \nu\}, \text{ for } 1 \leq j \leq d.
\]

According to (2.3.1), if \( \beta^*_j = 0 \), we have \( |\beta^d_j| \leq C(\sqrt{\log d/n} + sk \log d/n) \) with high probability. The following theorem characterizes the estimation error, \( \|T_\nu(\beta^d) - \beta^*\|_2 \), and divide and conquer error, \( \|T_\nu(\beta^d) - T_\nu(\hat{\beta}^d)\|_2 \), of the thresholded estimator \( T_\nu(\beta^d) \).

**Theorem 2.3.1.** Under the linear model (2.2.2), suppose Conditions 2.2.1 and 2.2.2 are fulfilled and choose \( \lambda \asymp \sqrt{k \log d/n} \), \( \vartheta_1 \asymp \sqrt{k \log d/n} \) and \( \vartheta_2 n^{-1/2} = o(1) \). Take the parameter of the hard threshold operator in (2.3.2) as \( \nu = C_0 \sqrt{\log d/n} \) for some sufficiently large constant \( C_0 \). If the number of subsamples satisfies \( k = O(\sqrt{n/(s^2 \log d)}) \), for large enough \( d \) and \( n \), we have with probability \( 1 - c/d \),
\[
\|T_\nu(\beta^d) - T_\nu(\hat{\beta}^d)\|_2 \leq C \frac{s^{3/2}k \log d}{n}, \quad \|T_\nu(\beta^d) - \beta^*\|_\infty \leq C \sqrt{\frac{\log d}{n}}
\]
and
\[
\|T_\nu(\beta^d) - \beta^*\|_2 \leq C \frac{s \log d}{n}.
\]
Remark 2.3.2. In fact, in the proof of Theorem 2.3.1, we show that if the thresholding parameter \( \nu \) satisfies \( \nu \geq \| \beta^d - \beta^* \|_\infty \), we have \( \| T_\nu(\beta^d) - \beta^* \|_2 \leq 2 \sqrt{2s} \cdot \nu \); it is for this reason that we choose \( \nu \approx \sqrt{\log d/n} \). Unfortunately, the constant is difficult to choose in practice. In the following paragraphs we propose a practical method to select the tuning parameter \( \nu \).

Let \(( M^{(j)} X^{(j)})^\top \) denote the transposed \( \ell^{th} \) row of \( M^{(j)} X^{(j)} \). Inspection of the proof of Theorem 2.2.1 reveals that the leading term of \( \sqrt{n} \| \beta^d - \beta^* \|_\infty \) satisfies

\[
T_0 = \max_{1 \leq \ell \leq d} \frac{1}{\sqrt{k}} \sum_{j=1}^k \frac{1}{\sqrt{n_k}} (M^{(j)} X^{(j)})_{\ell}^\top (\hat{\epsilon}^{(j)} \circ \xi^{(j)}),
\]

where \( \hat{\epsilon}^{(j)} \in \mathbb{R}^{nk} \) is an estimator of \( \epsilon^{(j)} \) such that for any \( i \in I_j \), \( \hat{\epsilon}^{(j)}_i = Y^{(j)}_i - X^{(j)}_i \hat{\beta}(D_j) \), and \( \xi^{(j)} \) is a subvector of \( \{ \xi_i \}_{i=1}^n \) with indices in \( I_j \). Recall that “\( \circ \)” denotes the Hadamard product. The \( \alpha \)-quantile of \( W_0 \) conditioning on \( \{ Y_i, X_i \}_{i=1}^n \) is defined as \( c_{W_0}(\alpha) = \inf \{ t \mid \mathbb{P}(W_0 \leq t \mid Y, X) \geq \alpha \} \). We estimate \( c_{W_0}(\alpha) \) by Monte-Carlo and thus choose \( \nu_0 = c_{W_0}(\alpha)/\sqrt{n} \). This choice ensures

\[
\| T_{\nu_0}(\beta^d) - \beta^* \|_2 = O_\mathbb{P}(\sqrt{s \log d/n}),
\]

which coincides with the \( \ell_2 \) convergence rate of the Lasso.

Remark 2.3.3. Lemma 2.3.1 and Theorem 2.3.1 show that if the number of sub-samples satisfies \( k = o(\sqrt{n}/(s^2 \log d)) \), \( \| \beta^d - \hat{\beta}^d \|_\infty = o_\mathbb{P}(\sqrt{\log d/n}) \) and \( \| T_\nu(\beta^d) - \beta^* \|_2 \leq 2 \sqrt{2s} \cdot \nu \) for this reason that we choose \( \nu \approx \sqrt{\log d/n} \). Unfortunately, the constant is difficult to choose in practice. In the following paragraphs we propose a practical method to select the tuning parameter \( \nu \).

Let \(( M^{(j)} X^{(j)})^\top \) denote the transposed \( \ell^{th} \) row of \( M^{(j)} X^{(j)} \). Inspection of the proof of Theorem 2.2.1 reveals that the leading term of \( \sqrt{n} \| \beta^d - \beta^* \|_\infty \) satisfies

\[
T_0 = \max_{1 \leq \ell \leq d} \frac{1}{\sqrt{k}} \sum_{j=1}^k \frac{1}{\sqrt{n_k}} (M^{(j)} X^{(j)})_{\ell}^\top (\hat{\epsilon}^{(j)} \circ \xi^{(j)}),
\]

where \( \hat{\epsilon}^{(j)} \in \mathbb{R}^{nk} \) is an estimator of \( \epsilon^{(j)} \) such that for any \( i \in I_j \), \( \hat{\epsilon}^{(j)}_i = Y^{(j)}_i - X^{(j)}_i \hat{\beta}(D_j) \), and \( \xi^{(j)} \) is a subvector of \( \{ \xi_i \}_{i=1}^n \) with indices in \( I_j \). Recall that “\( \circ \)” denotes the Hadamard product. The \( \alpha \)-quantile of \( W_0 \) conditioning on \( \{ Y_i, X_i \}_{i=1}^n \) is defined as \( c_{W_0}(\alpha) = \inf \{ t \mid \mathbb{P}(W_0 \leq t \mid Y, X) \geq \alpha \} \). We estimate \( c_{W_0}(\alpha) \) by Monte-Carlo and thus choose \( \nu_0 = c_{W_0}(\alpha)/\sqrt{n} \). This choice ensures

\[
\| T_{\nu_0}(\beta^d) - \beta^* \|_2 = O_\mathbb{P}(\sqrt{s \log d/n}),
\]

which coincides with the \( \ell_2 \) convergence rate of the Lasso.
\( T_v(\hat{\beta}^d)\|_2 = o_p(\sqrt{s \log d/n}) \), and thus the error incurred by the divide and conquer procedure is negligible compared to the statistical minimax rate. The reason for this contraction phenomenon is that \( \hat{\beta}^d \) and \( \beta^d \) share the same leading term in their Taylor expansions around \( \beta^* \). The difference between them is only the difference of two remainder terms which has a smaller order than the leading term. We uncover a similar phenomenon in the low dimensional case covered in Appendix A.1 in the Supplementary Material. However, in the low dimensional case \( \ell_2 \) norm consistency is automatic while the high dimensional case requires an additional thresholding step to guarantee sparsity and, consequently, \( \ell_2 \) norm consistency.

2.3.2 The high-dimensional generalized linear model

We generalize the DC estimation of the linear model to GLM. Recall that \( \hat{\beta}^d(D_j) \) is the de-biased estimator defined in (2.2.10) and the aggregated estimator is \( \bar{\beta}^d = k^{-1} \sum_{j=1}^k \hat{\beta}^d(D_j) \). We still denote \( \hat{\beta}^d = \hat{\beta}^d(\bigcup_{j=1}^k D_j) \). The next lemma bounds the error incurred by splitting the sample and the statistical rate of convergence of \( \bar{\beta}^d \) in terms of the infinity norm.

**Lemma 2.3.2.** Consider the generalized linear model (2.1.7) with canonical link. Under Conditions 2.1.1, 2.2.3 and 2.2.4, for \( \hat{\beta}^\lambda \) with \( \lambda \asymp \sqrt{k \log d/n} \), we have with probability \( 1 - c/d \), there exists a constant \( C > 0 \), such that

\[
\|\bar{\beta}^d - \hat{\beta}^d\|_\infty \leq C \frac{(s \lor s_1)k \log d}{n}, \|\bar{\beta}^d - \beta^*\|_\infty \leq C \left( \sqrt{\frac{\log d}{n}} + \frac{(s \lor s_1)k \log d}{n} \right).
\]

**Remark 2.3.4.** The term \( \sqrt{\log d/n} \) in above is the estimation error of \( \|\hat{\beta}^d - \beta^*\|_\infty \), while the error term \( (s \lor s_1)k \log d/n \) is attributable to the distance between \( \bar{\beta}^d \) and \( \hat{\beta}^d \).
Applying a similar thresholding step as in the linear model, we quantify the $\ell_2$-norm estimation error, $\|T_\nu(\hat{\beta}^d) - \beta^*\|_2$ and the distance between the divide and conquer estimator and full sample estimator $\|T_\nu(\bar{\beta}^d) - T_\nu(\tilde{\beta}^d)\|_2$.

**Theorem 2.3.2.** For the GLM (2.1.7), under Conditions 2.1.1 - 2.2.4, choose $\lambda \approx \sqrt{k \log d/n}$ and $\lambda_v \approx \sqrt{k \log d/n}$. Take the parameter of the hard threshold operator in (2.3.2) as $\nu = C_0 \sqrt{\log d/n}$ for some sufficiently large constant $C_0$. If the number of subsamples satisfies $k = O\left(\sqrt{n/((s \vee s_1)^2 \log d)}\right)$, for large enough $d$ and $n$, we have with probability $1 - c/d$,

$$\|T_\nu(\bar{\beta}^d) - T_\nu(\tilde{\beta}^d)\|_2 \leq C\frac{(s \vee s_1) s^{1/2} k \log d}{n}, \quad \|T_\nu(\bar{\beta}^d) - \beta^*\|_\infty \leq C\sqrt{\frac{\log d}{n}} \quad (2.3.3)$$

and $\|T_\nu(\bar{\beta}^d) - \beta^*\|_2 \leq C\sqrt{s \log d/n}$.

**Remark 2.3.5.** As in the case of the linear model, Theorem 2.3.2 reveals that the loss incurred by the divide and conquer procedure is negligible compared to the statistical minimax estimation error provided $k = o\left(\sqrt{n/(s \vee s_1)^2 \log d}\right)$.

A similar proof strategy to that of Theorem 2.3.2 allows us to construct an estimator of $\Theta^*_vv$ that achieves the required consistency with the scaling of Corollary 2.2.2. Our estimator is $\tilde{\Theta}_{vv} := [T_\zeta(\Theta)]_{vv}$, where $\Theta = k^{-1} \sum_{j=1}^k \tilde{\Theta}^{(j)}$ and $T_\zeta(\cdot)$ is the thresholding operator defined in equation (2.3.2) with $\zeta = C_1 \sqrt{\log d/n}$ for some sufficiently large constant $C_1$.

**Corollary 2.3.1.** Under the conditions and scaling of Theorem 2.2.2, $|\tilde{\Theta}_{vv} - \Theta^*_vv| = o_P(1)$.

Substituting this estimator in Corollary 2.2.2 delivers a practically implementable test statistic based on $k = o\left((s \vee s_1) \log d)^{-1} \sqrt{n}\right)$ subsamples.

**Remark 2.3.6.** Notice that point estimation requires less stringent scaling of $k$ than hypothesis testing in both the linear and generalized linear models. This is because
the testing and estimation require different rates for the higher order term $\Delta$ in the decomposition

$$\sqrt{n}(\beta^d - \beta^*) = Z + \Delta,$$

where $Z$ is the leading term contributing to the asymptotic normality of $\sqrt{n}(\beta^d - \beta^*)$. For hypothesis testing, we need $\|\Delta/\sqrt{n}\|_\infty = o_P(1/\sqrt{n})$ to guarantee the asymptotic normality. For estimation, we need $\|\Delta/\sqrt{n}\|_\infty = o_P(\sqrt{\log d/n})$ to match the minimax rate of $\|\beta^d - \beta^*\|_\infty$. Therefore, the number of splits $k$ for testing is more stringent by a factor of $1/\sqrt{\log d}$ than in estimation.

2.4 Simulations

In this section, we illustrate and validate our theoretical findings through simulations. For hypothesis testing, we use QQ plots to compare the distribution of $p$-values for divide and conquer test statistics to their theoretical uniform distribution. We also investigate the estimated type I error and power of the divide and conquer tests. For estimation, we validate our claim in the beginning of Section 2.3 that the loss incurred by the divide and conquer strategy is negligible compared with the statistical error of the corresponding full sample estimator in the high dimensional case. Specifically, we compare the performance of the divide and conquer thresholding estimator of Section 2.3.1 with the full sample Lasso and the average Lasso over subsamples. An analogous empirical verification of the theory is performed for the low dimensional case as well; we put it in Appendixes A.3 and A.4 of the Supplementary Material.
Figure 2.1: QQ plots of the \( p \)-values of the Wald (A) and score (B) divide and conquer test statistics against the theoretical quantiles of the uniform \([0,1]\) distribution under the null hypothesis.

### 2.4.1 Results on hypothesis testing

We explore the probability of rejection of a null hypothesis of the form \( H_0 : \beta^*_1 = 0 \) when data \( (Y_i, X_i)_{i=1}^n \) are generated according to the linear model,

\[
Y_i = X_i^\top \beta^* + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2_\epsilon),
\]

where \( \sigma^2_\epsilon = 1 \) and

\[
\beta^* = (\beta^*_1, 0, \cdots, 0, 1, \cdots, 1)^\top, 
\]

where \( d = 5000 \) and \( s = 3 \). In each Monte Carlo replication, we split the initial sample of size \( n \) into \( k \) subsamples of size \( n/k \). In particular we choose \( n = 5000 \) and \( k \in \{1, 2, 5, 10, 20, 25, 40, 50, 100, 200, 500\} \). The number of Monte Carlo replications is 500. Using \( \hat{\beta}_{\text{Lasso}} \) as a preliminary estimator of \( \beta^* \), we construct Wald and Rao’s score test statistics as described in Sections 2.2.1 and 2.2.2 respectively. Panels (A) and (B) of Figure 2.1 are QQ plots of the \( p \)-values of the divide and conquer Wald and score test statistics under the null hypothesis against the theoretical
quantiles of the uniform $[0,1]$ distribution for eight different values of $k$. For both test constructions, the distributions of the $p$-values are close to uniform and remain so as we split the data set. When $k \geq 100$, the distribution of the corresponding $p$-values deviates from the uniform distribution visibly, as expected from the theory developed in Sections 2.2.1 and 2.2.2. Panel (A) of Figure 2.2 shows that, for both test constructions, when the number of splits $k \leq 50$, the empirical level of the test is close to both the nominal $\alpha = 0.05$ level and the level of the full sample oracle OLS estimator which knows the true support of $\beta^*$. On the other hand, the type I error increases dramatically when $k$ is larger than 50. This is consistent with asymptotic normality of the test statistics we established when $k$ is controlled appropriately. Panel (B) of Figure 2.2 displays the power of the test for two different signal strengths, $\beta_1^* = 0.05$ and 0.06. We see that the power for the Score and Wald tests improves when the signal strength goes from 0.05 to 0.06. In addition, we find that the power is high regardless of how large $k$ is. However, Figure 2.2(A) shows that the Type I error is large when $k$ is large, which makes the tests invalid. Therefore, these results illustrate that the Type I and II errors are controllable when the number of splits $k$ is relatively small. We also record the wall time for computation for these $k$’s in Table 2.1. The wall time is computed by taking the maximal time taken for each split and averaged over replications.

2.4.2 Results on estimation

In this section, we turn our attention to experimental validation of our divide and conquer estimation theory, focusing first on the low dimensional case and then on the high dimensional case.
Figure 2.2: (A) Estimated probabilities of type I error for the Wald and score tests as a function of $k$. (B) Estimated power with signal strength 0.05 and 0.06 for the Wald, and score tests as a function of $k$.  

Table 2.1: Computation time for the divide and conquer testing and estimation, where $k = 1$ corresponds to the non-splitting case and $k > 1$ corresponds to the distributed case. 

<table>
<thead>
<tr>
<th>$k$</th>
<th>1</th>
<th>2</th>
<th>5</th>
<th>10</th>
<th>20</th>
<th>25</th>
<th>50</th>
<th>100</th>
<th>200</th>
</tr>
</thead>
<tbody>
<tr>
<td>Score test (s)</td>
<td>364.39</td>
<td>73.22</td>
<td>35.09</td>
<td>23.61</td>
<td>23.56</td>
<td>20.78</td>
<td>24.13</td>
<td>37.53</td>
<td>64.67</td>
</tr>
<tr>
<td>Wald test (s)</td>
<td>426.23</td>
<td>68.95</td>
<td>19.66</td>
<td>10.09</td>
<td>6.70</td>
<td>5.71</td>
<td>3.88</td>
<td>2.60</td>
<td>1.91</td>
</tr>
<tr>
<td>$\mathcal{T}_\nu(\beta^d)(10^3s)$</td>
<td>61.50</td>
<td>30.00</td>
<td>7.92</td>
<td>6.58</td>
<td>4.48</td>
<td>2.94</td>
<td>2.64</td>
<td>2.11</td>
<td>1.66</td>
</tr>
<tr>
<td>Split Lasso (s)</td>
<td>89.18</td>
<td>32.02</td>
<td>34.57</td>
<td>6.47</td>
<td>4.87</td>
<td>4.16</td>
<td>2.56</td>
<td>1.92</td>
<td>2.64</td>
</tr>
</tbody>
</table>

The high dimensional linear model

We now consider the same setting of Section 2.4.1 with $n = 5000$, $d = 5000$ and $\beta^*_j = 10$ for all $j$ in the support of $\beta^*$. In this context, we analyze the performance of the thresholded averaged debiased estimator of Section 2.3.1. Figure 2.3(A) depicts the average over 100 Monte Carlo replications of $\|b - \beta^*\|_2$ for three different estimators: debiased divide-and-conquer $b = \mathcal{T}_\nu(\overline{\beta}^d)$, the Lasso estimator based on the whole sample $b = \hat{\beta}_{\text{Lasso}}$ and the estimator obtained by naively averaging the Lasso estimators from the $k$ subsamples $b = \overline{\beta}_{\text{Lasso}}$. The parameter $\nu$ is taken as $\nu = \sqrt{\log d/n}$ in the specification of $\mathcal{T}_\nu(\overline{\beta}^d)$. As expected, the performance of $\overline{\beta}_{\text{Lasso}}$ deteriorates sharply as $k$ increases. $\mathcal{T}_\nu(\overline{\beta}^d)$ outperforms $\hat{\beta}_{\text{Lasso}}$ as long as $k$ is not too large. This is expected because, for sufficiently large signal strength, both $\hat{\beta}_{\text{Lasso}}$ and
\( T_\nu(\tilde{\beta}^d) \) recover the correct support, however \( T_\nu(\tilde{\beta}^d) \) is unbiased for those \( \beta^*_j \) in the support of \( \beta^* \), whilst \( \hat{\beta}_{\text{Lasso}} \) is biased. Figure 2.3(B) shows the error incurred by the divide and conquer procedure \( \| T_\nu(\tilde{\beta}^d) - T_\nu(\hat{\beta}^d) \|_2 \) relative to the statistical error of the full sample estimator, \( \| T_\nu(\tilde{\beta}^d) - \beta^* \|_2 \), for four different scalings of \( k \). We observe that, with \( k \approx O(\sqrt{n/s^2 \log d}) \), the relative error incurred by the divide and conquer procedure can hardly converge. This is consistent with Theorem 2.3.1. Given the lower bound of statistical error of the full sample Lasso estimator \( \hat{\beta} \), From Theorem 2.3.1 we derive that

\[
\frac{E \| T_\nu(\tilde{\beta}^d) - T_\nu(\hat{\beta}^d) \|_2^2}{E \| T_\nu(\hat{\beta}^d) - \beta^* \|_2^2} \leq \frac{s^2 k^2 \log d}{n}.
\]

When \( k \approx O(\sqrt{n/s^2 \log d}) \), the righthand side is an \( O(1) \) term. Therefore the line with inverted triangles in Figure 2.3(B) implies that the statistical error rate we developed in Theorem 2.3.1 is tight. We also record the wall time for estimation computation for these \( k \)’s in Table 2.1. The wall time is computed by taking the maximal time taken for each splits and averaged over replications. We notice that the computation time decreases with \( k \) at first due to the parallel algorithm. However, for the score test and split Lasso, the time becomes increasing when \( k \) is large, this is because the computation time to aggregate results from different splits is no longer negligible for very large \( k \)’s.

## 2.5 Proofs

In this section, we present the proofs of the main theorems appearing in Sections 2.2 and 2.3. The statements and proofs of several auxiliary lemmas appear in the Supplementary Material. To simplify notation, we take \( \beta_\nu^H = 0 \) without loss of generality.
2.5.1 Proofs for Section 2.2.1

The proof of Theorem 2.2.1 relies on the following lemma, which bounds the probability that optimization problems in (2.2.4) are feasible.

Lemma 2.5.1. Assume \( \Sigma = \mathbb{E}(X_iX_i^\top) \) satisfies \( C_{\min} < \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq C_{\max} \) as well as \( \| \Sigma^{-1/2}X_1 \|_\psi_2 = \kappa \), then we have

\[
P \Bigg( \max_{j=1,\ldots,k} \| M^{(j)}\hat{\Sigma}^{(j)} - I \|_\text{max} \leq a \sqrt{\frac{\log d}{n}} \Bigg) \geq 1 - 2kd^{-c_2},
\]

where \( c_2 = \frac{a^2C_{\min}}{24e^2\kappa^2C_{\max}} - 2 \).

Proof. The proof is an application of the union bound in Lemma 6.2 of [60]. □

Proof of Theorem 2.2.1. For \( 1 \leq j \leq k \), let \( \sqrt{\hat{n}_k}(\hat{\beta}^{d_j} - \beta^*) = Z^{(j)} + \Delta^{(j)} \), where \( Z^{(j)} = \frac{1}{\sqrt{n}_k}M^{(j)}X^{(j)\top}e^{(j)} \). From Theorem A.6.1, we know that as long as \( m_v^{(j)\top} \hat{\Sigma}^{(j)} m_v^{(j)} \geq c > 0 \) holds uniformly for \( j = 1, \ldots, d \),

\[
\overline{\Delta} := \sqrt{n}_k^{-1} \sum_{j=1}^k \frac{\Delta_v^{(j)}}{Q^{(j)}} = o_P(1).
\]
Then we define
\[ V_n := \sqrt{n} \frac{1}{k} \sum_{j=1}^{k} Z_j^{(j)} \]$$ Q_j^{(j)} = \sum_{j=1}^{k} \sum_{i \in I_j} \xi_{iv}^{(j)} , \quad \text{where} \quad \xi_{iv}^{(j)} := \frac{m_v^{(j)^T} X_i^{(j)} \epsilon_i^{(j)}}{(n\bar{m}_v^{(j)^T} \bar{\Sigma}^{(j)} m_v^{(j)})^{1/2}}. $$

We now establish the asymptotic normality of $V_n$ by verifying the requirements of the Lindeberg-Feller central limit theorem [e.g. [62, Theorem 4.12]]. By the fact that $\epsilon_i$ is independent of $X$ for all $i$ and $E[\epsilon_i] = 0$,

\[ \mathbb{E}(\xi_{iv}) = \mathbb{E}(\mathbb{E}(\xi_{iv}^{(j)} \mid X)) = \mathbb{E}\left\{ \mathbb{E}\left[ m_v^{(j)^T} X_i^{(j)} \epsilon_i^{(j)} / (n\bar{m}_v^{(j)^T} \bar{\Sigma}^{(j)} m_v^{(j)})^{1/2} \mid X \right] \right\} = \mathbb{E}\left\{ (n\bar{m}_v^{(j)^T} \bar{\Sigma}^{(j)} m_v^{(j)})^{-1/2} m_v^{(j)^T} X_i^{(j)} \mathbb{E}(\epsilon_i^{(j)}) \right\} = 0. \]

By independence of $\{\epsilon_i\}_{i=1}^n$ and the definition of $\hat{\Sigma}^{(j)}$, we also have

\[ \text{Var}(\xi_{iv} | X) = \frac{1}{n} \sum_{j=1}^{k} \sum_{i \in I_j} \text{Var}(\xi_{iv}^{(j)} | X) = \frac{1}{n} \sum_{j=1}^{k} \bar{m}_v^{(j)^T} \bar{\Sigma}^{(j)} m_v^{(j)} - \sum_{i \in I_j} m_v^{(j)^T} X_i^{(j)} X_i^{(j)^T} m_v^{(j)} \text{Var}(\epsilon_i^{(j)} | X) = \sigma^2. \]

Therefore we have

\[ \text{Var}(V_n) = \mathbb{E}(\text{Var}(V_n | X)) + \text{Var}(\mathbb{E}(V_n | X)) = \sigma^2. \]

It only remains to verify the Lindeberg condition, i.e.,

\[ \lim_{k \to \infty} \lim_{n_k \to \infty} \frac{1}{\sigma^2} \sum_{j=1}^{k} \sum_{i \in I_j} \mathbb{E}\left[ (\xi_{iv}^{(j)})^2 \mathbb{1}\{|\xi_{iv}^{(j)}| > \epsilon \sigma\} \right] = 0, \quad \forall \epsilon > 0, \quad (2.5.1) \]

whose verification is relegated to the Appendix A.5 of the Supplementary Material. Finally we reach the conclusion by the Slutsky’s Theorem.
Proof of Corollary 2.2.1. Let $F_n := \{m_v^{(j)\top} \Sigma^{(j)} m_v^{(j)} \geq c > 0, j = 1, ..., k\}$, where $n$ is the total sample size. According to Theorem 2.2.1, when $F_n$ holds, we have

$$\lim_{n \to \infty} P(S_n \leq t \mid X) - \Phi(t) = 0.$$  

From the proof of Lemma 13 in [60], $\lim_{n \to \infty} P(F_n) = 1$. For any $t \in \mathbb{R}$ and $\delta > 0$, by applying dominating convergence Theorem to $\mathbb{1}\{|P(S_n \leq t \mid X) - \Phi(t)| > \delta\}$ and $F_n$ holds}, we have

$$\lim_{n \to \infty} P(\mathbb{1}\{|P(S_n \leq t \mid X) - \Phi(t)| > \delta\}) = 0.$$  

According to the dominate convergence theorem, since $P(S_n \leq t \mid X) \in [0, 1]$, we have

$$\lim_{n \to \infty} P(S_n \leq t) = \lim_{n \to \infty} E[P(S_n \leq t \mid X)] = E[\lim_{n \to \infty} P(S_n \leq t \mid X)] = \Phi(t).$$

Therefore, we complete the proof of the corollary.

The proofs of Theorem 2.2.2 and Corollary 2.2.2 are stated as an application of Lemmas A.5.6 and A.5.7 in the Supplementary Material, which apply under a more general set of requirements. We present the proof of Theorem 2.2.2 below and defer Corollary 2.2.2 to Appendix A.5 in the Supplementary Materials.

Proof of Theorem 2.2.2. We verify (A1)-(A4) of Lemma A.5.6. For (A1), decompose the object of interest as

$$\frac{1}{n_k} \|X^{(j)}\hat{\Theta}^{(j)}\|_{\max} = \frac{1}{n_k} \|X^{(j)}(\hat{\Theta}^{(j)} - \Theta^*)\|_{\max} + \frac{1}{n_k} \|X^{(j)}\Theta^*\|_{\max} = \Delta_1 + \Delta_2,$$
where $\Delta_1$ can be further decomposed and bounded by
\[
\frac{1}{n_k} \left\| X^{(j)} (\hat{\Theta}^{(j)} - \Theta^*) \right\| \leq \frac{1}{n_k} \max_{1 \leq i \leq n} \max_{1 \leq v \leq d} \left[ \left\| X_i^{(j)} \right\|_1 \right] \\
\leq \frac{1}{n_k} \max_{1 \leq i \leq n} \max_{1 \leq v \leq d} \left\| \hat{\Theta}^{(j)} - \Theta^* \right\|_1.
\]

We have
\[
\mathbb{P}(\Delta_1 > q/2) \leq \mathbb{P} \left( \max_{1 \leq v \leq d} \left\| \hat{\Theta}^{(j)} - \Theta^* \right\|_1 > \frac{n q}{kM} \right) < \psi
\]
and by Condition 2.2.4, $\psi = o(d^{-1}) = o(k^{-1})$ for any $q \geq 2CMs_1(k/n)^{3/2} \cdot \sqrt{\log d}$, a fortiori for $q$ a constant. Since $X_i$ is sub-Gaussian, a matching probability bound can easily be obtained for $\Delta_2$, thus we obtain
\[
\mathbb{P} \left( n_k^{-1} \left\| X^{(j)} \right\| \leq 2 \psi
\]
for $\psi = o(k^{-1})$. (A2) and (A3) of Lemma A.5.6 are applications of Lemmas A.5.3 and A.5.4 respectively. To establish (A4), observe that
\[
(\hat{\Theta}^{(j)}_V \nabla^2 \ell_{nk}^{(j)} (\hat{\beta}^\lambda(D_j)) - e_v) = \Delta_1 + \Delta_2 + \Delta_3,
\]
where $\Delta_1 = (\hat{\Theta}^{(j)}_V - \Theta^*_V) \nabla^2 \ell_{nk}^{(j)} (\hat{\beta}^\lambda(D_j))$, $\Delta_2 = \Theta^*_V \nabla^2 \ell_{nk}^{(j)} (\hat{\beta}^\lambda(D_j)) - \nabla^2 \ell_{nk}^{(j)} (\Theta^*_V)$ and $\Delta_3 = \Theta^*_V \nabla^2 \ell_{nk}^{(j)} (\Theta^*_V) - e_v$. We thus consider $|\Delta_{\ell}(\hat{\beta}^\lambda(D_j) - \beta^*)|$ for $\ell = 1, 2, 3$.
\[
|\Delta_2(\hat{\beta}^\lambda(D_j) - \beta^*)| = \left| \frac{1}{n_k} \sum_{i \in I_j} \Theta^*_V X_i \nabla^2 \ell_{nk}^{(j)} (\hat{\beta}^\lambda(D_j) - \beta^*) \left[ b''(X_i^\top \hat{\beta}^\lambda(D_j)) - b''(X_i^\top \beta^*) \right] \right| \\
\leq U_3 \max_{1 \leq i \leq n} \left| \Theta^*_V X_i \right| \left| \frac{1}{n_k} \left\| X(\beta^\lambda(D_j) - \beta^*) \right\|_2^2 \right.
\]
\[
\mathbb{P} \left( \left\| X(\hat{\beta}^\lambda(D_j) - \beta^*) \right\|_2^2 \geq n^{-1} sk \log(d/\delta) \right) < \delta \text{ by Lemma A.5.4, thus } \mathbb{P} \left( |\Delta_2 \cdot (\hat{\beta}^\lambda(D_j) - \beta^*)| > t \right) < \delta \text{ for } t = MU_3 n^{-1} sk \log(d/\delta). \text{ Invoking Hölder’s inequality,}
\]
Hoeffding’s inequality and Condition 2.1.1, we also obtain, for $t \approx n^{-1} sk \log(d/\delta)$,

$$
\mathbb{P} \left( |\Delta_3(\hat{\beta}^\lambda(D_j) - \beta^*)| > t \right) \\
\leq \mathbb{P} \left( \left\| \Theta_v^T \left( \frac{1}{nk} \sum_{i \in \mathcal{I}_j} b''(X_i^T \beta^*) X_i X_i^T \right) - e_v \right\|_2 \left\| \hat{\beta}^\lambda(D_j) - \beta^* \right\|_1 > t \right).
$$

Therefore $\mathbb{P} \left( |\Delta_2(\hat{\beta}^\lambda(D_j) - \beta^*)| > t \right) < 2\delta$. Finally, with $t \approx n^{-1}(s \lor s_1)k \log(d/\delta)$,

$$
\mathbb{P} \left( |\Delta_1(\hat{\beta}^\lambda(D_j) - \beta^*)| > t \right) \\
\leq \mathbb{P} \left( \frac{1}{nk} \left\| \sum_{i \in \mathcal{I}_j} X_i^T (\hat{\Theta}_v - \Theta_v) b''(X_i^T \hat{\beta}^\lambda(D_j)) \right\|_2 \left\| X^{(j)}(\hat{\beta}^\lambda(D_j) - \beta^*) \right\|_2 > t \right),
$$

hence $\mathbb{P} \left( |\Delta_1(\hat{\beta}^\lambda(D_j) - \beta^*)| > t \right) < 2\delta$. This follows because by Lemma A.5.4

$$
\mathbb{P} \left( \left\| \frac{1}{nk} X^{(j)}(\hat{\beta}^\lambda(D_j) - \beta^*) \right\|_2 \geq n^{-1/2} \sqrt{sk \log(d/\delta)} < \delta
$$

and by Lemma C.4 of [95],

$$
\mathbb{P} \left( \left\| \frac{1}{nk} \sum_{i \in \mathcal{I}_j} X_i^T (\hat{\Theta}_v - \Theta_v) b''(X_i^T \hat{\beta}^\lambda(D_j)) \right\|_2 \geq n^{-1/2} \sqrt{s_1 k \log(d/\delta)} < \delta.
$$

2.5.2 Proofs for Theorems in Section 2.2.2

The proof of Theorem 2.2.3 relies on several preliminary lemmas, collected in Appendix A.5 in the Supplementary Material. Without loss of generality we set $H_0 : \beta^*_v = 0$ to ease notation.

Proof of Theorem 2.2.3 Since $\bar{S}(0) = k^{-1} \sum_{j=1}^k \hat{S}^{(j)}(0, \hat{\beta}^\lambda_v(D_j))$, and (B1)-(B4) of Condition A.5.1 in the Supplementary Material are fulfilled under Conditions 2.2.3 and 2.1.1 by Lemma A.5.8 (see Appendix A.5 in the Supplementary Material). The
proof is now simply an application of Lemma A.5.11 in the Supplementary Material with $\beta^* = 0$ under the restriction of the null hypothesis.

Proof of Lemma 2.2.2. The proof is an application of Lemma A.5.14 in the Supplementary Material, noting that (B1)-(B5) of Condition A.5.1 in the Supplementary Material are fulfilled under Conditions 2.2.3 and 2.1.1 by Lemmas A.5.8 and A.5.9 in the Supplementary Material.

2.5.3 Proofs for Theorems in Section 2.3

Recall from Section 2.1 that for an arbitrary matrix $M$, $M_\ell$ denotes the transposed $\ell$th row of $M$ and $[M]_\ell$ denotes the $\ell$th column of $M$.

Proof of Theorem 2.3.1. By Lemma 2.3.1 and $k = O(\sqrt{n/(s^2 \log d)})$, there exists a sufficiently large $C_0$ such that for the event $\mathcal{E} := \{\|\beta^d - \beta^*\|_{\infty} \leq C_0 \sqrt{\log d/n}\}$, we have $\mathbb{P}(\mathcal{E}) \geq 1 - c/d$. We choose $\nu = C_0 \sqrt{\log d/n}$, which implies that, under $\mathcal{E}$, we have $\nu \geq \|\beta^d - \beta^*\|_{\infty}$.

Let $S$ be the support of $\beta^*$. The derivations in the remainder of the proof hold on the event $\mathcal{E}$. Observe $T_\nu(\beta^d_{Sc}) = 0$ as $\|\beta^d_{Sc}\|_{\infty} \leq \nu$. For $j \in S$, if $|\beta^*_j| \geq 2\nu$, we have $|\beta^d_j| \geq |\beta^*_j| - \nu \geq \nu$ and thus $|T_\nu(\beta^d_j) - \beta^*_j| = |\beta^d_j - \beta^*_j| \leq \nu$. While if $|\beta^*_j| < 2\nu$, $|T_\nu(\beta^d_j) - \beta^*_j| \leq |\beta^*_j| \vee |\beta^d_j - \beta^*_j| \leq 2\nu$. Therefore, on the event $\mathcal{E}$,

$$\|T_\nu(\beta^d) - \beta^*\|_2 = \|T_\nu(\beta^d_S) - \beta^*_S\|_2 \leq 2\sqrt{5}\nu$$

and $\|T_\nu(\beta^d) - \beta^*\|_{\infty} = \|T_\nu(\beta^d_S) - \beta^*_S\|_{\infty} \leq 2\nu$.

The statement of the theorem follows because $\nu = C_0 \sqrt{\log d/n}$ and $\mathbb{P}(\mathcal{E}) \geq 1 - c/d$. Following the same reasoning, on the event $\mathcal{E}' := \mathcal{E} \cup \{\|\beta^d - \beta^*\|_{\infty} \leq C_0 \sqrt{\log d/n}\}$, the support of $\beta^*$
\{\|\tilde{\beta}^d - \beta^d\|_\infty \leq C_0sk\log d/n\}, \text{ we have}

\[\|\mathcal{C}v(\beta^d) - \mathcal{C}v(\tilde{\beta}^d)\|_2 = \|\mathcal{C}v(\beta^d_S) - \mathcal{C}v(\tilde{\beta}^d_S)\|_2\]
\[\leq \|\beta^d_S - \tilde{\beta}^d_S\|_2 \leq \sqrt{s}\|\beta^d_S - \tilde{\beta}^d_S\|_\infty \leq Cs^{3/2}k\log d/n.\]

As Lemma 2.3.1 also gives \(P(\mathcal{E}') \geq 1 - c/d\), the proof is complete.

Proof of Corollary 2.3.1 By an analogous proof strategy to that of Theorem 2.3.2
\[\|[\mathcal{C}(\Theta)]_{ov} - \Theta^*_v\| = O_p(\sqrt{n^{-1}\log d}) = o_p(1) \text{ under the conditions of the Corollary provided } k = o((s \vee s_1)\log d)^{-1}\sqrt{n}.\]
Chapter 3

Distributed estimation of principal eigenspaces

3.1 Problem setup

Suppose we have $N$ i.i.d random samples $\{X_i\}_{i=1}^{N} \subseteq \mathbb{R}^d$ with $E X_1 = 0$ and covariance matrix $E(X_1 X_1^\top) = \Sigma$. By spectral decomposition, $\Sigma = V \Lambda V^\top$, where $\Lambda = \text{diag}(\lambda_1, \lambda_2, \cdots, \lambda_d)$ with $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$ and $V = (v_1, \cdots, v_d) \in \mathcal{O}_{d \times d}$. For a given $K \in [d]$, let $V_K = (v_1, \cdots, v_K)$. Our goal is to estimate $\text{Col}(V_K)$, i.e., the linear space spanned by the top $K$ eigenvectors of $\Sigma$. To ensure the identifiability of $\text{Col}(V_K)$, we assume $\Delta := \lambda_K - \lambda_{K+1} > 0$ and define $\kappa := \lambda_1 / \Delta$ to be the condition number. Let $r = r(\Sigma) := \text{Tr}(\Sigma) / \lambda_1$ be the effected rank of $\Sigma$.

The standard way of estimating $\text{Col}(V_K)$ is to use the top $K$ eigenspace of the sample covariance $\hat{\Sigma} = \frac{1}{N} \sum_{i=1}^{N} X_i X_i^\top$. Let $\hat{\Sigma} = \hat{V} \hat{\Lambda} \hat{V}^\top$ be spectral decomposition of $\hat{\Sigma}$, where $\hat{\Lambda} = \text{diag}(\hat{\lambda}_1, \cdots, \hat{\lambda}_d)$ with $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_d$ and $\hat{V} = (\hat{v}_1, \cdots, \hat{v}_d)$. We use the empirical top $K$ eigenspace $\text{Col}(\hat{V}_K)$, where $\hat{V}_K = (\hat{v}_1, \cdots, \hat{v}_K)$, to estimate the eigenspace $\text{Col}(V_K)$. To measure the statistical error, we adopt $\rho(\hat{V}_K, V_K) := \|\hat{V}_K \hat{V}_K^\top - V_K V_K^\top\|_F$, which is the Frobenius norm of the difference.
between projection matrices of two spaces and is a well-defined distance between linear subspaces. In fact, $\rho(\mathbf{V}_K, \hat{\mathbf{V}}_K)$ is equivalent to the so-called sin $\Theta$ distance. Denote the singular values of $\hat{\mathbf{V}}_K^\top \mathbf{V}_K$ by $\{\sigma_i\}_{i=1}^K$ in descending order. Recall that $\Theta(\hat{\mathbf{V}}_K, \mathbf{V}_K) = \text{diag}(\theta_1, \cdots, \theta_K)$, the principal angles between Col($\mathbf{V}_K$) and Col($\hat{\mathbf{V}}_K$), are defined as $\text{diag}(\cos^{-1} \sigma_1, \cdots, \cos^{-1} \sigma_K)$. Then we define $\sin \Theta(\hat{\mathbf{V}}_K, \mathbf{V}_K)$ to be $\text{diag}(\sin \theta_1, \cdots, \sin \theta_K)$. Note that

$$\rho^2(\mathbf{V}_K, \hat{\mathbf{V}}_K) = \|\mathbf{V}_K \mathbf{V}_K^\top\|_F^2 + \|\hat{\mathbf{V}}_K \hat{\mathbf{V}}_K^\top\|_F^2 - 2 \text{Tr}(\mathbf{V}_K \mathbf{V}_K^\top \hat{\mathbf{V}}_K \hat{\mathbf{V}}_K^\top) = 2K - 2\|\hat{\mathbf{V}}_K \mathbf{V}_K\|_F^2,$$

$$= 2 \sum_{i=1}^K (1 - \sigma_i^2) = 2 \sum_{i=1}^K \sin^2 \theta_i = 2\|\Theta(\hat{\mathbf{V}}_K, \mathbf{V}_K)\|_F^2. \quad (3.1.1)$$

Therefore, $\rho(\mathbf{V}_K, \hat{\mathbf{V}}_K)$ and $\|\sin \Theta(\mathbf{V}_K, \hat{\mathbf{V}}_K)\|_F$ are equivalent.

Now consider the estimation of top $K$ eigenspace under the distributed data setting, where our $N = m \cdot n$ samples are scattered across $m$ machines with each machine storing $n$ samples\footnote{Note that here for simplicity we assume the subsample sizes are homogeneous. We can easily extend our analysis to the case of heterogeneous sub-sample sizes with similar theoretical results.}. Application of standard PCA here requires data or covariance aggregation, thus leads to huge communication cost for high-dimensional big data. In addition, for the areas such as genetic, biomedical studies and customer services, it is hard to communicate raw data because of privacy and ownership concerns. To address these problems, we need to avoid naive data aggregation and design a communication-efficient and privacy-preserving distributed algorithm for PCA. In addition, this new algorithm should be statistically accurate in the sense that it enjoys the same statistical error rate as the full sample PCA.

Throughout the paper, we assume that all the random samples $\{\mathbf{X}_i\}_{i=1}^N$ are i.i.d sub-Gaussian. We adopt the definition of sub-Gaussian random vectors in \[70\] and \[102\] as specified below, where $M$ is assumed to be a constant. It is not hard to show that the following definition is equivalent to the definition $\|((\Sigma^{1/2})^\dagger \mathbf{X})\|_{\psi_2} \leq M$ used in \[123\], \[127\], and many other authors.
Definition 3.1.1. We say the random vector $X \in \mathbb{R}^d$ is sub-Gaussian if there exists $M > 0$ such that $\|u^\top X\|_{\psi_2} \leq M\sqrt{\mathbb{E}(u^\top X)^2}$, $\forall u \in \mathbb{R}^d$.

We emphasize here that the global i.i.d assumption on $\{X_i\}_{i=1}^{N}$ can be further relaxed. In fact, our statistical analysis only requires the following three conditions: (i) within each server $\ell$, data are i.i.d.; (ii) across different servers, data are independent; (iii) the covariance matrices of the data in each server $\{\Sigma^{(\ell)}\}_{\ell=1}^{m}$ share similar top $K$ eigenspaces. We will further study this heterogeneous regime in Section 5. To avoid future confusion, unless specified, we always assume i.i.d. data across servers.

3.2 Methodology

We now introduce our distributed PCA algorithm. For $\ell \in [m]$, let $\{X^{(\ell)}_i\}_{i=1}^{n}$ denote the samples stored on the $\ell$-th machine. We specify the distributed in Algorithm 1.

**Algorithm 1** Distributed PCA

1. On each server, compute locally the $K$ leading eigenvectors $\hat{V}^{(\ell)}_K = (\hat{v}_1^{(\ell)}, \cdots, \hat{v}_K^{(\ell)}) \in \mathbb{R}^{d \times k}$ of the sample covariance matrix $\hat{\Sigma}^{(\ell)} = (1/n) \sum_{i=1}^{n} X^{(\ell)}_i X^{(\ell)}_i^\top$. Send $\hat{V}^{(\ell)}_K$ to the central processor.

2. On the central processor, compute $\tilde{\Sigma} = (1/m) \sum_{\ell=1}^{m} \hat{V}^{(\ell)}_K \hat{V}^{(\ell)}_K^\top$, and its $K$ leading eigenvectors $\{\tilde{v}_j\}_{j=1}^{K}$. Output: $\tilde{V}_K = (\tilde{v}_1, \cdots, \tilde{v}_K) \in \mathbb{R}^{d \times K}$.

In other words, each server first calculates the top $K$ eigenvectors of the local sample covariance matrix, and then transmits these eigenvectors $\{\hat{V}^{(\ell)}_K\}_{\ell=1}^{m}$ to a central server, where the estimators get aggregated. This procedure has similar spirit as distributed estimation based on one-shot averaging in [38], [3], [75], among others. To see this, we recall the SDP formulation of the eigenvalue problem. Let $\hat{V}_K = (\hat{v}_1, \cdots, \hat{v}_K)$ contain the $K$ leading eigenvectors of $\hat{\Sigma} = \frac{1}{m} \sum_{\ell=1}^{m} \hat{\Sigma}^{(\ell)}$. Lemma ?? in
Section 8.2.2 asserts that \( \hat{P}_K = \hat{V}_K \hat{V}_K^\top \) solves the SDP:

\[
\min_{P \in S_{d \times d}} -\text{Tr}(P^\top \hat{\Sigma}) \quad \text{s.t.} \quad \text{Tr}(P) \leq K, \|P\|_{op} \leq 1, P \succeq 0.
\]

(3.2.1)

Here \( S_{d \times d} \) refers to the set of \( d \times d \) symmetric matrices. In the traditional setting, we have access to all the data, and \( \hat{P}_K \) is a natural estimator for \( V_K V_K^\top \). In the distributed setting, each machine can only access \( \hat{\Sigma}^{(\ell)} \). Consequently, it solves a local version of (3.2.1):

\[
\min_{P \in S_{d \times d}} -\text{Tr}(P^\top \hat{\Sigma}^{(\ell)}) \quad \text{s.t.} \quad \text{Tr}(P) \leq K, \|P\|_{op} \leq 1, P \succeq 0.
\]

(3.2.2)

The optimal solution is \( \hat{P}_K^{(\ell)} = \hat{V}_K^{(\ell)} \hat{V}_K^{(\ell)\top} \). Since the loss function in (3.2.1) is the average of local loss functions in (3.2.2), we can intuitively average the optimal solutions \( \hat{P}_K^{(\ell)} \) to approximate \( \hat{P}_K \). However, the average \( \frac{1}{m} \sum_{\ell=1}^m \hat{P}_K^{(\ell)} \) may no longer be a rank-\( K \) projection matrix. Hence a rounding step is needed, extracting the leading eigenvectors of that average to get a projection matrix.

Here is another way of understanding the aggregation procedure. Given a collection of estimators \( \{\hat{V}_K^{(\ell)}\}_{\ell=1}^m \subseteq O_{d \times K} \) and the loss \( \rho(\cdot, \cdot) \), we want to find the center \( U \in O_{d \times K} \) that minimizes the sum of squared losses \( \sum_{\ell=1}^m \rho^2(U, \hat{V}_K^{(\ell)}) \). Lemma ?? indicates that \( U = \tilde{V}_K \) is an optimal solution. Therefore, our distributed PCA estimator \( \tilde{V}_K \) is a generalized “center” of individual estimators.

It is worth noting that in this algorithm, we do not really need to compute \( \{\hat{\Sigma}^{(\ell)}\}_{\ell=1}^m \) and \( \hat{\Sigma} \). \( \{\hat{V}_K^{(\ell)}\}_{\ell=1}^m \) and \( \tilde{V}_K \) can be derived from top-\( K \) SVD of data matrices. This is far more expeditious than the entire SVD and highly scalable, by using, for example, the power method [50]. As regard to the estimation of the top eigenvalues of \( \Sigma \), we can send the aggregated eigenvectors \( \{\tilde{v}_j\}_{j=1}^K \) back to the \( m \) servers, where
each one computes \( \{ \lambda_j^{(t)} \}_{j=1}^K = \{ \tilde{v}_j^{\top} \Sigma_j^{(t)} \tilde{v}_j \}_{j=1}^K \). Then the central server collect all the eigenvalues and deliver the average eigenvalues \( \{ \tilde{\lambda}_j \}_{j=1}^K = \left\{ \frac{1}{m} \sum_{t=1}^m \lambda_j^{(t)} \right\}_{j=1}^K \) as the estimators of all eigenvalues.

As we can see, the communication cost of the proposed distributed PCA algorithm is of order \( O(mKd) \). In contrast, to share all the data or entire covariance, the communication cost will be of order \( O(md \min(n, d)) \). Since in most cases \( K = o(\min(n, d)) \), our distributed PCA requires much less communication cost than naive data aggregation.

### 3.3 Statistical error analysis

Algorithm 1 delivers \( \tilde{V}_K \) to estimate the top \( K \) eigenspace of \( \Sigma \). In this section we analyze the statistical error of \( \tilde{V}_K \), i.e., \( \rho(\tilde{V}_K, V_K) \). The main message is that \( \tilde{V}_K \) enjoys the same statistical error rate as the full sample counterpart \( \hat{V}_K \) as long as the subsample size \( n \) is sufficiently large.

We first conduct a bias and variance decomposition of \( \rho(\tilde{V}_K, V_K) \), which serves as the key step in establishing our theoretical results. Recall that \( \Sigma = (1/m) \sum_{\ell=1}^m \hat{V}_K^{(\ell)} \hat{V}_K^{(\ell)\top} \) and \( \tilde{V}_K \) consists of the top \( K \) eigenvectors of \( \tilde{\Sigma} \). Define \( \Sigma^* := \mathbb{E}(\hat{V}_K^{(\ell)} \hat{V}_K^{(\ell)\top}) \) and denote its top \( K \) eigenvectors by \( V_K^* = (v_1^*, \cdots, v_K^*) \in \mathbb{R}^{d \times K} \).

When the number of machines goes to infinity, \( \Sigma \) converges to \( \Sigma^* \), and naturally we expect \( \text{Col}(\tilde{V}_K) \) to converge to \( \text{Col}(V_K^*) \) as well. This line of thinking inspires us to decompose the statistical error \( \rho(\tilde{V}_K, V_K) \) into the following bias and sample variance terms:

\[
\rho(\tilde{V}_K, V_K) \leq \rho(\tilde{V}_K, V_K^*) + \rho(V_K^*, V_K). \tag{3.3.1}
\]

The first term is stochastic and the second term is deterministic. Here we elucidate on why we call \( \rho(\tilde{V}_K, V_K^*) \) the sample variance term and \( \rho(V_K^*, V_K) \) the bias term respectively.
1. Sample variance term $\rho(\widetilde{V}_K, V^*_K)$:

By Davis-Kahan’s Theorem (Theorem 2 in [131]) and (5.1.1), we have

$$
\rho(\widetilde{V}_K, V^*_K) \lesssim \frac{\|\tilde{\Sigma} - \Sigma^*\|_F}{\lambda_K(\Sigma^*) - \lambda_{K+1}(\Sigma^*)}.
$$

(3.3.2)

As we can see, $\rho(\widetilde{V}_K, V^*_K)$ depends on how the average $\tilde{\Sigma} = \frac{1}{m} \sum_{\ell=1}^m \hat{V}_K^{(\ell)} \hat{V}_K^{(\ell)\top}$ concentrates to its mean $\Sigma^*$. This explains why we call $\rho(\widetilde{V}_K, V^*_K)$ the sample variance term. We will show in the sequel that for sub-Gaussian random samples, $\{\|\hat{V}_K^{(\ell)} \hat{V}_K^{(\ell)\top} - \Sigma^*\|_F\}_{\ell=1}^m$ and $\|\tilde{\Sigma} - \Sigma^*\|_F$ are sub-exponential random variables and under appropriate regularity assumptions,

$$
\left\| \|\tilde{\Sigma} - \Sigma^*\|_F \right\|_{\psi_1} \lesssim \frac{1}{\sqrt{m}} \left\| \|\hat{V}_K^{(1)} \hat{V}_K^{(1)\top} - \Sigma^*\|_F \right\|_{\psi_1}.
$$

(3.3.3)

If we regard $\psi_1$-norm as a proxy for standard deviation, this result is a counterpart to the formula for the standard deviation of the sample mean under the context of matrix concentration. By (3.3.3), the average of projection matrices $\tilde{\Sigma}$ enjoys a similar square-root convergence, so does $\rho(\widetilde{V}_K, V^*_K)$.

2. Bias term $\rho(V^*_K, V_K)$:

The error $\rho(V^*_K, V_K)$ is deterministic and independent of how many machines we have, and is therefore called the bias term. We will show this bias term is exactly zero when the random sample has a symmetric innovation (to be defined later). In general, we will show that the bias term is negligible in comparison with the sample variance term when the number of nodes $m$ is not unreasonably large.

In the following subsections, we will analyze the sample variance term and bias term respectively and then combine these results to obtain the convergence rate for $\rho(\widetilde{V}_K, V_K)$. 

56
3.3.1 Analysis of the sample variance term

To analyze $\rho(\tilde{V}_K, V^*_K)$, as shown by \((3.3.2)\), we need to derive the order of the numerator $\|\tilde{\Sigma} - \Sigma^*\|_F$ and denominator $\lambda_K(\Sigma^*) - \lambda_{K+1}(\Sigma^*)$. We first focus on the matrix concentration term $\|\tilde{\Sigma} - \Sigma^*\|_F = \left\| \frac{1}{m} \sum_{\ell=1}^{m} (\hat{V}^{(\ell)}_K \hat{V}^{(\ell)\top}_K - \Sigma^*) \right\|_F$. Note that $\tilde{\Sigma} - \Sigma^*$ is an average of $m$ centered random matrices. To establish the correspondent concentration inequality, we first investigate each individual term in the average, i.e., $\hat{V}^{(\ell)}_K \hat{V}^{(\ell)\top}_K - \Sigma^*$ for $\ell \in [m]$. In the following lemma, we show that when random samples are sub-Gaussian, $\|\hat{V}^{(\ell)}_K \hat{V}^{(\ell)\top}_K - \Sigma^*\|_F$ is sub-exponential and we can give an explicit upper bound of its $\psi_1$-norm.

**Lemma 3.3.1.** Suppose that on the $\ell$-th server we have $n$ i.i.d. sub-Gaussian random samples $\{X_i\}_{i=1}^n$ in $\mathbb{R}^d$ with covariance matrix $\Sigma$. There exists a constant $C > 0$ such that when $n \geq r$, \[ \psi_1 \left\| \|\hat{V}^{(\ell)}_K \hat{V}^{(\ell)\top}_K - \Sigma^*\|_F \right\| \leq C \kappa \sqrt{\frac{K r}{n}}. \]

Note that here we use the Frobenius norm to measure the distance between two matrices. Therefore, it is equivalent to treat $\{\hat{V}^{(\ell)}_K \hat{V}^{(\ell)\top}_K \}_{\ell=1}^K$ and $\Sigma^*$ as $d^2$-dimensional vectors and apply the concentration inequality for random vectors to bound $\|\tilde{\Sigma} - \Sigma^*\|_F$. As we will demonstrate in the proof of Theorem 3.3.1, \[ \psi_1 \left\| \|\tilde{\Sigma} - \Sigma^*\|_F \right\| \lesssim \frac{1}{\sqrt{m}} \psi_1 \left\| \|\hat{V}^{(\ell)}_K \hat{V}^{(\ell)\top}_K - \Sigma^*\|_F \right\|.

With regard to $\lambda_K(\Sigma^*) - \lambda_{K+1}(\Sigma^*)$, when the individual node has enough samples, $\hat{V}^{(\ell)}_K$ and $V_K$ will be close to each other and so will $\Sigma^* = \mathbb{E}(\hat{V}^{(\ell)}_K \hat{V}^{(\ell)\top}_K)$ and $V_K V_K^\top$. Given $\lambda_K(V_K V_K^\top) = 1$ and $\lambda_{K+1}(V_K V_K^\top) = 0$, we accordingly expect $\lambda_K(\Sigma^*)$ and $\lambda_{K+1}(\Sigma^*)$ be separated by a positive constant as well.

All the arguments above lead to the following theorem on $\rho(\tilde{V}_K, V^*_K)$.

**Theorem 3.3.1.** Suppose $X_1, \cdots, X_N$ are i.i.d. sub-Gaussian random vectors in $\mathbb{R}^d$ with covariance matrix $\Sigma$ and they are scattered across $m$ machines. If $n \geq r$ and
\[ \| \Sigma^* - V_K V_K^T \|_{op} \leq 1/4, \]

then

\[ \left\| \rho(\tilde{V}_K, V_K^*) \right\|_{\psi_1} \leq C \kappa \sqrt{\frac{K r}{N}}, \]

where \( C \) is some universal constant.

### 3.3.2 Analysis of the bias term

In this section, we study the bias term \( \rho(V_K^*, V_K) \) in (3.3.1). We first focus on a special case where the bias term is exactly zero. For a random vector \( X \) with covariance \( \Sigma = V \Lambda V^T \), let \( Z = \Lambda^{-\frac{1}{2}} V^T X \). We say \( X \) has symmetric innovation if \( Z \overset{d}{=} (I_d - 2e_j e_j^T) Z, \quad \forall j \in [d] \). In other words, flipping the sign of one component of \( Z \) will not change the distribution of \( Z \). Note that if \( Z \) has density, this is equivalent to say that its density function has the form \( p(|z_1|, |z_2|, \cdots, |z_d|) \). All elliptical distributions centered at the origin belong to this family. In addition, if \( Z \) has symmetric and independent entries, \( X \) has also symmetric innovation. It turns out that when the random samples have symmetric innovation, \( \Sigma^* := \mathbb{E}(\hat{V}_K^{(l)} \hat{V}_K^{(l)\top}) \) and \( \Sigma \) share exactly the same set of eigenvectors. When we were finishing the paper, we noticed that [27] had independently established a similar result for the Gaussian case.

**Definition 3.3.1.** Let \( \mathcal{V} \) be a \( K \)-dimensional linear subspace of \( \mathbb{R}^d \). For a subspace estimator represented by \( \hat{V} \in \mathcal{O}_{d \times K} \), we say it is unbiased for \( \mathcal{V} \) if and only if the top \( K \) eigenspace of \( \mathbb{E}(\hat{V} \hat{V}^\top) \) is \( \mathcal{V} \).

If \( \hat{V}_K^{(l)} \) is unbiased for \( \text{Col}(V_K) \), then \( \rho(V_K^*, V_K) = 0 \) and we will only have the sample variance term in (3.3.1). In that case, aggregating \( \{\hat{V}_K^{(l)}\}_{l=1}^m \) reduces variance and yields a better estimator \( \tilde{V}_K \). Theorem 3.3.2 shows that this is the case so long as the distribution has symmetric innovation and the sample size is large enough.

**Theorem 3.3.2.** Suppose on the \( \ell \)-th server we have \( n \) i.i.d. random samples \( \{X_i\}_{i=1}^n \) with covariance \( \Sigma \). If \( \{X_i\}_{i=1}^n \) have symmetric innovation, then \( V^\top \Sigma^* V \) is diagonal,
i.e., $\Sigma^*$ and $\Sigma$ share the same set of eigenvectors. Furthermore, if $\|\Sigma^* - V_K V_K^T\|_{op} < 1/2$, then $\{\hat{V}_K^{(l)}\}_{l=1}^n$ are unbiased for $\text{Col}(V_K)$ and $\rho(V_K^*, V_K) = 0$.

It is worth pointing out that distributed PCA is closely related to aggregation of random sketches of a matrix [33, 117]. To approximate the subspace spanned by the $K$ leading left singular vectors of a large matrix $A \in \mathbb{R}^{d_1 \times d_2}$, we could construct a suitable random matrix $Y \in \mathbb{R}^{d_2 \times n}$ with $n \geq K$, and use the left singular subspace of $AY \in \mathbb{R}^{d_1 \times n}$ as an estimator. $AY$ is called a random sketch of $A$. It has been shown that to obtain reasonable statistical accuracy, $n$ can be much smaller than $\min(d_1, d_2)$ as long as $A$ is approximately low rank. Hence it is much cheaper to compute SVD on $AY$ than on $A$. When we want to aggregate a number of such subspace estimators, a smart choice of the random matrix ensemble for $Y$ is always preferable. It follows from Theorem 3.3.2 that if we let $Y$ have i.i.d. columns from a distribution with symmetric innovation (e.g., Gaussian distribution or independent entries), then the subspace estimators are unbiased, which facilitates aggregation.

Here we explain why we need the condition $\|\Sigma^* - V_K V_K^T\|_{op} < 1/2$ to achieve zero bias. First of all, the condition is similar to a bound on the “variance” of the random matrix $\hat{V}_K^{(l)}$ whose covariance $\Sigma^*$ is under investigation. As demonstrated above, with the symmetric innovation, $\Sigma^*$ has the same set of eigenvectors as $\Sigma$, but we still cannot guarantee that the top $K$ eigenvectors of $\Sigma^*$ match with those of $\Sigma$. For example, the $(K+1)$-th eigenvector of $\Sigma$ might be the $K$-th eigenvector of $\Sigma^*$. In order to ensure the top $K$ eigenspace of $\Sigma^*$ is exactly the same as that of $\Sigma$, we require $\hat{V}_K^{(l)}$ to not deviate too far from $V_K$ so that $\Sigma^*$ is close enough to $V_K V_K^T$. Both Theorems 3.3.1 and 3.3.2 require control of $\|\Sigma^* - V_K V_K^T\|_{op}$, which will be studied shortly.

For general distributions, the bias term is not necessarily zero. However, it turns out that when the subsample size is large enough, the bias term $\rho(V_K^*, V_K)$ is of high-order compared with the statistical error of $\hat{V}_K^{(l)}$ on the individual subsample. By the
decomposition \( (3.3.1) \) and Theorem \( 3.3.1 \) we can therefore expect the aggregated estimator \( \hat{V}_K \) to enjoy sharper statistical error rate than PCA on the individual subsample. In other words, the aggregation does improve the statistical efficiency. A similar phenomenon also appears in statistical error analysis of the average of the debiased Lasso estimators in [3] and [75]. Recall that in sparse linear regression, the Lasso estimator \( \hat{\beta} \) satisfies that 
\[
\| \hat{\beta} - \beta^* \|_2 = O_P(\sqrt{s \log d/n}),
\]
where \( \beta^* \) is the true regression vector, \( s \) is the number of nonzero coefficients of \( \beta^* \) and \( d \) is the dimension. The debiasing step reduces the bias of \( \hat{\beta} \) to the order \( O_P(s \log d/n) \), which is negligible when \( m \) is not too large, compared with the statistical error of \( \hat{\beta} \) and thus enables the average of the debiased Lasso estimators to enhance the statistical efficiency.

Below we present a second-order Davis-Kahan theorem that explicitly characterizes the first and second order errors on top \( K \) eigenspace due to matrix perturbation. This is a genuine generalization of the former high-order perturbation theories on a single eigenvector, e.g., Lemma 1 in [67] and Theorem 2 in [34]. By nature it is deterministic and serves as the core of the bias analysis in our work. Let us first introduce the notation.

Let \( A, \hat{A} \in \mathbb{R}^{d \times d} \) be symmetric matrices with eigenvalues \( \lambda_1 \geq \cdots \geq \lambda_d \), and \( \hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_d \), respectively. Let \( \{u_j\}_{j=1}^d, \{\hat{u}_j\}_{j=1}^d \) be two orthonormal bases of \( \mathbb{R}^d \) such that \( Au_j = \lambda_j u_j \) and \( \hat{A}\hat{u}_j = \hat{\lambda}_j \hat{u}_j \) for all \( j \in [d] \). Fix \( s \in \{0, 1, \cdots, d - K\} \) and assume that \( \Delta = \min\{\lambda_s - \lambda_{s+1}, \lambda_s + K - \lambda_{s+K+1}\} > 0 \), where \( \lambda_0 = +\infty \) and \( \lambda_{d+1} = -\infty \). Define \( U = (u_{s+1}, \cdots, u_{s+K}) \), \( \hat{U} = (\hat{u}_{s+1}, \cdots, \hat{u}_{s+K}) \) and \( H = \hat{U}^\top U \).

Let \( H = XDY^\top \) be its singular value decomposition and \( \hat{H} = X\hat{Y}^\top \). Define \( E = \hat{A} - A, S = \{s + 1, \cdots, s + K\}, G_j = \sum_{i \in S} (\lambda_i - \lambda_{s+j})^{-1} u_i u_i^\top \) for \( j \in [K] \), and

\[
f: \mathbb{R}^{d \times K} \to \mathbb{R}^{d \times K}, \quad (v_1, \cdots, v_K) \mapsto (-G_1 v_1, \cdots, -G_K v_K).
\]
Lemma 3.3.2. Assuming that \( \varepsilon = \sqrt{8K}\|\mathbf{E}\|_{op}/\Delta \leq 1/\sqrt{2} \), we have

\[
\min_{\mathbf{O} \in \mathcal{O}_{K \times K}} \|\hat{\mathbf{O}} - \mathbf{U}\|_F = \|\hat{\mathbf{U}}\mathbf{H} - \mathbf{U}\|_F \leq \|\hat{\mathbf{U}}\hat{\mathbf{U}}^\top - \mathbf{U}\mathbf{U}^\top\|_F \leq \varepsilon, \\
\|\hat{\mathbf{U}}\mathbf{H} - \mathbf{U} - f(\mathbf{E}\mathbf{U})\|_F \leq 5\varepsilon^2, \\
\|\hat{\mathbf{U}}\hat{\mathbf{U}}^\top - [\mathbf{U}\mathbf{U}^\top + f(\mathbf{E}\mathbf{U})\mathbf{U}^\top + \mathbf{U}(\mathbf{E}\mathbf{U})^\top]\|_F \leq 12\varepsilon^2. 
\] (3.3.4)

The first bound is just the standard Davis-Kahan inequality, see Theorem 2 in [131]. The next two investigate the difference between column space spanned by \( \hat{\mathbf{U}} \) and \( \mathbf{U} \). As in regular Taylor expansion, we decompose the difference into the leading linear terms and quadratic residuals with respect to the perturbation. Now we apply Lemma 3.3.2 to the context of principal eigenspace estimation. Let \( \mathbf{A} = \mathbf{\Sigma} \), \( \mathbf{E} = \hat{\mathbf{\Sigma}} - \mathbf{\Sigma} \) and \( \mathbf{S} = [\mathbf{K}] \). It thus follows that \( \mathbf{U} = \mathbf{V}_K \) and \( \hat{\mathbf{U}} = \hat{\mathbf{V}}_K \). Since \( \mathbf{EE} = 0 \), the expectation of the linear leading term \( \mathbb{E}(f(\mathbf{E}\mathbf{V}_K))\mathbf{V}_K^\top + \mathbf{V}_K f(\mathbf{E}\mathbf{V}_K)^\top) \) is then zero. Based on this fact, from the last inequality in (3.3.4) we can derive that the bias term \( \rho(\mathbf{V}_K^*, \mathbf{V}_K) \) will be a high-order term compared with the linear leading term. More rigorously, by Davis-Kahan’s Theorem, we can control the bias as follows:

\[
\rho(\mathbf{V}_K^*, \mathbf{V}_K) \leq 2\sqrt{2} \cdot \|\mathbf{\Sigma}^* - \mathbf{V}_K\mathbf{V}_K^\top\|_F = 2\sqrt{2} \cdot \|\mathbb{E}[\hat{\mathbf{V}}_K\hat{\mathbf{V}}_K^\top - \mathbf{V}_K\mathbf{V}_K^\top]\|_F. 
\] (3.3.5)

By using the fact that \( \mathbb{E}\mathbf{E} = 0 \) and \( f \) is linear, we have

\[
\rho(\mathbf{V}_K^*, \mathbf{V}_K) \leq 2\sqrt{2} \cdot \|\mathbb{E}[\hat{\mathbf{V}}_K\hat{\mathbf{V}}_K^\top - (\mathbf{V}_K\mathbf{V}_K^\top + f(\mathbf{E}\mathbf{V}_K)\mathbf{V}_K^\top + \mathbf{V}_K f(\mathbf{E}\mathbf{V}_K)^\top)]\|_F
\]

By Jensen’s inequality, the above is further bounded by

\[
2\sqrt{2} \cdot \mathbb{E}\|\hat{\mathbf{V}}_K\hat{\mathbf{V}}_K^\top - (\mathbf{V}_K\mathbf{V}_K^\top + f(\mathbf{E}\mathbf{V}_K)\mathbf{V}_K^\top + \mathbf{V}_K f(\mathbf{E}\mathbf{V}_K)^\top)\|_F
\]
Theorem 3.3.3. There are constants $C_1$ and $C_2$ such that when $n \geq r$,

$$\rho(V_K^*, V_K) \leq C_1 \| \Sigma^* - V_K V_K^T \|_F \leq C_2 \kappa^2 \sqrt{Kr/n}. $$

As a by-product, we get $\| \Sigma^* - V_K V_K^T \|_{op} \lesssim \kappa^2 \sqrt{Kr/n}$. Hence when $n \geq C \kappa^2 \sqrt{Kr}$ for some large enough $C$, the assumptions in Theorems 3.3.1 and 3.3.2 on $\| \Sigma^* - V_K V_K^T \|_{op}$ are guaranteed to hold.

### 3.3.3 Properties of distributed PCA

We now combine the results we obtained in the previous two subsections to derive the statistical error rate of $\tilde{V}_K$. We first present a theorem under the setting of global i.i.d. data and discuss its optimality.

**Theorem 3.3.4.** Suppose we have $N$ i.i.d. sub-Gaussian random samples with covariance $\Sigma$. They are scattered across $m$ servers, each of which stores $n$ samples. There exist constants $C, C_1, C_2, C_3$ and $C_4$ such that the followings hold when $n \geq C \kappa^2 \sqrt{Kr}$.

1. Symmetric innovation:

   $$\left\| \rho(\tilde{V}_K, V_K) \right\|_{\psi_1} \leq C_1 \kappa \sqrt{Kr/N}. \quad (3.3.6)$$

2. General distribution:

   $$\left\| \rho(\tilde{V}_K, V_K) \right\|_{\psi_1} \leq C_1 \kappa \sqrt{Kr/N} + C_2 \kappa^2 \sqrt{Kr/n}. \quad (3.3.7)$$

Furthermore, if we further assume $m \leq C_3 \bar{n}/(\kappa^2 r)$,

$$\left\| \rho(\tilde{V}_K, V_K) \right\|_{\psi_1} \leq C_4 \kappa \sqrt{Kr/N}. \quad (3.3.8)$$
As we can see, with appropriate scaling conditions on \( n, m \) and \( d \), \( \tilde{V}_K \) can achieve the statistical error rate of order \( \kappa \sqrt{Kr/N} \). The result is applicable to the whole sample or traditional PCA, in which \( m = 1 \). Hence the distributed PCA and the traditional PCA share the same error bound as long as the technical conditions are satisfied.

In the second part of Theorem 3.3.4, the purpose of setting restrictions on \( n \) and \( m \) is to ensure that the distributed PCA algorithm delivers the same statistical rate as the centralized PCA which uses all the data. In the boundary case where \( n \asymp \kappa^2 \sqrt{K} r \), the bias of the local empirical eigenspace is of constant order. Since our aggregation cannot kill bias, there is no hope to achieve the centralized rate unless the number of machines is of constant order so that the centralized PCA has constant error too. Besides, our result says that when \( n \) is large, we can tolerate more data splits (larger \( m \)) for achieving the centralized statistical rate.

We now illustrate our result through a simple spiked covariance model introduced by [61]. Assume that \( \Lambda = \text{diag}(\lambda, 1, \ldots, 1) \), where \( \lambda > 1 \), and we are interested in the first eigenvector of \( \Sigma \). Note that \( K = 1, r = \text{Tr}(\Sigma)/\|\Sigma\|_{op} = (\lambda + d - 1)/\lambda \asymp d/\lambda \) when \( \lambda = O(d) \), and \( \kappa = \lambda/(\lambda - 1) \asymp 1 \). It is easy to see from (3.3.6) or (3.3.8) that

\[
\left\| \rho(\tilde{V}_1, V_1) \right\|_{\psi_1} \lesssim \kappa \sqrt{\frac{r}{N}} \lesssim \sqrt{\frac{d}{N\lambda}}.
\]

Without loss of generality, we could always assume that the direction of \( \tilde{V}_1 \) is chosen such that \( \tilde{V}_1^T V_1 \geq 0 \), i.e. \( \tilde{V}_1 \) is aligned with \( V_1 \). Note that

\[
\rho^2(\tilde{V}_1, V_1) = \|\tilde{V}_1 \tilde{V}_1^T - V_1 V_1^T\|_F^2 = 2(1 - \tilde{V}_1^T V_1)(1 + \tilde{V}_1^T V_1) \\
\geq 2(1 - \tilde{V}_1^T V_1) = \|\tilde{V}_1 - V_1\|_2^2.
\]

Hence

\[
\mathbb{E}\|\tilde{V}_1 - V_1\|_2^2 \lesssim \left\| \rho(\tilde{V}_1, V_1) \right\|_{\psi_1}^2 \lesssim \frac{d}{N\lambda}. \tag{3.3.9}
\]
We now compare this rate with the previous results under the spiked model. In [98], the authors derived the $\ell_2$ risk of the empirical eigenvectors when random samples are Gaussian. It is not hard to derive from Theorem 3.3.1 therein that given $N$ i.i.d $d$-dimensional Gaussian samples, when $N, d$ and $\lambda$ go to infinity,

$$\mathbb{E}\|\hat{V}_1 - V_1\|_2^2 \approx \frac{d}{N\lambda},$$

where $\hat{V}_1$ is the empirical leading eigenvector with $\hat{V}_1^\top V_1 \geq 0$. We see from (3.3.9) that the aggregated estimator $\tilde{V}_1$ performs as well as the full sample estimator $\hat{V}_1$ in terms of the mean squared error. See [127] for generalization of the results for spiked covariance.

In addition, our result is consistent with the minimax lower bound developed in [15]. For $\lambda > 0$ and fixed $c \geq 1$, define

$$\Theta = \{ \Sigma \text{ is symmetric and } \Sigma \succeq 0 : \lambda + 1 \leq \lambda_K \leq \lambda_1 \leq c\lambda + 1, \lambda_j = 1 \text{ for } K + 1 \leq j \leq d \}.$$ 

Assume that $K \leq d/2$ and $1 \lesssim d/\lambda \lesssim N$. Theorem 8 in [15] shows that under the Gaussian distribution with $\Sigma \in \Theta$, the minimax lower bound of $\mathbb{E}\rho^2(\tilde{V}, V_K)$ satisfies

$$\inf_{\tilde{V}} \sup_{\Sigma \in \Theta} \mathbb{E}\rho^2(\tilde{V}, V_K) \gtrsim \min \left\{ K, (d - K), \frac{K(\lambda + 1)(d - K)}{N\lambda^2} \right\} \gtrsim \frac{Kd}{N\lambda} \quad (3.3.10)$$

Based on $r = \text{Tr}(\Sigma)/\|\Sigma\|_{op} \leq (cK\lambda + d)/(c\lambda + 1) \lesssim Kd/\lambda$ and $\kappa \leq c \lesssim 1$, our (3.3.6) gives an upper bound

$$\mathbb{E}\rho^2(\tilde{V}_1, V_1) \lesssim \kappa^2 \frac{Kr}{n} \lesssim \frac{Kd}{N\lambda},$$

which matches the lower bound in (3.3.10).

Although the upper bound $\kappa\sqrt{Kr/N}$ established in Theorem 3.3.4 is optimal in the minimax sense as discussed above, the non-minimax risk of empirical eigenvectors can be improved when the condition number $\kappa$ is large. See [125], [69] and [102] for
sharper results. We use (3.3.6) as a benchmark rate for the centralized PCA only for the sake of simplicity.

Notice that in Theorem 3.3.4, the prerequisite for $\tilde{V}_K$ to enjoy the sharp statistical error rate is a lower bound on the subsample size $n$, i.e.,

$$n \gtrsim \kappa^2 \sqrt{Kr}.$$

(3.3.11)

As in the remarks after Lemma 3.3.2, this is the condition we used to ensure closeness between $\Sigma^*$ and $V_K V_K^T$. It is natural to ask whether this required sample complexity is sharp, or in other words, is it possible for $\tilde{V}_K$ to achieve the same statistical error rate with a smaller sample size on each machine? The answer is no. The following theorem presents a distribution family under which $\text{Col}(\tilde{V}_K)$ is even perpendicular to $\text{Col}(V_K)$ with high probability when $n$ is smaller than the threshold given in (3.3.11). This means that having a smaller sample size on each machine is too uninformative such that the aggregation step completely fails in improving estimation consistency.

**Theorem 3.3.5.** Consider a Bernoulli random variable $W$ with $P(W = 0) = P(W = 1) = 1/2$, a Rademacher random variable $P(Y = 1) = P(Y = -1) = 1/2$, and a random vector $Z \in \mathbb{R}^{d-1}$ that is uniformly distributed over the $(d-1)$-dimensional unit sphere. For $\lambda \geq 2$, we say a random vector $X \in \mathbb{R}^d$ follows the distribution $D(\lambda)$ if

$$X \overset{d}{=} \begin{pmatrix} 1_{\{W=0\}} \sqrt{2\lambda Y} \\ 1_{\{W=1\}} \sqrt{2(d-1)} Z \end{pmatrix}.$$ 

Now suppose we have $\{X_i\}_{i=1}^N$ as $N$ i.i.d. random samples of $X$. They are stored across $m$ servers, each of which has $n$ samples. When $32 \log d \leq n \leq (d-1)/(3\lambda)$, we have

$$P(\tilde{V}_1 \perp V_1) \geq \begin{cases} 1 - d^{-1}, & \text{if } m \leq d^3, \\ 1 - e^{-d/2}, & \text{if } m > d^3. \end{cases}$$
It is easy to verify that $D(\lambda)$ is symmetric, sub-Gaussian and satisfies $\mathbb{E}X = 0$ and $\mathbb{E}(XX^\top) = \text{diag}(\lambda, 1, \cdots, 1)$. Besides, $\kappa = \lambda/(\lambda - 1) \geq 1$ and $r = (\lambda + d - 1)/\lambda = d/\lambda + 1 - \lambda^{-1} \simeq d/\lambda$ when $2 \leq \lambda \lesssim d$. According to (3.3.11), we require $n \gtrsim d/\lambda$ to achieve the rate as demonstrated in (3.3.6). Theorem 3.3.5 shows that if we have fewer samples than this threshold, the aggregated estimator $\tilde{V}_1$ will be perpendicular to the true top eigenvector $V_1$ with high probability. Therefore, our lower bound for the subsample size $n$ is sharp.

3.4 Extension to heterogeneous samples

We now relax global i.i.d. assumptions in the previous section to the setting of heterogeneous covariance structures across servers. Suppose data on the server $\ell$ has covariance matrix $\Sigma^{(\ell)}$, whose top $K$ eigenvalues and eigenvectors are denoted by $\{\lambda_k^{(\ell)}\}_{k=1}^K$ and $V^{(\ell)}_K = (v_1^{(\ell)}, \cdots, v_K^{(\ell)})$ respectively. We will study two specific cases of heterogeneous covariances: one requires all covariances to share exactly the same principal eigenspaces, while the other considers the heterogeneous factor models with common factor eigen-structures.

3.4.1 Common principal eigenspaces

We assume that $\{\Sigma^{(\ell)}\}_{\ell=1}^m$ share the same top $K$ eigenspace, i.e. there exists some $V_K \in \mathcal{O}_{d \times K}$ such that $V^{(\ell)}_K V^{(\ell)}_K \top = V_K V_K \top$ for all $\ell \in [m]$. The following theorem can be viewed as a generalization of Theorem 3.3.4.

**Theorem 3.4.1.** Suppose we have in total $N$ sub-Gaussian samples scattered across $m$ servers, each of which stores $n$ i.i.d. samples with covariance $\Sigma^{(\ell)}$. Assume that $\{\Sigma^{(\ell)}\}_{\ell=1}^m$ share the same top $K$ eigenspace. For each $\ell \in [m]$, let $S_\ell = \kappa_\ell \sqrt{K r_\ell N}$ and $B_\ell = \kappa_\ell^2 \sqrt{K r_\ell n}$, where $r_\ell := \text{Tr}(\Sigma^{(\ell)})/\lambda_1^{(\ell)}$ and $\kappa_\ell := \lambda_1^{(\ell)}/(\lambda_K^{(\ell)} - \lambda_{K+1}^{(\ell)})$. 

66
1. Symmetric innovation: There exist some positive constants $C$ and $C_1$ such that

$$\|\rho(\tilde{V}_K, V_K)\|_{\psi_1} \leq C_1 \sqrt{\frac{1}{m} \sum_{\ell=1}^{m} S^2_{\ell}}$$

(3.4.1)

so long as $n \geq C \sqrt{K} \max_{\ell \in [m]} (\kappa^2 r_{\ell})$.

2. General distribution: There exist positive constants $C_2$ and $C_3$ such that when $n \geq \max_{\ell \in [m]} r_{\ell}$,

$$\|\rho(\tilde{V}_K, V_K)\|_{\psi_1} \leq C_2 \sqrt{\frac{1}{m} \sum_{\ell=1}^{m} S^2_{\ell}} + C_3 \sum_{\ell=1}^{m} B^2_{(\ell)}.$$  

(3.4.2)

### 3.4.2 Heterogeneous factor models

Suppose on the server $\ell$, the data conform to a factor model as below.

$$X_{(\ell)}^{(i)} = B_{(\ell)} f_{(\ell)}^{(i)} + u_{(\ell)}^{(i)}, \quad i \in [n],$$

where $B_{(\ell)} \in \mathbb{R}^{d \times K}$ is the loading matrix, $f_{(\ell)}^{(i)} \in \mathbb{R}^K$ is the factor that satisfies $\text{Cov}(f_{(\ell)}^{(i)}) = I$ and $u_{(\ell)}^{(i)} \in \mathbb{R}^d$ is the residual vector. It is not hard to see that $\Sigma_{(\ell)} = \text{Cov}(X_{(\ell)}^{(i)}) = B_{(\ell)} B_{(\ell)}^T + \Sigma_u^{(\ell)}$, where $\Sigma_u^{(\ell)}$ is the covariance matrix of $u_{(\ell)}^{(i)}$.

Let $B_{(\ell)} B_{(\ell)}^T = V_{K}^{(\ell)} \Lambda_{K}^{(\ell)} V_{K}^{(\ell)^T}$ be the spectral decomposition of $B_{(\ell)} B_{(\ell)}^T$. We assume that there exists a projection matrix $P_K = V_K V_K^T$, where $V_K \in \mathcal{O}_{d \times K}$, such that $V_{K}^{(\ell)} V_{K}^{(\ell)^T} = P_K$ for all $\ell \in [m]$. In other words, $\{B_{(\ell)} B_{(\ell)}^T\}_{\ell=1}^{m}$ share the same top $K$ eigenspace. Given the context of factor models, this implies that the factors have similar impact on the variation of the data across servers. Our goal now is to recover $\text{Col}(V_K)$ by the distributed PCA approach, namely Algorithm 1.

Recall that $\hat{\Sigma}_{(\ell)} = \frac{1}{n} \sum_{i=1}^{n} X_{(\ell)}^{(i)} X_{(\ell)}^{(i)^T}$ is the sample covariance matrix on the $\ell$-th machine, and $\tilde{V}_{K}^{(\ell)} = (\tilde{v}_{(\ell)}^{1}, \ldots, \tilde{v}_{(\ell)}^{K}) \in \mathcal{O}_{d \times K}$ stores $K$ leading eigenvectors of $\hat{\Sigma}_{(\ell)}$. 

67
Define $\tilde{\Sigma} = \frac{1}{m} \sum_{\ell=1}^{m} \tilde{V}_K^{(\ell)} \tilde{V}_K^{(\ell)\top}$, and let $\tilde{V}_K \in \mathcal{O}_{d \times K}$ be the top $K$ eigenvectors of $\tilde{\Sigma}$. Below we present a theorem that characterizes the statistical performance of the distributed PCA under the heterogeneous factor models.

**Theorem 3.4.2.** For each $\ell \in [m]$, let $S_\ell = \kappa_\ell \sqrt{\frac{K r_\ell}{N}}$ and $B_\ell = \frac{\kappa_\ell^2 \sqrt{K r_\ell}}{n}$. There exist some positive constants $C_1$, $C_2$ and $C_3$ such that when $n \geq \max_{\ell \in [m]} r_\ell$,

$$
\|\rho(\tilde{V}_K, V_K)\|_{\psi_1} \leq C_1 \sqrt{\frac{1}{m} \sum_{\ell=1}^{m} S_\ell^2} + C_2 \frac{1}{m} \sum_{\ell=1}^{m} B_\ell + C_3 \frac{\sqrt{K}}{m} \sum_{\ell=1}^{m} \frac{\|\Sigma_u^{(\ell)}\|_{\text{op}} \lambda_K(A_K^{(\ell)})}{\Lambda_K(A_K^{(\ell)})}.
$$

(3.4.3)

The first two terms in the RHS of (3.4.3) are similar to those in (3.4.2), while the third term characterizes the effect of heterogeneity in statistical efficiency of $\tilde{V}_K$. When $\|\Sigma_u^{(\ell)}\|_{\text{op}}$ is small compared with $\lambda_K(A_K^{(\ell)})$ as in spiky factor models, $\Sigma_u^{(\ell)}$ can hardly distort the eigenspace Col($V_K$) and thus has little influence on the final statistical error of $\tilde{V}_K$.

### 3.5 Simulation study

In this section, we conduct Monte Carlo simulations to validate the statistical error rate of $\tilde{V}_K$ that is established in the previous section. We also compare the statistical accuracy of $\tilde{V}_K$ and its full sample counterpart $\hat{V}_K$, that is, the empirical top $K$ eigenspace based on the full sample covariance. The main message is that our proposed distributed estimator performs equally well as the full sample estimator $\hat{V}_K$ when the subsample size $n$ is large enough.

#### 3.5.1 Verification of the statistical error rate

Consider $\{x_i\}_{i=1}^{N}$ i.i.d. following $N(0, \Sigma)$, where $\Sigma = \text{diag}(\lambda, \lambda/2, \lambda/4, 1, \cdots, 1)$. Here the number of spiky eigenvalues $K = 3$ and $V_K = (e_1, e_2, e_3)$. We generate $m$ subsamples, each of which has $n$ samples, and run our proposed distributed PCA
algorithm (Algorithm 1) to calculate $\tilde{V}_K$. Since the centered multivariate Gaussian distribution is symmetric, according to Theorem 3.3.4 when $\lambda = O(d)$ we have

$$\|\rho(\tilde{V}_K, V_K)\|_{\psi_1} = O\left(\frac{C_1\|\Sigma\|_{op}}{\lambda_K - \lambda_{K+1}} \sqrt{Kr(\Sigma)} \right) = O\left(\sqrt{\frac{d}{mn\delta}}\right),$$

(3.5.1)

where $\delta := \lambda_K - \lambda_{K+1} = \lambda/4 - 1$. Now we provide numerical verification of the order of the number of servers $m$, the eigengap $\delta$, the subsample size $n$ and dimension $d$ in the statistical error.

Figure 3.1: Statistical error rate with respect to: (a) the dimension $d$ when $\lambda = 50$ and $n = 2000$; (b) the number of servers $m$ when $\lambda = 50$ and $n = 2000$; (c) the subsample size $n$ when $\lambda = 50$ and $m = 50$; (d) the eigengap $\delta$ when $d = 800$ and $n = 2000$.

Figure 3.1 presents four plots that demonstrate how $\rho(\tilde{V}_K, V_K)$ changes as $d$, $m$, $n$ and $\delta$ increases respectively. Each data point on the plots is based on 100 independent Monte Carlo simulations. Figure 3.1(a) demonstrates how $\rho(\tilde{V}_K, V_K)$ increases with respect to the increasing dimension $d$ when $\lambda = 50$ and $n = 2000$. Each line on the plot represents a fixed number of machines $m$. Figure 3.1(b) shows the decay rate of $\rho(\tilde{V}_K, V_K)$ as the number of servers $m$ increases when $\lambda = 50$ and $n = 2000$. Different lines on the plot correspond to different dimensions $d$. Figure
3.1(c) demonstrates how $\rho(\tilde{V}_K, V_K)$ decays as the subsample size $n$ increases when $\lambda = 50$ and $m = 50$. Figure 3.1(d) shows the relationship between $\rho(\tilde{V}_K, V_K)$ and the eigengap $\delta$ when $d = 800$ and $n = 2000$. The results from figs 3.1(a)-3.1(d) show that $\rho(\tilde{V}_K, V_K)$ is proportion to $d^{\frac{1}{2}}, m^{-\frac{1}{2}}, n^{-\frac{1}{2}}$ and $\delta^{-\frac{1}{2}}$ respectively when the other three parameters are fixed. These empirical results are all consistent with (3.5.1).

Figure 3.1 demonstrates the marginal relationship between $\rho(\tilde{V}_K, V_K)$ and the four parameters $m$, $n$, $d$ and $\delta$. Now we study their joint relationship. Inspired by (3.5.1), we consider a multiple regression model as follows:

$$\log(\rho(\tilde{V}_K, V_K)) = \beta_0 + \beta_1 \log(d) + \beta_2 \log(m) + \beta_3 \log(n) + \beta_4 \log(\delta) + \epsilon,$$  \hspace{1cm} (3.5.2)

where $\epsilon$ is the error term. We collect all the data points $(d,m,n,\delta,\rho(\tilde{V}_K, V_K))$ from four plots in Figure 3.1 to fit the regression model (3.5.2). The fitting result is that $\hat{\beta}_1 = 0.5043$, $\hat{\beta}_2 = -0.4995$, $\hat{\beta}_3 = -0.5011$ and $\hat{\beta}_4 = -0.5120$ with the multiple $R^2 = 0.99997$. These estimates are quite consistent with the theoretical results in (3.5.1). Moreover, Figure 3.2 plots all the observed values of $\log(\rho(\tilde{V}_K, V_K))$ against its fitted values by the linear model (3.5.2). We can see that the observed and fitted values perfectly match. It indicates that the multiple regression model (3.5.2) well explains the joint relationship between the statistical error and the four parameters $m$, $n$, $d$ and $\delta$.

### 3.5.2 The effects of splitting

In this section we investigate how the number of data splits $m$ affects the statistical performance of $\tilde{V}_K$ when the total sample size $N$ is fixed. Since $N = mn$, it is easy to see that the larger $m$ is, the smaller $n$ will be, and hence the less computational load there will be on each individual server. In this way, to reduce the time consumption of the distributed algorithm, we prefer more splits of the data. However, per the
assumptions of Theorem 3.3.4, the subsample size \( n \) should be large enough to achieve the optimal statistical performance of \( \tilde{V}_K \). This motivates us to numerically illustrate how \( \rho(\tilde{V}_K, V_K) \) changes as \( m \) increases with \( N \) fixed.

We adopt the same data generation process as described in the beginning of Section 6.1 with \( \lambda = 50 \) and \( N = 6000 \). We split the data into \( m \) subsamples where \( m \) is chosen to be all the factors of \( N \) that are less than or equal to 300. Figure 3.3 plots \( \rho(\tilde{V}_K, V_K) \) with respect to the number of machines \( m \). Each point on the plot is based on 100 simulations. Each line corresponds to a different dimension \( d \).

The results show that when the number of machines is not unreasonably large, or equivalently the number of subsample size \( n \) is not small, the statistical error does not depend on the number of machines when \( N \) is fixed. This is consistent with (3.5.1) where the statistical error rate only depends on the total sample size \( N = mn \). When the number of machines \( m \) is large (\( \log m \geq 5 \)), or the subsample size \( n \) is small, we observe slightly growing statistical error of the distributed PCA. This is aligned with

Figure 3.2: Observed and fitted values of \( \log(\rho(\tilde{V}_K, V_K)) \).
the required lower bound of $n$ in Theorem 3.3.4 to achieve the optimal statistical performance of $\tilde{V}_K$. Note that even when $m = 300$ ($\log(m) \approx 5.7$) and $n = 20$, our distributed PCA performs very well. This demonstrates that distributed PCA is statistically efficient as long as $m$ is within a reasonable range.

### 3.5.3 Comparison between distributed and full sample PCA

In this subsection, we compare the statistical performance of the following three methods:

1. Distributed PCA (DP)

2. Full sample PCA (FP), i.e., the PCA based on the all the samples

3. Distributed PCA with communication of five additional largest eigenvectors (DP5).

Here we explain more on the third method DP5. The difference between DP5 and DP is that on each server, DP5 calculates $\hat{V}_{K+5}^{(l)}$, the top $K + 5$ eigenvectors of $\Sigma^{(l)}$ and send them to the central server, and the central server computes the top $K$ eigenvectors of $(1/m) \sum_{l=1}^{m} \hat{V}_{K+5}^{(l)} \hat{V}_{K+5}^{(l)\top}$ as the final output. Intuitively, DP5 communicates more information of the covariance structure and is designed to guide the spill-over
effects of the eigenspace spanned by the top $K$ eigenvalues. In Figure 3.4, we compare the performance of all the three methods under various scenarios.

Figure 3.4: Comparison between DP, FP and DP5: (a) $m = 20, n = 2000$ and $\lambda = 50$; (b) $d = 1600, n = 1000$ and $\lambda = 30$; (c) $d = 800, m = 5$ and $\lambda = 30$; (d) $d = 1600, m = 10$ and $n = 500$.

From figures 3.4(a)-3.4(d), we can see that all the three methods have similar finite sample performance. This means that it suffices to communicate $K$ eigenvectors to enjoy the same statistical accuracy as the full sample PCA. For more challenging situations with large $p/(mn\delta)$ ratios, small improvements using FP are visible.
Chapter 4

High-dimensional robust low-rank matrix recovery

4.1 Model and methodology

4.1.1 Trace regression

In this paper, we consider the trace regression model. Suppose we have \(N\) matrices \(\{X_i \in \mathbb{R}^{d_1 \times d_2}\}_{i=1}^N\) and responses \(\{Y_i \in \mathbb{R}\}_{i=1}^N\). We say \(\{(Y_i, X_i)\}_{i=1}^N\) follow the trace regression model if

\[ Y_i = \langle X_i, \Theta^* \rangle + \epsilon_i, \]

where \(\langle X_i, \Theta^* \rangle := \text{Tr}(X_i^T \Theta^*)\), \(\Theta^* \in \mathbb{R}^{d_1 \times d_2}\) is the true coefficient matrix, \(\mathbb{E}X_i = 0\) and \(\{\epsilon_i\}_{i=1}^N\) are independent noises satisfying \(\mathbb{E}(\epsilon_i|X_i) = 0\). Note that here we do not assume \(\{X_i\}_{i=1}^N\) are independent to each other; nor do we assume \(\epsilon_i\) is independent to \(X_i\). Model (4.1.1) includes the following specific cases.

- **Linear regression**: \(d_1 = d_2 = d\), and \(\{X_i\}_{i=1}^N\) and \(\Theta^*\) are diagonal. Let \(x_i\) and \(\theta^*\) denote the vectors of diagonal elements of \(X_i\) and \(\Theta^*\) respectively, i.e., \(X_i = \text{diag}(x_{i1}, \ldots, x_{id})\) and \(\Theta^* = \text{diag}(\theta^*_{1}, \ldots, \theta^*_{d})\). Then, (4.1.1) reduces to
familiar linear model: \( Y_i = x_i^T \theta_i + \epsilon_i \). Having a low-rank \( \Theta^* \) is then equivalent to having a sparse \( \theta^* \).

- **Compressed sensing**: For matrix compressed sensing, entries of \( X_i \) jointly follow the Gaussian distribution or other ensembles. For vector compressed sensing, we can take \( X \) and \( \Theta^* \) as diagonal matrices.

- **Matrix completion**: \( X_i \) is a singleton, i.e., \( X_i = e_{j(i)} e_{k(i)}^T \) for \( 1 \leq j(i) \leq d_1 \) and \( 1 \leq k(i) \leq d_2 \). In other words, a random entry of the matrix \( \Theta \) is observed along with noise for each sample.

- **Multi-task learning**: The multi-task (reduced-rank) regression model is

\[
y_j = \Theta^*^T x_j + \epsilon_j, \quad j = 1, \ldots, n, \tag{4.1.2}
\]

where \( x_j \in \mathbb{R}^{d_1} \) is the covariate vector, \( y_j \in \mathbb{R}^{d_2} \) is the response vector, \( \Theta^* \in \mathbb{R}^{d_1 \times d_2} \) is the coefficient matrix and \( \epsilon_j \in \mathbb{R}^{d_2} \) is the noise with each entry independent to each other. See, for example, [66] and [122]. Each sample \((y_j, x_j)\) consists of \( d_2 \) responses and is equivalent to \( d_2 \) data points in (4.1.1), i.e., \( \{(Y_{(j-1)d_2+i} = y_{ji}, X_{(j-1)d_2+i} = x_j e_{i}^T)\}_{i=1}^{d_2} \). Therefore \( n \) samples in (4.1.2) correspond to \( N = n d_2 \) observations in (4.1.1).

In this paper, we impose rank constraint on the coefficient matrix \( \Theta^* \). Rank constraint can be viewed as a generalized sparsity constraint for two-dimensional matrices. For linear regression, rank constraint is equivalent to the sparsity constraint since \( \Theta^* \) is diagonal. The rank constraint reduces the effective number of parameters in \( \Theta^* \) and arises frequently in many applications. Consider the Netflix problem for instance, where \( \Theta^*_{ij} \) is the intrinsic score of film \( j \) given by customer \( i \) and we would like to recover the entire \( \Theta^* \) with only partial observations. Given that movies of similar types or qualities should receive similar scores from viewers, columns of \( \Theta^* \)
should share colinearity, thus delivering a low-rank structure of $\Theta^*$. The rationale of
the model can also be understood from the celebrated factor model in finance and
econometrics [47], which assumes that several market risk factors drive the returns of
a large panel of stocks. Consider $N \times T$ matrix $Y$ of $N$ stock returns (like movies)
over $T$ days (like viewers). These financial returns are driven by $K$ factors $F$ ($K \times T$
matrix, representing $K$ risk factors realized on $T$ days) with a loading matrix $B$
($N \times K$ matrix), where $K$ is much smaller than $N$ or $T$. The factor model admits
the following form:

$$Y = BF + E$$

where $E$ is idiosyncratic noise. Since $BF$ has a small rank $K$, $BF$ can be regarded as
the low-rank matrix $\Theta^*$ in the matrix completion problem. If all movies were rated by
all viewers in the Netflix problem, the ratings should also be modeled as a low-rank
matrix plus noise, namely, there should be several latent factors that drive ratings
of movies. The major challenge of the matrix completion problem is that there are
many missing entries.

Exact low-rank may be too stringent to model the real-world situations. Instead,
we consider near low-rank $\Theta^*$ satisfying

$$B_q(\Theta^*):= \sum_{i=1}^{d_1 \wedge d_2} \sigma_i(\Theta^*)^q \leq \rho,$$  \hspace{1cm} (4.1.3)

where $0 \leq q \leq 1$. Note that when $q = 0$, the constraint (4.1.3) is the exact rank
constraint. Restriction on $B_q(\Theta^*)$ ensures that the singular values decay fast enough;
it is more general and natural than the exact low-rank assumption. In the analysis,
we can allow $\rho$ to grow with dimensionality and sample size.

A popular method for estimating $\Theta^*$ is the penalized empirical loss that solves
$\hat{\Theta} \in \arg\min_{\Theta \in S} \mathcal{L}(\Theta) + \lambda_N \mathcal{P}(\Theta)$, where $S$ is a convex set in $\mathbb{R}^{d_1 \times d_2}$, $\mathcal{L}(\Theta)$ is the
loss function, $\lambda_N$ is the tuning parameter and $\mathcal{P}(\Theta)$ is a rank penalization function.
Most of the previous work, e.g., \cite{71} and \cite{90}, chose $\mathcal{L}(\Theta) = \sum_{1 \leq i \leq N} (Y_i - \langle \Theta, X_i \rangle)^2$ and $\mathcal{P}(\Theta) = \|\Theta\|_*$, and derived the rate for $\|\hat{\Theta} - \Theta^*\|_F$ under the assumption of sub-Gaussian or sub-exponential noise. However, the $\ell_2$ loss is sensitive to outliers and is unable to handle the data with moderately heavy or heavy tails.

### 4.1.2 Robustifying $\ell_2$ loss

We aim to accommodate heavy-tailed noise and design for the near low-rank matrix recovery by robustifying the traditional $\ell_2$ loss. We first notice that the $\ell_2$ risk can be expressed as

$$R(\Theta) = \mathbb{E} \mathcal{L}(\Theta) = \mathbb{E} (Y_i - \langle \Theta, X_i \rangle)^2$$

$$= \mathbb{E} Y_i^2 - 2\langle \Theta, \mathbb{E} Y_i X_i \rangle + \text{vec}(\Theta)^\top \mathbb{E} \left( \text{vec}(X_i) \text{vec}(X_i)^\top \right) \text{vec}(\Theta)$$

$$\equiv \mathbb{E} Y_i^2 - 2\langle \Theta, \Sigma_{XX} \rangle + \text{vec}(\Theta)^\top \Sigma_{XX} \text{vec}(\Theta). \quad (4.1.4)$$

Ignoring $\mathbb{E} Y_i^2$, if we substitute $\Sigma_{YX}$ and $\Sigma_{XX}$ by their corresponding sample covariances, we recover the empirical $\ell_2$ loss. This inspires us to define a generalized $\ell_2$ loss as follows.

$$\mathcal{L}(\Theta) = -\langle \hat{\Sigma}_{YX}, \Theta \rangle + \frac{1}{2} \text{vec}(\Theta)^\top \hat{\Sigma}_{XX} \text{vec}(\Theta), \quad (4.1.5)$$

where $\hat{\Sigma}_{YX}$ and $\hat{\Sigma}_{XX}$ are estimators of $\mathbb{E} Y_i X_i$ and $\mathbb{E} \text{vec}(X_i) \text{vec}(X_i)^\top$ respectively.

In this paper, we study the following M-estimator of $\Theta^*$ with the generalized $\ell_2$ loss:

$$\hat{\Theta} \in \text{argmin}_{\Theta \in \mathcal{S}} -\langle \hat{\Sigma}_{YX}, \Theta \rangle + \frac{1}{2} \text{vec}(\Theta)^\top \hat{\Sigma}_{XX} \text{vec}(\Theta) + \lambda N \|\Theta\|_* \quad (4.1.6)$$

where $\mathcal{S}$ is a convex set in $\mathbb{R}^{d_1 \times d_2}$. To handle heavy-tailed noise and design, we need to employ robust estimators $\hat{\Sigma}_{YX}$ and $\hat{\Sigma}_{XX}$. For ease of presentation, we always first consider the case where the design is sub-Gaussian and the response is heavy-tailed,
and then further allow the design to have heavy-tailed distribution if it is appropriate for the specific problem setup.

We now introduce the robust covariance estimators to be plugged in (4.1.6) by the principle of truncation, or more generally shrinkage. The intuition is that shrinkage reduces sensitivity of the estimator to the heavy-tailed corruption. However, shrinkage induces bias. Our theories revolve around finding appropriate shrinkage level so as to ensure the induced bias is not too large and the final statistical error rate is sharp. Different problem setups have different forms of \( \hat{\Sigma}_{YX} \) and \( \hat{\Sigma}_{XX} \), but the principle of shrinkage of data is universal. For the linear regression, matrix compressed sensing and matrix completion, in which the response is univariate, \( \hat{\Sigma}_{YX} \) and \( \hat{\Sigma}_{XX} \) take the following forms:

\[
\hat{\Sigma}_{YX} = \hat{\Sigma}_{\bar{Y}\bar{X}} = \frac{1}{N} \sum_{i=1}^{N} \bar{Y}_i \bar{X}_i \quad \text{and} \quad \hat{\Sigma}_{XX} = \hat{\Sigma}_{\bar{X}\bar{X}} = \frac{1}{N} \sum_{i=1}^{N} \text{vec}(\bar{X}_i)\text{vec}(\bar{X}_i)^\top, \quad (4.1.7)
\]

where tilde notation means truncated versions of the random variables if they have heavy tails and equals the original random variables (truncation threshold is infinite) if they have light tails.

For the multi-task regression, similar idea continues to apply. However, writing (4.1.2) in the general form of (4.1.1) requires adaptation of more complicated notation. We choose

\[
\hat{\Sigma}_{YX} = \frac{1}{N} \sum_{i=1}^{n} \sum_{j=1}^{d_2} \bar{Y}_{ij} \bar{x}_i e_j^\top = \frac{1}{d_2} \hat{\Sigma}_{\bar{X}\bar{Y}} \quad \text{and} \quad \hat{\Sigma}_{XX} = \frac{1}{N} \sum_{i=1}^{n} \text{vec}(\bar{x}_i e_j^\top)\text{vec}(\bar{x}_i e_j^\top)^\top = \frac{1}{d_2} \text{diag}(\hat{\Sigma}_{\bar{X}\bar{X}}, \cdots, \hat{\Sigma}_{\bar{X}\bar{X}}), \quad (4.1.8)
\]

where

\[
\hat{\Sigma}_{\bar{X}\bar{Y}} = \frac{1}{n} \sum_{i=1}^{n} \bar{x}_i \bar{y}_i^\top \quad \text{and} \quad \hat{\Sigma}_{\bar{X}\bar{X}} = \frac{1}{n} \sum_{i=1}^{n} \bar{x}_i \bar{x}_i^\top,
\]
and \( \tilde{y}_i \) and \( \tilde{x}_i \) are again transformed versions of \( y_i \) and \( x_i \). The tilde notation means shrinkage for heavy-tailed variables and identity mapping (no shrinkage) for light-tailed variables. The factor \( d_2^{-1} \) is due to the fact that \( n \) independent samples under model (4.1.2) are treated as \( N = nd_2 \) samples in (4.1.1). As we shall see, under only bounded moment assumptions of the design and noise, the proposed truncated or shrinkage covariance enjoys desired convergence rate to its population counterpart. This leads to a sharp \( M \)-estimator \( \hat{\Theta} \), whose statistical error rates match those established in [90] and [91] under the setting of sub-Gaussian design and noise.

Note that using truncated or shrinkage covariance in the generalized \( \ell_2 \) loss is equivalent to evaluating the traditional \( \ell_2 \) loss on the truncated or shrunk data. Nevertheless, instead of directly analyzing the quadratic loss with shrunk data, we study the robust covariance first and then derive the error rate of the corresponding \( M \)-estimator. This analytical framework is modular and allows other potential robust covariance estimators to be plugged in, for instance those based on Kendall’s tau ([41]), median of means ([88]), etc. As long as the recruited \( \hat{\Sigma}_{YX} \) and \( \hat{\Sigma}_{XX} \) satisfy the set of sufficient conditions given by Theorem 1, the corresponding \( \hat{\Theta} \) achieves the desired convergence. Similar idea was recently studied by [80] in estimating the high-dimensional precision matrix. The authors proposed to plug in appropriately chosen robust covariance estimators into graphical Lasso and CLIME and established sharp error bounds for the corresponding estimators of the precision matrix.

Finally, we conjecture that minimizing Huber loss, Tukey’s biweight loss or other robust but maybe non-convex losses ([79], [36]) with nuclear norm regularization can also potentially achieve the nearly minimax optimal rate. However, these papers typically focus on high-dimensional sparse regression and so far we have not seen any statistical guarantee for these methods under the trace regression for the low-rank recovery. Besides, our method is simple to implement with guaranteed optimization efficiency, while the other methods lack reliable and fast algorithms with convergence.

79
guarantee. Since our method is equivalent to applying the standard method to the truncated or shrunk data, the optimization is still a least square problem with nuclear-norm penalization. This is amenable to efficient algorithms such as the Peaceman-Rachford splitting method (PRSM) as described in Section 5.

4.2 Main results

In this section, we derive the statistical error rate of $\hat{\Theta}$ defined by (4.1.6). We always assume $d_1, d_2 \geq 2$ and $\rho > 1$ in (4.1.3). We first present the following general theorem that gives the estimation errors $\|\hat{\Theta} - \Theta^*\|_F$ and $\|\hat{\Theta} - \Theta^*\|_*$.

**Theorem 4.2.1.** Define $\hat{\Delta} = \hat{\Theta} - \Theta^*$, where $\Theta^*$ satisfies $B_q(\Theta^*) \leq \rho$. Suppose $\text{vec}(\hat{\Delta})^T \hat{\Sigma}_{XX} \text{vec}(\hat{\Delta}) \geq \kappa_L \|\hat{\Delta}\|_F^2$, where $\kappa_L$ is a positive constant that does not depend on $\hat{\Delta}$. Choose $\lambda_N \geq 2\|\hat{\Sigma}_{YX} - \text{mat}(\hat{\Sigma}_{XX} \text{vec}(\Theta^*))\|_{op}$. Then we have for some constants $C_1$ and $C_2$,

$$\|\hat{\Delta}\|_F^2 \leq C_1 \rho \left(\frac{\lambda_N}{\kappa_L}\right)^{2-q} \quad \text{and} \quad \|\hat{\Delta}\|_* \leq C_2 \rho \left(\frac{\lambda_N}{\kappa_L}\right)^{1-q}.$$

First of all, the above result is deterministic and nonasymptotic. As we can see from the theorem above, the statistical performance of $\hat{\Theta}$ relies on the restricted eigenvalue (RE) property of $\hat{\Sigma}_{XX}$, which was first studied by [9]. When the design is sub-Gaussian, we choose $\hat{\Sigma}_{XX}$ to be the traditional sample covariance, whose RE property has been well established (e.g., [106, 90, 91]). We will specify these results when we need them in the sequel. When the design only satisfies bounded moment conditions, we choose $\hat{\Sigma}_{XX} = \hat{\Sigma}_{XX}^\odot$ to be the sample covariance of shrunk data. We show that with appropriate level of shrinkage, $\hat{\Sigma}_{XX}^\odot$ still retains the RE property, thus satisfying the conditions of the theorem.

Secondly, the conclusion of the theorem says that $\|\hat{\Delta}\|_F^2$ and $\|\hat{\Delta}\|_*$ are proportional to $\lambda_N^{2-q}$ and $\lambda_N^{1-q}$ respectively, but we require $\lambda_N \geq \|\hat{\Sigma}_{YX} - \text{mat}(\hat{\Sigma}_{XX} \text{vec}(\Theta^*))\|_{op}$. This implies that $\|\hat{\Sigma}_{YX} - \text{mat}(\hat{\Sigma}_{XX} \text{vec}(\Theta^*))\|_{op}$ is crucial to the statistical error of $\hat{\Theta}$. In
the following subsections, we will derive the rate of \( \| \hat{\Sigma}_{YX} - \text{mat}(\hat{\Sigma}_{XX}\text{vec}(\Theta^*)) \|_{op} \) for all the aforementioned specific problems with only bounded moment conditions on the response, and in some cases also on the design. Under such weak assumptions, we show that the proposed robust M-estimator possesses the same rates as those presented in [90, 91] with sub-Gaussian assumptions on the design and noise.

Finally, we emphasize one key technical contribution of our work: the bias-and-variance tradeoff through tuning of the truncation level \( \tau \). As we explained above, \( \| \hat{\Sigma}_{YX} - \text{mat}(\hat{\Sigma}_{XX}\text{vec}(\Theta^*)) \|_{op} \) is crucial to the estimation accuracy of \( \hat{\Theta} \). In our analysis, we decompose \( \hat{\Sigma}_{YX} - \text{mat}(\hat{\Sigma}_{XX}\text{vec}(\Theta^*)) \) into the following three terms:

\[
\hat{\Sigma}_{YX} - \text{mat}(\hat{\Sigma}_{XX}\text{vec}(\Theta^*)) = \hat{\Sigma}_{YX} - E[\hat{\Sigma}_{YX}] + E[\hat{\Sigma}_{YX}] - E[\text{mat}(\hat{\Sigma}_{XX}\text{vec}(\Theta^*))]
\]

\[
\quad + E[\text{mat}(\hat{\Sigma}_{XX}\text{vec}(\Theta^*)) - \text{mat}(\hat{\Sigma}_{XX}\text{vec}(\Theta^*))].
\]

Choosing \( \hat{\Sigma}_{YX} \) and \( \hat{\Sigma}_{XX} \) to be the truncated or shrinkage sample covariance, we will show that the truncation level \( \tau \) only contributes to high-order terms in both concentration terms above. This allows us to strike a perfect balance between the bias term and the concentration terms and establish the optimal rate for \( \| \hat{\Sigma}_{YX} - \text{mat}(\hat{\Sigma}_{XX}\text{vec}(\Theta^*)) \|_{op} \). In contrast, if we simply treat the truncated response as a sub-Gaussian random variable bounded by \( \tau \), \( \tau \) will contribute to the leading terms in the concentration bounds, which implies sub-optimal results. This observation also inspired us to construct the \( \ell_4 \)-norm shrinkage sample covariance and establish its (near) minimax optimal rate in Section 4. This new robust covariance estimator is employed in the multi-tasking regression with heavy-tailed data and leads to a minimax optimal MLE of \( \Theta^* \).
4.2.1 Linear model

For the linear regression problem, $\Theta^*$ and $\{X_i\}_{i=1}^N$ are $d \times d$ diagonal matrices. We denote the diagonals of $\Theta^*$ and $\{X_i\}_{i=1}^N$ by $\theta^*$ and $\{x_i\}_{i=1}^N$ respectively for ease of presentation. The optimization problem in (4.1.6) reduces to

$$\hat{\theta} \in \arg\min_{\theta \in \mathbb{R}^d} - \hat{\Sigma}_{YX}^\top \theta + \frac{1}{2} \theta^\top \hat{\Sigma}_{XX} \theta + \lambda_N \|\theta\|_1,$$

(4.2.1)

where $\hat{\Sigma}_{YX} = \hat{\Sigma}_{YX}^\top = N^{-1} \sum_{i=1}^N \tilde{Y}_i \tilde{x}_i$, $\hat{\Sigma}_{XX} = \hat{\Sigma}_{XX}^\top = N^{-1} \sum_{i=1}^N \tilde{x}_i \tilde{x}_i^\top$. When the design is sub-Gaussian, we only need to truncate the response. Therefore, we choose $\tilde{Y}_i = \tilde{Y}_i(\tau) = sgn(Y_i)(|Y_i| \land \tau)$ and $\tilde{x}_i = x_i$, for some threshold $\tau$. When the design is heavy-tailed, we choose $\tilde{Y}_i(\tau) = sgn(Y_i)(|Y_i| \land \tau_1)$ and $\tilde{x}_{ij} = sgn(x_{ij})(|x_{ij}| \land \tau_2)$, where $\tau_1$ and $\tau_2$ are both predetermined threshold values. To avoid redundancy, we will not repeat stating these choices in lemmas or theorems in this subsection.

To establish the statistical error rate of $\hat{\theta}$ in (4.2.1), in the following lemma, we derive the rate of $\|\hat{\Sigma}_{YX} - \text{mat}(\hat{\Sigma}_{XX} \text{vec}(\Theta^*))\|_{op}$ in (4.1.6) for the sub-Gaussian design and bounded-moment (polynomial tail) design respectively. Note here that

$$\|\hat{\Sigma}_{YX} - \text{mat}(\hat{\Sigma}_{XX} \text{vec}(\Theta^*))\|_{op} = \|\hat{\Sigma}_{YX} - \hat{\Sigma}_{XX} \theta^*\|_{\infty}.$$

Lemma 4.2.1. Uniform convergence of cross covariance.

(a) **Sub-Gaussian design.** Consider the following conditions:

(C1) $\{x_i\}_{i=1}^N$ are i.i.d. sub-Gaussian vectors with $\|x_i\|_{\psi_2} \leq \kappa_0 < \infty$, $\mathbb{E}x_i = 0$ and $\lambda_{\min}(\mathbb{E}x_i x_i^\top) \geq \kappa_\mathcal{L} > 0$;

(C2) $\forall i = 1, ..., N$, $\mathbb{E}|Y_i|^{2k} \leq M < \infty$ for some $k > 1$. 82
Choose $\tau \asymp \sqrt{N/\log d}$. For any $\delta > 0$, there exists a constant $\gamma_1 > 0$ such that as long as $\log d/N < \gamma_1$, we have

$$P\left( \left\| \hat{\Sigma}_{Yx}(\tau) - \hat{\Sigma}_{xx}\theta^* \right\|_\infty \geq \nu_1 \sqrt{\frac{\delta \log d}{N}} \right) \leq 2d^{1-\delta},$$

(4.2.2)

where $\nu_1$ is a universal constant.

(b) **(Bounded moment design)** Consider instead the following set of conditions:

(C1') $\|\theta^*\|_1 \leq R < \infty$;

(C2') $E|x_{ij}|^4 \leq M < \infty$, $1 \leq j \leq d$;

(C3') $\forall i = 1, ..., N, E|Y_i|^4 \leq M < \infty$.

Choose $\tau_1, \tau_2 \asymp (N/\log d)^\frac{1}{4}$. For any $\delta > 0$, it holds that

$$P\left( \left\| \hat{\Sigma}_{Yx}(\tau_1, \tau_2) - \hat{\Sigma}_{xx}(\tau_2)\theta^* \right\|_{\max} > \nu_2 \sqrt{\frac{\delta \log d}{N}} \right) \leq 2d^{1-\delta},$$

where $\nu_2$ is a universal constant.

**Remark 4.2.1.** If we choose $\hat{\Sigma}_{Yx}$ and $\hat{\Sigma}_{xx}$ to be the sample covariance, i.e., $\hat{\Sigma}_{Yx} = \Sigma_{Yx} = \frac{1}{N} \sum_{i=1}^{N} Y_i x_i$ and $\hat{\Sigma}_{xx} = \Sigma_{xx} = \frac{1}{N} \sum_{i=1}^{N} x_i x_i^\top$, Corollary 2 of [93] showed that under the sub-Gaussian noise and design,

$$\| \Sigma_{Yx} - \Sigma_{xx}\theta^* \|_\infty = O_P\left( \sqrt{\frac{\log d}{N}} \right).$$

This is the same rate as what we achieved under only the bounded moment conditions on response and design.

Next we establish the restricted strong convexity of the proposed robust $\ell_2$ loss.

**Lemma 4.2.2.** Restricted strong convexity.
(a) **Sub-Gaussian design.** Under Condition (C1) of Lemma 4.2.1, it holds for certain constants $C_1, C_2$ and any $\eta_1 > 1$ that

$$
P\left( v^\top \hat{\Sigma}_{xx} v \geq \frac{1}{2} v^\top \Sigma_{xx} v - \frac{C_1 \eta_1 \log d}{N} \|v\|_2^2, \forall v \in \mathbb{R}^d \right) \geq 1 - \frac{d^{1-\eta_1}}{3} - 2d \exp(-C_2 N).
$$

(4.2.3)

(b) **Bounded moment design.** If $x_i$ satisfies Condition (C2') of Lemma 4.2.1, then it holds for some constant $C_3 > 0$ and any $\eta_2 > 2$ that

$$
P\left( v^\top \hat{\Sigma}_{xx}(\tau_2) v \geq v^\top \Sigma_{xx} v - C_3 \eta_2 \sqrt{\frac{\log d}{N}} \|v\|_1^2, \forall v \in \mathbb{R}^d \right) \leq d^{2-\eta_2},
$$

(4.2.4)

as long as $\tau_2 \approx (N/\log d)^{\frac{1}{4}}$.

**Remark 4.2.2.** Comparing the results we get for sub-Gaussian design and heavy-tailed design, we see that the coefficients before $\|v\|_1^2$ are different. Under the sub-Gaussian design, that coefficient is of the order $\log d/N$, while under the heavy-tailed design, the coefficient is of order $\sqrt{\log d/N}$. This difference leads to different scaling requirements for $N, d$ and $\rho$ in the sequel. As we shall see, the heavy-tailed design requires stronger scaling conditions to retain the same statistical error rate as the sub-Gaussian design for the linear model.

Finally we derive the statistical error rate of $\hat{\theta}$ as defined in (4.2.1).

**Theorem 4.2.2.** Assume $\sum_{i=1}^d |\theta^*_i|^q \leq \rho$, where $0 \leq q \leq 1$.

(a) **Sub-Gaussian design:** Suppose Conditions (C1) and (C2) in Lemma 4.2.1 hold. For any $\delta > 0$, choose $\tau \approx \sqrt{N/\log d}$ and $\lambda_N = 2\nu_1 \sqrt{\delta \log d/N}$, where $\nu_1$ and $\delta$ are the same as in part (a) of Lemma 4.2.1. There exist positive constants $\{C_i\}_{i=1}^3$ such that as long as $\rho(\log d/N)^{1-\frac{q}{2}} \leq C_1$, it holds that

$$
P\left( \|\hat{\theta}(\tau, \lambda_N) - \theta^*\|_2^2 > C_2 \rho \left( \frac{\delta \log d}{N} \right)^{1-\frac{q}{2}} \right) \leq 3d^{1-\delta}
$$

84
and
\[ P\left(\|\hat{\theta}(\tau, \lambda_N) - \theta^*\|_1 > C_3\rho\left(\frac{\delta \log d}{N}\right)^{\frac{1}{2-q}}\right) \leq 3d^{1-\delta}. \]

(b) **Bounded moment design:** For any $\delta > 0$ choose $\tau_1, \tau_2 \approx (N/\log d)^{\frac{1}{4}}$ and $\lambda_N = 2\nu_2\sqrt{\delta \log d/N}$, where $\nu_2$ and $\delta$ are the same as in part (b) of Lemma 4.2.1 Under Conditions (C1'), (C2') and (C3'), there exist constants $\{C_i\}_{i=4}$ such that as long as $\rho(\log d/N)^{1-\frac{q}{2}} \leq C_4$, we have
\[ P\left(\|\hat{\theta}(\tau_1, \tau_2, \lambda_N) - \theta^*\|_2^2 > C_5\rho\left(\frac{\delta \log d}{N}\right)^{\frac{1}{2-q}}\right) \leq 3d^{1-\delta} \]
and
\[ P\left(\|\hat{\theta}(\tau_1, \tau_2, \lambda_N) - \theta^*\|_1 > C_6\rho\left(\frac{\delta \log d}{N}\right)^{\frac{1}{2-q}}\right) \leq 3d^{1-\delta}. \]

**Remark 4.2.3.** Under both sub-Gaussian and heavy-tailed design, our proposed $\hat{\theta}$ achieves the minimax optimal rate of $\ell_2$ norm established by [100]. However, the difference lies in the scaling requirement on $N$, $d$ and $\rho$. For sub-Gaussian design, we require $\rho(\log d/N)^{1-\frac{q}{2}} \leq C_1$, whereas for heavy-tailed design we need $\rho(\log d/N)^{1-\frac{q}{2}} \leq C_4$. Under the typical high-dimensional regime that $d \gg N \gg \log d$, the former is weaker. Therefore, heavy-tailed design requires stronger scaling than sub-Gaussian design to achieve the optimal statistical rates.

### 4.2.2 Matrix compressed sensing

For the matrix compressed sensing problem, since the design is chosen by users, we only consider the most popular design: the Gaussian design. We thus keep the original design matrix and only truncate the response. In (4.1.7), choose $\tilde{Y}_i = sgn(Y_i)(|Y_i|^{\tau})$
and \( \tilde{X}_i = X_i \), then we have

\[
\hat{\Sigma}_{YX} = \hat{\Sigma}_{YX}(\tau) = \frac{1}{N} \sum_{i=1}^{N} sgn(Y_i)(|Y_i| \wedge \tau)X_i,
\]

\[
\hat{\Sigma}_{XX} = \frac{1}{N} \sum_{i=1}^{N} \text{vec}(X_i)\text{vec}(X_i)^\top.
\]

(4.2.5)

The following lemma quantifies the rate of \( \| \hat{\Sigma}_{YX} - \text{mat} (\hat{\Sigma}_{XX} \text{vec}(\Theta^*)) \|_{op}. \) Note that here \( \hat{\Sigma}_{YX} - \text{mat} (\hat{\Sigma}_{XX} \text{vec}(\Theta^*)) = \hat{\Sigma}_{YX}(\tau) - \frac{1}{N} \sum_{i=1}^{N} \langle X_i, \Theta^* \rangle X_i. \)

Lemma 4.2.3. Consider the following conditions:

(C1) \( \{ \text{vec}(X_i) \}_{i=1}^{N} \) are i.i.d. sub-Gaussian vectors with \( \| \text{vec}(X_i) \|_{\psi_2} \leq \kappa_0 < \infty, \)

\[
\mathbb{E}X_i = 0 \quad \text{and} \quad \lambda_{\min}(\mathbb{E}\text{vec}(X_i)\text{vec}(X_i)^\top) \geq \kappa_L > 0.
\]

(C2) \( \forall i = 1, ..., N, \ \mathbb{E}|Y_i|^{2k} \leq M < \infty \) for some \( k > 1. \)

There exists a constant \( \gamma > 0 \) such that as long as \( (d_1 + d_2)/N < \gamma, \) it holds that

\[
P\left( \| \hat{\Sigma}_{YX}(\tau) - \frac{1}{N} \sum_{i=1}^{N} \langle X_i, \Theta^* \rangle X_i \|_{op} \geq \nu \sqrt{\frac{d_1 + d_2}{N}} \right) \leq \eta \exp(- (d_1 + d_2)), \quad (4.2.6)
\]

where \( \tau \approx \sqrt{N/(d_1 + d_2)} \) and \( \nu \) and \( \eta \) are constants.

Remark 4.2.4. For the sample covariance \( \bar{\Sigma}_{YX} = \frac{1}{N} \sum_{i=1}^{N} Y_i X_i, \) [90] showed that when the noise and design are sub-Gaussian,

\[
\| \bar{\Sigma}_{YX} - \frac{1}{N} \sum_{i=1}^{N} \langle X_i, \Theta^* \rangle X_i \|_{op} = \| \frac{1}{N} \sum_{i=1}^{N} \epsilon_i X_i \|_{op} = O_P(\sqrt{(d_1 + d_2)/N}).
\]

Lemma 4.2.3 shows that \( \hat{\Sigma}_{YX}(\tau) \) achieves the same rate for response with just bounded moments.

The following theorem gives the statistical error rate of \( \hat{\Theta} \) in (4.1.6).
Theorem 4.2.3. Suppose Conditions (C1) and (C2) in Lemma 4.2.3 hold and $B_q(\Theta^*) \leq \rho$. We further assume that $\text{vec}(X_i)$ is Gaussian. Choose $\tau \asymp \sqrt{N/(d_1 + d_2)}$ and $\lambda_N = 2\nu \sqrt{(d_1 + d_2)/N}$, where $\nu$ is the same as in Lemma 4.2.3. There exist constants $\{C_i\}_{i=1}^4$ such that once $\rho \left((d_1 + d_2)/N\right)^{1/2} \leq C_1$, we have

$$P\left(\|\hat{\Theta}(\tau, \lambda_N) - \Theta^*\|_F^2 \geq C_2 \rho \left(\frac{d_1 + d_2}{N}\right)^{1 - \frac{3}{2}}\right) \leq \eta \exp(-(d_1 + d_2)) + 2 \exp(-N/32),$$

and

$$P\left(\|\hat{\Theta}(\tau, \lambda_N) - \Theta^*\|_\ast \geq C_3 \rho \left(\frac{d_1 + d_2}{N}\right)^{1 - \frac{1}{2}}\right) \leq \eta \exp(-(d_1 + d_2)) + 2 \exp(-N/32),$$

where $\eta$ is the same constant as in Lemma 4.2.3.

Remark 4.2.5. The Frobenius norm rate here is identical to the rate established under sub-Gaussian noise in [90]. When $q = 0$, $\rho$ is the upper bound of the rank of $\Theta^*$ and the rate of convergence depends only on $\rho(d_1 + d_2) \asymp \rho(d_1 \lor d_2)$, the effective number of independent parameters in $\Theta^*$, rather than the ambient dimension $d_1 \ast d_2$.

Remark 4.2.6. The rate we derived in Theorem 4.2.3 is minimax optimal. Denote max$(d_1, d_2)$ by $d$ and define

$$\Psi(N, d, r, \rho) := \min\left(\frac{rd}{N}, \rho\left(\frac{d}{N}\right)^{1 - \frac{1}{2}}\right),$$

Under the restricted isometry condition, [103] gives a minimax lower bound on the statistical rate of recovering a low-rank matrix under trace regression:

$$\inf_{\Theta} \sup_{\Theta^* \in B_q(\rho), \text{rank}(\Theta^*) \leq r} P\left(\|\hat{\Theta} - \Theta^*\|_F^2 \geq C\Psi(N, d, r, \rho)\right) \geq c > 0,$$

87
where \( C \) is a constant independent of \( N, d, \rho \) and \( c \) is a universal positive constant.

When \( r \) is very large or \( q = 0 \), \( \rho(d/N)^{1-\frac{q}{2}} \) becomes the dominant term in \( \Psi(N,d,r,\rho) \).

This matches the upper bound we achieved in Theorem 4.2.3.

4.2.3 Matrix completion

In this section, we consider the matrix completion problem with heavy-tailed noises. Under a conventional setting, \( X_i \) is a singleton, \( \|\Theta^*\|_{\max} = O(1) \) and \( \|\Theta^*\|_F = O(\sqrt{d_1d_2}) \). If we rescale the original model as

\[
Y_i = \langle X_i, \Theta^* \rangle + \epsilon_i = \langle \sqrt{d_1d_2}X_i, \Theta^* / \sqrt{d_1d_2} \rangle + \epsilon_i
\]

and treat \( \sqrt{d_1d_2}X_i \) as the new design \( \tilde{X}_i \) and \( \Theta^* / \sqrt{d_1d_2} \) as the new coefficient matrix \( \tilde{\Theta}^* \), then \( \|\tilde{X}_i\|_F = O(\sqrt{d_1d_2}) \) and \( \|\tilde{\Theta}^*\|_F = O(1) \). Therefore, by rescaling, we can assume without loss of generality that \( \Theta^* \) satisfies \( \|\Theta^*\|_F \leq 1 \) and \( X_i \) is uniformly sampled from \( \{\sqrt{d_1d_2} \cdot e_j e_k^\top \}_{1 \leq j \leq d_1, 1 \leq k \leq d_2} \).

For the matrix completion problem, in order to achieve consistent estimation, we require the true coefficient matrix \( \Theta^* \) not to be overly spiky, i.e., \( \|\Theta^*\|_{\max} \leq R\|\Theta^*\|_F / \sqrt{d_1d_2} \leq R/\sqrt{d_1d_2} \). We put a similar constraint in seeking the corresponding M-estimator:

\[
\hat{\Theta} \in \arg\min_{\|\Theta\|_{\max} \leq R/\sqrt{d_1d_2}} -\langle \hat{\Sigma}_Y X(\tau), \Theta \rangle + \frac{1}{2} \vec{\text{vec}}(\Theta)^\top \hat{\Sigma}_XX \vec{\text{vec}}(\Theta) + \lambda_N\|\Theta\|_s. \quad (4.2.7)
\]

This spikiness condition is proposed by [91] and it is required by the matrix completion problem per se instead of our robust estimation.

To derive robust estimation in matrix completion problem, we choose \( \tilde{Y}_i = \text{sgn}(Y_i)(|Y_i| \wedge \tau) \) and \( \tilde{X}_i = X_i \) in (4.1.7). Then, \( \hat{\Sigma}_Y X \) and \( \hat{\Sigma}_XX \) are given by (4.1.5). Note that the design \( X_i \) here takes the singleton form, which leads to different scaling and consistency rates from the setting of matrix compressed sensing.
Lemma 4.2.4. Under the following conditions:

(C1) \( \| \Theta^* \|_F \leq 1 \) and \( \| \Theta^* \|_{\text{max}} \leq R/\sqrt{d_1 d_2} \), where \( 0 < R < \infty \);

(C2) \( X_i \) is uniformly sampled from \( \{ \sqrt{d_1 d_2} \cdot e_j e_k^\top \} \)

(C3) \( \forall i = 1, \ldots, N, \E(\epsilon_i^2 | X_i))^k \leq M < \infty \), where \( k > 1 \);

there exists a constant \( \gamma > 0 \) such that for any \( \delta > 0 \), as long as \( (d_1 \lor d_2) \log(d_1 + d_2)/N < \gamma \),

\[
P\left( \| \bar{\Sigma}_{YX}(\tau) - \frac{1}{N} \sum_{i=1}^{N} \langle X_i, \Theta^* \rangle X_i \|_{\text{op}} > \nu \sqrt{\frac{\delta(d_1 \lor d_2) \log(d_1 + d_2)}{N}} \right) \leq 2(d_1 + d_2)^{1-\delta},
\]

where \( \tau \propto \sqrt{N/((d_1 \lor d_2) \log(d_1 + d_2))} \) and \( \nu \) is a universal constant.

Remark 4.2.7. Again, for \( \bar{\Sigma}_{YX} = \frac{1}{N} \sum_{i=1}^{N} Y_i X_i \), \cite{[1]} proved that \( \| \bar{\Sigma}_{YX} - \frac{1}{N} \sum_{i=1}^{N} \langle X_i, \Theta^* \rangle X_i \|_{\text{op}} = O_P(\sqrt{(d_1 + d_2) \log(d_1 + d_2)/N}) \) for sub-exponential noise. Compared with this result, Lemma 4.2.4 achieves the same rate of convergence. By Jensen’s inequality, condition (C3) is implied by \( \E\epsilon_i^{2k} \leq M < \infty \).

Now we present the following theorem on the statistical error of \( \hat{\Theta} \) defined in (4.2.7).

Theorem 4.2.4. Suppose that the conditions of Lemma 4.2.4 hold. Consider \( B_q(\Theta^*) \leq \rho \) with \( \| \Theta^* \|_{\text{max}}/\| \Theta^* \|_F \leq R/\sqrt{d_1 d_2} \). For any \( \delta > 0 \), choose

\[
\tau \propto \sqrt{N/((d_1 \lor d_2) \log(d_1 + d_2))} \quad \text{and} \quad \lambda_N = 2\nu \sqrt{\delta(d_1 \lor d_2) \log(d_1 + d_2)/N}
\]

and assume \( (d_1 \lor d_2) \log(d_1 + d_2)/N < \gamma \), where \( \nu \) and \( \gamma \) are the same as in Lemma 4.2.4

There exist universal constants \( \{ C_i \}_{i=1}^4 \) such that with probability at least
\[1 - 2(d_1 + d_2)^{1-\delta} - C_1 \exp(-C_2(d_1 + d_2))\] we have
\[
\|\hat{\Theta}(\tau, \lambda N) - \Theta^*\|_F^2 \leq C_3 \max \left\{ \rho \left( \frac{\delta R^2(d_1 + d_2) \log(d_1 + d_2)}{N} \right)^{1-q} \frac{R^2}{N} \right\}
\]
and
\[
\|\hat{\Theta}(\tau, \lambda N) - \Theta^*\|_* \leq C_4 \max \left\{ \rho \left( \frac{\delta R^2(d_1 + d_2) \log(d_1 + d_2)}{N} \right)^{1-q} \frac{\rho R^{2-2q}}{N^{1-q}}^{\frac{1}{2q}} \right\},
\]
where \(\hat{\Theta}\) is defined in (4.2.7).

**Remark 4.2.8.** We claim that the rate we derived in Theorem 4.2.4 is minimax optimal up to a logarithmic factor and a trailing term. Theorem 3 in [91] shows that for matrix completion problems,
\[
\inf_{\hat{\Theta} \in S} \sup_{B_q(\Theta)} E\|\hat{\Theta} - \Theta^*\|_F^2 \geq c \min \left( \rho \left( \frac{d}{N} \right)^{1-\frac{q}{2}}, \frac{d^2}{N} \right),
\]
where \(c\) is some constant. As commented therein, as long as \(\rho = O((d/N)^{\frac{q}{2}} d)\), \(\rho(d/N)^{1-q/2}\) is the dominant term and this is what we established in Theorem 4.2.4 up to a logarithmic factor and a small trailing term.

**4.2.4 Multi-task learning**

Before presenting the theoretical results, we first simplify (4.1.6) under the setting of multi-task regression. According to (4.1.8), (4.1.6) can be reduced to the following form:
\[
\hat{\Theta} \in \arg\min_{\Theta \in \mathcal{S}} \frac{1}{d_2} \left( - (\hat{\Sigma}_{\tilde{x}\tilde{y}}, \Theta) + \frac{1}{n} \sum_{i=1}^{n} \|\Theta^T \tilde{x}_i\|_2^2 \right) + \lambda_N \|\Theta\|_*.
\]
(4.2.9)

Recall here that \(n\) is the sample size in terms of (4.1.2) and \(N = d_2 n\). We also have
\[
\hat{\Sigma}_{XY} = \text{mat}(\hat{\Sigma}_{XX} \text{vec}(\Theta^*)) = (\hat{\Sigma}_{\tilde{x}\tilde{y}} - \hat{\Sigma}_{\tilde{x}\tilde{x}} \Theta^*)/d_2.
\]
Under the sub-Gaussian design, we only need to shrink the response vector $y_i$. We choose for $\tilde{x}_i = x_i$ and $\tilde{y}_i = (\|y_i\|_2 \wedge \tau) y_i / \|y_i\|_2$, where $\tau$ is some threshold value that depends on $n, d_1$ and $d_2$. In other words, we keep the original design, but shrink the Euclidean norm of the response. Note that when $y_i$ is one-dimensional, the shrinkage reduces to the truncation $y_i(\tau) = \text{sgn}(y_i)(|y_i| \wedge \tau)$. When the design is only of bounded moments, we need to shrink both the design vector $x_i$ and response vector $y_i$ by their $\ell_4$ norm instead, i.e., we choose $\tilde{x}_i = (\|x_i\|_4 \wedge \tau_1) x_i / \|x_i\|_4$ and $\tilde{y}_i = (\|y_i\|_4 \wedge \tau_2) y_i / \|y_i\|_4$, where $\tau_1$ and $\tau_2$ are two thresholds. Here shrinking based on the fourth moment is uncommon, but it is crucial; in fact it accelerates the convergence rate of the induced bias term so that it can attain the desired statistical error rate. The details can be found in the proofs. Again, we will omit stating these choices in the following lemmas and theorems to ease presentation.

**Lemma 4.2.5.** Convergence rate of gradient of the robustified quadratic loss.

(a) **Sub-Gaussian design.** Under the following conditions:

(C1) $\lambda_{\max}(E y_i y_i^\top) \leq R < \infty$;

(C2) $\{x_i\}_{i=1}^n$ are i.i.d. sub-Gaussian vectors with $\|x_i\|_{\psi_2} \leq \kappa_0 < \infty$, $E x_i = 0$ and $\lambda_{\min}(E x_i x_i^\top) \geq \kappa_L > 0$.

(C3) $\forall i = 1, ..., n$, $j_1, j_2 = 1, ..., d_2$ and $j_1 \neq j_2$, $\epsilon_{ij_1} \perp \epsilon_{ij_2} | x_i$, and $\forall j = 1, ..., d_1$,

$E (E(\epsilon_{ij}^2 | x_i))^k \leq M < \infty$, where $k > 1$;

there exists some constant $\gamma > 0$ such that if $(d_1 + d_2) \log(d_1 + d_2)/n < \gamma$, we have for any $\delta > 0$,

$$P\left(\|\tilde{\Sigma}_{\hat{x}\hat{y}}(\tau) - \tilde{\Sigma}_{\hat{x}\hat{x}} \Theta^*\|_{op} \geq \sqrt{\frac{(\nu_1 + \delta)(d_1 + d_2) \log(d_1 + d_2)}{n}}\right) \leq 2(d_1 + d_2)^{1-\eta_1 \delta},$$

where $\tau \asymp \sqrt{n/((d_1 + d_2) \log(d_1 + d_2))}$ and $\nu_1$ and $\eta_1$ are universal constants.
(b) **Bounded moment design.** Consider Condition (C2') that for any \( \mathbf{v} \in S^{d_1-1} \), 
\[ \mathbb{E}((\mathbf{v}^\top \mathbf{x}_i)^4) \leq M < \infty. \]
Under Conditions (C1), (C2') and (C3), it holds for any \( \delta > 0 \)
\[
P\left( \|\hat{\Sigma}_{xy}(\tau_1, \tau_2) - \hat{\Sigma}_{x\hat{x}}(\tau_1) \Theta^*\|_{op} \geq \sqrt{\frac{(\nu_2 + \delta)(d_1 + d_2) \log(d_1 + d_2)}{n}} \right)
\leq 2(d_1 + d_2)^{1-\eta_2 \delta},
\]
where \( \tau_1, \tau_2 \asymp \left( n/((d_1+d_2) \log(d_1+d_2)) \right)^{\frac{1}{4}} \) and \( \nu_2 \) and \( \eta_2 \) are universal constants.

**Remark 4.2.9.** When the noise and design are sub-Gaussian, [90] used the covering argument to show that for regular sample covariance matrices \( \Sigma_{xy} \) and \( \Sigma_{xx} \),
\[
\|\Sigma_{xy} - \Sigma_{xx} \Theta^*\|_{op} = \frac{1}{n} \sum_{j=1}^{n} \epsilon_j x_j^\top \|_{op} = O_P(\sqrt{(d_1 + d_2)/n}).
\]

Lemma 4.2.5 shows that up to a logarithmic factor, the shrinkage sample covariance achieves nearly the same rate of convergence for noise and design with only bounded moments.

Finally we establish the statistical error rate for the low-rank multi-task learning.

**Theorem 4.2.5.** Statistical error rate for multi-task learning. Assume \( B_q(\Theta^*) \leq \rho \).

(a) **Sub-Gaussian design.** Suppose that Conditions (C1), (C2) and (C3) in Lemma 4.2.5 hold. For any \( \delta > 0 \), choose
\[
\tau \asymp \sqrt{n/((d_1 + d_2) \log(d_1 + d_2))}
\]
and
\[
\lambda_N = \frac{2}{d_2} \sqrt{(\nu_1 + \delta)(d_1 + d_2) \log(d_1 + d_2)/n},
\]
where \( \nu_1 \) is the same as in Lemma 4.2.5. There exist constants \( \gamma_1, \gamma_2 > 0 \) such that if \( (d_1 + d_2) \log(d_1 + d_2)/n < \gamma_1 \) and \( d_1 + d_2 \geq \gamma_2 \), then with probability at
least $1 - 3(d_1 + d_2)^{1-m\delta}$ we have

$$\|\hat{\Theta}(\tau, \lambda_N) - \Theta^*\|_F^2 \leq C_1\rho \left( \frac{(\nu_1 + \delta)(d_1 + d_2) \log(d_1 + d_2)}{n} \right)^{1-\frac{q}{2}}$$

and

$$\|\hat{\Theta}(\tau, \lambda_N) - \Theta^*\|_* \leq C_2\rho \left( \frac{(\nu_1 + \delta)(d_1 + d_2) \log(d_1 + d_2)}{n} \right)^{1-\frac{q}{2}},$$

where $C_1$ and $C_2$ are universal constants and $\eta_1$ is the same as in Lemma 4.2.5.

(b) **Bounded moment design.** Suppose instead that Conditions (C1), (C2') and (C3) in Lemma 4.2.5 hold. For any $\delta > 0$, choose

$$\tau_1, \tau_2 \asymp \left( n/((d_1 + d_2) \log(d_1 + d_2)) \right)^{\frac{1}{4}}$$

and

$$\lambda_N = \frac{2}{d_2} \sqrt{(\nu_2 + \delta)(d_1 + d_2) \log(d_1 + d_2)/n},$$

where $\nu_2$ is the same as in Lemma 4.2.5. There exist constants $\gamma_3, \gamma_4 > 0$ such that if $(d_1 + d_2) \log(d_1 + d_2)/n < \gamma_3$ and $d_1 + d_2 \geq \gamma_4$, then with probability at least $1 - 3(d_1 + d_2)^{1-m\delta}$,

$$\|\hat{\Theta}(\tau_1, \tau_2, \lambda_N) - \Theta^*\|_F^2 \leq C_3\rho \left( \frac{(\nu_2 + \delta)(d_1 + d_2) \log(d_1 + d_2)}{n} \right)^{1-\frac{q}{2}}$$

and

$$\|\hat{\Theta}(\tau_1, \tau_2, \lambda_N) - \Theta^*\|_* \leq C_4\rho \left( \frac{(\nu_2 + \delta)(d_1 + d_2) \log(d_1 + d_2)}{n} \right)^{1-\frac{q}{2}},$$

where $C_3$ and $C_4$ are universal constants and $\eta_2$ is the same as in Lemma 4.2.5.

**Remark 4.2.10.** According to (A.39) in Appendix A, the multi-task regression model satisfies the lower bound part of the RI condition in [104] with $\nu \asymp \sqrt{d_2}$. Substituting
\[ \nu \approx \sqrt{d_2}, \Delta = \rho^2 / \nu, \ M = \max(d_1, d_2) \] and \[ N = nd_2 \] into Theorem 5 in [104] yields that
\[ \Psi(n, r, d_1, d_2, \rho) = \frac{1}{d_2} \min \left( \frac{r \max(d_1, d_2)}{n}, \rho \left( \frac{\max(d_1, d_2)}{n} \right)^{1 - \frac{q}{2}}, \rho^\frac{1}{2} \right). \]

Note that therein \( C(\alpha, \nu) \approx d_2 \). Therefore we have
\[ \inf_{\hat{\Theta}} \sup_{\Theta^* \in B_q, \text{rank}(\Theta^*) \leq \tau} P \left( \| \hat{\Theta} - \Theta^* \|_F^2 \geq C d_2 \Psi(n, r, d_1, d_2, \rho) \right) \geq c > 0, \]
where \( C \) and \( c \) are constants. Under regular scaling assumptions, the dominant term in the minimax rate is \( \rho(\max(d_1, d_2)/n)^{1 - q/2} \), which matches our upper bound in Theorem 4.2.5 up to a logarithmic factor.

### 4.3 Robust covariance estimation

In multi-task regression, the error bound derivation of \( \| \hat{\Sigma}_{\tilde{x} \tilde{y}}(\tau_1, \tau_2) - \hat{\Sigma}_{\tilde{x} \tilde{x}}(\tau_1) \Omega^* \|_{op} \) sheds light on applying the \( \ell_4 \)-norm shrinkage for robust covariance estimation. This topic is of its own interest, so we emphasize this interesting finding with a separate section. Here we formulate the problem and the result, whose proof is relegated to Appendix A.

Suppose we have \( n \) i.i.d. \( d \)-dimensional random vectors \( \{x_i\}_{i=1}^n \) with \( \mathbb{E}x_i = 0 \). Our goal is to estimate the covariance matrix \( \Sigma = \mathbb{E}(x_i x_i^\top) \) when the distribution of \( \{x_i\}_{i=1}^n \) has only bounded fourth moment. For any \( \tau \in \mathbb{R}^+ \), let \( \tilde{x}_i := (\|x_i\|_4 \wedge \tau)x_i / \|x_i\|_4 \), where \( \| \cdot \|_4 \) is the \( \ell_4 \)-norm. We propose the following shrinkage sample covariance to estimate \( \Sigma \).
\[ \tilde{\Sigma}_n(\tau) = \frac{1}{n} \sum_{i=1}^n \tilde{x}_i \tilde{x}_i^\top. \]

The following theorem establishes the statistical error rate of \( \tilde{\Sigma}_n(\tau) \) with exponential deviation bound.
**Theorem 4.3.1.** Suppose $\mathbb{E}(v^T x_i)^4 \leq R$ for any $v \in \mathcal{S}^{d-1}$, then it holds that for any $\delta > 0$,
\[
P \left( \|\tilde{\Sigma}_n(\tau) - \Sigma\|_{\text{op}} \geq \sqrt{\frac{\delta R d \log\frac{d}{n}}{n}} \right) \leq d^{1-C_\delta},
\]
where $\tau \simeq \left(\frac{n R}{\delta \log d}\right)^{1/4}$ and $C$ is a universal constant. Unlike the bounded moment conditions in the previous section, $R$ here can go to infinity with certain rates.

Below we also present a lower bound result, showing that our shrinkage sample covariance is minimax optimal up to a logarithmic factor.

**Theorem 4.3.2.** Define $\Sigma_v := vv^T + I$. Suppose $\{x_i\}_{i=1}^n$ are i.i.d. $d$-dimensional random vectors with mean zero and covariance $\Sigma_v$. When $d \geq 34$, it holds that
\[
\inf_{\tilde{\Sigma}} \max_{\|v\|_2 = 1} P \left( \|\tilde{\Sigma} - \Sigma_v\|_{\text{op}} \geq \frac{1}{48} \sqrt{\frac{6d}{n}} \right) \geq \frac{1}{3}.
\]

We have several comments for the proposed shrinkage sample covariance. First of all, to understand its behavior, we compare it with the sample covariance $\Sigma_n$ under the Gaussian data setting. Suppose $\{x_i\}_{i=1}^n$ are i.i.d. Gaussian vectors with $\mathbb{E}x_i = 0$ and $\mathbb{E}x_i x_i^T = I$. Then we have $\|x_i\|_4^4 = \sum_{j=1}^d x_{ij}^4 = O_P(d)$. On the other hand, $\sup_{v \in \mathcal{S}^{d-1}} \mathbb{E}(v^T x_i)^4 = 3$ and thus $\tau^4 \asymp n/\log d$. Therefore, depending on whether $n$ is greater or smaller than $d \log d$, one would expect either all the vectors or none of them to be shrunk, and the shrinkage sample covariance to be either equal to the sample covariance or close to a multiple of it, with scaling of order $\tau/d^{1/4}$. In other words, in the low-dimensional regime, there is no need to shrink Gaussian random vectors for covariance estimation, while in the high-dimensional regime, we need to shrink the sample covariance matrix towards zero.

Note that $\tilde{\Sigma}_n(\tau)$ outperforms the sample covariance $\Sigma_n$ even with Gaussian samples if the dimension is high. According to Theorem 5.39 in [123], $\|\Sigma_n - \Sigma\|_{\text{op}} = O_P(\sqrt{d/n} \vee (d/n))$. When $d/n$ is large, the $d/n$ term will dominate $\sqrt{d/n}$, thus
delivering statistical error of order $d/n$ for Gaussian sample covariance. However, our shrinkage sample covariance always attains order $\sqrt{d \log d/n}$ regardless of relationship between the dimension and the sample size. Theorem 4.3.2 shows that the minimax optimal rate of estimating the covariance in terms of the operator norm is $\sqrt{d/n}$. Hence, the shrinkage sample covariance is minimax optimal up to a logarithmic factor, whereas traditional sample covariance is not in high dimensions. Therefore, shrinkage overcomes not only heavy-tailed corruption, but also curse of dimensionality. In Section 4.4.5 we conduct simulations to further illustrate this point.

If we are concerned with error bound in terms of the elementwise max norm, we need to naturally apply elementwise truncation rather than the $\ell_4$-norm shrinkage. Define $\mathbf{x}_i$ such that $\mathbf{x}_{ij} = \text{sgn}(x_{ij})(|x_{ij}| \wedge \tau)$ for $1 \leq j \leq d_1$ and $\mathbf{\tilde{\Sigma}}_n(\tau) = n^{-1} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^\top$. It is not hard to derive that with $\tau \asymp (n / \log d)^{1/4}$, $\|\mathbf{\tilde{\Sigma}}_n(\tau) - \Sigma\|_{\text{max}} = O_P(\sqrt{\log d/n})$ as in [36]. Note that with this choice of $\tau$, $\mathbf{\tilde{\Sigma}}_n(\tau)$ will equal to $\mathbf{\Sigma}_n$ with high probability when data are Gaussian. To see this, again suppose $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are i.i.d. Gaussian zero mean and identity covariance. The maximum of $|x_{ij}|$ would be sharply concentrated around a term of order $\sqrt{\log(nd)}$. Therefore, if $n \geq \log^2(n d) \log(d)$, with high probability the threshold would not be reached, one would have $\mathbf{x}_i = \mathbf{x}_i$ for all $i$ and $\mathbf{\tilde{\Sigma}}_n = \mathbf{\Sigma}_n$. Further regularization can be applied to $\mathbf{\Sigma}_n$ if the true covariance is sparse. See, for example, [86, 8, 72, 19, 18, 39], among others.

Finally, in comparison with the robust covariance matrix estimator given in [36], our estimator here is positive semidefinite and is very simple to implement. Our concentration inequality here is on the operator norm, while their result is on the element-wise max norm.
4.4 Simulation study

In this section, we first compare the numerical performance of the robust procedure and standard procedure in linear regression, matrix compressed sensing, matrix completion and multi-task regression. For linear regression, note that the objective (4.2.1) is just the LASSO problem. We compare our approach with some other alternatives when the noise follows the t-distribution. For each setup of the other three problems, we investigate three noise settings: log-normal noise, truncated Cauchy noise and Gaussian noise. They represent heavy-tailed asymmetric distributions, heavy-tailed symmetric distributions and light-tailed symmetric distributions. Besides, in the simulations we assume the design matrices are known to be sub-Gaussian. In practice, nevertheless, it is not hard to check whether a random variable is sub-Gaussian. Recall that the $\psi_2$-norm of a random variable $X$ is defined as

$$\|X\|_{\psi_2} := \sup_{k \in \mathbb{N}} \frac{(\mathbb{E}|X|^k)^{\frac{1}{k}}}{\sqrt{k}}.$$  

Given a real dataset, we can empirically evaluate $(\mathbb{E}|X|^k)^{\frac{1}{k}}/\sqrt{k}$ for $k = 1, 2, \ldots$ and see how this ratio changes as $k$ increases. If we observe significant blow-up, we will believe that $X$ has a heavier tail than sub-Gaussian random variables, and we need to apply truncation or shrinkage on the design accordingly. Furthermore, the tuning of the truncation level in our procedure is adaptive to the tail of the data distribution. The truncation or shrinkage level $\tau$ is usually tuned by cross validation to deliver best performance. The choice of $\tau$ will automatically adapt to the tail of the data: $\tau$ will be large if the tail is light, while $\tau$ will be small if the tail is heavy.

All the robust procedures proposed in our work are very easy to implement: we only need to truncate or shrink the data appropriately, and then apply the standard procedure to the transformed data. As for choosing the regularization and threshold parameters, we refer to the rates in [90], [91] and our theories, and we only tune
the constants before the rate to optimize the numerical performance. Thereby we can highlight the potential of our robust methodologies and verify the established statistical rates and optimal order of threshold values. In practice, however, these parameters need to be chosen carefully by cross validation to deliver good numerical results.

The main message is that with appropriate regularization and truncation or shrinkage, our robust procedure outperforms the standard one under the setting with bounded moment noise, and it performs equally well as the standard procedure under the Gaussian noise. The simulations are based on 100 independent Monte Carlo replications. Besides presenting the numerical results, we also elaborate the optimization algorithms we exploited, which might be of interest to readers concerned with implementations.

In addition, we compare the numerical performance of the regular sample covariance and shrinkage sample covariance as proposed in (4.3.1) in estimating the true covariance. We choose $d/n = 0.2, 0.5, 1$ and for each ratio, we let $n = 100, 200, ..., 500$. Simulation results show superiority of the shrinkage sample covariance over the regular sample covariance under both Gaussian noise and $t_3$ noise. Therefore, the shrinkage not only overcomes the heavy-tailed corruption, but also mitigates the curse of high dimensionality.

4.4.1 Linear model

In this section we focus on the high-dimensional sparse linear regression with heavy-tailed noise. Besides comparing LASSO based on truncated response and standard LASSO, we also investigate Huber loss minimization with $\ell_1$-regularization and the median of means approach. The result indicates that our approach and the Huber approach have nearly the same performance and they significantly outperform the other two. For clarity, below we elucidate the specific procedures of these four methods.
and the scaling of tuning parameters. Consider the following linear model:

\[ y_i = x_i^\top \theta^* + \epsilon_i, \quad i = 1, \ldots, N, \]

where \( x_i \in \mathbb{R}^{100} \) and \( x_i \sim N(0, .25 \cdot I) \), \( \theta^* = (1, 1, 1, 0, \ldots, 0)^\top \), \( \epsilon_i \sim t_{2.1} \).

1. LASSO with truncated response: We solve

\[
\hat{\theta} = \arg\min_{\theta \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^n (\tilde{y}_i - x_i^\top \theta)^2 + \lambda \|\theta\|_1,
\]

where \( \tilde{y}_i := \text{sgn}(y_i) \min(|y_i|, \tau) \), \( \lambda \asymp \sqrt{\log d/n} \) and \( \tau \asymp \sqrt{n/\log d} \).

2. Huber loss minimization with \( \ell_1 \)-regularization: Define the Huber loss as follows.

\[
h_{\tau}(x) := \begin{cases} 
-\tau(x + \tau) + \frac{1}{2} \tau^2 & x < -\tau \\
\frac{1}{2} x^2 & |x| \leq \tau \\
\tau(x - \tau) + \frac{1}{2} \tau^2 & x > \tau 
\end{cases}
\]

We solve the following optimization problem:

\[
\hat{\theta} = \arg\min_{\theta \in \mathbb{R}^{100}} \frac{1}{n} \sum_{i=1}^n h_{\tau}(y_i - \theta^\top x_i) + \lambda \|\theta\|_1,
\]

where \( \tau \asymp \sqrt{n/\log d} \) and \( \lambda \asymp \sqrt{\log d/n} \).

3. Median of means: Randomly and evenly divide the whole dataset into \( K \) subsets \( \{D_k\}_{k=1}^K \) and calculate a standard LASSO estimator \( \hat{\theta}^{(k)} \) on each sub-dataset \( D_k \). Then we take the geometric median of \( \{\hat{\theta}^{(k)}\}_{k=1}^K \) as the final estimator, which is mathematically defined as

\[
\hat{\theta} := \arg\min_{\theta \in \mathbb{R}^d} \sum_{k=1}^K \|\theta - \hat{\theta}^{(k)}\|_2.
\]
$K$ is chosen to be of order $\log n$.

4. Standard LASSO:

$$\hat{\theta} := \arg\min_{\theta \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} (y_i - x_i^\top \beta)^2 + \lambda \|\beta\|_1,$$

where $\lambda \asymp \sqrt{\log d/n}$.

We let $N$ grow geometrically from 100 to 1,000 and compare $\log(\|\hat{\theta} - \theta^*\|_2)$ using four different approaches based on 500 independent Monte Carlo simulations. The constants before the rates of the tuning parameters are tuned for optimal performance.

From the figure above, we can see that the truncation approach and the Huber approach have nearly the same performance and they significantly outperform the median of means and standard LASSO. Note that the median of means and standard method have exactly the same performance because the optimal number of sample splits $K$ turns out to be always 1 in this simulation study.

4.4.2 Compressed sensing

We first specify the parameters in the true model: $Y = \langle X_i, \Theta^* \rangle + \epsilon_i$. In the simulation, we set $d_1 = d_2 = d$ and construct $\Theta^*$ to be $\sum_{i=1}^{5} v_i v_i^\top$, where $v_i$ is the $i$th top
Figure 4.1: Statistical errors of $\ln \| \hat{\Theta} - \Theta^* \|_F$ v.s. logarithmic sample size $\ln N$ for different dimensions $d$ in matrix compressed sensing.

The eigenvector of the sample covariance of 100 i.i.d. centered Gaussian random vectors with covariance $I_d$, so that $\text{rank}(\Theta^*) = 5$ and $\|\Theta^*\|_F \approx \sqrt{5}$. The design matrix $X_i$ has i.i.d. standard Gaussian entries. The noise distributions are characterized as follows:

- **Log-normal**: $\epsilon_i \overset{d}{=} (Z - \mathbb{E}Z)/50$, where $\ln Z \sim N(0, \sigma^2)$ and $\sigma^2 = 6.25$;

- **Truncated Cauchy**: $\epsilon_i \overset{d}{=} \min(Z, 10^3)/10$, where $Z$ follows Cauchy distribution;

- **Gaussian**: $\epsilon_i \sim N(0, \sigma^2)$, where $\sigma^2 = 0.25$.

The constants above are chosen to ensure appropriate signal-to-noise ratio for better presentation. We present the numerical results in Figure 4.1. As we can observe from the plots, the robust estimator has much smaller statistical error than the standard estimator under the heavy-tailed noise, i.e., the log-normal and truncated Cauchy noise. When $d = 40$ or 60, robust procedures deliver sharper estimation as the sample size increases, while the standard procedure does not necessarily do so under the heavy-tailed noise. Under the setting of Gaussian noise, the robust estimator has nearly the same statistical performance as the standard one, which shows that it does not hurt to use the robust procedure under the light-tailed noise setting.
As for the implementation, we exploit the contractive Peaceman-Rachford splitting method (PRSM) to solve the compressed sensing problem. Here we briefly introduce the general scheme of the contractive PRSM for clarity. The contractive PRSM is for minimizing the summation of two convex functions under linear constraint:

$$\min_{x \in \mathbb{R}^{p_1}, y \in \mathbb{R}^{p_2}} \ f_1(x) + f_2(y),$$

subject to \( C_1x + C_2y - c = 0, \) \hfill (4.4.1)

where \( C_1 \in \mathbb{R}^{p_3 \times p_1}, \ C_2 \in \mathbb{R}^{p_3 \times p_2} \) and \( c \in \mathbb{R}^{p_3}. \) The general iteration scheme of the contractive PRSM is

$$\begin{align*}
\mathbf{x}^{(k+1)} &= \arg\min_x \left\{ f_1(x) - (\rho^{(k)})^\top (C_1x + C_2y^{(k)} - c) \\
&\quad + \frac{\beta}{2} \|C_1x + C_2y^{(k)} - c\|_2^2 \right\}, \\
\rho^{(k+\frac{1}{2})} &= \rho^{(k)} - \alpha \beta (C_1x^{(k+1)} + C_2y^{(k)} - c), \\
\mathbf{y}^{(k+1)} &= \arg\min_y \left\{ f_2(y) - (\rho^{(k+\frac{1}{2})})^\top (C_1x^{(k+1)} + C_2y - c) \\
&\quad + \frac{\beta}{2} \|C_1x^{(k+1)} + C_2y - c\|_2^2 \right\}, \\
\rho^{(k+1)} &= \rho^{(k+\frac{1}{2})} - \alpha \beta (C_1x^{(k+1)} + C_2y^{(k+1)} - c),
\end{align*}$$

where \( \rho \in \mathbb{R}^{p_3} \) is the Lagrangian multiplier, \( \beta \) is the penalty parameter and \( \alpha \) is the relaxation factor (33). Since the parameter of interest in our work is \( \Theta^* \), now we use \( \theta_x \in \mathbb{R}^{d_1 d_2} \) and \( \theta_y \in \mathbb{R}^{d_1 d_2} \) to replace \( x \) and \( y \) respectively in (4.4.2). By substituting

$$f_1(\theta_x) = \frac{1}{N} \sum_{i=1}^{N} (Y_i - \langle \text{vec}(X_i), \theta_x \rangle)^2, \quad f_2(\theta_y) = \lambda \|\text{mat}(\theta_y)\|_\star, \quad C_1 = I, \quad C_2 = -I \quad \text{and} \quad c = 0$$

into (4.4.2), we can obtain the PRSM algorithm for the compressed sensing problem. Let \( X \) be a \( N \)-by-\( d_1 d_2 \) matrix whose rows are i.i.d. random designs \( \{\text{vec}(X_i)\}_{i=1}^{N} \) and
\( Y \) be the \( N \)-dimensional response vector. Then we have the following iteration scheme specifically for compressed sensing.

\[
\begin{cases}
\theta_x^{(k+1)} = (2X^T X/N + \beta \cdot I)^{-1}(\beta \cdot \theta_y^{(k)} + \rho^{(k)} + 2X^T Y/N), \\
\rho^{(k+1/2)} = \rho^{(k)} - \alpha \beta (\theta_x^{(k+1)} - \theta_y^{(k)}), \\
\theta_y^{(k+1)} = \text{vec}(S_{\lambda/\beta}(\text{mat}(\theta_x - \rho^{(k+1/2)}/\beta))), \\
\rho^{(k+1)} = \rho^{(k+1/2)} - \alpha \beta (\theta_x^{(k+1)} - \theta_y^{(k+1)}),
\end{cases}
\] (4.4.3)

where we choose \( \alpha = 0.9 \) and \( \beta = 1 \) according to [33] and [55], \( \rho \in \mathbb{R}^{d_1 \times d_2} \) is the Lagrangian multiplier and \( S_\tau(z) \) is the singular value soft-thresholding function for matrix version of \( z \in \mathbb{R}^{d_1 \times d_2} \). To be more specific, let \( Z = \text{mat}(z) \in \mathbb{R}^{d_1 \times d_2} \) and \( Z = U\Lambda V^\top = U \diag(\lambda_1, \ldots, \lambda_r) V^\top \) be its singular value decomposition. Then \( S_\tau(z) = \text{vec}(U \diag((\lambda_1 - \tau)_+, (\lambda_2 - \tau)_+, \ldots, (\lambda_r - \tau)_+) V^\top) \), where \((x)_+ = \max(x, 0)\).

The algorithm stops if \( \|\theta_x - \theta_y\|_2 \) is smaller than some predetermined threshold, and returns \( \text{mat}(\theta_y) \) as the final estimator of \( \Theta^* \).

### 4.4.3 Matrix completion

We again set \( d_1 = d_2 = d \) and construct \( \Theta^* \) as \( \sum_{i=1}^{5} v_i v_i^\top/\sqrt{5} \), where \( v_i \) is the \( i \)th top eigenvector of the sample covariance of 100 i.i.d. centered Gaussian random vectors with identity covariance. Each design matrix \( X_i \) takes the singleton form, which is uniformly sampled from \( \{e_j e_k^\top\}_{1 \leq j, k \leq d} \). The noise distributions are

- Log-normal: \( \epsilon_i \overset{d}{=} (Z - EZ)/250 \), where \( \ln Z \sim \mathcal{N}(0, \sigma^2) \) and \( \sigma^2 = 9 \);
- Truncated Cauchy: \( \epsilon_i \overset{d}{=} \min(Z, 10^3)/16 \), where \( Z \) follows Cauchy distribution;
- Gaussian: \( \epsilon_i \sim \mathcal{N}(0, \sigma^2) \), where \( \sigma^2 = 0.25 \).

Again, the constants above are set for an appropriate signal-to-noise ratio for better presentation. We present the numerical results in Figure 4.2. Analogous to the
matrix compressed sensing, we can observe from the figure that compared with the standard procedure, the robust procedure has significantly smaller statistical error in estimating $\Theta^*$ under the log-normal and truncated Cauchy noise. Under Gaussian noise, the robust procedure has nearly the same statistical performance as the standard procedure.

To solve the matrix completion problem in (4.2.7), we adapt the ADMM method inspired by [48]. They propose to recover the matrix by minimizing the square loss plus both nuclear norm and matrix max-norm penalizations under the entrywise max-norm constraint. By simply setting the penalization coefficient for the matrix max-norm to be zero, we can apply their algorithm to our problem. Let $L, R, W \in \mathbb{R}^{(d_1+d_2) \times (d_1+d_2)}$, which are variables in our algorithm. Define $\Theta^n, \Theta^s \in \mathbb{R}^{d_1 \times d_2}$ such that $\Theta^n_{ij} = \sum_{t=1}^{N} 1_{\{X_t = e_i e_j^\top\}}$ and $\Theta^s_{ij} = \sum_{t=1}^{N} Y_t 1_{\{X_t = e_i e_j^\top\}}$, so $\Theta^n$ and $\Theta^s$ are constants given the data. Below we present the iteration scheme of our algorithm to solve the matrix completion problem (4.2.7).
\[
L^{(k+1)} = \Pi_{S_{d_1+d_2}^+} \{ R^{(k)} - \rho^{-1}(W^{(k)} + \lambda_N I) \},
\]
\[
C = \begin{pmatrix}
C_{11} & C_{12} \\
C_{21} & C_{22}
\end{pmatrix} = L^{(k+1)} + W^{(k)}/\rho,
\]
\[
R_{ij}^{(k+1)} = \Pi_{[-R,R]} \{(\rho C_{ij}^{12} + 2\Theta_{ij}^*/N)/(\rho + 2\Theta_{ij}^n/N)\}, \quad 1 \leq i \leq d_1, 1 \leq j \leq d_2 \quad (4.4.4)
\]
\[
R^{(k+1)} = \begin{pmatrix}
C_{11} & R_{12} \\
(R_{12})^\top & C_{22}
\end{pmatrix},
\]
\[
W^{(k+1)} = W^{(k)} + \gamma \rho (L^{(k+1)} - R^{(k+1)}).
\]

In the algorithm above, \(\Pi_{S_{d_1+d_2}^+}(\cdot)\) is the projection operator onto the space of positive semidefinite matrices \(S_{d_1+d_2}^+\), \(\rho\) is the penalization parameter which we set to be 0.1 in our simulation and \(\gamma\) is the step length which is typically set to be 1.618 according to [48]. We omit the stopping criteria for this algorithm here since it is complex. Readers who are interested in the implementation details as well as the derivation of this algorithm can refer to [48]. Once the stopping criteria is satisfied, the algorithm returns \(R^{12}\) as the final estimator of \(\Theta^*\).

### 4.4.4 Multi-task regression

We choose \(d_1 = d_2 = d\) and construct \(\Theta^*\) as in Section 4.4.2. The design vectors \(\{x_i\}_{i=1}^n\) are i.i.d. \(\mathcal{N}(0, I_d)\). The noise distributions are characterized as follows:

- **Log-normal**: \(\epsilon_i \sim (Z - \mathbb{E}Z)/50, \text{ where } \ln Z \sim \mathcal{N}(0, \sigma^2) \text{ and } \sigma^2 = 4;\)
- **Truncated Cauchy**: \(\epsilon_i \sim \text{min}(Z, 10^4)/10, \text{ where } Z \text{ follows Cauchy distribution};\)
- **Gaussian**: \(\epsilon_i \sim \mathcal{N}(0, \sigma^2), \text{ where } \sigma^2 = 0.25.\)
Figure 4.3: Statistical errors of \( \log \| \hat{\Theta} - \Theta^* \|_F \) v.s. logarithmic sample size \( \log N \) for different dimensions \( d \) in multi-task regression.

We present the numerical results in Figure 4.3. Similar to the two examples before, the robust procedure has sharper accuracy in estimating \( \Theta^* \) under two heavy-tailed noises, while maintains competitive under Gaussian noise.

For multi-task regression, we exploit the contractive PRSM method again. Let \( X \) be the \( n \)-by-\( d_1 \) design matrix and \( Y \) be the \( n \)-by-\( d_2 \) response matrix. Following the general iteration scheme (4.4.2), we can develop the iteration steps for the multi-task regression.

\[
\begin{align*}
\Theta^{(k+1)}_x &= (2X^T X / n + \beta \cdot I)^{-1} (\beta \cdot \Theta^{(k)}_y + \rho^{(k)} + 2X^T Y / n), \\
\rho^{(k+\frac{1}{2})} &= \rho^{(k)} - \alpha \beta (\Theta^{(k+1)}_x - \Theta^{(k)}_y), \\
\Theta^{(k+1)}_y &= S_{\lambda/\beta} (\Theta_x - \rho^{(k+\frac{1}{2})} / \beta), \\
\rho^{(k+1)} &= \rho^{(k+\frac{1}{2})} - \alpha \beta (\Theta^{(k+1)}_x - \Theta^{(k+1)}_y),
\end{align*}
\]

(4.4.5)

where \( \Theta_x, \Theta_y, \rho \in \mathbb{R}^{d_1 \times d_2} \) and \( S_r(\cdot) \) is the same singular value soft thresholding function as in (4.4.3). Analogous to the compressed sensing, we choose \( \alpha = 0.9 \) and \( \beta = 1 \). As long as \( \| \Theta_x - \Theta_y \|_F \) is smaller than some predetermined threshold, the iteration stops and the algorithm returns \( \Theta_y \) as the final estimator of \( \Theta^* \).
4.4.5 Covariance estimation

In this subsection, we investigate the statistical error of the shrinkage sample covariance $\tilde{\Sigma}_n(\tau)$ proposed in Section 4.3 compared with the regular sample covariance $\Sigma_n$. We consider two common distributions: Gaussian and Student’s $t_3$ random samples. The dimension is set to be proportional to sample size, i.e., $d/n = \alpha$ with $\alpha$ being 0.2, 0.5, 1. $n$ will range from 100 to 500 for each case. Regardless of how large the dimension $d$ is, the true covariance $\Sigma$ is always set to be a diagonal matrix with the first diagonal element equal to 4 and all the other diagonal elements equal to 1. The statistical errors are measured in terms of the operator norm, and our simulation is based on 1,000 independent Monte Carlo replications. We present the results in Figure 4.4.

As we can see, for Gaussian samples, as long as we fix $d/n$, the statistical error of both $\Sigma_n$ and $\tilde{\Sigma}_n(\tau)$ does not change, which is consistent with Theorem 5.39 in [123] and Theorem 4.3.1 in our paper. Also, the higher the dimension is, the more significant the superiority of $\tilde{\Sigma}_n(\tau)$ is over $\Sigma_n$. This validates our remark after The-
orem 4.3.1 that the shrinkage ameliorates the impact of dimensionality. Even for Gaussian data, shrinkage is meaningful and provides significant improvement. For $t_3$ distribution, since it is heavy-tailed, the regular sample covariance does not maintain constant statistical error for a fixed $d/n$; instead the error increases as the sample size increases. In contrast, our shrinkage sample covariance still retains stable performance and enjoys much higher accuracy. This strongly supports the sharp statistical error rate we derived for $\tilde{\Sigma}_n(\tau)$ in Theorem.
Chapter 5

Taming the heavy-tailed features

5.1 Problem setup

In this paper we consider the corrupted generalized linear model (CGLM). Recall the definition of the standard GLM with the canonical link. Suppose we have \( n \) samples \( \{(y_i, x_i)\}_{i=1}^{n} \), where \( y_i \) is the response and \( x_i \) is the feature vector. Under the GLM with the canonical link, the response follows the distribution

\[
    f_n(y; \mathbf{X}, \boldsymbol{\beta}^*) = \prod_{i=1}^{n} f(y_i; \eta_i^*) = \prod_{i=1}^{n} \left\{ c(y_i) \exp \left[ \frac{y_i \eta_i^* - b(\eta_i^*)}{\phi} \right] \right\},
\]

(5.1.1)

where \( y = (y_1, \ldots, y_n)^\top, \mathbf{X} = (x_1, \ldots, x_n)^\top, \boldsymbol{\beta}^* \in \mathbb{R}^d \) is a regression vector, \( \eta_i^* = x_i^\top \boldsymbol{\beta}^* \) and \( \phi > 0 \) is the dispersion parameter. The negative log-likelihood corresponding to (5.1.1) is given, up to an affine transformation, by

\[
    \ell_n(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^{n} -y_i x_i^\top \boldsymbol{\beta} + b(x_i^\top \boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^{n} -y_i \eta_i + b(\eta_i) = \frac{1}{n} \sum_{i=1}^{n} \ell_i(\boldsymbol{\beta}),
\]

(5.1.2)

and the gradient and Hessian of \( \ell_n(\boldsymbol{\beta}) \) are respectively

\[
    \nabla \ell_n(\boldsymbol{\beta}) = -\frac{1}{n} \sum_{i=1}^{n} (y_i - b'(x_i^\top \boldsymbol{\beta}^*)) x_i
\]

(5.1.3)
\[ \nabla^2 \ell_n(\beta) = \frac{1}{n} \sum_{i=1}^{n} b''(x_i^\top \beta^*) x_i x_i^\top, \quad (5.1.4) \]

For ease of notations, we write the empirical hessian \( \nabla^2 \ell_n(\beta) \) as \( H_n(\beta) \) and \( \mathbb{E}(b''(x_i^\top \beta) x_i x_i^\top) \) as \( H(\beta) \).

Now we consider further an extra random corruption on the response \( y_i \). Suppose we can only observe corrupted responses \( z_i = y_i + \epsilon_i \), where \( \epsilon_i \) is a random noise.

We emphasize here that the introduction of \( \epsilon_i \) significantly improves the flexibility of the original GLM. The new model now embraces many more real-world problems with complex structures, e.g., the linear regression model with heavy-tailed noise, the logistic regression with mislabeled samples and so forth.

To handle the heavy-tailed features and corruptions, we propose to shrink or clip the data \( \{(z_i, x_i)\}_{i=1}^{n} \) first and feed them to the log-likelihood (5.1.2) to calculate MLE. More rigorously, our negative log-likelihood is evaluated on the shrunk data \( \{\tilde{z}_i, \tilde{x}_i\}_{i=1}^{n} \) as follows.

\[ \tilde{\ell}_n(\beta) := \frac{1}{n} \sum_{i=1}^{n} \tilde{z}_i \tilde{x}_i^\top \beta + b(\tilde{x}_i^\top \beta). \quad (5.1.5) \]

We denote the hessian matrix of \( \tilde{\ell}_n(\beta) \) by \( \tilde{H}_n(\beta) \) and its population version \( \mathbb{E}\tilde{H}_n(\beta) \) by \( \tilde{H}(\beta) \). In the next section, we will elucidate the specific shrinkage and clipping methods in both the low-dimensional and high-dimensional regimes and explicitly derive the statistical error rates of the MLE based on \( \tilde{\ell}_n(\beta) \).

### 5.2 Main results

From now on in this chapter, we refer to some quantities as **constants** if they are independent of the sample size \( n \), the dimension \( d \), and the sparsity \( s \) of \( \beta^* \) in the high-dimensional regime.
5.2.1 Low-dimensional regime

The standard MLE estimator is defined as \( \hat{\beta} := \arg\min_{\beta \in \mathbb{R}^d} \ell_n(\beta) \), where \( \ell_n(\cdot) \) is characterized as in (5.1.2). It is widely known that under genuine GLM with bounded features, \( \hat{\beta} \) enjoys \( \sqrt{d/n} \)-consistency to the true parameter \( \beta^* \) in terms of the Euclidean norm. However, when the feature vectors have only bounded moments, there is no guarantee of \( \sqrt{d/n} \)-consistency any more, let alone further perturbation \( \epsilon_i \) on the response. To reduce the disruption due to heavy-tailed data, we apply the following \( \ell_4 \)-norm shrinkage to feature vectors:

\[
\tilde{x}_i := \frac{\min(\|x_i\|_4, \tau_1)}{\|x_i\|_4} \cdot x_i
\]

and also clipping to the response:

\[
\tilde{z}_i := \min(|z_i|, \tau_2)
\]

before MLE calculation, where \( \tau_1 \) and \( \tau_2 \) are predetermined thresholds. Clipping on the response is natural; when \( |z_i| \) is abnormally large, clipping reduces its magnitude to prevent potential wild corruptions by \( \epsilon_i \). Here we explain more on why we shrink features in terms of the \( \ell_4 \)-norm rather than other norms. The \( \ell_4 \)-norm shrinkage has been proven to be successful in low-dimensional covariance estimation in [46]. Theorem 6 therein shows that when data have only bounded fourth moment, the \( \ell_4 \)-norm shrinkage sample covariance enjoys operator-norm convergence rate of order \( O_p(\sqrt{d \log d/n}) \) to the population covariance matrix. This inspires us to apply similar \( \ell_4 \)-norm shrinkage to heavy-tailed features to ensure that the empirical hessian \( \tilde{H}_n(\beta) \) is close to its population version \( H(\beta) \) and thus well-behaved. We shall rigorize this argument in Lemma [5.2.1] later.
After data shrinkage and clipping, we minimize the negative log-likelihood based on the new data \( \{\tilde{z}_i, \tilde{x}_i\}_{i=1}^n \) with respect to \( \beta \) to derive the M-estimator. Specifically, define
\[
\tilde{\ell}_n(\beta) := \frac{1}{n} \sum_{i=1}^n -\tilde{z}_i \tilde{x}_i^\top \beta + b(\tilde{x}_i^\top \beta)
\]
and we choose \( \tilde{\beta} := \arg\min_{\beta \in \mathbb{R}^d} \tilde{\ell}_n(\beta) \) to estimate \( \beta^* \). In the sequel, we will show that \( \|\tilde{\beta} - \beta^*\|_2 = O_P(\sqrt{d \log n/n}) \) with exponential exception probability. One crucial step in this statistical error analysis is to establish a uniform strong convexity of \( \tilde{\ell}(\beta) \) over \( \beta \in B_2(\beta^*, r) \) up to some small tolerance term, where \( r > 0 \) is small. The following lemma rigorously justifies this point.

**Lemma 5.2.1.** Suppose the following conditions hold: (1) \( b''(x_i^\top \beta^*) \leq M < \infty \) always holds and \( \forall \, \omega > 0, \exists \, m(\omega) > 0 \) such that \( b''(\eta) \geq m(\omega) > 0 \) for \( |\eta| \leq \omega \); (2) \( \mathbb{E}x_i = 0, \lambda_{\min}(\mathbb{E}x_i x_i^\top) \geq \kappa_0 > 0 \) and \( \mathbb{E}(v^\top x_i)^4 \leq R < \infty \) for all \( v \in S^{d-1} \); (3) \( \|\beta^*\|_2 \leq L < \infty \). Choose the shrinkage threshold \( \tau_1 \asymp (n/\log n)^{-\frac{1}{4}} \). For any \( 0 < r < 1 \) and \( t > 0 \), as long as \( \sqrt{d \log n/n} \) is sufficiently small, it holds with probability at least \( 1 - 2 \exp(-t) \) that for all \( \Delta \in \mathbb{R}^d \) such that \( \|\Delta\|_2 \leq r \),
\[
\delta \tilde{\ell}_n(\beta^* + \Delta; \beta^*) \geq \kappa \|\Delta\|_2^2 - Cr^2\left(\sqrt{\frac{r}{n}} + \sqrt{\frac{d}{n}}\right),
\]
where \( \kappa \) and \( C \) are some constants.

**Remark 5.2.1.** Here we explain the conditions for Lemma 5.2.1. Condition (1) assumes that the response from the GLM has bounded variance and is non-degenerate when \( \eta \) is bounded. Note here that we do not assume a uniform lower bound of \( b''(\eta) \). \( m(\omega) \) is allowed to decay to zero as \( \omega \to \infty \). Condition (2) says that the population covariance matrix of the design vector \( x_i \) is positive definite and \( x_i \) has bounded fourth moment. Condition (3) is natural; it holds if we have \( \text{var}(x_i^\top \beta^*) < \infty \) and \( \lambda_{\min}(\mathbb{E}x_i x_i^\top) \geq \kappa_0 > 0 \). Note that the ordinary least square (OLS) estimator has been shown to enjoy consistency under similar bounded fourth moment conditions.
Theorem 1 later establishes similar results for CGLM that embraces a much broader family of loss functions.

Remark 5.2.2. When deriving the statistical rate of \( \tilde{\beta} \) in Theorem 1 later, we will let the radius of the local neighborhood \( r \) decay to zero so that the tolerance term \( r^2(\sqrt{t/n} + \sqrt{d/n}) \) is negligible.

Given Lemma 5.2.1 we are now ready to derive the statistical error rate \( \| \tilde{\beta} - \beta^* \|_2 \).

Theorem 5.2.1. Suppose the conditions of Lemma 5.2.1 hold. We further assume that (1) \( \mathbb{E} z_i^4 \leq M_1 < \infty \); (2) \( \| \mathbb{E}[\epsilon_i x_i] \|_2 \leq M_2 \sqrt{d/n} \) for some constant \( M_2 \). Choose \( \tau_1, \tau_2 \approx (n / \log n)^{\frac{1}{4}} \). There exists a constant \( C > 0 \) such that for any \( \xi > 1 \),

\[
P\left( \| \tilde{\beta} - \beta^* \|_2 \geq C\xi \sqrt{\frac{d\log n}{n}} \right) \leq 3n^{1-\xi}.
\]

Remark 5.2.3. Condition 1 requires merely bounded fourth moments of the response from CGLM. Condition 2 requires the additional corruption to be nearly uncorrelated with the design, which is trivially satisfied if \( \mathbb{E}(\epsilon_i | x_i) = 0 \).

Sometimes the covariance between \( \epsilon_i \) and \( x_i \) does not vanish as \( n \) and \( d \) grow. For example, in binary logistic regression with a random corruption \( \epsilon_i \), suppose

\[
P(\epsilon_i = -1 | y_i = 1) = p, \quad P(\epsilon_i = 0 | y_i = 1) = 1 - p, \quad P(\epsilon_i = 1 | y_i = 0) = p, \quad P(\epsilon_i = 0 | y_i = 0) = 1 - p.
\]

where \( p < 0.5 \). In other words, we flip the genuine label \( y_i \) with probability \( p \). Then we have

\[
\mathbb{E}\epsilon_i x_i = \mathbb{E}(\epsilon_i x_i \cdot 1_{\{y_i=0\}}) + \mathbb{E}(\epsilon_i x_i \cdot 1_{\{y_i=1\}}) = p\mathbb{E}(x_i(1_{\{y_i=0\}} - 1_{\{y_i=1\}})) = 2p\mathbb{E}(x_i \cdot 1_{\{y_i=0\}}).
\]

The last equality holds because we assume \( \mathbb{E}x_i = 0 \). Therefore, \( \mathbb{E}\epsilon_i x_i \propto p \) and if \( p \) does not decay, neither will \( \mathbb{E}\epsilon_i x_i \). [89] solves this noisy label problem through
minimizing weighted negative log-likelihood

\[ \beta^w := \argmin_{\beta \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \ell^w(x_i, z_i; \beta) \]

\[ = \argmin_{\beta \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \frac{(1-p)\ell(x_i, z_i; \beta) - p \cdot \ell(x_i, 1-z_i; \beta)}{1-2p}. \]  

(5.2.2)

Lemma 1 therein shows that \( \mathbb{E}_{\epsilon_i} \ell^w(x_i, z_i) = \ell(x_i, y_i) \). This implies that when the sample size is sufficiently large, minimizing the weighted negative log-likelihood above is similar to minimizing the negative log-likelihood with true labels. In the presence of heavy-tailed features, however, we propose to replace \( x_i \) with the \( \ell_4 \)-norm shrunk feature \( \tilde{x}_i \), i.e., we recruit

\[ \beta^w := \argmin_{\beta \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \ell^w(\tilde{x}_i, z_i; \beta) = \frac{1}{n} \sum_{i=1}^{n} \frac{(1-p)\ell(\tilde{x}_i, z_i; \beta) - p \cdot \ell(\tilde{x}_i, 1-z_i; \beta)}{1-2p}. \]

(5.2.3)

to estimate the regression vector \( \beta^* \). The following corollary establishes the statistical error rate of \( \beta^w \) with exponential deviation bound.

**Corollary 5.2.1.** Under the logistic regression with a random corruption \( \epsilon_i \) satisfying (5.2.1), choose \( \tau_1 \approx (n/\log n)^{\frac{1}{4}} \). Under the conditions of Lemma 5.2.1, it holds for some constant \( C \) and any \( \xi > 1 \) such that as long as \( \sqrt{d \log d/n} \) is sufficiently small,

\[ \mathbb{P}(\|\beta^w - \beta^*\|_2 \geq C\xi \sqrt{\frac{d \log n}{n}}) \leq 2n^{1-\xi}. \]

**Remark 5.2.4.** Here we do not need to truncate the response by \( \tau_2 \), because in logistic regression the response is always bounded.

### 5.2.2 High-dimensional regime

In this section, we consider the regime where the dimension \( d \) grows much faster than the sample size \( n \). Recall that the standard \( \ell_1 \)-regularized MLE of the regression
vector $\mathbf{\beta}^*$ under the GLM is

$$\hat{\mathbf{\beta}} := \arg\min_{\mathbf{\beta} \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \left( -y_i \mathbf{x}_i^T \mathbf{\beta} + b(\mathbf{x}_i^T \mathbf{\beta}) \right) + \lambda \| \mathbf{\beta} \|_1,$$

(5.2.4)

where $(y_i, \mathbf{x}_i)$ comes from the genuine GLM and $\lambda > 0$ is the tuning parameter. [93] shows that $\|\hat{\mathbf{\beta}} - \mathbf{\beta}^*\|_2 = O_p(\sqrt{s \log d/n})$ under GLM when $\{\mathbf{x}_i\}_{i=1}^n$ are sub-Gaussian. However, in the presence of heavy-tailed features $\mathbf{x}_i$ and corruptions $\epsilon_i$, the statistical accuracy of $\hat{\mathbf{\beta}}$ might deteriorate if we directly evaluate the log-likelihood (5.2.4) on $\{(z_i, \mathbf{x}_i)\}_{i=1}^n$. In this section, we aim to develop a robust $\ell_1$-regularized MLE for the regression vector $\mathbf{\beta}^*$. Define the clipped feature vector $\tilde{\mathbf{x}}_i$ such that for any $j \in [d]$,

$$\tilde{x}_{ij} := \min(|x_{ij}|, \tau_1)x_{ij}/|x_{ij}|$$

and clipped response

$$\tilde{z}_i = \min(|z_i|, \tau_2)z_i/|z_i|.$$

We propose to evaluate the negative log-likelihood on the truncated data.

$$\tilde{\ell}_n(\mathbf{\beta}) = \frac{1}{n} \sum_{i=1}^{n} -\tilde{z}_i \tilde{\mathbf{x}}_i^T \mathbf{\beta} + b(\tilde{\mathbf{x}}_i^T \mathbf{\beta}).$$

Again we denote the hessian matrix of $\tilde{\ell}_n(\mathbf{\beta})$ by $\tilde{\mathbf{H}}_n(\mathbf{\beta})$ and $\mathbb{E}(\tilde{\mathbf{H}}_n(\mathbf{\beta}))$ by $\tilde{\mathbf{H}}(\mathbf{\beta})$. We study the following $\mathcal{M}$-estimator as the robust estimator of $\mathbf{\beta}^*$.

$$\tilde{\mathbf{\beta}} = \arg\min_{\mathbf{\beta} \in \mathbb{R}^d} \tilde{\ell}_n(\mathbf{\beta}) + \lambda \| \mathbf{\beta} \|_1.$$

For $\mathcal{S} \subset [d]$ and $|\mathcal{S}| = s$, define the restricted cone $C(\mathcal{S}) := \{ \mathbf{v} \in \mathbb{R}^d : \| \mathbf{v}_{\mathcal{S}^c} \|_1 \leq 3\| \mathbf{v}_\mathcal{S} \|_1 \}$. By Lemma 1 in [93], when $\lambda > 2 \left\| \nabla \tilde{\ell}_n(\mathbf{\beta}) \right\|_\infty$, $\tilde{\mathbf{\beta}} - \mathbf{\beta}^* \in C(\mathcal{S})$, which is a crucial property that gives rise to statistical consistency of $\tilde{\mathbf{\beta}}$ under high-dimensional
regimes. Therefore, in the following we first present a lemma that characterizes the order of $\|\nabla_\beta \tilde{\ell}_n(\beta^*)\|_{\max}$.

**Lemma 5.2.2.** Under the following conditions: (1) $b''(x_i^T \beta^*) \leq M < \infty$ always holds and $\forall \omega > 0$, $\exists m(\omega) > 0$ such that $b''(\eta) \geq m(\omega) > 0$ for $|\eta| \leq \omega$; (2) $E x_{ij} = 0$, $E x_{ij}^2 x_{ik}^2 \leq R < \infty$ for all $1 \leq j, k \leq d$; (3) $E z_i^4 \leq M_1$ and $E \epsilon_i^4 \leq M_1$; (4) $\|\beta^*\|_1 \leq L < \infty$; (5) $|E \epsilon_i x_{ij}| \leq M_2/\sqrt{n}$ for some universal constant $M_2 < \infty$ and all $1 \leq j \leq d$. With $\tau_1, \tau_2 \approx (n/\log d)^{\frac{1}{4}}$, it holds for some constant $C$ and all $\xi > 1$ that

$$P(\|\nabla \tilde{\ell}(\beta^*)\|_{\infty} \geq C \xi \sqrt{\frac{\log d}{n}}) \leq 2d^{1-\xi}.$$

**Remark 5.2.5.** In this lemma we show that $\|\nabla \tilde{\ell}_n(\beta^*)\|_{\infty} = O_P(\sqrt{\log d/n})$. In the sequel we will choose $\lambda \approx \sqrt{\log d/n}$ to achieve the minimax optimal rate for $\tilde{\beta}$.

Another requirement for statistical guarantee of $\tilde{\beta}$ is the restricted strong convexity (RSC) of $\tilde{\ell}_n$, which is first formulated in [93]. RSC ensures that $\tilde{\ell}_n(\beta)$ is “not too flat”, so that if $|\tilde{\ell}_n(\beta) - \tilde{\ell}_n(\beta^*)|$ is small, then $\tilde{\beta}$ and $\beta^*$ are close. In high-dimensional sparse linear regression, RSC is implied by the restricted eigenvalue (RE) condition ([9], [119], etc.), a widely studied and acknowledged condition for statistical error analysis of the Lasso estimator.

Unlike the quadratic loss in linear regression, the negative log-likelihood $\tilde{\ell}_n(\beta)$ has its hessian matrix $H_n(\beta)$ depend on $\beta$, which creates technical difficulty of verifying its RSC. Generally speaking, RSC condition does not hold uniformly over all $\beta \in \mathbb{R}^d$. This motivates us to study localized RSC (LRSC), i.e., RSC with $\beta$ constrained within a small neighborhood of $\beta^*$. This idea was first explored by [40] and [112]. Specifically, we say a loss function $L(\beta)$ satisfies LRSC($\beta^*, r, S, \kappa, \tau_L$) if for all $\Delta \in C(S) \cap B_2(0, r)$,

$$\delta L(\beta^* + \Delta; \beta^*) \geq \kappa \|\Delta\|_2^2 - \tau_L,$$

116
where \( \tau_c \) is a small tolerance term. Later we will see that this localized version of RSC suffices for establishing the statistical rate of \( \beta^* \). The following lemma verifies the LRSC of \( \hat{\ell}_n(\beta) \).

**Lemma 5.2.3.** Suppose the conditions of Lemma 5.2.2 hold. Let \( S \) be the true support of \( \beta^* \) with \( |S| = s \). Assume that for any \( v \in \mathbb{R}^d \) such that \( v \in C(S) \) and \( \|v\|_2 = 1 \), \( 0 < \kappa_0 \leq v^T \mathbb{E}(x_i x_i^T)v \leq \kappa_1 < \infty \). Set \( \tau_1 \approx (n/\log d)^{\frac{1}{4}} \). For any \( 0 < r < 1 \) and \( t > 0 \), as long as \( s^2 \log d/n \) is sufficiently small, it holds with probability at least \( 1 - 2 \exp(-t) \) that for all \( \Delta \in \mathbb{R}^d \) such that \( \|\Delta\|_2 \leq r \) and \( \Delta \in C(S) \),

\[
\delta \hat{\ell}_n(\beta^* + \Delta; \beta^*) \geq \kappa \|\Delta\|_2^2 - C_0 r^2 \left( \sqrt{\frac{t}{n}} + \sqrt{\frac{s \log d}{n}} \right),
\]

where \( \kappa \) and \( C_0 \) are some constants.

**Remark 5.2.6.** This lemma is the high-dimensional analogue of Lemma 5.2.1. Similarly, in the sequel we will let \( r \) converge to zero when analyzing the statistical rate of \( \hat{\beta} \) so that the tolerance term \( r^2 \left( \sqrt{t/n} + \sqrt{s \log d/n} \right) \) is negligible.

The lemma above together with Lemma 5.2.2 underpins the statistical guarantee of \( \hat{\beta} \) as follows.

**Theorem 5.2.2.** Under the assumptions of Lemma 5.2.2 and 5.2.3 choose \( \lambda = 2C\xi \sqrt{\log d/n} \) and \( \tau_1, \tau_2 \approx (n/\log d)^{\frac{1}{4}} \), where \( \xi \) and \( C \) are the same constants as in Lemma 5.2.2. It holds for some constant \( C_1 \) that

\[
P\left( \|\hat{\beta} - \beta^*\|_2 \geq C_1 \xi \sqrt{\frac{s \log d}{n}} \right) \leq 4d^{1-\xi}.
\]
5.3 Numerical study

5.3.1 High-dimensional sparse linear regression

We first consider the high-dimensional sparse linear model $y_i = x_i^T \beta^* + \epsilon_i$. We set $d = 1000$, $n = 100, 200, 500, 1000, 5000, 10000$ and $\beta^* = (1, 1, 1, 1, 0, \cdots, 0)^T$. Recall that in the high-dimensional regime, we propose elementwise clipping on both the heavy-tailed features and responses. In Figure 2, we compare estimation error of the $\ell_1$-regularized least squares estimators based on clipped data and original data under standard Gaussian features and $t_{4.1}$ features respectively. All feature vectors $\{x_i\}_{i=1}^n$ are i.i.d. generated and within each $x_i$, all dimensions $\{x_{ij}\}_{j=1}^d$ are i.i.d. $\{\epsilon_i\}_{i=1}^n$ are i.i.d. noises that are independent to the features and we adjust the magnitude of the noise such that $SD(\epsilon_i) = 5$ regardless of the distribution it conforms to. The clipping threshold levels $\tau_1, \tau_2$ and the regularization parameter $\lambda$ are selected by cross-validation. The plot is based on 1,000 independent Monte Carlo simulations. From Figure 2, we first observe that under both light-tailed and heavy-tailed features, the heavier tail $\epsilon_i$ has, the more data clipping improves the statistical accuracy. More
Figure 5.2: Statistical Error of MLE based on minimizing $\tilde{\ell}_n(\beta)$ with 10% mislabeled data

importantly, the benefit from data clipping is much more significant in the presence of heavy-tailed features, which justifies our conjecture and theories.

5.3.2 Logistic regression with mislabeled data

In this subsection we consider the logistic regression with mislabeled data as characterized by (5.2.1). We minimize the weighted negative log-likelihood to derive $\hat{\beta}^w$ and $\tilde{\beta}^w$ as described in (5.2.2) and (5.2.3) to estimate the regression vector $\beta^*$ and compare their performances. All the samples are independently generated and all dimensions of features are independent to each other as well. The tuning parameters $\lambda$ and $\tau_1$ are chosen based on cross validation. We investigate both the low-dimensional and high-dimensional regimes.

- In the low-dimensional regime, we set $d = 10, n = 100, 200, 500, 1000, 2000, 5000, 10000$, $\beta^* = (0.5, \cdots, 0.5, -0.5, \cdots, -0.5)$, $p = 0.1$. The left panel of Figure 5.2 compares $\|\hat{\beta}^w - \beta^*\|_2$ and $\|\tilde{\beta}^w - \beta^*\|_2$ under $t_{2.1}, t_{4.1}$ and Gaussian features. We can observe that $\tilde{\beta}^w$ significantly outperforms $\hat{\beta}^w$ under $t_{2.1}$ and $t_{4.1}$ features, and they perform equally well when features are Gaussian. This perfectly matches our intuition and supports our theories.
• In the high-dimensional regime, we apply elementwise clipping to \( x_i \) to derive \( \tilde{\beta}^w \). We set \( d = 100, n = 50, 100, 250, 500, 1000, 2500, 5000 \), \( \beta^* = (1, 1, -1, 0, \ldots, 0) \), \( p = 0.1 \). As shown in the right panel of Figure 5.2, \( \tilde{\beta}^w \) enjoys sharper statistical accuracy than \( \hat{\beta}^w \) under all the three feature distributions. The outstanding performance of \( \tilde{\beta}^w \) under the Gaussian feature scenario is particularly surprising. We conjecture that feature clipping here downsizes \( \| \nabla \tilde{\ell}_n^w(\beta^*) \|_\infty \) and thus leads to more effective regularization.

5.3.3 Experiments on MNIST

We extract deep features of all images of the digits 4 and 9 in the MNIST dataset through a pre-trained convolutional neural network, which has 0.8% testing error in recognizing 0 to 9. Readers can refer to Deep MNIST Tutorial \(^1\) by Google for the details of the architecture of the neural nets. We aim to use the extracted deep features of images to classify 4’s and 9’s with artificial mislabelling. We randomly flip the true labels with probability \( p \) and minimize the weighted negative log-likelihood \( \tilde{\ell}_n^w(\beta) = (1/n) \sum_{i=1}^n \ell_i^w(\tilde{x}_i, z_i; \beta) \) as characterized in (5.2.3) to estimate the regression vector. We repeat the procedure for 100 times and compare the average performance of the resulting MLE based on original deep features \( \{x_i\}_{i=1}^n \) and shrunk features \( \{x_i/\|x_i\|_4 \cdot \min(\|x_i\|_4, 0.7)\} \). The result is presented in Table 5.1. We discover that feature shrinkage robustifies the MLE so well such that the performance is insensitive to the mislabelling probability.

\(^1\) https://www.tensorflow.org/get_started/mnist/pros
<table>
<thead>
<tr>
<th>Mislabeled Prob</th>
<th>Original Features</th>
<th>Shrunk Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>0%</td>
<td>0.5%</td>
<td>0.8%</td>
</tr>
<tr>
<td>10%</td>
<td>1.0%</td>
<td>0.8%</td>
</tr>
<tr>
<td>20%</td>
<td>1.4%</td>
<td>0.8%</td>
</tr>
<tr>
<td>40%</td>
<td>3.1%</td>
<td>0.9%</td>
</tr>
</tbody>
</table>

Table 5.1: Testing classification error using original deep features and shrunk deep features
Chapter 6

Future research

Recent years have witnessed broad and close collaborations across statistics, operations research and computer science to conquer the challenges in the big data era. My future research will lie at the intersection of these three areas. My focus will be designing efficient and reliable statistical inference methods under modern data settings and developing statistical properties of M-estimators generated from complex optimization problems.

6.1 Distributed statistical inference via the generalized jackknife

In distributed statistical inference, the bias of local estimators is the bottleneck for effective aggregation. In Chapter 2, we debiased the local Lasso estimator, reducing its bias to be a high-order term. Hence, the averaged debiased Lasso estimator achieved a sharper statistical rate than the local Lasso estimator. Similarly, in Chapter 3, we showed that the bias of the local top eigenvectors is a high-order term, so that the aggregation step could further improve statistical efficiency. However, when the local sample size is small, or the number of data splits is large, this high-order bias will
hurt the statistical rate of the aggregated estimator. This is the reason we imposed
upper bounds on the number of data splits for our distributed procedure to achieve
full-sample efficiency.

One way to relax the upper bound on the number of data splits is through the
generalized jackknife, a classical debiasing technique that was first proposed by [108].
Simply speaking, instead of naively averaging all the local estimators, we can take
a weighted average where the weights are tuned such that the high-order bias are
cancelled in the aggregation and variance of the aggregated estimator is minimized.
We can thus remove our restriction on the number of data splits while still achieving
full-sample efficiency.

6.2 Statistical inference over data streams

In contrast to static data sets, data streams are generated continuously and of un-
bounded length. They naturally arise in telecommunications, network traffic engi-
neering, medical monitoring, high-frequency trading, etc. Under these scenarios, it is
often desirable to maintain some synopsis statistics based on recent stream history;
stale data points should be discounted or discarded in calculating these statistics.
One goal of my future research is to design novel data stream procedures for statis-
tical inference with strong statistical guarantees, high computational efficiency and
low memory cost.

To give a flavor of this problem, let \( x_t \) be the data point received at time \( t \).
Consider an empirical risk function \( \ell(\beta; \{x_t\}_{t=T-t_0+1}) \), which is only evaluated on
the most recent session from \( t = T - t_0 + 1 \) to \( T \). Define the M-estimator
\( \hat{\beta}_T := \arg\min_{\beta} \ell(\beta; \{x_t\}_{t=T-t_0+1}) \). Given a fixed length \( t_0 \), the goal is to maintain \( \{\hat{\beta}_T\}_{T=t_0+1}^\infty \)
or their well approximations without repetitively solving the empirical risk minimiza-
tion problem for each \( T \).
6.3 Non-convex statistical optimization

Non-convex optimization is ubiquitous in statistical learning problems, e.g., principle component analysis, high-dimensional sparse learning, mixture models, neural nets and so forth. Given possibly poor behavior of local optima or saddle points in the non-convex problem, people turn to their convex relaxations and study the statistical properties of the resulting M-estimator. Recent research, however, shows that many non-convex M-estimators outperform their convex-relaxed counterparts in empirical performance. Since the non-convex M-estimation is far less understood than the convex case, my future research plan is to study the statistical properties of the non-convex M-estimators and design fast algorithms to solve the corresponding optimization problems.

Many non-convex problems exhibit local strong convexity within a small neighborhood of the true parameter. This inspires a two-stage procedure that first seeks a warm initializer that enters this strongly convex region, and then leverages this local property by gradient descent for a refined statistical rate. My short-term goal is to characterize the local landscape of the non-convex problems so as to achieve sharp statistical rate through empirical risk minimization. The technical tools we built for analyzing local strong convexity in Chapter 5 are potentially powerful in achieving this goal.
Appendix A

Proofs for Chapter 2

A.1 The low-dimensional linear model

As mentioned earlier, the infinity norm bound derived in Lemma 2.3.1 can be used to do model selection, after which the selected support can be shared across all the local agents. We significantly reduce the dimension of the problem as we only need to refit the data on the selected model. The remaining challenge is to implement the divide and conquer strategy in the low dimensional setting, which is also of independent interest. Here we focus on the linear model, while the generalized linear model is covered in Appendix A.2.

In this section $d$ still stands for dimension, but in contrast with the rest of this paper in which $d \gg n$, here we consider $d < n$. More specifically, we consider the linear model (2.2.2) with $d < n$ and i.i.d sub-Gaussian noise $\{\epsilon_i\}_{i=1}^n$. It is well known that the ordinary least square (OLS) estimator of $\beta^*$ is defined as $\hat{\beta} = (X^\top X)^{-1}X^\top Y$. In the massive data setting, the communication cost of estimating and inverting covariance matrices is very high (order $O(kd^2)$). However, as pointed out by [28], this
estimator exactly coincides with the DC estimator,

\[ \hat{\beta} = \left( \sum_{j=1}^{k} X^{(j)^\top}X^{(j)} \right)^{-1} \sum_{j=1}^{k} X^{(j)^\top}Y^{(j)}. \]

In this section, we study the DC strategy to approximate \( \hat{\beta} \) with the communication cost only \( O(kd) \), which implies that we can only communicate \( d \)-dimensional vectors.

The OLS estimator based on the subsample \( D_j \) is defined as \( \hat{\beta}(D_j) = (X^{(j)^\top}X^{(j)})^{-1}X^{(j)^\top}Y^{(j)}. \) In order to estimate \( \beta^* \), a simple and natural idea is to take the average of \( \{\hat{\beta}(D_j)\}_{j=1}^{k} \), which we denote by \( \overline{\beta} \). The question is whether this estimator preserves the statistical error as \( \hat{\beta} \). The following theorem gives an upper bound of the gap between \( \overline{\beta} \) and \( \hat{\beta} \), and shows that this gap is negligible compared with the statistical error of \( \hat{\beta} \) as long as \( k \) is not too large.

Here we give intuitive discussion on the source of efficiency loss. According to proof of Theorem A.1.1, we have

\[ \overline{\beta} - \hat{\beta} = \frac{1}{k} \sum_{j=1}^{k} \left( (X^{(j)^\top}X^{(j)}/n_k)^{-1} - (X^\top X/n)^{-1} \right) X^{(j)^\top}e^{(j)}/n_k. \]

Since \( \{X^{(j)^\top}e^{(j)}/n_k\}_{j=1}^{k} \) are homogeneous and independent to each other conditional on \( X \), the efficiency loss incurred by the DC procedure, i.e., the gap \( \overline{\beta} - \hat{\beta} \), is characterized by the difference between \( (\frac{1}{k} \sum_{j=1}^{k} \frac{1}{n_k}X^{(j)^\top}X^{(j)})^{-1} \) and \( (\frac{1}{k} \sum_{j=1}^{k} \frac{1}{n_k}X^{(j)^\top}X^{(j)})^{-1} \).

The rate of \( \overline{\beta} - \hat{\beta} \) is studied in detail in subsequent theorems.

**Theorem A.1.1.** Consider the linear model (2.2.2). Suppose Conditions 2.2.1 and 2.2.2 hold and \( \{\epsilon_i\}_{i=1}^{n} \) are i.i.d sub-Gaussian random variables with \( \|\epsilon_i\|_{\psi_2} \leq \sigma_1 \). If the number of subsamples satisfies \( k = O(nd/(d \vee \log n)^2) \), then for sufficiently large \( n \) and \( d \) it follows that

\[ \|\overline{\beta} - \hat{\beta}\|_2 = O_P\left( \frac{\sqrt{k}(d \vee \log n)}{n} \right), \quad \|\overline{\beta} - \beta^*\|_2 = O_P\left( \sqrt{d/n} \right). \]  

(A.1.1)
Remark A.1.1. By taking $k = o(nd/(d \vee \log n)^2)$, the loss incurred by the divide and conquer procedure, i.e., $\|\hat{\beta} - \beta\|_2$, converges at a faster rate than the statistical error of the full sample estimator $\hat{\beta}$. In another independent work [103], the authors also reveal a similar phenomenon under the broad family of generative linear models. They show that when $k = o(n/d)$, $E\|\beta - \beta^\ast\|_2^2/E\|\hat{\beta} - \beta^\ast\|_2^2 \to 1$. In other words, there is no first-order loss by divide and conquer.

We now take a different viewpoint by returning to the high dimensional setting of Section 2.3.1 ($d \gg n$) and applying Theorem A.1.1 in the context of a refitting estimator. In this refitting setting, the sparsity $s$ of Lemma 2.3.1 becomes the dimension of a low dimensional parameter estimation problem on the selected support. Our refitting estimator is defined as

\[
\bar{\beta} := \frac{1}{k} \sum_{j=1}^{k} (X^{(j)}_S X^{(j)}_S)^{-1} X^{(j)}_S Y^{(j)}, \tag{A.1.2}
\]

where $\hat{S} := \{j : |\beta^\ast_j| > 2C\sqrt{\log d/n}\}$ and $C$ is the same constant as in (2.3.1).

Corollary A.1.1. Suppose $\beta^\ast_{\min} > 2C\sqrt{\log d/n}$, where $\beta^\ast_{\min} := \min_{1 \leq j \leq d} |\beta^\ast_j|$ and $C$ is the same constant as in (2.3.1). Define the full sample oracle estimator as $\hat{\beta}^\circ = (X_S^\top X_S)^{-1} X_S^\top Y$, where $S$ is the true support of $\beta^\ast$. If $k = O(\sqrt{n/(s^2 \log d)})$, then for sufficiently large $n$ and $d$ we have

\[
\|\bar{\beta} - \hat{\beta}^\circ\|_2 = O_p\left(\frac{\sqrt{k} \log n}{n}\right), \quad \|\bar{\beta} - \beta^\ast\|_2 = O_p\left(\frac{\sqrt{s}}{n}\right). \tag{A.1.3}
\]

We see from Corollary A.1.1 that $\bar{\beta}$ achieves the oracle rate when the minimum signal strength is not too weak and the number of subsamples $k$ is not too large.
A.2 The low-dimensional generalized linear model

The next theorem quantifies the gap between \( \bar{\beta} \) and \( \hat{\beta} \), where \( \bar{\beta} \) is the average of subsampled GLM estimators and \( \hat{\beta} \) is the full sample GLM estimator. In Theorem \[A.2.1\], \( \| \bar{\beta} - \hat{\beta} \|_2 \) is the distance between the divide and conquer estimator and the full sample estimator, while \( \| \bar{\beta} - \beta^* \|_2 \) is the estimation error on each machine.

**Theorem A.2.1.** Under Condition 2.2.3, if \( k = O(\sqrt{n}/(d \vee \log n)) \), then we have for sufficiently large \( d \) and \( n \),

\[
\| \bar{\beta} - \hat{\beta} \|_2 = O_p\left(\frac{k\sqrt{d}(d \vee \log n)}{n}\right), \quad \| \bar{\beta} - \beta^* \|_2 = O_p\left(\sqrt{d/n}\right). \tag{A.2.1}
\]

**Remark A.2.1.** In analogy to Theorem \[A.1.1\], by constraining the growth rate of the number of subsamples according to \( k = o(\sqrt{n}/(d \vee \log n)) \), the error incurred by the divide and conquer procedure, i.e., \( \| \bar{\beta} - \hat{\beta} \|_2 \) decays at a faster rate than that of the statistical error of the full sample estimator \( \hat{\beta} \).

We notice a recent independent work \[78\] on distributed estimation under curved exponential families with fixed dimensions. They propose a KL-divergence-based combination method to aggregate MLEs from multiple data repositories and show that it can achieve the best possible approximation to the global MLE given the entire dataset. In the future work, it will be interesting to extend their approach to the GLM setting and characterize the statistical error rate of the correspondent distributed estimator.

The less stringent scaling of \( k \) in the low dimensional linear model relative to the generalized linear model comes from the fact that the Hessian matrix depends on the estimator of \( \beta^* \) in the GLM. This results in a larger variance relative to the linear model. Figures 3(A) and 4(A) indicate that the deduced scaling is sharp for both cases.

As in the linear model, Lemma \[2.3.2\] together with Theorem \[A.2.1\] allow us to study the theoretical properties of a refitting estimator for the high dimensional GLM.
Estimation on the estimated support set is again a low dimensional problem, thus the $d$ of Theorem A.2.1 corresponds to the $s$ of Lemma 2.3.2 in this refitting setting.

The refitted GLM estimator is defined as

$$
\mathcal{B}^r = \frac{1}{k} \sum_{j=1}^{k} \hat{\beta}^r(D_j),
$$

(A.2.2)

where $\hat{\beta}^r(D_j) = \arg\min_{\beta \in R^d, \beta_{S_c} = 0} \ell_n^{(j)}(\beta)$ and $\hat{S} := \{j : |\beta_j| > 2C \sqrt{\log d/n}\}$. The following corollary quantifies the statistical rate of $\mathcal{B}^r$.

**Corollary A.2.1.** Suppose $\beta^*_{\min} > 2C \sqrt{\log d/n}$, where $\beta^*_{\min} := \min_{1 \leq j \leq d} |\beta_j^*|$ and $C$ is the same constant as Lemma 2.3.2. Define the full sample oracle estimator as $\hat{\beta}^o = \arg\min_{\beta \in R^d, S_{c} = 0} \ell_n(\beta)$, where $S$ is the true support of $\beta^*$. If $k = O\left(\sqrt{n/(s \vee s_1^2 \log d)}\right)$, then for sufficiently large $n$ and $d$ we have

$$
\|\mathcal{B}^r - \hat{\beta}^o\|_2 = O_P\left(\frac{k \sqrt{s} \log n}{n}\right), \quad \|\mathcal{B}^r - \beta^*\|_2 = O_P\left(\sqrt{s/n}\right).
$$

(A.2.3)

We thus see that $\mathcal{B}^r$ achieves the oracle rate when the minimum signal strength is not too weak and the number of subsamples $k$ is not too large.

**A.3 Simulation for the low-dimensional linear model**

All $n \times d$ entries of the design matrix $X$ are generated as i.i.d. standard normal random variables and the errors $\{\epsilon_i\}_{i=1}^n$ are i.i.d. standard normal as well. The true regression vector $\beta^*$ satisfies $\beta_j^* = 10/\sqrt{d}$ for $j = 1, \ldots, d/2$ and $\beta_j^* = -10/\sqrt{d}$ for $j > d/2$, which guarantees that $\|\beta^*\|_2 = 10$. Then we generate the response variable $\{Y_i\}_{i=1}^n$ according to the model (2.2.2). Denote the full sample ordinary least-squares estimator and the divide and conquer estimator by $\hat{\beta}$ and $\mathcal{B}$ respectively. Figure
Figure A.1: (A) The ratio between the loss of the divide and conquer procedure and the statistical error of the estimator based on the whole sample with $d = \sqrt{n}/2$ and different growth rates of $k$. (B) Statistical error of the DC estimator against $\log k$.

(A) illustrates the change in the ratio $\| \beta - \hat{\beta} \|^2 / \| \hat{\beta} - \beta^* \|^2$ as the sample size increases, where $k$ assumes three different growth rates and $d = \sqrt{n}/2$. Figure A.1(B) focuses on the relationship between the statistical error of $\beta$ and $\log k$ under three different scalings of $n$ and $d$. All the data points are obtained based on average over 100 Monte Carlo replications.

As Figure A.1(A) demonstrates, when $k = O(n^{1/3})$, $O(n^{1/4})$ or $O(1)$, the ratio decreases with ever faster rates, which is consistent with the argument of Remark A.1.1 that the ratio goes to zero when $k = o(n/d) = o(\sqrt{n})$. When $k = O(\sqrt{n})$, however, we observe that the ratio is essentially constant, which suggests the rate we derived in Theorem A.1.1 is sharp. We also report the wall time of our proposed distributed approach and the naive average Lasso in Table 2.1. The time is computed in a similar way as the testing part. A comparison between these two approaches reveals the heavy computation incurred by debiasing. However, we observe that as the splits grow, the time consumption on individual data sample decreases since the local problem size becomes smaller, which mitigates the time complexity problem if we have a parallel computing system.
From Figure A.1(B), we see that when $k$ is not large, the statistical error of $\tilde{\beta}$ is very small because the loss incurred by the divide and conquer procedure is negligible compared to the statistical error of $\hat{\beta}$. However, when $k$ is larger than a threshold, there is a surge in the statistical error, since the loss of the divide and conquer begins to dominate the statistical error of $\hat{\beta}$. We also notice that the larger the ratio $n/d$, the larger the threshold of $\log k$, which is again consistent with Remark A.1.1.

### A.4 Simulation for the low-dimensional logistic regression

In logistic regression, given covariates $X$, the response $Y|X \sim \text{Ber}(\eta(X))$, where $\text{Ber}(\eta)$ denotes the Bernoulli distribution with expectation $\eta$ and

$$
\eta(X) = \frac{1}{1 + \exp(-X^\top \beta^*)}.
$$

We see that $\text{Ber}(\eta(X))$ is in exponential dispersion family canonical form (2.1.7) with $b(\theta) = \log(1 + e^\theta)$, $\phi = 1$ and $c(y) = 1$. The use of the canonical link,

$$
\eta(X) = \frac{1}{1 + e^{-\eta(X)}},
$$

leads to the simplification $\theta(X) = X^\top \beta^*$.

In our Monte Carlo experiments, all $n \times d$ entries of the design matrix $X$ are generated as i.i.d. standard normal random variables. The true regression vector $\beta^*$ satisfies $\beta^*_j = 1/\sqrt{d}$ for $j \leq d/2$ and $\beta^*_j = -1/\sqrt{d}$ for $j > d/2$, which guarantees that $\|\beta^*\|_2 = 1$. Finally, we generate the response variables $\{Y_i\}_{i=1}^n$ according to $\text{Ber}(\eta(X))$. Figure A.2(A) illustrates the change of the ratio $\|\tilde{\beta} - \beta\|_2/\|\beta - \beta^*\|_2$ as the sample size increases, where $k$ assumes three different growths rates and $d = 20$. Figure A.2(B) focuses on the relationship between the statistical error of $\tilde{\beta}$ and $\log k$. 
Figure A.2: (A) The ratio between the loss of the divide and conquer procedure and the statistical error of the estimator based on the whole sample when \(d = 20\). (B) Statistical error of the DC estimator under three different scalings of \(n\) and \(d\). All the data points are obtained based on an average over 100 Monte Carlo replications.

Figure A.2 reveals similar phenomena to those revealed in Figure A.1 of the previous subsection. More specifically, Figure A.2(A) shows that when \(k = O(n^{1/3})\), \(O(n^{1/4})\) or \(O(1)\), the ratio decreases with even faster rates, which is consistent with the argument of Remark A.2.1 that the ratio converges to zero when \(k = o(\sqrt{n}/d) = o(\sqrt{n})\). When \(k = O(\sqrt{n})\), however, we observe that the ratio remains essentially constant when \(\log n\) is large, which suggests the rate we derived in Theorem A.1.1 is sharp.

As for Figure A.2(B), we again observe that the statistical error of \(\tilde{\beta}\) is very small when \(k\) is sufficiently small, but grows fast when \(k\) becomes large. The reasoning is the same as in the linear model, i.e. when \(k\) is large, the loss incurred by the divide and conquer procedure is non-negligible as compared with the statistical error of \(\|\tilde{\beta}\|_2\). In addition, as Figure A.2(B) reveals, the larger is \(\sqrt{n}/d\), the larger the threshold of \(k\), which is again consistent with the threshold rate pointed out in Remark A.2.1.
A.5 Auxiliary lemmas and theorems for testing

In this section, we provide the proofs of the technical lemmas and theorems for the divide and conquer hypothesis testing.

Proof of Theorem 2.2.1. It remains to verify the Lindeberg’s Condition for (2.5.1). By Lemma A.5.1

\[ |\xi^{(j)}_{iv}| \leq n^{-1/2}c_{nk}^{-1}|m_v^{(j)\top}X_i^{(j)}||\epsilon^{(j)}_i| \leq n^{-1/2}c_{nk}^{-1}\vartheta_2|\epsilon^{(j)}_i|, \]

where \( \lim \inf_{nk} c_{nk} = c_\infty > 0 \), hence the event \( \{|\xi^{(j)}_{iv}| > \epsilon \sigma\} \) is contained in the event \( \{|\epsilon^{(j)}_i| > \epsilon \sigma c_{nk} \vartheta_2^{-1}\sqrt{n}\} \) and we have

\[
\frac{1}{\sigma^2} \sum_{j=1}^k \sum_{i \in I_j} \mathbb{E} \left[ (\xi^{(j)}_{iv})^2 \mathbb{1}\{|\xi^{(j)}_{iv}| > \epsilon \sigma\} \right] X \\
\leq \frac{1}{\sigma^2} \sum_{j=1}^k \sum_{i \in I_j} \mathbb{E} \left[ (\xi^{(j)}_{iv})^2 \mathbb{1}\{|\epsilon^{(j)}_i| > \epsilon \sigma c_{nk} \vartheta_2^{-1}\sqrt{n}\} \right] \\
= \frac{1}{\sigma^2} \sum_{j=1}^k \frac{1}{n_k} \sum_{i \in I_j} \left( m_v^{(j)\top}X_i^{(j)}\right)^2 \mathbb{E} \left[ (\epsilon^{(j)}_i)^2 \mathbb{1}\{|\epsilon^{(j)}_i| > \epsilon \sigma c_{nk} \vartheta_2^{-1}\sqrt{n}\} \right] \\
= \frac{1}{\sigma^2} \sum_{j=1}^k \frac{1}{n_k} \sum_{i \in I_j} \left( m_v^{(j)\top}\hat{\Sigma}m_v^{(j)}\right) \mathbb{E} \left[ (\epsilon^{(j)}_i)^2 \mathbb{1}\{|\epsilon^{(j)}_i| > \epsilon \sigma c_{nk} \vartheta_2^{-1}\sqrt{n}\} \right].
\]

Taking expectation with respect to \( X \) on both sides above yields that

\[
\frac{1}{\sigma^2} \sum_{j=1}^k \sum_{i \in I_j} \mathbb{E} \left[ (\xi^{(j)}_{iv})^2 \mathbb{1}\{|\xi^{(j)}_{iv}| > \epsilon \sigma\} \right] \\
\leq \frac{1}{\sigma^2} \mathbb{E} \left[ (\epsilon^{(j)}_i)^2 \mathbb{1}\{|\epsilon^{(j)}_i| > \epsilon \sigma c_{nk} \vartheta_2^{-1}\sqrt{n}\} \right].
\]

Let \( \delta = \epsilon \sigma c_{nk} \vartheta_2^{-1}\sqrt{n} \). Then, for any \( \eta > 0 \),

\[
\mathbb{E} \left[ (\epsilon^{(j)}_i)^2 \mathbb{1}\{|\epsilon^{(j)}_i| > \delta\} \right] \leq \mathbb{E} \left[ (\epsilon^{(j)}_i)^2 \frac{1}{\delta^\eta} \mathbb{1}\{|\epsilon^{(j)}_i| > \delta\} \right] \leq \delta^{-\eta} \mathbb{E} \left[ |\epsilon^{(j)}_i|^{2+\eta} \right]. \quad (A.5.1)
\]

133
Since $\vartheta_2 n^{-1/2} = o(1)$ by the statement of the theorem, the choice $\eta = 2$ delivers

$$
\frac{1}{\sigma^2} \lim_{k \to \infty} \lim_{n_k \to \infty} \sum_{j=1}^{k} \sum_{i \in I_j} \mathbb{E}\left[ (\xi_{iv}^{(j)})^2 1\{|\xi_{iv}^{(j)}| > \epsilon \sigma\} \right]
\leq \lim_{k \to \infty} \lim_{n_k \to \infty} k^{-1} n^{-1} \vartheta_2 c_n \epsilon^{-2} \sigma^{-2} \mathbb{E}\left( (\epsilon_{iv}^{(j)})^4 \right) = 0 \quad (A.5.2)
$$

by the bounded forth moment assumption. By the law of iterated expectations, all conditional results hold in unconditional form as well. Hence, $\overline{V}_n \sim N(0, \sigma^2)$ by the Lindeberg-Feller central limit theorem. $\square$

Proof of Corollary 2.2.2. We verify (A5)-(A9) of Lemma A.5.7 in the Supplementary Material. (A5) is satisfied because $\hat{\Theta}_{iv}$ is consistent under the required scaling by the statement of the corollary. (A6) is satisfied by Condition 2.2.4. To verify (A7), first note that $\nabla \ell_i(\beta^*) = (b'(X_i^\top \beta^*) - Y_i)X_i$. According to Lemma A.5.2 in the Supplementary Material, we know that conditional on $X$, $b'(X_i^\top \beta^*) - Y_i$ is a sub-Gaussian random variable. Therefore Lemma A.6.6 in the Supplementary Material delivers

$$
\mathbb{P}\left( \left\| \frac{1}{n} \sum_{j=1}^{k} \sum_{i \in I_j} \nabla \ell_i(\beta^*) \right\|_\infty > t \mid X \right) \leq d \exp \left( 1 - \frac{ct^2}{nM^2} \right),
$$

which implies that with probability $1 - c/d$,

$$
\left\| \sum_{j=1}^{k} \sum_{i \in I_j} \nabla \ell_i(\beta^*) \right\|_\infty = C \sqrt{n \log d} \quad (A.5.3)
$$

It only remains to verify (A8). Let $\xi_{iv}^{(j)} = \Theta_{iv}^* \nabla \ell_i^{(j)}(\beta^*) / \sqrt{n \Theta_{iv}^*}$. By the definition of the log likelihood,

$$
\mathbb{E}[\xi_{iv}^{(j)}] = \frac{\Theta_{iv}^* \mathbb{E}[\nabla \ell_i^{(j)}(\beta^*)]}{(n \Theta_{iv}^*)^{1/2}} = 0
$$
and by independence of \( \{(Y_i, X_i)\}_{i=1}^{n} \),

\[
\begin{align*}
\text{Var}(\sum_{j=1}^{k} \sum_{i \in I_j} \xi_{iv}^{(j)}) &= \sum_{j=1}^{k} \sum_{i \in I_j} \text{Var}(\xi_{iv}^{(j)}) = \sum_{j=1}^{k} \sum_{i \in I_j} \mathbb{E}[(\xi_{iv}^{(j)})^2] \\
&= \frac{1}{n} \sum_{i=1}^{n} (\Theta_{iv}^*)^{-1} \Theta_{v}^* \mathbb{E}[(\nabla \ell_i(\beta^*))^2] \\
&= \frac{1}{n} \sum_{i=1}^{n} (\Theta_{iv}^*)^{-1} \Theta_{v}^* \mathbb{E}[\xi_{iv}^{(j)}] = 1.
\end{align*}
\]

By Condition 2.2.3, \( \theta_{\min} > 0 \), the event \( \{|\xi_{iv}^{(j)}| > \epsilon\} \) coincides with the event \( \{|\Theta_{iv}^* \nabla \ell_i(\beta^*)| > \epsilon \sqrt{\theta_{\min}} n\} = \{|\Theta_{iv}^* X_i (Y_i - b'(X_i^T \beta^*))| > \epsilon \sqrt{\theta_{\min}} n\}. \) Furthermore, since \( |\Theta_{iv}^* X_i| \leq M \) by Condition 2.2.4, this event is contained in the event \( \{|Y_i - b'(X_i^T \beta^*)| > \delta\} \), where \( \delta = \epsilon \sqrt{\theta_{\min}} n/M. \) By an analogous calculation to that of equation (A.5.1) in the Supplementary Material, we have

\[
\mathbb{E}[(Y_i - b'(X_i^T \beta^*))^2 \mathbb{1} \{|Y_i - b'(X_i^T \beta^*)| > \delta\}|X] \leq \delta^{-\eta} \mathbb{E}[(Y_i - b'(X_i^T \beta^*))^{2+\eta}|X].
\]

Hence, setting \( \eta = 2 \) and noting that \( \mathbb{E}[(Y_i - b'(X_i^T \beta^*))^{2+\eta}|X] \leq C \sqrt{2 + \eta \phi U_2} \) by Lemma A.5.2 in the Supplementary Material, it follows that

\[
\begin{align*}
\lim_{k \to \infty} \lim_{n_k \to \infty} \sum_{j=1}^{k} \sum_{i \in I_j} \mathbb{E}[(\xi_{iv}^{(j)})^2 \mathbb{1} \{|\xi_{iv}^{(j)}| > \epsilon\}] \\
&\leq \theta_{\min}^{-1} \lim_{k \to \infty} \lim_{n_k \to \infty} n^{-1} \sum_{j=1}^{k} \sum_{i \in I_j} \Theta_{v}^* \mathbb{E}[X_i X_i^T] \Theta_{v}^* \delta^{-2} \\
&\leq \theta_{\min}^{-1} \lim_{k \to \infty} \lim_{n_k \to \infty} M^3 s_1^2/(n \epsilon^2 \theta_{\min}) = 0, \tag{A.5.4}
\end{align*}
\]

where the last inequality follows because \( \|\Sigma\|_{\max} = \|\mathbb{E}[X_i X_i^T]\|_{\max} < M^2 \) by Condition 2.2.3. Similarly, we have for any \( \epsilon > 0 \),

\[
\epsilon^{-3} \lim_{k \to \infty} \lim_{n_k \to \infty} \sum_{j=1}^{k} \sum_{i \in I_j} \mathbb{E}[(\xi_{iv}^{(j)})^3 \mathbb{1} \{|\xi_{iv}^{(j)}| > \epsilon\}] = 0.
\]

135
Applying the self-normalized Berry-Essen inequality, we complete the proof of this corollary. □

Lemma A.5.1. Under Condition 2.2.2, \((m_v^{(j)})^\top \widehat{\Sigma}_v^{(j)})^{-1/2} \geq c_{n_k}\) for any \(j \in \{1, \ldots, k\}\) and for any \(v \in \{1, \ldots, d\}\), where \(c_{n_k}\) satisfies \(\liminf_{n_k \to \infty} c_{n_k} = c_\infty > 0\).

Proof. The proof appears in the proof of Lemma B1 of [140]. □

Lemma A.5.2. Under the GLM (2.1.7), we have

\[
\mathbb{E} \exp(t(Y - \mu(\theta))) = \exp(\phi^{-1}(b(\theta + t\phi) - b(\theta) - \phi t b'(\theta))),
\]

and typically when there exists \(U > 0\) such that \(b''(\theta) < U\) for all \(\theta \in \mathbb{R}\), we will have

\[
\mathbb{E} \exp(t(Y - \mu(\theta))) \leq \exp\left(\frac{\phi U t^2}{2}\right),
\]

which implies that \(Y\) is a sub-Gaussian random variable with variance proxy \(\phi U\).

Proof.

\[
\begin{align*}
\mathbb{E} \exp(t(Y - \mu(\theta))) & = \int_{-\infty}^{+\infty} c(y) \exp\left(\frac{y\theta - b(\theta)}{\phi}\right) \exp(t(y - \mu(\theta))) dy \\
& = \int_{-\infty}^{+\infty} c(y) \exp\left(\frac{(\theta + t\phi)y - (b(\theta) + \phi t b'(\theta))}{\phi}\right) dy \\
& = \int_{-\infty}^{+\infty} c(y) \exp\left(\frac{(\theta + t\phi)y - b(\theta + t\phi) + b(\theta + t\phi) - (b(\theta) + \phi t b'(\theta))}{\phi}\right) dy \\
& = \exp\left(\phi^{-1}(b(\theta + t\phi) - b(\theta) - \phi t b'(\theta))\right).
\end{align*}
\]

When \(b''(\theta) < U\), the mean value theorem gives

\[
\mathbb{E} \exp(t(Y - \mu(\theta))) = \exp\left(\frac{b''(\theta)\phi^2 t^2}{2\phi}\right) \leq \exp\left(\frac{\phi U t^2}{2}\right).
\]

□
Proof of Lemma 2.2.1. We first show that, for any $j \in \{1, \ldots, k\}$, $|\hat{\sigma}^2(D_j) - \sigma^2| = o_P(k^{-1})$. To this end, letting

$$
\hat{\epsilon}_i = Y_i^{(j)} - X_i^{(j)^T} \hat{\theta}^\lambda(D_j) = Y_i^{(j)} - X_i^{(j)^T} \beta^* - X_i^{(j)^T} (\hat{\theta}^\lambda(D_j) - \beta^*) ,
$$

we write

$$
|\hat{\sigma}^2(D_j) - \sigma^2| = \left| \frac{1}{nk} \sum_{i \in I_j} \hat{\epsilon}_i^2 - \sigma^2 \right| \leq \Delta_1^{(j)} + 2 \Delta_2^{(j)} + \Delta_3^{(j)},
$$

$$
\Delta_1^{(j)} := \left| \frac{1}{nk} \sum_{i \in I_j} \epsilon_i^2 - \sigma^2 \right|, \quad \Delta_2^{(j)} := \left| (\hat{\theta}^\lambda(D_j) - \beta^*) \left( \frac{1}{nk} \sum_{i \in I_j} X_i^{(j)} \epsilon_i^{(j)} \right) \right| \quad \text{and}
$$

$$
\Delta_3^{(j)} := \left| (\hat{\theta}^\lambda(D_j) - \beta^*)^T \left( \frac{1}{nk} \sum_{i \in I_j} X_i^{(j)} X_i^{(j)^T} \right) (\hat{\theta}^\lambda(D_j) - \beta^*) \right|
$$

$$
= \| X^{(j)} (\hat{\theta}^\lambda(D_j) - \beta^*) \|_2^2 / nk = O_P(\lambda^2 s)
$$

by Theorem 6.1 of [14]. Hence, with $\lambda = C \sigma^2 \sqrt{k \log d/n}$, $\Delta_3^{(j)} = o_P(1)$ for $k = o((s \log d)^{-1})$, a fortiori for $k = o((s \log d)^{-1} \sqrt{n})$. Letting

$$
\Delta_2^{(j)} = \| \hat{\theta}^\lambda(D_j) - \beta^* \|_1 \left\| \frac{1}{nk} \sum_{i \in I_j} X_i^{(j)} \epsilon_i^{(j)} - \mathbb{E} [X_i^{(j)} \epsilon_i^{(j)}] \right\|_\infty ,
$$

$$
\Delta_2^{(j)} = \| \hat{\theta}^\lambda(D_j) - \beta^* \|_1 \left\| \mathbb{E} [X_i^{(j)} \epsilon_i^{(j)}] \right\|_\infty .
$$

We obtain the bound

$$
\Delta_2^{(j)} = \left| (\hat{\theta}^\lambda(D_j) - \beta^*) \left( \frac{1}{nk} \sum_{i \in I_j} X_i^{(j)} \epsilon_i^{(j)} - \mathbb{E} [X_i^{(j)} \epsilon_i^{(j)}] \right) + \mathbb{E} [X_i^{(j)} \epsilon_i^{(j)}] \right|
$$

$$
\leq \Delta_2^{(j)} + \Delta_2^{(j)} .
$$
By the statement of the Lemma, \( \mathbb{E} [X_i^{(j)} \epsilon_i^{(j)}] = \mathbb{E} [X_i^{(j)} \mathbb{E} [\epsilon_i^{(j)} | X_i^{(j)}]] = 0 \), hence \( \Delta_{22}^{(j)} = 0 \), while by the central limit theorem and Theorem 6.1 of [13],

\[
\Delta_{21}^{(j)} \leq O_P(\lambda s) O_P(n_k^{-1/2}).
\]

We conclude \( \Delta_2^{(j)} = O_P(\lambda s n_k^{-1/2}) \), and with \( \lambda \asymp \sigma^2 \sqrt{k \log d/n} \), \( \Delta_2^{(j)} = o(1) \) with \( k = o(n(s \log d)^{-2/3}) \), a fortiori for \( k = o(\sqrt{n}(s \log d)^{-1}) \). Finally, noting that \( \sigma^2 = \mathbb{E} [\epsilon_i^{(j)}] \), \( \Delta_1^{(j)} = O_P(n_k^{-1/2}) = o_P(1) \) by the central limit theorem. Combining the bounds, we obtain \( |\hat{\sigma}^2(D_j) - \sigma^2| = o_P(1) \) for any \( j \in \{1, \ldots, k\} \) and therefore \( |\hat{\sigma}^2 - \sigma^2| \leq k^{-1} \sum_{j=1}^k |\hat{\sigma}^2(D_j) - \sigma^2| = o_P(1). \)

\[\Box\]

**Lemma A.5.3.** Under Condition [2.2.3] we have for any \( \beta, \beta' \in \mathbb{R}^d \) and any \( i = 1, \ldots, n \), \( |\ell''_i(X_i^\top \beta) - \ell''_i(X_i^\top \beta')| \leq K_i |X_i^\top (\beta - \beta')| \), where \( 0 < K_i < \infty \).

**Proof.** By the canonical form of the generalized linear model (equation (2.1.8)),

\[
|\ell''_i(X_i^\top \beta) - \ell''_i(X_i^\top \beta')| = |b''(X_i^\top \beta) - b''(X_i^\top \beta')| \leq |b''(\eta)||X_i^\top (\beta - \beta')|
\]

by the mean value theorem, where \( \eta \) lies in a line segment between \( X_i^\top \beta \) and \( X_i^\top \beta' \). 

\( |b''(\eta)| < U_3 < \infty \) by Condition [2.2.3] for any \( \eta \), hence the conclusion follows with \( K_i = U_3 \) for all \( i \).

\[\Box\]

**Lemma A.5.4.** Under Conditions [2.1.6] and [2.1.1] (i), we have for any \( \delta \in (0, 1) \) such that \( \delta^{-1} \ll d \),

\[
\mathbb{P} \left( \frac{1}{n} \|X(\beta^\lambda - \beta^*)\|_2^2 \geq s \frac{\log(d/\delta)}{n} \right) < \delta.
\]

**Proof.** Decompose the object of interest as

\[
\frac{1}{n} \|X(\beta^\lambda - \beta^*)\|_2^2 = (\beta^\lambda - \beta^*)^\top (\tilde{\Sigma} - \Sigma)(\beta^\lambda - \beta^*) + (\beta^\lambda - \beta^*)^\top \Sigma(\beta^\lambda - \beta^*)
\]

\[
\leq \|\tilde{\Sigma} - \Sigma\|_{\text{max}} \|\beta^\lambda - \beta^*\|_1^2 + \lambda_{\text{max}}(\Sigma) \|\beta^\lambda - \beta^*\|_2^2.
\]

138
This gives rise to the tail probability bound

\[
\mathbb{P}\left(\frac{1}{n}\|X(\hat{\beta}^\lambda - \beta^*)\|_2^2 > t\right) \\
\leq \mathbb{P}\left(\|\hat{\Sigma} - \Sigma\|_{\text{max}}\|\hat{\beta}^\lambda - \beta^*\|_1^2 > \frac{t}{2}\right) + \mathbb{P}\left(\lambda_{\text{max}}(\Sigma)\|\hat{\beta}^\lambda - \beta^*\|_2^2 > \frac{t}{2}\right). \tag{A.5.5}
\]

Let \( M := \left\{ \|\hat{\Sigma} - \Sigma\|_{\infty} \leq M \right\} \). Since \( \{X_i\}_{i=1}^n \) is bounded, it is sub-Gaussian as well. Suppose \( \|X_i\|_{\psi_2} < \kappa \), then by Lemma A.6.3 we have,

\[
\mathbb{P}(M^c) \leq \sum_{p,q=1}^d \mathbb{P}(|\hat{\Sigma}_{pq}^{(j)} - \Sigma_{pq}| > M) \leq d^2 \exp\left(-Cn \cdot \min\left\{ \frac{M^2}{\kappa^4}, \frac{M}{\kappa^2} \right\} \right),
\]

where \( C \) is a constant. Hence taking \( M = n^{-1} \log(d/\delta) \),

\[
\mathbb{P}(M^c) \leq d^2 \exp\left\{-Cn \min\left\{ \frac{(\log(d/\delta))^2}{\kappa^4 n^2}, \frac{(\log(d/\delta))^2}{\kappa^2 n} \right\} \right\}
\]

and the right hand side is less than \( \delta \) for \( \delta^{-1} \ll d \). Thus by Condition 2.1.1 the first term on the right hand side of equation (A.5.5) is

\[
\mathbb{P}\left(\|\hat{\Sigma} - \Sigma\|_{\text{max}}\|\hat{\beta}^\lambda - \beta^*\|_1^2 \gtrsim \frac{s \log(d/\delta)}{n} \right) < 2\delta.
\]

Furthermore, by Condition 2.2.3 (i), the second term on the right hand side of equation (A.5.5) is

\[
\mathbb{P}\left(\lambda_{\text{max}}(\Sigma)\|\hat{\beta}^\lambda - \beta^*\|_2^2 \gtrsim C_{\text{max}} \frac{s \log(d/\delta)}{n} \right) \leq \delta.
\]

Taking \( t \) as the dominant term, \( t \asymp C_{\text{max}} n^{-1} s \log(d/\delta) \), yields the result. \( \square \)

**Lemma A.5.5.** Under Condition 2.2.3 we have for any \( i = 1, \ldots, n \),

\[
|b''(X_i^\top \beta_1) - b''(X_i^\top \beta_2)| \leq MU_3\|\beta_1 - \beta_2\|_1,
\]

139
and if we consider the sub-Gaussian design instead, we have

$$\mathbb{P}\left(\left\|b''(X_i^\top \beta_1) - b''(X_i^\top \beta_2)\right\| \geq hU_3 \|\beta_1 - \beta_2\|_1\right) \leq nd \exp\left(1 - \frac{Ch^2}{s_1^2}\right).$$

**Proof.** For the bounded design, by Condition 2.2.3 (iii), we have

$$|b''(X_i^\top \beta_1) - b''(X_i^\top \beta_2)| \leq U_3 |X_i^\top (\beta_1 - \beta_2)| \leq MU_3 \|\beta_1 - \beta_2\|_1.$$  

For the sub-Gaussian design, denote the event \(\{\max_{1 \leq i \leq n, 1 \leq j \leq d} |X_{ij}| \leq h\}\) by \(C\), where \(\kappa\) is a positive constant. Then it follows that,

$$\mathbb{P}(C^c) \leq nd \exp\left(1 - \frac{Ch^2}{s_1^2}\right),$$

where \(C\) is a constant. Since on the event \(C\), \(|b''(X_i^\top \beta_1) - b''(X_i^\top \beta_2)| \leq hU_3 \|\beta_1 - \beta_2\|_1\), we reach the conclusion. \(\square\)

**Remark A.5.1.** For the sub-Gaussian design, in order to let the tail probability go to zero, \(h \gg \log((n \vee d))\).

**Lemma A.5.6.** Suppose, for any \(k \ll d\) satisfying \(k = o((s \vee s_1) \log d)^{-1} \sqrt{n})\), the following conditions are satisfied. (A1) \(\mathbb{P}\left(n_k^{-1} \|X^{(j)} \hat{\Theta}^{(j)}\|_{\max} \geq H\right) \leq \xi\), where \(H\) is a constant and \(\xi = o(k^{-1})\). (A2) For any \(\beta, \beta' \in \mathbb{R}^d\) and for any \(i \in \{1, \ldots, n\}\), \(|\ell''_i(X_i^\top \beta) - \ell''_i(X_i^\top \beta')| \leq K_i |X_i^\top (\beta - \beta')|\) with \(\mathbb{P}(K_i > h) \leq \psi\) for \(\psi = o(k^{-1})\) and \(h = O(1)\). (A3) \(\mathbb{P}\left(n_k^{-1} \|X^{(j)} (\hat{\beta}_i^\lambda - \beta^\star)\|_2^2 \gtrsim n^{-1} sk \log(d/\delta)\right) \leq \delta\). (A4) \(\mathbb{P}\left(\max_{1 \leq v \leq d} \left|\left(\hat{\Theta}_v^{(j)\top} \nabla \ell^{(j)}_{nk}(\hat{\beta}^\lambda (D_j)) - e_v\right)(\hat{\beta}_i^\lambda (D_j) - \beta^\star)\right| \gtrsim n^{-1} sk \log(d/\delta)\right) < \delta\).

Then

$$\beta_v^d - \beta_v^\star = -\frac{1}{k} \sum_{j=1}^k \hat{\Theta}_v^{(j)\top} \nabla \ell^{(j)}_{nk}(\beta^\star) + o_P(n^{-1/2}).$$

for any \(1 \leq v \leq d\).
Proof of Lemma A.5.6.  $\overline{\beta}_v^d - \beta_v^* = k^{-1} \sum_{j=1}^{k} (\widehat{\beta}_v(D_j) - \beta_v^*)$. By the definition of $\widehat{\beta}_v(D_j)$,

$$
\widehat{\beta}_v(D_j) - \beta_v^* = \hat{\beta}_v^\lambda(D_j) - \beta_v^* - \Theta_v^{(j)\top} \nabla \ell_{n_k}^{(j)}(\hat{\beta}_v^\lambda(D_j)).
$$

Consider a mean value expansion of $\nabla \ell_{n_k}^{(j)}(\hat{\beta}_v^\lambda(D_j))$ around $\beta^*$:

$$
\nabla \ell_{n_k}^{(j)}(\hat{\beta}_v^\lambda(D_j)) = \nabla \ell_{n_k}^{(j)}(\beta^*) + \nabla^2 \ell_{n_k}^{(j)}(\beta^*) (\hat{\beta}_v^\lambda(D_j) - \beta^*),
$$

where $\beta_{\alpha} = \alpha \hat{\beta}_v^\lambda(D_j) + (1 - \alpha) \beta^*$, $\alpha \in [0, 1]$. So

$$
\frac{1}{k} \sum_{j=1}^{k} \hat{\beta}_v^d(D_j) - \beta_v^* = -\frac{1}{k} \sum_{j=1}^{k} \Theta_v^{(j)\top} \nabla \ell_{n_k}^{(j)}(\beta^*) - \Delta,
$$

where $\Delta = \frac{1}{k} \sum_{j=1}^{k} (\Theta_v^{(j)\top} \nabla^2 \ell_{n_k}^{(j)}(\beta_{\alpha}) - e_v)(\hat{\beta}_v^\lambda(D_j) - \beta^*)$. Note that $|\Delta| \leq \frac{1}{k} \sum_{j=1}^{k} (|\Delta_1^{(j)}| + |\Delta_2^{(j)}|)$, where

$$
|\Delta_1^{(j)}| = \left| (\Theta_v^{(j)\top} \nabla^2 \ell_{n_k}^{(j)}(\hat{\beta}_v^\lambda(D_j)) - e_v)(\hat{\beta}_v^\lambda(D_j) - \beta^*) \right|.
$$

By (A4) of the lemma, for $t > n^{-1} sk \log(d/\delta)$,

$$
P \left( \sum_{j=1}^{k} |\Delta_1^{(j)}| > kt \right) \leq P \left( \bigcup_{j=1}^{k} |\Delta_1^{(j)}| > t \right) \leq \sum_{j=1}^{k} P(|\Delta_1^{(j)}| > t) < k \delta.
$$

Substituting $\delta = o(k^{-1})$ in the expression for $t$ and noting that $k \ll d$, we obtain

$$
k^{-1} \sum_{j=1}^{k} \Delta_1^{(j)} = o_P(n^{-1/2}) \text{ for } k = o((s \log d)^{-1} \sqrt{n}).
$$

By (A2),

$$
|\Delta_2^{(j)}| = \left| \Theta_v^{(j)\top} \left( \nabla^2 \ell_{n_k}^{(j)}(\beta_{\alpha}) - \nabla^2 \ell_{n_k}^{(j)}(\hat{\beta}_v^\lambda(D_j)) \right) (\hat{\beta}_v^\lambda(D_j) - \beta^*) \right|
$$

$$
= \left| \frac{1}{n_k} \sum_{i \in I_j} \Theta_v^{(j)\top} X_i X_i^{\top} (\hat{\beta}_v^\lambda(D_j) - \beta^*) (\ell''_i(X_i^{\top} \beta_{\alpha}) - \ell''_i(X_i^{\top} \hat{\beta}_v^\lambda(D_j))) \right|
$$

$$
\leq \left( \max_{1 \leq i \leq n} K_i \right) \left( \frac{1}{n_k} \|X^{(j)}\|_{\text{max}} \right) \left( \frac{1}{n_k} X^{(j)} \Theta_v^{(j)\top} \|X^{(j)}\|_{\text{max}} \right) \frac{1}{n_k} X^{(j)} (\hat{\beta}_v^\lambda(D_j) - \beta^*) \|^2_2,
$$

141
Further decomposing the first term, we have

\[ P\left(\left| \sum_{j=1}^{k} \Delta_{2,j}^{(j)} \right| > kt \right) \leq P\left(\bigcup_{j=1}^{k} \left| \Delta_{2,j}^{(j)} \right| > t \right) \leq \sum_{j=1}^{k} P\left(\left| \Delta_{2,j}^{(j)} \right| > t \right) < k(\psi + \delta + \xi). \]

Substituting \( \delta = o(k^{-1}) \) in the expression for \( t \) and noting that \( k \ll d \), we obtain

\[ k^{-1} \sum_{j=1}^{k} \Delta_{2,j}^{(j)} = o_{P}(n^{-1/2}) \] for \( sk \log(d/\delta) = o(\sqrt{n}) \), i.e. for \( k = o((s \log d)^{-1}\sqrt{n}) \).

Combining these two results delivers \( \Delta = o_{P}(n^{-1/2}) \) for \( k = o((s \log d)^{-1}\sqrt{n}) \). \( \square \)

**Lemma A.5.7.** Suppose, in addition to Conditions (A1)-(A5) of Lemma A.5.6 (A5) \( |\Theta_{vv} - \Theta_{vv}^{*}| = o_{P}(1) \) for all \( v \in \{1, \ldots, d\} \); (A6) \( 1/\Theta_{vv}^{*} = O(1) \) for all \( v \in \{1, \ldots, d\} \); (A7) \( \|_{v} \sum_{1 \leq j \leq k} \sum_{i \in D_{j}} \nabla \ell_{i}(\beta^{*})\|_{\infty} = O_{P}(\sqrt{n \log d}) \); (A8) For each \( v \in \{1, \ldots, d\} \), letting \( \xi_{iv}^{(j)} = \Theta_{v}^{*T} \nabla \ell_{i}^{(j)}(\beta^{*})/\sqrt{n\Theta_{vv}^{*}}, \) \( E[\xi_{iv}^{(j)}] = 0, \) \( \text{Var}(\sum_{j=1}^{k} \sum_{i \in D_{j}} \xi_{iv}^{(j)}) = 1 \) and, for all \( \epsilon > 0, \)

\[ \lim_{k \to \infty} \lim_{n_{k} \to \infty} \sum_{j=1}^{k} \sum_{i \in D_{j}} E[\xi_{iv}^{(j)}]^{2} I\{\xi_{iv}^{(j)} > \epsilon\} = 0. \]  

(A.5.6)

Then under \( H_{0} : \beta_{v}^{*} = \beta_{v}^{H} \), taking \( k = o((s \lor s_{1}) \log d)^{-1}\sqrt{n}) \) delivers \( \overline{S_{n}} \sim N(0, 1) \), where \( \overline{S_{n}} \) is defined in equation (2.2.14).

**Proof.** Rewrite equation (2.2.14) as

\[ \overline{S_{n}} = \sqrt{n^{-1}} \sum_{j=1}^{k} \left[ \frac{\hat{\beta}_{v}^{d} - \beta_{v}^{H}}{(\Theta_{vv}^{*})^{1/2}} + \frac{\hat{\beta}_{v}^{d} - \beta_{v}^{H}}{(\Theta_{vv}^{*})^{1/2}} \left( \frac{(\Theta_{vv}^{*})^{1/2}}{\Theta_{vv}^{*}} - 1 \right) \right] \]

\[ = \sum_{j=1}^{k} \sum_{i \in I_{j}} (\Delta_{1,i}^{(j)} + \Delta_{2,i}^{(j)}), \quad \text{where} \]

\[ \Delta_{1,i}^{(j)} = \frac{\Theta_{v}^{(j)}T \nabla \ell_{i}^{(j)}(\beta^{*})}{(n\Theta_{vv}^{*})^{1/2}}, \quad \Delta_{2,i}^{(j)} = \frac{\Theta_{v}^{(j)}T \nabla \ell_{i}^{(j)}(\beta^{*})}{(n\Theta_{vv}^{*})^{1/2}} \left( \frac{(\Theta_{vv}^{*})^{1/2}}{\Theta_{vv}^{*}} - 1 \right). \]

Further decomposing the first term, we have

\[ \sum_{j=1}^{k} \sum_{i \in I_{j}} \Delta_{1,i}^{(j)} = \sum_{j=1}^{k} \sum_{i \in I_{j}} \xi_{iv}^{(j)} + \Delta, \quad \text{where} \]

\[ \Delta = \sum_{j=1}^{k} \sum_{i \in I_{j}} \left( \Theta_{v}^{(j)} - \Theta_{v}^{*} \right)^{T} \nabla \ell_{i}(\beta^{*}) (n\Theta_{vv}^{*})^{1/2}. \]
and \( \sum_{j=1}^{k} \sum_{i \in I_j} \xi_{i,v}^{(j)} \sim N(0,1) \) by the Lindeberg-Feller central limit theorem. Then by Hölder’s inequality, Condition 2.2.4 and Assumption (A6) and (A7),

\[
|\Delta| \leq \max_{1 \leq j \leq k} \| \hat{\Theta}_v^{(j)} - \Theta_v^* \|_1 \frac{\| \sum_{j=1}^{k} \sum_{i \in I_j} \nabla \ell_i(\beta^*) \|_\infty}{(n \Theta_v^*)^{1/2}} = O_p\left( s_1 \sqrt{k \log d} \right) O_p(\sqrt{\log d}) = o_p(1),
\]

where the last equation holds with the choice of \( k = o((s_1 \log d)^{-1} \sqrt{n}) \). Letting \( \overline{\Delta}^{(j)} = (\Theta_v^*)^{1/2} - \overline{\Theta}_{vv}^{1/2} \) we have

\[
\sum_{j=1}^{k} \sum_{i \in I_j} \Delta_{21,i}^{(j)} = \sum_{j=1}^{k} \sum_{i \in I_j} \left( \frac{\Theta_v^* \nabla \ell_i(\beta^*)}{(\Theta_v^*)^{1/2}} \overline{\Delta}^{(j)} + (\hat{\Theta}_v^{(j)} - \Theta_v^*)^\top \nabla \ell_i(\beta^*) \right) \overline{\Delta}^{(j)}
\]

where \( \left| \sum_{j=1}^{k} \sum_{i \in I_j} \Delta_{21,i}^{(j)} \right| \leq \left| \sum_{j=1}^{k} \sum_{i \in I_j} \xi_{i,v}^{(j)} \right| (\overline{\Theta}_{vv}^{1/2} - (\Theta_v^*)^{1/2}) \). Since \( \Theta_v^* \geq 0, \overline{\Theta}_{vv}^{1/2} = |\overline{\Theta}_{vv}|^{1/2} = |\overline{\Theta}_{vv} - \Theta_v^* + \Theta_v^*|^{1/2} \leq |\overline{\Theta}_{vv} - \Theta_v^*|^{1/2} + (\Theta_v^*)^{1/2} \). Similarly

\[
(\Theta_v^*)^{1/2} = |\Theta_v^*|^{1/2} = |\overline{\Theta}_{vv} - \Theta_v^* + \overline{\Theta}_{vv}|^{1/2} \leq |\Theta_v^* - \overline{\Theta}_{vv}|^{1/2} + |\overline{\Theta}_{vv}|^{1/2},
\]

yielding \( |\overline{\Theta}_{vv}^{1/2} - (\Theta_v^*)^{1/2}| \leq |\overline{\Theta}_{vv} - \Theta_v^*|^{1/2} \) and consequently, by assumption (A5),

\[
|\overline{\Delta}^{(j)}| = |\overline{\Theta}_{vv}^{1/2} - (\Theta_v^*)^{1/2}| = o_p(1).
\]

Invoking (A9) and the Lindeberg-Feller CLT, \( \left| \sum_{j=1}^{k} \sum_{i \in I_j} \Delta_{21,i}^{(j)} \right| = o_p(1) \). Similarly

\[
\left| \sum_{j=1}^{k} \sum_{i \in I_j} \Delta_{22,i}^{(j)} \right| \leq \max_{1 \leq j \leq k} \| \hat{\Theta}_v^{(j)} - \Theta_v^* \|_1 \left| \overline{\Delta}^{(j)} \right| \left( \Theta_v^* \Theta_v^* \right)^{-1/2} \sum_{j=1}^{k} \sum_{i \in I_j} \xi_{i,v}^{(j)} \right| = o_p(1).
\]

Combining all terms in the decomposition \( \left\lfloor A.5.7 \right\rfloor \) delivers the result.
(B1)-(B5) of Condition A.5.1 are used in the proofs of subsequent lemmas.

**Condition A.5.1.** (B1) \( \|w^*\|_1 \lesssim s_1, \|J^*\|_{\text{max}} < \infty \) and for any \( \delta \in (0,1) \),

\[
\mathbb{P}\left( \|\hat{\beta}^{\lambda}_{-v} - \beta^*_v\|_1 \gtrsim n^{-1/2} s \sqrt{\log(d/\delta)} \right) < \delta
\]

and

\[
\mathbb{P}\left( \|\hat{w} - w^*\|_1 \gtrsim n^{-1/2} s_1 \sqrt{\log(d/\delta)} \right) < \delta.
\]

(B2) For any \( \delta \in (0,1) \),

\[
\mathbb{P}\left( \|\nabla - v_{n,\ell_n}(\beta^*_v; \beta^*_v)\|_{\infty} \gtrsim n^{-1/2} \sqrt{\log(d/\delta)} \right) < \delta.
\]

(B3) Suppose \( \hat{\beta}^{\lambda}_{-v} \) satisfies (B1). Define

\[
H_v := (\nabla^2_{v,-v} \ll_n(\beta^*_v; \beta^*_v) - \hat{w}^T \nabla^2_{v,-v} \ll_n(\beta^*_v; \beta^*_v)) \cdot (\hat{\beta}^{\lambda}_{-v} - \beta^*_v).
\]

Then for \( \beta_{-v,\alpha} = \alpha \beta^*_v + (1 - \alpha) \hat{\beta}^{\lambda}_{-v} \) and for any \( \delta \in (0,1) \),

\[
\mathbb{P}\left( \sup_{\alpha \in [0,1]} |H_v| \gtrsim s_1 s \frac{\log(d/\delta)}{n} \right) < \delta.
\]

(B4) There exists a constant \( C > 0 \) such that \( C < I_{\theta|\gamma} < \infty \), and for \( v^* = (1, -w^* \top)^\top \), it holds that

\[
\frac{\sqrt{n} v^* \top \nabla \ell_n(\beta^*_v; \beta^*_v)}{\sqrt{v^* \top J^* v^*}} \sim N(0,1).
\]

(B5) For any \( \delta \), if there exists an estimator \( \tilde{\beta} = (\tilde{\beta}_v, \tilde{\beta}_v^\top) \) satisfying \( \|\tilde{\beta} - \beta^*\|_1 \leq C s n^{-1/2} \sqrt{\log(d/\delta)} \) with probability > \( 1 - \delta \), then

\[
\mathbb{P}\left( \|\nabla^2 \ell_n(\tilde{\beta}) - J^*\|_{\text{max}} \gtrsim n^{-1/2} \sqrt{\log(d/\delta)} \right) < \delta.
\]
The proof of Theorem 2.2.3 is an application of Lemma A.5.11. To apply this Lemma, we must first verify (B1) to (B4) of Condition A.5.1. We do this in Lemma A.5.8.

**Lemma A.5.8.** Under the requirements of Theorem 2.2.3, (B1) - (B4) of Condition A.5.1 are fulfilled.

**Proof. Verification of (B1).** As stated in Theorem 2.2.3, \( \|w^*\|_1 = O(s_1) \) and \( \|J^*\|_{\text{max}} < \infty \) by part (i) of Condition 2.2.3. The rest of (B1) follows from the proof of Lemma C.3 of [95].

**Verification of (B2).** Let \( X_i = (Q_i, Z^\top_i) \). Since \( \|\nabla_\gamma \ell_n(\beta^*)\|_{\infty} = \left\| -\frac{1}{n} \sum_{i=1}^{n} (Y_i - b'(X_i^\top \beta^*))Z_i \right\|_{\infty} \), since the product of a sub-Gaussian random variable and a bounded random variable is sub-Gaussian, and since \( E[\nabla_\gamma \ell_n(\beta^*)] = 0 \), we have by Condition 2.2.3, Bernstein’s inequality and the union bound

\[
P(\|\nabla_\gamma \ell_n(\beta^*)\|_{\infty} > t) < (d - 1) \exp\{-nt^2/M^2 \sigma^2_b\}.
\]

Setting \( 2(d - 1) \exp\{-nt^2/M^2 \sigma^2_b\} = \delta \) and solving for \( t \) delivers the result.

**Verification of (B3)** Let \( \beta^*_\alpha = (\theta^*, \gamma^\alpha) \) and decompose the object of interest as

\[
|\left( \nabla^2_{v,-v} \ell_n(\beta^*_v, \beta^*_{-v,\alpha}) - \hat{w}^\top \nabla^2_{v,-v} \ell_n(\beta^*_v, \beta^*_{-v,\alpha}) (\hat{\beta}^\lambda - \beta^*_{-v}) \right| \leq \sum_{t=1}^{5} |\Delta_t|, \quad (A.5.8)
\]

where the terms \( \Delta_1 - \Delta_5 \) are given by \( \Delta_1 = \nabla^2_{v,-v} \ell_n(\beta^*_\alpha) - \nabla^2_{v,-v} \ell_n(\beta^*), \)

\[
\Delta_2 = \nabla^2_{v,-v} \ell_n(\beta^*) - w^* \nabla^2_{v,-v} J^*_{-v,-v}, \quad \Delta_4 = w^* \nabla^2_{v,-v} \ell_n(\beta^*) - \nabla^2_{v,-v} \ell_n(\beta^*),
\]

\[
\Delta_3 = w^* \nabla^2_{v,-v} \ell_n(\beta^*), \quad \Delta_5 = (w^* - \hat{w}) \nabla^2_{v,-v} \ell_n(\beta^*_\alpha).
\]
We have the following bounds

\[ |\Delta_1| = \left| \frac{1}{n} \sum_{i=1}^{n} Z_i Z_i^\top (\hat{\beta}_v^\lambda - \beta_v^*) \left( \ell''_i (X_i^\top \beta_n^*) - \ell''_i (X_i^\top \beta^*) \right) \right| \]

\[ \leq \max_{1 \leq i \leq n} K_i \max_{1 \leq i \leq n} \|X_i\|_\infty \left| \frac{1}{n} Z \hat{\beta}_v - \beta_v^* \right|_2^2, \]

\[ |\Delta_2| \leq \left\| \nabla^2_{v,v} \ell_n (\hat{\beta}^\lambda) - J_{v,v}^* \right\|_\infty \hat{\beta}_{v,v} - \beta_{v,v}^* 1, \]

\[ |\Delta_3| \leq \|w\|_1 |J_{v,v}^* - \nabla^2_{v,v} \ell_n (\beta^*)| \max \|\hat{\beta}_{v,v} - \beta_{v,v}^*\|_1, \]

\[ |\Delta_4| = \left| w^\top \left( \nabla^2_{v,v} \ell_n (\beta^*) - \nabla^2_{v,v} \ell_n (\beta_v^*) \right) (\hat{\gamma}^\lambda - \lambda^*) \right| \]

\[ \leq \max_{1 \leq i \leq n} K_i \|w\|_1 \left| \frac{1}{n} Z \hat{\beta}_v - \beta_v^* \right|_2^2, \]

and \[ |\Delta_5| \leq \|w^* - \hat{w}\|_1 \|\nabla_{v,v} \ell_n (\beta_v^*)\|_{\text{max}} \|\hat{\beta}_v - \beta_v^*\|_1. \]

Let \( \epsilon = \delta / 5 \). Then by Condition 2.2.3 and Lemma A.5.4

\[ \mathbb{P} \left( |\Delta_1| \geq s_1 \log(d/\epsilon) \frac{n}{n} \right) < \epsilon \quad \text{and} \quad \mathbb{P} \left( |\Delta_4| \geq s_1 \log(d/\epsilon) \frac{n}{n} \right) < \epsilon. \]

Noting the \( \beta^* \) itself satisfies the requirements on \( \hat{\beta} \) in (B5), Lemma A.5.9 and Condition 2.1.1 together give

\[ \mathbb{P} \left( |\Delta_2| \geq s_1 \log(d/\epsilon) \frac{n}{n} \right) < \epsilon \quad \text{and} \quad \mathbb{P} \left( |\Delta_3| \geq s_1 \log(d/\epsilon) \frac{n}{n} \right) < \epsilon. \]

By (B1) verified above and noting that

\[ \left\| \nabla^2_{v,v} \ell_n (\beta_v^*) \right\|_{\text{max}} \leq \left\| \nabla^2_{v,v} \ell_n (\beta_v^*) - \nabla^2_{v,v} \ell_n (\beta^*) \right\|_{\text{max}} + \left\| \nabla^2_{v,v} \ell_n (\beta^*) \right\|_{\text{max}}, \]

146
the proof of Lemma A.5.9 delivers \( P\left( |\Delta_5| \gtrsim s_1 s \log(d/\epsilon)/n \right) < \epsilon \). Combining the bounds, we finally have

\[
P\left( \sup_{\alpha \in [0,1]} H_v \gtrsim s_1 s \frac{\log(d/\delta)}{n} \right) < \delta.
\]

**Verification of (B4).** See [95], proof of Lemma C.2. \( \square \)

In the following lemma, we verify (B5) under the same conditions.

**Lemma A.5.9.** Under Conditions 2.2.3 and 2.1.1, (B5) of Condition A.5.1 is fulfilled.

**Proof.** We obtain a tail probability bound for \( \Delta_1 \) and \( \Delta_2 \) in the decomposition

\[
\|\nabla^2 \ell_n(\hat{\beta}) - J^*\|_{\max} \leq \|\nabla^2 \ell_n(\hat{\beta}) - \nabla^2 \ell_n(\beta^*)\|_{\max} + \|\nabla^2 \ell_n(\beta^*) - J^*\|_{\max} = \Delta_1 + \Delta_2.
\]

For the control over \( \Delta_1 \), note that by Condition 2.2.3 (ii) and (iii),

\[
|[\nabla^2 \ell_n(\beta^*)]_{jk}| \leq |b''(X_i^\top \beta^*)| |X_{ij}X_{ik}| \leq U_3 M^2.
\]

Hence Hoeffding’s inequality and the union bound deliver

\[
P(\Delta_2 > t) = P\left( \|\nabla^2 \ell_n(\beta^*) - J^*\|_{\max} > t \right) \leq 2d^2 \exp\left\{ -\frac{nt^2}{8U_2^2 M^4} \right\}. \tag{A.5.9}
\]

For the control over \( \Delta_1 \), we have by Lemma A.5.5

\[
|[\nabla^2 \ell_n(\hat{\beta}) - \nabla^2 \ell_n(\beta^*)]_{jk}| = |(b''(X_i^\top \hat{\beta}) - b''(X_i^\top \beta^*)) X_{ij}X_{ik}|
\leq M^3 U_3 \|\hat{\beta} - \beta^*\|_1 \leq M^3 U_3 s \sqrt{n^{-1} \log(d/\delta)}
\]
with probability \(> 1 - \delta\). Hoeffding’s inequality and the union bound again deliver

\[
P(\Delta_1 > t) = P\left( \|\nabla_n^2 \ell_n(\beta) - \nabla_n^2 \ell_n(\beta^*)\| \max > t \right) \\
\leq 2d^2 \exp \left\{ - \frac{n^2t^2}{8U_3^2M^6s^2 \log(d/\delta)} \right\}.
\]

(A.5.10)

Combining the bounds from equations (A.5.9) and (A.5.10) we have

\[
P\left( \|\nabla \ell(\beta) - J^*\| \max > t \right) \\
\leq 2d^2 \left( \exp \left\{ - \frac{nt^2}{8U_3^2M^4} \right\} + \exp \left\{ - \frac{n^2t^2}{8U_3^2M^6s^2 \log(d/\delta)} \right\} \right).
\]

Setting each term equal to \(\delta/2\), solving for \(t\) and ignoring the relative magnitude of constants, we have \(t = U_3 \max\{n^{-1} \log(d/\delta), n^{-1/2} \sqrt{\log(d/\delta)}\} = U_3 n^{-1/2} \log(d/\delta)\), thus verifying (B5).

\(\square\)

**Lemma A.5.10.** For each \(j \in \{1, \ldots, k\}\), let \(\beta_{-v,\alpha_j} = \alpha_j \beta_{-v}^\lambda(D_j) + (1 - \alpha_j) \beta_{-v}^\star\), for some \(\alpha_j \in [0,1]\), where \(\beta_{-v}^\lambda(D_j)\) is defined in equation (2.1.2). Define

\[
\Delta_1^{(j)} = (\tilde{w}(D_j) - w^*)^\top \nabla_{-v} \ell_{nk}(\beta^*, \beta_{-v}^*) \quad \text{and} \\
\Delta_2^{(j)} = (\nabla_{-v} \ell_{nk}^{(j)}(\beta_{-v}^*, \beta_{-v,\alpha_j}) - \tilde{w}^\top \nabla_{-v} \ell_{nk}(\beta_{-v,\alpha_j}) - \tilde{w}^\top \nabla_{-v} \ell_{nk}(\beta_{-v}^*, \beta_{-v}) \beta_{-v}^\lambda - \beta_{-v}^\star).
\]

Under (B1) - (B3) of Condition A.5.1, \(k^{-1} \sum_{j=1}^k \Delta_1^{(j)} = o_p(n^{-1/2})\) and \(k^{-1} \sum_{j=1}^k \Delta_2^{(j)} = o_p(n^{-1/2})\) whenever \(k \ll d\) is chosen to satisfy \(k = o((s_1 \log d)^{-1} \sqrt{n})\).

**Proof.** By Hölder’s inequality,

\[
|\Delta_1^{(j)}| = |(w^* - \tilde{w}(D_j))^\top \nabla_{-v} \ell_{nk}(\beta_{-v}^*, \beta_{-v}^*)| \\
\leq \|\tilde{w}(D_j) - w^*\|_1 \|\nabla_{-v} \ell_{nk}(\beta_{-v}^*, \beta_{-v}^*)\|_{\infty},
\]

hence, for any \(t\),

\[
\{ |\Delta_1^{(j)}| > t \} \subseteq \{ \|\tilde{w}(D_j) - w^*\|_1 \|\nabla_{-v} \ell_{nk}(\beta_{-v}^*, \beta_{-v}^*)\|_{\infty} > t \}.
\]

148
Take $t = vq$ where $v = Cn^{-1/2}s_1\sqrt{k\log(d/\delta)}$ and $q = Cn^{-1/2}\sqrt{k\log(d/\delta)}$. Define two events $E_1$ and $E_2$ as following.

$$E_1 := \{\|\hat{w}(D_j) - w^*\|_1\|\nabla_{-v}^\ell(v,j)(\beta^*_v, \beta^*_{-v})\|_\infty > vq\},$$

$$E_2 := \{\|\hat{w}(D_j) - w^*\|_1 \leq 1\}.$$

Then we obtain that $P(E_1) = P(E_1 \cap E_2) + P(E_1 \cap E_2^c)$, which is not greater than $2\delta$ by (B1) and (B2) of Condition A.5.1. Hence the union bound delivers

$$P\left(\sum_{j=1}^k \Delta_1(j) > kvq\right) \leq P\left(\bigcup_{j=1}^k \{\Delta_1(j) > vq\}\right) \leq \sum_{j=1}^k P\left(\Delta_1(j) > vq\right) \leq 2k\delta = o(1)$$

for $\delta = o(k^{-1})$. Taking $\delta = k^{-1}$ for $\alpha > 0$ arbitrarily small in the definition of $v$ and $q$, the requirement is $ks_1 \log d = o(\sqrt{n})$ and $ks_1 \log k = o(\sqrt{n})$ for $\alpha > 0$ arbitrarily small. Since $k \ll d$, $k^{-1} \sum_{j=1}^k \Delta_1(j) = o_P(n^{-1/2})$ with $k = o((s_1 \log d)^{-1}\sqrt{n})$. Next, consider $|\Delta_2(j)| \leq \sup_{\alpha \in [0,1]} |G_v|$, where

$$G_v := (\nabla_{v,-v}^\ell(v,j)(\beta^*_v, \beta^*_{-v, \alpha}) - \hat{w}^\top \nabla_{v,-v}^\ell(v,j)(\beta^*_v, \beta^*_{-v, \alpha})) (\hat{\beta}^\lambda_v(D_j) - \beta^*_{-v}).$$

By (B3) of Condition A.5.1 $P(\Delta_2(j) \geq t) < \delta$ for $t \asymp s_1n^{-1}k\log(d/\delta)$, hence, proceeding in an analogous fashion to in the control over $k^{-1} \sum_{j=1}^k \Delta_1(j)$, we obtain

$$P\left(\sum_{j=1}^k \Delta_2(j) > kt\right) \leq P\left(\bigcup_{j=1}^k \Delta_2(j) > t\right) \leq \sum_{j=1}^k P\left(\Delta_2(j) > t\right) \leq k\delta = o(1)$$

for $\delta = o(k^{-1})$. Hence $k^{-1} \sum_{j=1}^k \Delta_2^{(j)} = o_P(n^{-1/2})$ with $k = o((s_1 \log d)^{-1} \cdot n^{3/2})$. Since $(s_1 \log d)^{-1}\sqrt{n} = o((s_1 \log d)^{-1}n^{3/2})$, $k^{-1} \sum_{j=1}^k (\Delta_1^{(j)} + \Delta_2^{(j)}) = o_P(n^{-1/2})$ requires $k = o((s_1 \log d)^{-1}\sqrt{n})$. \hfill \Box
Lemma A.5.11. Under (B1) - (B4) of Condition A.5.1 with \(k \ll d\) chosen to satisfy the scaling \(k = o\left((s \lor s_1) \log d\right)^{-1} \sqrt{n}\),

\[
\frac{1}{k} \sum_{j=1}^{k} \tilde{S}^{(j)}(\beta^*_v, \tilde{\beta}^\lambda_{-v}(D_j)) = \frac{1}{k} \sum_{j=1}^{k} S^{(j)}(\beta^*_v, \beta^*_u) + o_P(n^{-1/2}) \quad \text{and}
\]

\[
\lim_{n \to \infty} \sup_t \mathbb{P}\left(\frac{1}{k} \sum_{j=1}^{k} S^{(j)}(\beta^*_v, \beta^*_u) < t\right) - \Phi(t) \to 0.
\]

Proof. Recall

\[
S^{(j)}(\beta^*_v, \beta^*_u) = \nabla_v \ell n_k^j(\beta^*_v, \beta^*_u) - w^* \nabla_v \ell n_k^j(\beta^*_v, \beta^*_u).
\]

Through a mean value expansion of \(\tilde{S}^{(j)}(\beta^*_v, \tilde{\beta}^\lambda_{-v}(D_j))\) around \(\beta^*_v\), we have for each \(j \in \{1, \ldots, k\}\),

\[
\tilde{S}^{(j)}(\beta^*_v, \tilde{\beta}^\lambda_{-v}(D_j)) = \nabla_v \ell n_k^j(\beta^*_v, \tilde{\beta}^\lambda_{-v}(D_j)) - \hat{w}(D_j)\nabla_v \ell n_k^j(\beta^*_v, \tilde{\beta}^\lambda_{-v}(D_j))
\]

\[
= S^{(j)}(\beta^*_v, \beta^*_u) + \Delta_1^{(j)} + \Delta_2^{(j)},
\]

for some \(\beta_{-v, \alpha} = \alpha \tilde{\beta}_{-v}(D_j) + (1 - \alpha) \beta^*_v\), where

\[
\Delta_1^{(j)} = (w^* - \hat{w}(D_j))\nabla_v \ell n_k^j(\beta^*_v, \beta^*_u)
\]

\[
\Delta_2^{(j)} = \left[ \nabla^2_{v, -v} \ell n_k^j(\beta^*_v, \beta_{-v, \alpha}) - \hat{w}(D_j)\nabla^2_{v, -v} \ell n_k^j(\beta^*_v, \beta_{-v, \alpha}) \right](\tilde{\beta}^\lambda_{-v}(D_j) - \beta^*_v).
\]

Here \(h_v = \nabla^2_{v, -v} \ell n_k^j(\beta^*_v, \beta_{-v, \alpha}) - \hat{w}(D_j)\nabla^2_{v, -v} \ell n_k^j(\beta^*_v, \beta_{-v, \alpha})\). It follows that

\[
\frac{1}{k} \sum_{j=1}^{k} \tilde{S}^{(j)}(\beta^*_v, \tilde{\beta}^\lambda_{-v}(D_j)) = \frac{1}{k} \sum_{j=1}^{k} S^{(j)}(\beta^*_v, \beta^*_u) + \frac{1}{k} \sum_{j=1}^{k} \left(\Delta_1^{(j)} + \Delta_2^{(j)}\right)
\]

\[
= \frac{1}{k} \sum_{j=1}^{k} S^{(j)}(\theta^*, \gamma^*) + o_P(n^{-1/2}) \tag{A.5.11}
\]

150
by Lemma A.5.10 whenever \( k = o((s_1 \log d)^{-1} \sqrt{n}) \). Observe

\[
\sqrt{n} \left( k^{-1} \sum_{j=1}^{k} S^{(j)}(\beta_v^*, \beta_{-v}^*) \right) = \sqrt{n}(1, -w^*)^T \left( \frac{1}{k} \sum_{j=1}^{k} \nabla \phi^{(j)}_{n_k}(\beta_v^*, \beta_{-v}^*) \right) \quad \text{and}
\]

\[
J^*_{v|-v} = (1, -w^*)^T J(1, -w^*)^T.
\]

So \( \sqrt{n} \frac{1}{k} \sum_{j=1}^{k} S^{(j)}(\beta_v^*, \beta_{-v}^*) \sim N(0, J^*_{v|-v}) \) by Condition (B4). Similar to Corollary 2.2.2 we apply the Berry-Essen inequality to show that

\[
\sup_t |\mathbb{P}(\sqrt{n} \frac{1}{k} \sum_{j=1}^{k} S^{(j)}(\beta_v^*, \beta_{-v}^*) < t) - \Phi(t)| \to 0.
\]

\[\square\]

**Lemma A.5.12.** Under Condition (B1), for any \( \delta \in (0, 1) \),

\[
\mathbb{P} \left( \|w - w^*\|_1 > Cn^{-1/2}s_1 \sqrt{k \log(d/\delta)} \right) < k\delta,
\]

\[
\mathbb{P} \left( \|\beta_{-v} - \beta_{-v}^*\|_1 > Cn^{-1/2} s \sqrt{k \log(d/\delta)} \right) < k\delta.
\]

**Proof.** Set \( t = Cs_1 \sqrt{n^{-1}(k \log(d/\delta))} \) and note

\[
\mathbb{P} \left( \| \sum_{j=1}^{k} (\hat{w}(D_j) - w^*) \|_1 > kt \right) \leq \sum_{j=1}^{k} \mathbb{P} \left( \| \hat{w} - w^* \|_1 > t \right).
\]

by the union bound. Then by Condition (B1),

\[
\mathbb{P} \left( \| \hat{w} - w^* \|_1 > Cn^{-1/2}s_1 \sqrt{k \log(d/\delta)} \right) < k\delta.
\]

The proof of the second bound is analogous, setting \( t = Cs \sqrt{n^{-1}(k \log(d/\delta))} \). \[\square\]

**Lemma A.5.13.** Suppose (B5) of Condition A.5.1 is satisfied. For any \( \delta \), if there exists an estimator \( \hat{\beta} = (\hat{\beta}_v^T, \hat{\beta}_{-v}^T)^T \) satisfying \( \| \hat{\beta} - \beta^* \|_1 \leq Cs \sqrt{n^{-1} \log(d/\delta)} \) with
probability $1 - \delta$, then

$$
\mathbb{P} \left( \left\| \frac{1}{k} \sum_{j=1}^{k} \nabla^2 \ell^{(j)}(\tilde{\beta}) - J^* \right\|_{\max} > Cn^{-1/2}k\sqrt{\log(d/\delta)} \right) < k\delta.
$$

**Proof.** The proof follows from (B5) in Condition A.5.1 via an analogous argument to that of Lemma A.5.12, taking $t = C\sqrt{n^{-1}(k \log(d/\delta))}$. \qed

**Lemma A.5.14.** Suppose (B1)-(B5) of Condition A.5.1 are fulfilled. Then for any $k \ll d$ satisfying $k = o\left(\left(\left(s \lor s_1\right) \log d\right)^{-1} \sqrt{n}\right)$, $|\tilde{J}_{\theta|_v} - J^*_{v|_v}| = o_P(1)$.

**Proof.** Recall that $J^*_{v|_v} = J^*_{v,v} - J^*_{v, -v} J^*_{-v,v} J^*_{-v,v}$ and

$$
J_{v|_v} = \frac{1}{k} \sum_{j=1}^{k} \left( \nabla_{v,v} \ell^{(j)}(\beta_{v_k}, \beta_{-v}) - w^T \nabla^2_{v,v} \ell^{(j)}(\beta_{v_k}, \beta_{-v}) \right),
$$

so $|J_{v|_v} - J^*_{v|_v}| = \delta_1 + \delta_2$, where

$$
\Delta_1 = \frac{1}{k} \sum_{j=1}^{k} \left( \nabla_{v,v} \ell^{(j)}(\beta_{v_k}, \beta_{-v}) - J^*_{v,v} \right)
$$

and

$$
\Delta_2 = \left| w^T \left( \frac{1}{k} \sum_{j=1}^{k} \nabla^2_{v,v} \ell^{(j)}(\beta_{v_k}, \beta_{-v}) - w^* \cdot J^*_{v,v} \right) \right|.
$$

Let $\tilde{\beta} = (\beta_{v}, \beta_{-v})$ and note that $\|\tilde{\beta} - \beta^*\|_1$ satisfies the clause in (B5) of Condition A.5.1 by Lemma A.5.12 when $k = o\left(\left(\left(s \lor s_1\right) \log d\right)^{-1} \sqrt{n}\right)$. Hence $\Delta_1 = o_P(1)$ by Lemma A.5.13

$$
\Delta_2 \leq \left| (\overline{w} - w^*)^T \left( \frac{1}{k} \sum_{j=1}^{k} \nabla^2_{v,v} \ell^{(j)}(\beta_{v_k}, \beta_{-v}) - J^*_{v,v} \right) \right|_{\Delta_{21}}
$$

$$
+ \left| (\overline{w} - w^*)^T J^*_{v,v} \right|_{\Delta_{22}}
$$

$$
+ \left| w^*^T \left( \frac{1}{k} \sum_{j=1}^{k} \nabla^2_{v,v} \ell^{(j)}(\beta_{v_k}, \beta_{-v}) - J^*_{v,v} \right) \right|_{\Delta_{23}}.
$$

152
By the fact that \( \|J^*\|_{\text{max}} < \infty \) and \( \|w^*\|_1 \leq Cs_1 \) by (B1) of Condition A.5.1, an application of Lemmas A.5.12 and A.5.13 delivers

\[
\begin{align*}
\Delta_{21} & \leq \|\overline{w} - w^*\|_1 \frac{1}{k} \sum_{j=1}^{k} \nabla^2_{-v,v} \ell_{nk}(\beta^d, \overline{\beta}_{-v}) - J^*_{-v,v} \|_{\infty} = o_p(1), \\
\Delta_{22} & \leq \|\overline{w} - w^*\|_1 \|J^*_{-v,v}\|_{\infty} = o_p(1), \\
\Delta_{23} & \leq \frac{1}{k} \sum_{j=1}^{k} \nabla^2_{-v,v} \ell_{nk}(\beta^d, \overline{\beta}_{-v}) - J^*_{-v,v} \|_{\infty} \|w^*\|_1 = o_p(1)
\end{align*}
\]

for \( k = o((s_1 \log d)^{-1}n) \), a fortiori for \( k = o((s \vee s_1) \log d)^{-1} \sqrt{n} \). Hence \( |J_{v|-v} - J^*_{v|-v}| = o_p(1) \).

\[\boxed{}\]

\section*{A.6 Auxiliary lemmas for estimation}

In this section, we provide the proofs of the technical lemmas and theorems for the divide and conquer estimation. Using Lemma 2.5.1 we can derive Lemma A.6.1, which serves as a crucial step in establishing Lemma 2.3.1.

\begin{lemma}
Suppose Conditions 2.2.1 and 2.2.2 are fulfilled. Let \( \lambda \asymp \sqrt{k \log d/n} \) and \( \vartheta_1 \asymp \sqrt{k \log d/n} \). With \( k = o((s \log d)^{-1} \sqrt{n}) \), \( \sqrt{n}(\overline{\beta}^d - \beta^*) = Z + \Delta \), where

\[
Z = \frac{1}{\sqrt{k}} \sum_{j=1}^{k} \frac{1}{\sqrt{n_k}} M^{(j)} X^{(j)} \epsilon^{(j)} \quad \text{and} \quad \|\Delta\|_{\infty} = o_p(1).
\]

\end{lemma}

\begin{proof}[Proof of Lemma A.6.1] For notational convenience, we write \( \hat{\beta}_{\text{Lasso}}^{\lambda}(D_j) \) simply as \( \hat{\beta}^{\lambda}(D_j) \). Decompose \( \overline{\beta}^d - \beta^* \) as

\[
\begin{align*}
\overline{\beta}^d - \beta^* &= \frac{1}{k} \sum_{j=1}^{k} \left( \hat{\beta}^{\lambda}(D_j) - \beta^* + \frac{1}{n_k} M^{(j)} X^{(j)} (\beta^* - \hat{\beta}^{\lambda}(D_j)) \right) \\
&\quad + \frac{1}{k} \sum_{j=1}^{k} \frac{1}{n_k} M^{(j)} X^{(j)} \epsilon^{(j)} \\
&= \frac{1}{k} \sum_{j=1}^{k} (I - M^{(j)} \hat{\Sigma}^{(j)}) \left( \hat{\beta}^{\lambda}(D_j) - \beta^* \right) + \frac{1}{k} \sum_{j=1}^{k} \frac{1}{n_k} M^{(j)} X^{(j)} \epsilon^{(j)},
\end{align*}
\]

153
hence $\sqrt{n}(\beta^d - \beta^*) = Z + \Delta$, where

$$Z = \frac{1}{\sqrt{k}} \sum_{j=1}^{k} \frac{1}{\sqrt{n_k}} M^{(j)} X^{(j)}^T \epsilon^{(j)}$$

and

$$\Delta = \frac{1}{\sqrt{n}} \sum_{j=1}^{k} (I - M^{(j)} \hat{\Sigma}^{(j)}) \left( \hat{\beta}^\lambda (D_j) - \beta^* \right).$$

Defining $\Delta^{(j)} = (I - M^{(j)} \hat{\Sigma}^{(j)}) \left( \hat{\beta}^\lambda (D_j) - \beta^* \right)$, we have

$$\|\Delta^{(j)}\|_\infty \leq \|\Delta^{(j)}\|_1 \leq \|M^{(j)} \hat{\Sigma}^{(j)} - I\|_{\max} \|\hat{\beta}^\lambda (D_j) - \beta^*\|_1$$

by Hölder’s inequality, where $\|I - M^{(j)} \hat{\Sigma}^{(j)}\|_{\max} \leq \vartheta_1$ by the definition of $M^{(j)}$ and, for $\lambda = C\sigma^2 \sqrt{\log d/n_k}$,

$$\mathbb{P} \left( \|\hat{\beta}^\lambda (D_j) - \beta^*\|_1^2 > C \frac{s^2 \log(2d)}{n_k} + t \right) \leq \exp \left( - \frac{cn_k t}{s^2 \sigma^2} \right) \quad (A.6.1)$$

by [14]. We thus bound the expectation of the $\ell_1$ loss by

$$\mathbb{E} \left[ \|\hat{\beta}^\lambda (D_j) - \beta^*\|_1^2 \right] \leq \frac{2Cs^2 \log(2d)}{n_k} + \int_0^\infty \exp \left( - \frac{cn_k t}{s^2 \sigma^2} \right) dt \quad (A.6.2)$$

Define the event $\mathcal{E}^{(j)} := \{ \|\hat{\beta}^\lambda (D_j) - \beta^*\|_1 \leq s \sqrt{C \log(2d)/n_k} \}$ for $j = 1, \ldots, k$. Then

$$\|\Delta^{(j)}\|_\infty \leq \Delta_1^{(j)} + \Delta_2^{(j)} + \Delta_3^{(j)}$$

where

$$\Delta_1^{(j)} = \|M^{(j)} \hat{\Sigma}^{(j)} - I\|_{\max} \|\hat{\beta}^\lambda (D_j) - \beta^*\|_1 \mathbb{1} \{ \mathcal{E}^{(j)} \}$$

$$\Delta_2^{(j)} = \|M^{(j)} \hat{\Sigma}^{(j)} - I\|_{\max} \|\hat{\beta}^\lambda (D_j) - \beta^*\|_1 \mathbb{1} \{ \mathcal{E}^{(j)^c} \}$$

$$\Delta_3^{(j)} = \mathbb{E}[\|M^{(j)} \hat{\Sigma}^{(j)} - I\|_{\max} \|\hat{\beta}^\lambda (D_j) - \beta^*\|_1 \mathbb{1} \{ \mathcal{E}^{(j)^c} \}]$$

and

$$\Delta_3^{(j)} = \mathbb{E}[\|M^{(j)} \hat{\Sigma}^{(j)} - I\|_{\max} \|\hat{\beta}^\lambda (D_j) - \beta^*\|_1].$$
Consider $\Delta_1^{(j)}$, $\Delta_2^{(j)}$ and $\Delta_3^{(j)}$ in turn. By Hoeffding’s inequality, we have for any $t > 0$,

$$\mathbb{P}\left(\frac{1}{k} \sum_{j=1}^k \Delta_1^{(j)} > t\right) \leq \exp\left( -\frac{n_k k t^2}{s^2 \vartheta_1^2 \log(2d)} \right) \leq \exp\left( -\frac{n_k k t^2}{C s^2 \log^2(2d)} \right). \tag{A.6.3}$$

By Markov’s inequality,

$$\mathbb{P}\left(\frac{1}{k} \sum_{j=1}^k \Delta_2^{(j)} > t\right) \leq \frac{\sum_{j=1}^k \mathbb{E}[\Delta_2^{(j)}]}{kt} \leq 2t^{-1} \mathbb{E}\left[\|M^{(j)} \bar{\Sigma}^{(j)} - I\|_{\max} | \hat{\beta}^{(j)}(D_j) - \beta^*|_1 \mathbb{1}\{E^{(j)c}\}\right] \tag{A.6.4}$$

$$\leq 2t^{-1} \vartheta_1 \sqrt{\mathbb{E}\left[\|\hat{\beta}^{(j)}(D_j) - \beta^*\|_1^2\right]} \mathbb{P}(E^{(j)c}) \leq C t^{-1} \sqrt{\frac{\log d}{n_k}} \cdot \frac{s^2 \log(2d)}{n_k} d^{-c} \leq C t^{-1} s n_k^{-1} d^{-c/2} \log d, \tag{A.6.5}$$

where the penultimate inequality follows from Jensen’s inequality. Finally, by Jensen’s inequality again,

$$\frac{1}{k} \sum_{j=1}^k \Delta_3^{(j)} = \mathbb{E}\left[\|M^{(j)} \hat{\Sigma}^{(j)} - I\|_{\max} | \hat{\beta}^{(j)}(D_j) - \beta^*|_1 \right] \leq \vartheta_1 \sqrt{\mathbb{E}\left[\|\hat{\beta}^{(j)}(D_j) - \beta^*\|_1^2\right]} \leq C s \log d \frac{d}{n_k}. \tag{A.6.6}$$

Combining (A.6.3), (A.6.5) and (A.6.6),

$$\mathbb{P}\left(\|\Delta\|_\infty > 3C \sqrt{n} \cdot \frac{s \log d}{n_k}\right) \leq \sum_{u=1}^3 \mathbb{P}\left(\frac{1}{k} \sum_{j=1}^k \Delta^{(j)} > C \sqrt{n} \cdot \frac{s \log d}{n_k}\right) \leq \exp(-ckn) + d^{-c/2} \to 0, \tag{A.6.7}$$

and taking $k = o\left(\left(s \log d\right)^{-1} \sqrt{n}\right)$ delivers $\|\Delta\|_\infty = o_p(1)$. □
Proof of Theorem A.1.1.

\[ \beta - \hat{\beta} = \frac{1}{k} \sum_{j=1}^{k} ((X^{(j)})^\top X^{(j)})^{-1}(X^{(j)})^\top Y^{(j)} - (X^\top X)^{-1}X^\top Y \]

\[ = \frac{1}{k} \sum_{j=1}^{k} \left( \frac{(X^{(j)})^\top X^{(j)}/n_k}{(X^\top X/n)^{-1}} - (X^\top X/n)^{-1}X^{(j)}^\top e^{(j)}/n_k \right) \]

\[ = \frac{1}{k} \sum_{j=1}^{k} \left( \frac{(X^{(j)})^\top X^{(j)}/n_k}{(X^\top X/n)^{-1}} - \Sigma^{-1} \right) X^{(j)}^\top e^{(j)}/n_k \]

\[ + \left( \Sigma^{-1} - (X^\top X/n)^{-1} \right) X^\top e/n. \quad (A.6.8) \]

For simplicity, denote \( X^{(j)}^\top X^{(j)}/n_k \) by \( S^{(j)}_X \), \( X^\top X/n \) by \( S_X \), \( (S_X^{-1})^{-1} - (\Sigma)^{-1} \) by \( D_1^{(j)} \) and \( (\Sigma)^{-1} - S_X^{-1} \) by \( D_2 \). For any \( \tau \in \mathbb{R} \), define an event \( \mathcal{E}^{(j)} = \{ \| (S_X^{-1})^{-1}\|_2 \leq 2/C_{\min} \} \cap \{ \| S_X - \Sigma \|_2 \leq (\delta_1 \vee \delta_2^1) \} \) for all \( j = 1, \ldots, k \), where \( \delta_1 = C_1 \sqrt{d/n_k} + \tau/\sqrt{n_k} \), and an event \( \mathcal{E} = \{ \| (S_X^{-1})^{-1}\|_2 \leq 2/C_{\min} \} \cap \{ \| S_X - \Sigma \|_2 < (\delta_2 \vee \delta_2^2) \} \), where \( \delta_2 = C_1 \sqrt{d/n} + \tau/\sqrt{n} \). Note that by Lemma A.6.2 and A.6.5, the probability of both \( (\mathcal{E}^{(j)})^c \) and \( \mathcal{E}^c \) are very small. In particular

\[ \mathbb{P}(\mathcal{E}^c) \leq \exp(-cn) + \exp(-c_1 \tau^2) \quad \text{and} \quad \mathbb{P}((\mathcal{E}^{(j)})^c) \leq \exp(-cn/k) + \exp(-c_1 \tau^2). \]

Then, letting \( \mathcal{E}_0 := \bigcap_{j=1}^{k} \mathcal{E}^{(j)} \), an application of the union bound and Lemma A.6.9 delivers

\[ \mathbb{P}\left( \| \beta - \hat{\beta} \|_2 > t \right) \leq \mathbb{P}\left( \left\{ \| \frac{1}{k} \sum_{j=1}^{k} (X^{(j)}D_1^{(j)})^\top e^{(j)}/n_k \|_2 > t/2 \right\} \cap \mathcal{E}_0 \right) \]

\[ + \mathbb{P}\left( \left\{ \| XD_2^\top e/n \|_2 > t/2 \right\} \cap \mathcal{E} \right) + \mathbb{P}(\mathcal{E}_0^c) + \mathbb{P}(\mathcal{E}^c) \]

\[ \leq 2 \exp\left( d \log(6) - \frac{t^2 C_{\min}^3 n_k}{32 C_3 \sigma_1^2 \delta_1^2} \right) + k \exp(-cn/k) + (k + 1) \exp(-c_1 \tau^2). \]
When \( d \to \infty \) and \( \log n = o(d) \), choose \( \tau = \sqrt{d/c_1} \) and \( \delta_1 = O(\sqrt{kd/n}) \). Then there exists a constant \( C \) such that

\[
P\left( \| \mathbf{\bar{b}} - \hat{\beta} \|_2 > C \frac{\sqrt{kd}}{n} \right) \leq (k + 3) \exp(-d) + k \exp(- \frac{cn}{k}).
\]

Otherwise choose \( \tau = \sqrt{\log n/c_1} \) and \( \delta_1 = O(\sqrt{k \log n/n}) \). Then there exists a constant \( C \) such that

\[
P\left( \| \mathbf{\bar{b}} - \hat{\beta} \|_2 > C \frac{\sqrt{k \log n}}{n} \right) \leq \frac{k + 3}{n} + k \exp(- \frac{cn}{k}).
\]

Overall, we have

\[
P\left( \| \mathbf{\bar{b}} - \hat{\beta} \|_2 > C \frac{\sqrt{k (d \vee \log n)}}{n} \right) \leq ck \exp(-(d \vee \log n)) + k \exp(-cn/k),
\]

which leads to the final conclusion.

Proof of Corollary A.1.1. Define an event

\[
\mathcal{E} = \{ \| \mathbf{\bar{b}}^d - \beta^* \|_\infty \leq 2C \sqrt{\log d/n} \},
\]

then by the condition on the minimal signal strength and Lemma 2.3.1 for some constant \( C' \) we have

\[
P\left( \| \mathbf{\bar{b}}^o - \hat{\beta}^o \|_2 > C' \frac{\sqrt{k(s \vee \log n)}}{n} \right)
\]

\[
\leq P\left( \left\{ \| \mathbf{\bar{b}}^o - \hat{\beta}^o \|_2 > C' \frac{\sqrt{k(s \vee \log n)}}{n} \right\} \cap \mathcal{E} \right) + P(\mathcal{E}^c)
\]

\[
\leq P\left( \left\{ \| \mathbf{\bar{b}}^o - \hat{\beta}^o \|_2 > C' \frac{\sqrt{k(s \vee \log n)}}{n} \right\} \cap \mathcal{E} \right) + c/d
\]

\[
\leq ck \exp(-(s \vee \log n)) + k \exp(-cn/k) + c/d.
\]

157
where $\overline{\beta} = \frac{1}{k} \sum_{j=1}^{k} (X_S^{(j)\top} X_S^{(j)})^{-1} X_S^{(j)\top} Y^{(j)}$, which is the average of the oracle estimators on the subsamples. Then the conclusion can be easily validated.

\[ \frac{\partial^2 \ell_n(\beta)}{\partial \beta} = \frac{1}{n} X^\top D(X\beta)X, \quad S_X := \frac{1}{n} X^\top X. \]

For any $j = 1, \ldots, k$, $\hat{\beta}^{(j)}$ satisfies

$$\nabla \ell_{nk}^{(j)}(\hat{\beta}^{(j)}) = \frac{1}{nk} X^{(j)\top} (Y^{(j)} - \mu(X^{(j)} \hat{\beta}^{(j)})) = 0.$$

Through a Taylor expansion of the left hand side at the point $\beta = \beta^*$, we have

$$\frac{1}{nk} X^{(j)\top} (Y^{(j)} - \mu(X^{(j)} \beta^*)) = S^{(j)}(\hat{\beta}^{(j)} - \beta^*) - r^{(j)} = 0,$$

where the remainder term $r^{(j)}$ is a $d$ dimensional vector with $g$th component

$$r_g^{(j)} = \frac{1}{6nk} (\hat{\beta}^{(j)} - \beta^*)^\top \nabla^2 \ell_{nk}[X_{g}^{(j)\top} X^{(j)} \mu(X^{(j)} \beta)](\hat{\beta}^{(j)} - \beta^*)$$

$$= \frac{1}{6nk} (\hat{\beta}^{(j)} - \beta^*)^\top X^{(j)\top} \text{diag}\{X_{g}^{(j)\top} \circ \mu''((X^{(j)} \hat{\beta}^{(j)}))\} X^{(j)}(\hat{\beta}^{(j)} - \beta^*),$$

where $\tilde{\beta}^{(j)}$ is in a line segment between $\hat{\beta}^{(j)}$ and $\beta^*$. It therefore follows that

$$\hat{\beta}^{(j)} = \beta^* + (S^{(j)})^{-1}[X^{(j)\top} (Y^{(j)} - \mu(X^{(j)} \beta^*)) + nk r^{(j)}].$$

A similar equation holds for the global MLE $\hat{\beta}$:

$$\hat{\beta} = \beta^* + S^{-1}[X^\top (Y - \mu(X \beta^*)) + nr],$$

Proof of Theorem A.2.1. The following notation is used throughout the proof.

$S(\beta) := \nabla^2 \ell_n(\beta) = \frac{1}{n} X^\top D(X\beta)X, \quad S^{(j)} := \frac{1}{nk} X^{(j)\top} X^{(j)}.$

$S(\beta) := \nabla^2 \ell_{nk}(\beta) = \frac{1}{nk} X^{(j)\top} D(X^{(j)} \beta)X^{(j)}, \quad S_X := \frac{1}{n} X^\top X.$
where for \( g = 1, \ldots, d \),
\[
    r_g = \frac{1}{6n} (\hat{\beta} - \beta^*)^\top X^\top \text{diag}\{X_g \circ \mu''((X\tilde{\beta}^{(j)})})X(\hat{\beta} - \beta^*).
\]

Therefore we have
\[
    \frac{1}{k} \sum_{j=1}^k \hat{\beta}^{(j)} - \beta = \frac{1}{k} \sum_{j=1}^k \{(S^{(j)})^{-1} - \Sigma^{-1}\} X^{(j)^\top} (Y^{(j)} - \mu(X^{(j)}\beta^*))
\]
\[
    - \{S^{-1} - \Sigma^{-1}\} X^\top (Y - \mu(X\beta^*)) + R = B + R,
\]

where \( R = (1/k) \sum_{j=1}^k (S^{(j)})^{-1} r^{(j)} - S^{-1} r \). We next derive stochastic bounds for \( \|B\|_2 \) and \( \|R\|_2 \) respectively, but to study the appropriate threshold, we introduce the following events with probability that approaches one under appropriate scaling. For \( j = 1, \ldots, k \) and \( \kappa, \tau, t > 0 \),
\[
    \mathcal{E}^{(j)} := \{(S^{(j)})^{-1}_2 \leq 2/C_{\min}\} \cap \left\{ \frac{\|S^{(j)} - \Sigma\|_2}{\delta_1 \vee \delta_2^2} \leq 1 \right\} \cap \left\{ \|S^{(j)}_x\|_2 \leq 2C_{\max}\right\};
\]
\[
    \mathcal{E} := \{\|S^{-1}\|_2 \leq 2/L_{\min}\} \cap \left\{ \|S - \Sigma\|_2 \leq (\delta_2 \vee \delta_2^2) \right\} \cap \left\{ \|S_x\|_2 \leq 2C_{\max}\right\};
\]
\[
    \mathcal{F}^{(j)} := \left\{ \|\hat{\beta}^{(j)} - \beta^*\|_2 > t \right\}, \quad \mathcal{F} := \{\|\hat{\beta} - \beta^*\|_2 > t \};
\]

where \( \delta_1 = C_1 \sqrt{d/n_k + \tau/\sqrt{n_k}} \) and \( \delta_2 = C_1 \sqrt{d/n_k + \tau/\sqrt{n}} \). Denote the intersection of all the above events by \( \mathcal{A} \). Note that Condition 2.2.3 implies that \( \sqrt{b'(X_i^\top \beta)}X_i \) are i.i.d. sub-Gaussian vectors, so by Lemmas A.6.2, A.6.5, A.6.4 and A.6.11 we have
\[
    \mathbb{P}(\mathcal{A}^c) \leq (2k + 1) \exp \left(-\frac{cn}{k}\right) + (k + 1) \exp(-c_1 \tau^2)
\]
\[
    + 2k \exp \left(d \log 6 - \frac{nC_{\min}^2 L_{\min}^2 t^2}{2^{11} C_{\max} U_2 \phi k}\right)
\]

We first consider the bounded design, i.e., Condition 2.2.3 (ii). In order to bound \( \|R\|_2 \), we first derive an upper bound for \( r^{(j)}_g \). Under the event \( \mathcal{A} \), by Lemma A.5.5
we have
\[
\max_{1 \leq g \leq d, 1 \leq j \leq k} r_g^{(j)} \leq \frac{1}{3} MU_3 C_{\text{max}}^2 t^2 \quad \text{and} \quad \max_{1 \leq g \leq d} r_g \leq \frac{1}{3} MU_3 C_{\text{max}} t^2.
\]

It follows that, under \(A\),
\[
\|R\|_2 \leq \frac{2}{3} M \sqrt{dU_3 C_{\text{max}}} t^2. \quad (A.6.9)
\]

Note that \(B\) is very similar to the RHS of Equation (A.6.8). Now we use essentially the same proof strategy as in the OLS part to bound \(\|B\|_2\). Following similar notations as in OLS, we denote \((S^{(j)})^{-1} - \Sigma^{-1}\) by \(D_1^{(j)}\), \(S^{-1} - \Sigma^{-1}\) by \(D_2\), \(Y^{(j)} - \mu(X^{(j)}\beta^*)\) by \(\epsilon^{(j)}\) and \(Y - \mu(X\beta^*)\) by \(\epsilon\). For concision, we relegate the details of the proof to Lemma \(A.6.10\) which delivers the following stochastic bound on \(\|B\|_2\).

\[
P(\{\|B\|_2 > t_1\} \cap A) \leq 2 \exp \left( d \log(6) - \frac{C_{\min}^4 L_{\min}^2 n t_1^2}{128 \phi U_1 C_{\text{max}} (\delta_1 \vee \delta_1^2)^2} \right). \quad (A.6.10)
\]

Combining Equation (A.6.10) with (A.6.9) leads us to the following inequality.

\[
P \left( \|\beta - \hat{\beta}\|_2 > \frac{2}{3} M \sqrt{dU_3 C_{\text{max}}} t^2 + t_1 \right) \leq (2k + 1) \exp \left( -\frac{cn}{k} \right) + (k + 1) \exp(-c_1 \tau^2) + (k + 1) \exp \left( d \log 6 - \frac{C_{\min}^2 L_{\min}^2 n t_1^2}{211 C_{\text{max}} U_2 \phi k} \right)
\]
\[
+ 2 \exp \left( d \log 6 - \frac{C_{\min}^4 L_{\min}^2 t_1^2}{128 \phi U_2 C_{\text{max}} (\delta_1 \vee \delta_1^2)^2} \right).
\]

Choose \(t = t_1 = \sqrt{d/n_k}\) and, when \(d \gg \log n\), choose \(\tau = \sqrt{d/c_1}\) and \(\delta_1 = O(\sqrt{kd/n})\). Then there exists a constant \(C > 0\) such that

\[
P \left( \|\beta - \hat{\beta}\|_2 > \frac{C k d^{3/2}}{n} \right) \leq (2k + 1) \exp(-\frac{cn}{k}) + 2(k + 1) \exp(-d).
\]
When it is not true that \( d \gg \log n \), choose \( \tau = \sqrt{\log n/c_1} \) and \( \delta = O(\sqrt{k \log n/n}) \). Then there exists a constant \( C > 0 \) such that

\[
\mathbb{P}\left( \|\hat{\beta} - \beta\|_2 > C \frac{k \sqrt{d \log n}}{n} \right) \leq (2k + 1) \exp\left(-\frac{cn}{k}\right) + \frac{k + 3}{n}.
\]

Overall, we have

\[
\mathbb{P}\left( \|\hat{\beta} - \hat{\beta}^o\|_2 > C \frac{k \sqrt{d \vee (d \log n)}}{n} \right) \leq c k \exp\left(-cn/k\right) + C k \exp\left(-c d \vee \log n\right),
\]

which leads to the final conclusion.

**Proof of Corollary A.2.1.** Define an event

\[
\mathcal{E} = \{\|\hat{\beta}^d - \beta^*\|_{\infty} \leq 2C \sqrt{\log d/n}\},
\]

then by the conditions of Corollary A.2.1 and results of Lemma 2.3.2 and Theorem A.2.1

\[
\mathbb{P}\left( \|\hat{\beta} - \hat{\beta}^o\|_2 > C' \frac{k \sqrt{s \vee \log n}}{n} \right) \leq \mathbb{P}(\{\|\hat{\beta} - \hat{\beta}^o\|_2 > C' \frac{k \sqrt{s \vee \log n}}{n}\} \cap \mathcal{E}) + \mathbb{P}(\mathcal{E}^c)
\]

\[
\leq \mathbb{P}(\{\|\hat{\beta}^o - \hat{\beta}^o\|_2 > C' \frac{k \sqrt{s \vee \log n}}{n}\} \cap \mathcal{E}) + c/d
\]

\[
\leq c k \exp(-s \vee \log n) + k \exp(-cn/k) + c/d.
\]

where \( \hat{\beta}^o = \frac{1}{k} \sum_{j=1}^k \hat{\beta}_j(D_j), \hat{\beta}^o(D_j) = \operatorname{argmax}_{\beta \in \mathbb{R}^d, \beta_{sc}=0} \ell(j)(\beta) \) and \( C' \) is a constant. Then it is not hard to see that the final conclusion is true.

**Proof of Lemma 2.3.1.** According to Lemma A.6.1, we have \( \sqrt{n}(\hat{\beta}^d - \beta^*) = \mathbf{Z} + \Delta \), where \( \mathbf{Z} = \frac{1}{\sqrt{k}} \sum_{j=1}^k \frac{1}{\sqrt{n}} \mathbf{M}^{(j)} \mathbf{X}^{(j)} \mathbf{\hat{\epsilon}}^{(j)} \). In (A.6.7), we prove that \( \|\Delta\|_{\infty}/\sqrt{n} \leq C s k \log d/n \) with probability larger than \( 1 - \exp(-ckn) - d^{-c/2} \geq 1 - c_1/d \) for some constant \( c_1 \). Since \( \hat{\beta}^{d} \) is a special case of \( \hat{\beta}^d \) when \( k = 1 \), we also have \( \sqrt{n}(\hat{\beta}^{d} -
\( \beta^* = Z + \Delta_1 \), where (A.6.7) gives \( \|\Delta\|_\infty / \sqrt{n} \leq Cs \log d/n \). Therefore, we have \( \|\beta - \hat{\beta}\|_\infty \leq Cs \log d/n \) with high probability.

It only remains to bound the rate of \( \|Z\|_\infty / \sqrt{n} \). By Condition 2.2.2, conditioning on \( \{X_i\}_{i=1}^n \), we have for any \( \ell = 1, \ldots, d \),

\[
\mathbb{P}\left( \frac{|Z\ell|}{\sqrt{n}} > t \Big| \{X_i\}_{i=1}^n \right) = \mathbb{P}\left( \frac{1}{n} \sum_{j=1}^k M_\ell^{(j)\top} X^{(j)\top} \epsilon^{(j)} > t \Big| \{X_i\}_{i=1}^n \right) \\
\leq 2 \exp\left( -\frac{cnt^2}{\kappa^2 Q_\ell} \right),
\]

where \( \kappa \) is the variance proxy of \( \epsilon \) defined in Condition 2.2.2 and

\[
Q_\ell = \frac{1}{n} \sum_{j=1}^k \|X^{(j)}M_\ell^{(j)\top}\|_2^2.
\]

Let \( Q_{\text{max}} = \max_{1 \leq \ell \leq d} Q_\ell \). Applying the union bound to (A.6.11), we have

\[
\mathbb{P}\left( \frac{\|Z\|_\infty}{\sqrt{n}} > t \Big| \{X_i\}_{i=1}^n \right) \leq \mathbb{P}\left( \max_{1 \leq \ell \leq d} |Z\ell|/\sqrt{n} > t \Big| \{X_i\}_{i=1}^n \right) \\
\leq \sum_{\ell=1}^d \mathbb{P}\left( |Z\ell|/\sqrt{n} > t \Big| \{X_i\}_{i=1}^n \right) \leq 2d \exp\left( -\frac{cnt^2}{\kappa^2 Q_{\text{max}}} \right),
\]

Let \( t = \sqrt{2 \kappa^2 Q_{\text{max}} \log d/(cn)} \), then with conditional probability \( 1 - 2/d \),

\[
\|Z\|_\infty / \sqrt{n} \leq \sqrt{\kappa^2 Q_{\text{max}} \log d/(cn)}.
\]

The last step is to bound \( Q_{\text{max}} \). By the definition of \( Q_\ell \), we have

\[
Q_\ell = \frac{1}{k} \sum_{j=1}^k M_\ell^{(j)\top} \Sigma^{(j)} M_\ell^{(j)} \leq \frac{1}{k} \sum_{j=1}^k [\Omega]\ell^{\top} \hat{\Sigma}^{(j)} [\Omega]_\ell \\
= \frac{1}{k} \sum_{j=1}^k \frac{1}{n} \sum_{i \in D_j} (X_i^{\top}[\Omega]_\ell)^2 = \frac{1}{n} \sum_{i=1}^n (X_i^{\top}[\Omega]_\ell)^2,
\]

162
where \( \Omega = \Sigma^{-1} \). The inequality is due to the fact that \( M^{(j)}_\ell \) is the minimizer in (2.2.4). By condition (2.2.2) and the connection between sub-Gaussian and subexponential distributions, the random variable \((X_i^\top \Omega_\ell)^2\) satisfies

\[
\sup_{q \geq 1} q^{-1}(\mathbb{E}|(X_i^\top \Omega_\ell)^2|^q)^{1/q} \leq 4\kappa^2\Omega_{\ell\ell}.
\]

Therefore, by Bernstein’s inequality for subexponential random variables, we have

\[
P\left(\left|\frac{1}{n} \sum_{i=1}^n (X_i^\top \Omega_\ell)^2 - \mathbb{E}[X_i^\top \Omega_\ell]^2\right| > t\right) \leq 2 \exp\left(-c\left(\frac{nt^2}{16\kappa^4\Omega_{\ell\ell}^2}\right) \wedge \left(\frac{nt}{4\kappa^2\Omega_{\ell\ell}}\right)\right).
\]

Applying the union bound again, we have

\[
P\left(\max_{1 \leq \ell \leq d} \left|\frac{1}{n} \sum_{i=1}^n (X_i^\top \Omega_\ell)^2 - \mathbb{E}[X_i^\top \Omega_\ell]^2\right| > 8\kappa^2\Omega_{\ell\ell} \sqrt{\frac{\log d}{cn}}\right) \leq 2 \delta / d.
\]

Therefore, with probability \(1 - 2/d\), there exist a constant \(C_1\) such that

\[
Q_{\max} = \max_{1 \leq \ell \leq d} Q_\ell \leq \max_{1 \leq \ell \leq d} \left(\frac{1}{n} \sum_{i=1}^n (X_i^\top \Omega_\ell)^2 - \mathbb{E}[X_i^\top \Omega_\ell]^2\right) + \mathbb{E}[X_i^\top \Omega_\ell]^2
\]

\[
\leq 8\kappa^2\Omega_{jj} \sqrt{\frac{\log d}{cn}} + \Omega_{jj} \leq C_1,
\]

where the last inequality is due to Condition 2.2.1. By (A.6.12), we have with probability \(1 - 4/d\), \(\|Z\|_\infty / \sqrt{n} \leq \sqrt{\kappa^2C_1 \log d/(cn)}\). Combining this with the result on \(\|\Delta\|_\infty\) delivers the rate in the lemma.
**Proof of Lemma 2.3.2.** The strategy of proving this lemma is similar to the proof of Lemma 2.3.1. In the proof of Lemma A.5.6 and Theorem 2.2.2 we have shown that

\[
(\bar{\beta}^d - \beta^*) = -\frac{1}{k} \sum_{j=1}^{k} \tilde{\Theta}^{(j)\top} \nabla \ell_n^{(j)}(\beta^*) + \frac{1}{k} \sum_{j=1}^{k} \Delta_j,
\]

where the remainder term for each \( j \) is

\[
\Delta_j = \left( I - \tilde{\Theta}^{(j)\top} \frac{1}{n_k} \sum_{i \in I_j} b''(\tilde{\eta}_i) X_i X_i^\top \right) \left( \hat{\beta}^{\lambda}(D_j) - \beta^* \right)
\]

and \( \tilde{\eta}_i = tX_i^\top \beta^* + (1-t)X_i^\top \hat{\beta}^{\lambda}(D_j) \) for some \( t \in (0, 1) \). We bound \( \Delta_j \) by decomposing it into three terms:

\[
\| \Delta_j \|_\infty \leq \left\| (I - \Theta^* \frac{1}{n_k} \sum_{i \in I_j} b''(X_i^\top \beta^*) X_i X_i^\top) (\hat{\beta}^{\lambda}(D_j) - \beta^*) \right\|_\infty
\]

\[
+ \left\| \Theta^* \frac{1}{n_k} \sum_{i \in I_j} (b''(X_i^\top \hat{\beta}^{\lambda}(D_j)) - b''(X_i^\top \beta^*)) X_i X_i^\top) (\hat{\beta}^{\lambda}(D_j) - \beta^*) \right\|_\infty
\]

\[
+ \left\| (\tilde{\Theta}^{(j)} - \Theta^*) \frac{1}{n_k} \sum_{i \in I_j} b''(X_i^\top \hat{\beta}^{\lambda}(D_j)) X_i X_i^\top \right\|_\infty \left( \hat{\beta}^{\lambda}(D_j) - \beta^* \right) \|_\infty.
\]

By Hoeffding’s inequality and Condition 2.2.1, the first term is bounded by

\[
|I_1| \leq \left\| I - \Theta^* \frac{1}{n_k} \sum_{i \in I_j} b''(X_i^\top \beta^*) X_i X_i^\top \right\|_\infty \left\| \hat{\beta}^{\lambda}(D_j) - \beta^* \right\|_1 \leq C_{sk} \frac{\log d}{n}, \quad (A.6.14)
\]
with probability $1 - c/d$. By Condition 2.2.3 (iii), Condition 2.2.4 (iv) and Lemma A.5.4, we have with probability $1 - c/d$,

\[ |I_2| \leq \max_i \| \Theta^* X_i \|_\infty \frac{1}{n_k} \sum_{i \in I_j} U_3 [X_i(\hat{\beta}^\lambda(D_j) - \beta^*)]^2 \leq C \frac{k \log d}{n}. \quad (A.6.15) \]

Finally, we bound $I_3$ by with probability $1 - c/d$,

\[
|I_3| \leq \frac{\sqrt{U_2}}{n_k} \sqrt{\sum_{i \in I_j} b''(X_i^\top \hat{\beta}^\lambda(D_j)) [X_i^\top (\hat{\Theta}^(j) - \Theta^*])^2} \sqrt{\sum_{i \in I_j} [X_i(\hat{\beta}^\lambda(D_j) - \beta^*)]^2} \leq C \frac{(s_1 \vee s) k \log d}{n}, \quad (A.6.16)
\]

where the last inequality is due to Lemma A.5.4 and Lemma C.4 of [95].

Combining (A.6.14) - (A.6.16) and applying the union bound, we have

\[
\bigg\| \frac{1}{k} \sum_{j=1}^k \Delta_j \bigg\|_\infty \leq \max_j \| \Delta_j \|_\infty = O_P \left( \frac{(s_1 \vee s) k \log d}{n} \right). \]

Therefore, we only need to bound the infinity norm of the leading term $T$. By Condition 2.2.4 and equation (A.5.3), we have with probability $1 - c/d$,

\[
\max_{1 \leq j \leq k} \max_{1 \leq v \leq d} \| \hat{\Theta}^{(j)}_v - \Theta^*_v \|_1 \leq C s_1 \sqrt{\log d/n} \quad \text{and} \quad \bigg\| \frac{1}{k} \sum_{j=1}^k \nabla \ell_{nk}^{(j)}(\beta^*) \bigg\|_\infty \leq C \sqrt{\log d/n}. \quad (A.6.17)
\]

This, together with Condition 2.2.3 and Condition 2.2.4 give the bound,

\[
\|T\|_\infty \leq \left( M \max_{v,j} \| \hat{\Theta}^{(j)}_v - \Theta^*_v \|_1 + \max_i \| X_i^\top \Theta^* \|_\infty \right) \| \frac{1}{k} \sum_{j=1}^k \nabla \ell_{nk}^{(j)}(\beta^*) \|_\infty \leq C \left( \sqrt{\log d/n} + \frac{s_1 \log d}{n} \right),
\]

165
with probability $1 - c/d$. Since $\tilde{\beta}^d$ is a special case of $\overline{\beta}^d$ when $k = 1$, the proof of the lemma is complete.

We borrow the following two lemmas on concentration inequalities from [124] for the proof later.

**Lemma A.6.2.** Suppose $X$ is a $n \times d$ matrix that has independent sub-Gaussian rows $\{X_i\}_{i=1}^n$. Denote $\mathbb{E}(X_iX_i^\top)$ by $\Sigma$, then we have

$$
\mathbb{P} \left( \left\| \frac{1}{n} X^\top X - \Sigma_X \right\|_2 \geq (\delta \lor \delta^2) \right) \leq \exp(-c_1 t^2),
$$

where $t \geq 0$, $\delta = C_1 \sqrt{d/n} + t/\sqrt{n}$ and $C_1$ and $c_1$ are both constants depending only on $\|X_i\|_{\psi_2}$.

**Proof.** See Theorem 5.39 and Remark 5.40 in [124].

**Lemma A.6.3.** (Bernstein-type inequality) Let $X_1, \ldots, X_n$ be independent centered sub-exponential random variables, and $M = \max_{1 \leq i \leq n} \|X_i\|_{\psi_1}$. Then for every $a = (a_1, \ldots, a_n) \in \mathbb{R}^n$ and every $t \geq 0$, we have

$$
\mathbb{P} \left( \sum_{i=1}^n a_iX_i \geq t \right) \leq \exp \left( -C_2 \min \left( \frac{t^2}{M^2 \|a\|_2^2}, \frac{t}{M \|a\|_\infty} \right) \right).
$$

**Proof.** See Proposition 5.16 in [124].

**Lemma A.6.4.** Suppose $X$ is a $n \times d$ matrix that has independent sub-gaussian rows $\{x_i\}_{i=1}^n$. If $\lambda_{\text{max}}(\Sigma) \leq C_{\text{max}}$ and $d \ll n$, then for all $M > C_{\text{max}}$, there exists a constant $c > 0$ such that when $n$ and $d$ are sufficiently large,

$$
\mathbb{P} \left( \left\| \frac{1}{n} X^\top X \right\|_2 \geq M \right) \leq \exp(-cn).
$$
Proof. Apply Lemma A.6.2 with $t = \sqrt{cn/c_1}$, where $(\sqrt{c/c_1} \lor c/c_1) < M - C_{\max}$, and it follows that

$$
\mathbb{P}\left( \frac{1}{n} \mathbf{X}^\top \mathbf{X} - \Sigma \geq (\delta \lor \delta^2) \right) \leq \exp(-cn).
$$

Since $d \ll n$, we obtain $(\delta \lor \delta^2) \to \sqrt{c/c_1}$, which completes the proof. \qed

**Lemma A.6.5.** Suppose $\mathbf{X}$ is a $n \times d$ matrix that has independent sub-Gaussian rows $\{\mathbf{X}_i\}_{i=1}^n$. $\mathbb{E}\mathbf{X}_i = \mathbf{0}$, $\lambda_{\min}(\Sigma) \geq C_{\min} > 0$ and $d \ll n$. For all $m < C_{\min}$, there exists a constant $c > 0$ such that when $n$ and $d$ are sufficiently large,

$$
\mathbb{P}\left( \left\| \left( \frac{1}{n} \mathbf{X}^\top \mathbf{X} \right)^{-1} \right\|_2 \geq \frac{1}{m} \right) = \mathbb{P}\left( \lambda_{\min} \left( \frac{1}{n} \mathbf{X}^\top \mathbf{X} \right) \leq m \right) \leq \exp(-cn).
$$

**Proof.** It is easy to check the following inequality. For any two symmetric and semi-definite $d \times d$ matrices $A$ and $B$, we have

$$
\lambda_{\min}(A) \geq \lambda_{\min}(B) - \|A - B\|_2,
$$

because for any vector $\mathbf{x}$ satisfying $\|\mathbf{x}\|_2 = 1$, we have

$$
\|A\mathbf{x}\|_2 = \|B\mathbf{x} + (A - B)\mathbf{x}\|_2 \geq \|B\mathbf{x}\|_2 - \|(A - B)\mathbf{x}\|_2 \geq \lambda_{\min}(B) - \|A - B\|_2.
$$

Then it follows that

$$
\mathbb{P}\left( \left\| \left( \frac{1}{n} \mathbf{X}^\top \mathbf{X} \right)^{-1} \right\|_2 \geq \frac{1}{m} \right) = \mathbb{P}\left( \lambda_{\min} \left( \frac{1}{n} \mathbf{X}^\top \mathbf{X} \right) \leq m \right)
\leq \mathbb{P}\left( \left\| \frac{1}{n} \mathbf{X}^\top \mathbf{X} - \Sigma \right\|_2 \geq C_{\min} - m \right) \leq \exp(-cn),
$$

where $c$ satisfies $(\sqrt{c/c_1} \lor c/c_1) < C_{\min} - m$ and the last inequality is an application of Lemma A.6.2 with $t = \sqrt{cn/c_1}$.

\qed
Lemma A.6.6. (Hoeffding-type Inequality). Let \(X_1, \ldots, X_n\) be independent centered sub-Gaussian random variables, and let \(K = \max_i \|X_i\|_{\psi_2}\). Then for every \(a = (a_1, \ldots, a_n) \in \mathbb{R}^n\) and every \(t > 0\), we have
\[
P\left(\left\|\sum_{i=1}^n a_i X_i\right\| \geq t\right) \leq e \cdot \exp\left(-\frac{ct^2}{K^2\|a\|_2^2}\right).
\]

Lemma A.6.7. (Sub-exponential is sub-Gaussian squared). A random variable \(X\) is a sub-Gaussian if and only if \(X^2\) is sub-exponential. Moreover,
\[
\|X\|_{\psi_2}^2 \leq \|X^2\|_{\psi_1} \leq 2\|X\|_{\psi_2}^2.
\]

Proof. See Lemma 5.14 in [124].

Lemma A.6.8. Let \(X_1, \ldots, X_n\) be independent centered sub-Gaussian random variables. Let \(\kappa = \max_i \|X_i\|_{\psi_2}\) and \(\sigma^2 = \max_i \mathbb{E}X_i^2\). Suppose \(\sigma^2 > 1\), then we have
\[
P\left(\frac{1}{n} \sum_{i=1}^n X_i^2 > 2\sigma^2\right) \leq \exp\left(-C_2 \frac{\sigma^2 n}{\kappa^2}\right).
\]

Proof. Combining Lemma A.6.3 and Lemma A.6.7 yields the result.

Lemma A.6.9. Following the same notation as in the beginning of Proof of Theorem A.1.1,
\[
P\left(\left\|\frac{1}{k} \sum_{j=1}^k (X^{(j)} D^{(j)}_1)^T \epsilon^{(j)}\right\|_2 > \frac{t}{2}\right) \cap \mathcal{E}_0 \right) \leq 6^d \exp\left(-\frac{t^2 C_{\min}^3 n}{32C_3^2s_1^2(\delta_1 \lor \delta_2^3)^2}\right)
\]
and
\[
P\left(\left\|XD_2\right\|_2 > t/2\right) \cap \mathcal{E} \right) \leq \exp\left(d \log(6) - \frac{t^2 C_{\min}^3 n}{32C_3^2s_1^2(\delta_2 \lor \delta_2^3)^2}\right).
\]
Proof.

\[
\mathbb{E} \left( \exp \left( \lambda (D_1^{(j)}v)^\top (X^{(j)}\epsilon / n_k) \right) \mid X^{(j)} \right) = \prod_{i=1}^{n_k} \mathbb{E} \left( \exp \left( \left( \frac{\lambda X_i^{(j)}}{n_k} \right)^\top (D^{(j)}v) \epsilon_i \right) \mid X^{(j)} \right) \leq \exp \left( C_3 \lambda^2 s_1^2 \sum_{i=1}^{n_k} \frac{A_i^{(j)}}{n} \right),
\]

(A.6.18)

\[
\mathbb{E} \left( \exp \left( \lambda (D_2v)^\top (X^\top \epsilon / n) \right) \mid X \right) = \prod_{i=1}^{N} \mathbb{E} \left( \exp \left( \left( \lambda X_i / N \right)^\top (D_2v) \epsilon_i \right) \mid X \right) \leq \exp \left( C_3 \lambda^2 s_1^2 \sum_{i=1}^{N} A_i^2 / n^2 \right),
\]

(A.6.19)

where we write \( A_i^{(j)} \) and \( A_i \) in place of \( (X_i^{(j)})^\top D_1^{(j)}v \) and \( (X_i)^\top D_2v \) respectively \( C_3 \) is an absolute constant, and the last inequality holds because \( \epsilon_i \) are sub-Gaussian.

Next we provide an upper bound on \( \sum_{i=1}^{n_k} (A_i^{(j)})^2 \) and \( \sum_{i=1}^{n} A_i^2 \). Note that

\[
\sum_{i=1}^{n} (A_i^{(j)})^2 = v^\top D_1^{(j)}X^\top XD_1^{(j)}v
\]

\[
= v^\top ((S_X^{(j)})^{-1} - (\Sigma)^{-1})n_kS_X^{(j)}((S_X^{(j)})^{-1} - (\Sigma)^{-1})v
\]

\[
= n_k v^\top \Sigma^{-1}(\Sigma - S_X^{(j)})(S_X^{(j)})^{-1}(\Sigma - S_X^{(j)})\Sigma^{-1}v,
\]

and similarly,

\[
\sum_{i=1}^{n} A_i^2 = nv^\top \Sigma^{-1}(\Sigma - S_X)(S_X)^{-1}(\Sigma - S_X)\Sigma^{-1}v.
\]

For any \( \tau \in \mathbb{R}, \) define the event \( \mathcal{E}^{(j)} = \{ \|(S_X^{(j)})^{-1}\|_2 \leq 2/C_{\min} \} \cap \{ \|S_X^{(j)} - \Sigma\|_2 \leq (\delta_1 \lor \delta_1^2) \} \) for all \( j = 1, \ldots, k, \) where \( \delta_1 = C_1 \sqrt{d/n_k} + \tau / \sqrt{n}, \) and the event \( \mathcal{E} = \{ \|(S_X)^{-1}\|_2 \leq 2/C_{\min} \} \cap \{ \|S_X - \Sigma\|_2 < (\delta_2 \lor \delta_2^2) \} \), where \( \delta_2 = C_1 \sqrt{d/n} + \tau / \sqrt{n}. \) On \( \mathcal{E}^{(j)} \) and \( \mathcal{E}, \) we have respectively

\[
\sum_{i=1}^{n_k} (A_i^{(j)})^2 \leq \frac{2n_k}{C_{\min}^3} (\delta_1 \lor \delta_1^2)^2 \text{ and } \sum_{i=1}^{n} A_i^2 \leq \frac{2n}{C_{\min}^3} (\delta_2 \lor \delta_2^2)^2.
\]
Therefore from Equation (A.6.18) and (A.6.19) we obtain

$$\mathbb{E} \left( \exp(\lambda(D_1^j v)^\top (X^{(j)^\top} \epsilon^{(j)}/n_k)) \mathbb{1}\{\mathcal{E}^{(j)}\} \right) \leq \exp \left( \frac{2C_3 \lambda^2 s_1^2}{C_{\min}^3 n_k} (\delta_1 \lor \delta_1^2)^2 \right)$$

and

$$\mathbb{E} \left( \exp(\lambda(D_2 v)^\top (X^\top \epsilon/n)) \mathbb{1}\{\mathcal{E}\} \right) \leq \exp \left( \frac{2C_3 \lambda^2 s_1^2}{C_{\min}^3 n} (\delta_2 \lor \delta_2^2)^2 \right).$$

In addition, according to Lemma A.6.2 and A.6.5 the probability of both $\mathcal{E}^{(j)}$ and $\mathcal{E}'$ are very small. More specifically,

$$\mathbb{P}(\mathcal{E}^c) \leq \exp(-cn) + \exp(-c_1 \tau^2) \quad \text{and} \quad \mathbb{P}(\mathcal{E}'^c) \leq \exp(-cn/k) + \exp(-c_1 \tau^2).$$

Let $\mathcal{E}_0 := \bigcap_{j=1}^k \mathcal{E}^{(j)}$. An application of the Chernoff bound trick leads us to the following inequality.

$$\mathbb{P} \left( \left\{ \frac{1}{k} \sum_{j=1}^k (D_1^j v)^\top (X^{(j)^\top} \epsilon^{(j)}/n_k) > t/2 \right\} \cap \mathcal{E}_0 \right)$$

$$\leq \exp(-\lambda t/2) \prod_{j=1}^k \mathbb{E} \left( \exp \left( \frac{\lambda}{k} (D_1^j v)^\top (X^{(j)^\top} \epsilon^{(j)}/n_k) \right) \mathbb{1}\{\mathcal{E}^{(j)}\} \right)$$

$$\leq \exp \left( -\lambda t/2 + \frac{2C_3 \lambda^2 s_1^2}{C_{\min}^3 n} (\delta_1 \lor \delta_1^2)^2 \right).$$

Minimize the right hand side by $\lambda$, then we have

$$\mathbb{P} \left( \left\{ \frac{1}{k} \sum_{j=1}^k (D_1^j v)^\top (X^{(j)^\top} \epsilon^{(j)}/n_k) > t/2 \right\} \cap \mathcal{E}_0 \right) \leq \exp \left( \frac{t^2 C_3^3 n}{32 C_3^3 s_1^2 (\delta_1 \lor \delta_1^2)^2} \right).$$
Consider the $1/2$–net of $\mathbb{R}^p$, denoted by $\mathcal{N}(1/2)$. Again it is known that $|\mathcal{N}(1/2)| < 6^p$. Using the maximal inequality, we have

\[
\mathbb{P}\left(\left\{ \left\| \frac{1}{k} \sum_{j=1}^{k} (X^{(j)}D^{(j)}_1)^\top \epsilon^{(j)}/n_k \right\|_2 > t/2 \right\} \cap \mathcal{E}_0 \right) \\
= \sup_{\|v\|_2 = 1} \mathbb{P}\left(\left\{ \frac{1}{k} \sum_{j=1}^{k} (D^{(j)}_1 v)^\top (X^{(j)}^\top \epsilon^{(j)})/n_k > t/2 \right\} \cap \mathcal{E}_0 \right) \\
\leq \sup_{v \in \mathcal{N}(1/2)} \mathbb{P}\left(\left\{ \frac{1}{k} \sum_{j=1}^{k} (D^{(j)}_1 v)^\top (X^{(j)}^\top \epsilon^{(j)})/n_k > t/4 \right\} \cap \mathcal{E}_0 \right) \\
\leq \exp\left( d \log(6) - \frac{t^2 C_3^3 n}{32 C_3 s^2 (\delta_1 \vee \delta_2^2)^2} \right).
\]

Proceeding in an analogous fashion, we obtain

\[
\mathbb{P}\left(\left\{ \left\| (XD_2)^\top \epsilon/n \right\|_2 > t/2 \right\} \cap \mathcal{E} \right) \leq \exp\left( d \log(6) - \frac{t^2 C_3^3 n}{32 C_3 s^2 (\delta_2 \vee \delta_2^2)^2} \right).
\]

Lemma A.6.10. Following the same notation as in the proof of Theorem A.2.1,

\[
\mathbb{P}(\|B\|_2 > t_1 \cap \mathcal{A}) \leq 2 \exp\left( d \log(6) - \frac{C_4^4 L_{\min}^2 n t_1^2}{128 \phi U_2 C_{\max}(\delta_1 \vee \delta_1^2)^2} \right).
\]

Proof. By Lemma A.5.2 for any $\lambda \in \mathbb{R}$ and $v$ such that $\|v\|_2 = 1$, we have

\[
\mathbb{E}\left( \exp(\lambda(D^{(j)}_1 v)^\top (X^{(j)}^\top \epsilon^{(j)}/n_k)) \mid X^{(j)} \right) \\
= \prod_{i=1}^{n_k} \mathbb{E}\left( \exp((\lambda X^{(j)}_i n_k)^\top (D^{(j)}_1 v) \epsilon_i) \mid X^{(j)} \right) \leq \exp\left( \phi U_2 \sum_{i=1}^{n_k} (A^{(j)}_i)^2/n_k^2 \right)
\]

171
and

\[
\mathbb{E} \left( \exp(\lambda (D_2 v)\top (X^\top \epsilon/n)) \mid X \right) = \prod_{i=1}^{n} \mathbb{E} \left( \exp((\lambda X_i/n)\top (D_2 v)\epsilon_i) \mid X \right) \leq \exp \left( \phi U \lambda^2 \sum_{i=1}^{n} A_i^2/n^2 \right),
\]

where we write \( A_i^{(j)} \) and \( A_i \) in place of \((X_i^{(j)}\top D_1^{(j)} v)\) and \((X_i \top D_2 v)\) respectively. Next we give a upper bound on \( \sum_{i=1}^{n_k} (A_i^{(j)})^2 \) and \( \sum_{i=1}^{n} A_i^2 \). Note that

\[
\sum_{i=1}^{n_k} (A_i^{(j)})^2 = v\top D_1^{(j)} X\top XD_1^{(j)} v
\]

\[
= v\top ((S^{(j)})^{-1} - \Sigma^{-1}) n S_X ((S^{(j)})^{-1} - \Sigma^{-1}) v
\]

\[
= n v\top \Sigma^{-1} (\Sigma - S^{(j)}) (S^{(j)})^{-1} S_X (S^{(j)})^{-1} (\Sigma - S^{(j)}) \Sigma^{-1} v.
\]

Similarly,

\[
\sum_{i=1}^{n} A_i^2 = n v\top \Sigma^{-1} (\Sigma - S) S^{-1} S_X S^{-1} (\Sigma - S) \Sigma^{-1} v.
\]

On \( \mathcal{E}^{(j)} \) and \( \mathcal{E} \), we have respectively

\[
\sum_{i=1}^{n_k} (A_i^{(j)})^2 \leq \frac{8C_{\max} n_k}{C_{\min}^4 L_{\min}^2} (\delta_1 \lor \delta_1^2)^2 \quad \text{and} \quad \sum_{i=1}^{n} A_i^2 \leq \frac{8C_{\max} n}{C_{\min}^4 L_{\min}^2} (\delta_2 \lor \delta_2^2)^2.
\]

Then it follows that

\[
\mathbb{E} \left( \exp(\lambda (D_1^{(j)} v)\top (X^{(j)}\top \epsilon^{(j)}/n_k)) \mathbb{1}\{\mathcal{E}^{(j)}\} \right) \leq \exp \left( \frac{8\phi U C_{\max} \lambda^2}{C_{\min}^4 L_{\min}^2 n_k} (\delta_1 \lor \delta_1^2)^2 \right)
\]

and

\[
\mathbb{E} \left( \exp(\lambda (D_2 v)\top (X^\top \epsilon/n)) \mathbb{1}\{\mathcal{E}\} \right) \leq \exp \left( \frac{8\phi U C_{\max} \lambda^2}{C_{\min}^4 L_{\min}^2 n} (\delta_2 \lor \delta_2^2)^2 \right).
\]

Now we follow exactly the same steps as in the OLS part. Denote \( \cap_{j=1}^{k} \mathcal{E}_j \) by \( \mathcal{E}_0 \). An application of the Chernoff bound technique and the maximal inequality leads us to
the following inequality.

\[
\mathbb{P}\left( \left\{ \left\| \frac{1}{k} \sum_{j=1}^{k} (X^{(j)} D^{(j)}_1)^\top \epsilon^{(j)} / n_k \right\|_2 > t/2 \right\} \cap \mathcal{E}_0 \right) 
\leq \exp \left( d \log(6) - \frac{C_{\min}^4 L_{\min}^2 n t^2}{128 \phi U_2 C_{\max}(\delta_1 \vee \delta_2^2)^2} \right).
\]

and

\[
\mathbb{P}\left( \left\{ \left\| (X D_2)^\top \epsilon / n \right\|_2 > t/2 \right\} \cap \mathcal{E} \right) \leq \exp \left( d \log(6) - \frac{C_{\min}^4 L_{\min}^2 n t^2}{128 \phi U_2 C_{\max}(\delta_1 \vee \delta_2^2)^2} \right).
\]

We have thus derived an upper bound for \( \|B\|_2 \) that holds with high probability. Specifically,

\[
\mathbb{P}(\|B\|_2 > t_1 \cap A) \leq \mathbb{P}\left( \left\{ \left\| \frac{1}{k} \sum_{j=1}^{k} (X^{(j)} D_1^{(j)})^\top \epsilon^{(j)} / n_k \right\|_2 > t_1/2 \right\} \cap \mathcal{E}_0 \right) 
+ \mathbb{P}\left( \left\{ \left\| \frac{(X D_2)^\top \epsilon}{n} \right\|_2 > t_1/2 \right\} \cap \mathcal{E} \right) \leq 2 \cdot 6^d \exp \left( - \frac{C_{\min}^4 L_{\min}^2 n t_1^2}{128 \phi U_2 C_{\max}(\delta_1 \vee \delta_2^2)^2} \right).
\]

Lemma A.6.11. Under Condition 2.2.3 for \( \tau \leq L_{\min} / (8 M C_{\max} U_3 \sqrt{d}) \) and sufficiently large \( n \) and \( d \) we have

\[
\mathbb{P}(\| \hat{\beta} - \beta^* \|_2 > \tau) \leq \exp \left( d \log 6 - \frac{n C_{\min}^2 L_{\min}^2 \tau^2}{2^{11} C_{\max} U_2 \phi} \right) + 2 \exp(-cn).
\]

Proof. The notation is that introduced in the proof of Theorem A.2.1. We further define \( \Sigma(\beta) := \mathbb{E}(b^\top (X^\top \beta) XX^\top) \) as well as the event \( \mathcal{H} := \{ \ell_n(\beta^*) > \max_{\beta \in \partial \mathcal{B}_\tau} \ell_n(\beta) \} \), where \( \mathcal{B}_\tau = \{ \beta : \| \beta - \beta^* \|_2 \leq \tau \} \). Note that as long as the event \( \mathcal{H} \) holds, the MLE falls in \( \mathcal{B}_\tau \), therefore the proof strategy involves showing that \( \mathbb{P}(\mathcal{H}) \) approaches 1 at
certain rate. By the Taylor expansion,

\[ \ell_n(\beta) - \ell_n(\beta^*) = (\beta - \beta^*)^\top v - \frac{1}{2}(\beta - \beta^*)^\top S(\bar{\beta})(\beta - \beta^*) = A_1 + A_2, \]

where \( S(\beta) = (1/n)X^\top D(X\beta)X \), \( \bar{\beta} \) is some vector between \( \beta \) and \( \beta^* \), \( v = (1/n)X^\top(Y - \mu(X\beta^*)) \), \( A_1 = (\beta - \beta^*)^\top v - (1/2)(\beta - \beta^*)^\top S(\beta^*)(\beta - \beta^*) \) and \( A_2 = -(1/2)(\beta - \beta^*)^\top (S(\bar{\beta}) - S(\beta^*))(\beta - \beta^*) \).

Define the event \( \mathcal{E} := \{\lambda_{\min}[S(\beta^*)] \geq L_{\min}/2\} \), where \( L_{\min} \) is the same constant in Condition 2.2.3. Note that by Condition 2.2.3 (ii), \( \sqrt{b''(X_i^\top \beta)X_i} \) is a sub-Gaussian random vector. Then by Condition 2.2.3 (iii) and Lemma A.6.5 for sufficiently large \( n \) and \( d \) we have \( \mathbb{P}(\mathcal{E}^c) \leq \exp(-cn) \). Therefore on the event \( \mathcal{E} \),

\[ A_1 \leq \tau(\|v\|_2 - \frac{L_{\min}}{4}\tau). \]

We next show that, under an appropriate choice of \( \tau \), \( |A_2| < L_{\min}\tau^2/8 \) with high probability. We first consider Condition 2.2.3 (ii). Define \( \mathcal{F} := \{\|X^\top X/n\|_2 \leq 2C_{\max}\} \). By Lemma A.6.4 we have \( \mathbb{P}(\mathcal{F}^c) \leq \exp(-cn) \). By Lemma A.5.5 on the event \( \mathcal{F} \), we have

\[ A_2 \leq \max_{1 \leq i \leq n} |b''(X_i^\top \bar{\beta}) - b''(X_i^\top \beta^*)|C_{\max}\tau^2 \]

\[ \leq MU_3\sqrt{d}\|\bar{\beta} - \beta^*\|_2 \cdot C_{\max}\tau^2 \]

\[ \leq MC_{\max}U_3\sqrt{d}\tau^3 \leq \frac{L_{\min}\tau^2}{8}, \]

where the last inequality holds if we choose \( \tau \leq L_{\min}/(8MC_{\max}U_3\sqrt{d}) \). Now we obtain the following probabilistic upper bound on \( \mathcal{H}^c \), which we later prove to be negligible.

\[ \mathbb{P}(\mathcal{H}^c) \leq \mathbb{P}(\mathcal{H}^c \cap \mathcal{E} \cap \mathcal{F}) + \mathbb{P}(\mathcal{E}^c) + \mathbb{P}(\mathcal{F}^c) \]

\[ \leq \mathbb{P}\left(\left\{ \|v\|_2 \geq \frac{L_{\min}\tau}{8}\right\} \cap \mathcal{E} \cap \mathcal{F} \right) + \mathbb{P}(\mathcal{E}^c) + \mathbb{P}(\mathcal{F}^c). \] (A.6.20)
Since each component of \( v \) is a weighted average of i.i.d. random variables, the effect of concentration tends to make \( \|v\|_2 \) very small with large probability, which inspires us to study the moment generating function and apply the Chernoff bound technique. By Lemma A.5.2 for any constant \( u \in \mathbb{R}^d, \|u\|_2 = 1 \) and let \( a_i = u^\top X_i \), then we have for any \( t \in \mathbb{R}, \)

\[
E(\exp(t\langle u, v \rangle) | X) = \prod_{i=1}^{n} E\left(\exp\left(\frac{ta_i}{n}(Y_i - \mu_{X_i^\top \beta})\right) | X \right) \\
\leq \exp\left(\frac{\phi U_2 t^2}{2n^2} \sum_{i=1}^{n} a_i^2 \right) \\
= \exp\left(\frac{\phi U_2 t^2}{2n} \cdot u^\top X^\top X u \right).
\]

It follows that

\[
E \exp(t\langle u, v \rangle 1_{\{E \cap F\}}) \leq \exp\left(\frac{\phi C_{\max} U_2 t^2}{2n} \right).
\]

By the Chernoff bound technique, we obtain

\[
P(\{\langle u, v \rangle > \epsilon \} \cap E \cap F) \leq \exp\left(-\frac{n\epsilon^2}{8C_{\max} U_2 \phi}\right).
\]

Consider a 1/2—net of \( \mathbb{R}^d \), denoted by \( \mathcal{N}(1/2) \). Since

\[
\|v\|_2 = \max_{\|u\|_2 = 1} \langle u, v \rangle \leq 2 \max_{u \in \mathcal{N}(1/2)} \langle u, v \rangle,
\]

it follows that

\[
P(\{\|v\|_2 > \frac{L_{\min} \tau}{8}\} \cap E \cap F) \leq \mathbb{P}\left(\left\{ \max_{u \in \mathcal{N}(1/2)} \langle u, v \rangle > \frac{L_{\min} \tau}{16} \right\} \cap E \cap F \right) \\
\leq 6^d \exp\left(\frac{-nL_{\min}^2 \tau^2}{2^{10} \phi C_{\max} U_2}\right) \\
= \exp\left(d \log 6 - \frac{nC_{\min}^2 L_{\min}^2 \tau^2}{2^{11} \phi C_{\max} U_2} \right).
\]

Finally combining the result above with Equation (A.6.20) delivers the conclusion. \( \square \)
Remark A.6.1. Simple calculation shows that when $d = o(\sqrt{n})$, $\|\hat{\beta} - \beta^*\|_2 = O_p(\sqrt{d/n})$. When $d$ is a fixed constant, $\|\hat{\beta} - \beta^*\|_2 = O_p(\sqrt{1/n})$. 
Appendix B

Proofs for Chapter 3

B.1 Proof of Lemma 3.3.1

Proof. It follows from concentration of sample covariance matrix (Lemma B.10.1) that \( \| \hat{\Sigma}^{(1)} - \Sigma \|_{\psi_1} \lesssim \lambda_1 \sqrt{r/n} \). By the variant of Davis-Kahan theorem in [131],

\[
\rho(\hat{V}_K^{(1)}, V_K) = \| \hat{V}_K^{(1)} \hat{V}_K^{(1)^\top} - V_K V_K^\top \|_F = \sqrt{2} \sin \Theta(\hat{V}_K^{(1)}, V_K) \lesssim \sqrt{K} \| \hat{\Sigma}^{(1)} - \Sigma \|_2 / \Delta.
\]

Hence

\[
\| \rho(\hat{V}_K^{(1)}, V_K) \|_{\psi_1} \lesssim \sqrt{K} \| \hat{\Sigma}^{(1)} - \Sigma \|_2 / \Delta \lesssim \kappa \sqrt{K r/n}.
\]

By Jensen’s inequality,

\[
\| \Sigma^* - V_K V_K^\top \|_F = \| \mathbb{E}(\hat{V}_K^{(1)} \hat{V}_K^{(1)^\top}) - V_K V_K^\top \|_F \leq \mathbb{E} \| \hat{V}_K^{(1)} \hat{V}_K^{(1)^\top} - V_K V_K^\top \|_F
\]

\[
= \mathbb{E} \rho(\hat{V}_K^{(1)}, V_K) \leq \| \rho(\hat{V}_K^{(1)}, V_K) \|_{\psi_1}.
\]
Therefore,
\[
\left\| \| \hat{\mathbf{V}}_K^{(1)} \hat{\mathbf{V}}_K^{(1)\top} - \Sigma^* \|_F \right\|_{\psi_1} \leq \left\| \| \hat{\mathbf{V}}_K^{(1)} \hat{\mathbf{V}}_K^{(1)\top} - \mathbf{V}_K \mathbf{V}_K^\top \|_F \right\|_{\psi_1} + \| \Sigma^* - \mathbf{V}_K \mathbf{V}_K^\top \|_F \\
\leq 2 \left\| \rho(\hat{\mathbf{V}}_K^{(1)}, \mathbf{V}_K) \right\|_{\psi_1} \lesssim \kappa \sqrt{\frac{Kr}{n}}.
\]

\[\Box\]

**B.2 Proof of Theorem 3.3.1**

Proof. When \( \| \Sigma^* - \mathbf{V}_K \mathbf{V}_K^\top \|_{\text{op}} < 1/4 \), the Weyl’s inequality forces \( \lambda_K(\Sigma^*) > \frac{3}{4} \) and \( \lambda_{K+1}(\Sigma^*) < \frac{1}{4} \). The Theorem 2 in [131] yields

\[
\rho(\tilde{\mathbf{V}}_K, \mathbf{V}_K^*) = \sqrt{2} \sin \Theta(\tilde{\mathbf{V}}_K, \mathbf{V}_K^*) \lesssim \frac{\| \tilde{\Sigma} - \Sigma^* \|_F}{\lambda_K(\Sigma^*) - \lambda_{K+1}(\Sigma^*)} \lesssim \| \tilde{\Sigma} - \Sigma^* \|_F. \tag{B.2.1}
\]

When \( n \geq r \), Lemma [B.10.2] and Lemma 3.3.1 imply that

\[
\left\| \| \tilde{\Sigma} - \Sigma^* \|_F \right\|_{\psi_1} = \left\| \frac{1}{m} \sum_{\ell=1}^m \hat{\mathbf{V}}_K^{(\ell)} \hat{\mathbf{V}}_K^{(\ell)\top} - \Sigma^* \right\|_F \lesssim \frac{1}{\sqrt{m}} \left\| \| \hat{\mathbf{V}}_K^{(1)} \hat{\mathbf{V}}_K^{(1)\top} - \Sigma^* \|_F \right\|_{\psi_1} \\
\lesssim \kappa \sqrt{\frac{Kr}{N}}.
\]

Combining the two inequalities above finishes the proof.

\[\Box\]

**B.3 Proof of Theorem 3.3.2**

Proof. Choose \( j \in [d] \) and let \( \mathbf{D}_j = \mathbf{I} - 2\mathbf{e}_j \mathbf{e}_j^\top \). Let \( \mathbf{\Sigma} = \mathbf{V} \mathbf{\Lambda} \mathbf{V}^\top \) be the spectral decomposition of \( \mathbf{\Sigma} \). Assume that \( \hat{\lambda} \) is an eigenvalue of the sample covariance \( \hat{\mathbf{\Sigma}} = (1/n) \sum_{i=1}^n \mathbf{X}_i \mathbf{X}_i^\top \) and \( \hat{\mathbf{\nu}} \in \mathbb{S}^{d-1} \) is the correspondent eigenvector that satisfies \( \hat{\mathbf{\Sigma}} \hat{\mathbf{\nu}} = \hat{\lambda} \hat{\mathbf{\nu}} \).
Define $Z_i = \Lambda^{-\frac{1}{2}} V^T X_i$ and $\hat{S} = (1/n) \sum_{i=1}^{n} Z_i Z_i^T$. Note that $\hat{\Sigma} = V \Lambda^{\frac{1}{2}} \hat{S} \Lambda^{\frac{1}{2}} V^T$. Consider the matrix $\hat{\Sigma} = V \Lambda^{\frac{1}{2}} D_j \hat{S} D_j \Lambda^{\frac{1}{2}} V^T$. By the sign symmetry, $\hat{\Sigma}$ and $\Sigma$ are identically distributed. It is not hard to verify that $\hat{\Sigma}$ also has an eigenvalue $\hat{\lambda}$ with the correspondent eigenvector being $V D_j V^T \hat{v}$. Denote the top $K$ eigenvectors of $\hat{\Sigma}$ by $\hat{V}_K = (\hat{v}_1, \cdots, \hat{v}_K)$ and the top $K$ eigenvectors of $\Sigma$ by $V_K$. Therefore we have

$$V^T E(\hat{V}_K \hat{V}_K^T) V = V^T E(V_K V_K^T) V = V^T V D_j V^T E(V_K V_K^T) V D_j V^T V = D_j V^T E(\hat{V}_K \hat{V}_K^T) V D_j V^T V.$$ 

Since the equation above holds for all $j \in [d]$, we can reach the conclusion that $V^T E(\hat{V}_K \hat{V}_K^T) V$ is diagonal, i.e., $\Sigma^* E(\hat{V}_K \hat{V}_K^T)$ and $\Sigma$ share the same set of eigenvectors.

Suppose that $\|\Sigma^* - V_K V_K^T\|_2 < 1/2$. As demonstrated above, for any $k \in [K]$, the $k$th column of $V_K$, which we denote by $v_k$, should be an eigenvector of $\Sigma^*$. Note that

$$\|\Sigma^* v_k\|_2 = \|(\Sigma^* - V_K V_K^T + V_K V_K^T) v_k\|_2 \geq 1 - \|\Sigma^* - V_K V_K^T\|_2 > 1 - \frac{1}{2} = \frac{1}{2}.$$ 

With regard to $\Sigma^*$, the correspondent eigenvalue of $v_k$ must be greater than $1/2$. Denote any eigenvector of $\Sigma$ that is not in $\{v_k\}_{k=1}^{K}$ by $u$, then analogously,

$$\|\Sigma^* u\|_2 = \|(\Sigma^* - V_K V_K^T + V_K V_K^T) u\|_2 \leq \|\Sigma^* - V_K V_K^T\|_2 < \frac{1}{2}.$$ 

For $\Sigma^*$, the correspondent eigenvalue of $u$ is smaller than $1/2$. Therefore, the top $K$ eigenspace of $\Sigma^*$ is exactly $\text{Col}(V_K)$, and $\rho(V^*_K, V_K) = 0$. 

\[\square\]
B.4 Proof of Lemma 3.3.2

Proof. The first claim follows from Theorem 2 in [131] and our Lemma B.11.1. To attack the second one, we divide the difference:

\[
\hat{U}H - U - f(EU) = [P_\perp \hat{U}H - f(EU)] + P_\perp \hat{U}(\hat{H} - H) + (P\hat{U}H - U) \quad (B.4.1)
\]

and conquer the terms separately. Lemma B.11.1 forces \( \|D - I\|_F = \|H - \hat{H}\|_F \leq \|U^T \hat{U}U - UU^T\|_F \leq \varepsilon^2 \). Then

\[
\|P_\perp \hat{U}(\hat{H} - H)\|_F \leq \varepsilon^2 \quad \text{and} \quad \|P\hat{U}H - U\|_F = \|U^T \hat{U}H - U\|_F = \|UYDX^TXY^T - U\|_F = \|UY(D - I)Y^T\|_F \leq \varepsilon^2. \quad (B.4.2)
\]

By triangle’s inequality, it remains to show that \( \|P_\perp \hat{U}H - f(EU)\|_F \leq 3\varepsilon^2 \). Define \( \Lambda = \text{diag}(\lambda_{s+1}, \ldots, \lambda_{s+K}) \), and \( L(V) = AV - VA \) for \( V \in \mathbb{R}^{d \times K} \). Note that \( L(v_1, \ldots, v_K) = ((A - \lambda_{s+1}I)v_1, \ldots, (A - \lambda_{s+K}I)v_K) \), and \( G_j(A - \lambda_{s+j}I) = P_\perp \) holds for all \( j \in [K] \). As a result, \( f(L(V)) = -P_\perp V \) for any \( V \in \mathbb{R}^{d \times K} \). This motivates us to work on \( L(\hat{U}H) \) in order to study \( P_\perp \hat{U}H \).

Let \( \hat{A} = \text{diag}(\hat{\lambda}_{s+1}, \ldots, \hat{\lambda}_{s+K}) \). By definition, \( \hat{A} \hat{U} = \hat{U} \hat{A} \) and

\[
L(\hat{U}H) = A \hat{U}H - \hat{U}HA
= (A - \hat{A})\hat{U}H + (\hat{A} \hat{U} - \hat{U} \hat{A})H + \hat{U}(\hat{A} - A)H + \hat{U}(AH - HA) \quad (B.4.3)
= -E\hat{U}H + \hat{U}(\hat{A} - A)H + \hat{U}(AH - HA).
\]

Now we study the images of these three terms under the linear mapping \( f \). First, the bounded property \( \|f(\cdot)\|_F \leq \Delta^{-1}\|\cdot\|_F \) implies that

\[
\|f(E\hat{U}H) - f(EU)\|_F \leq \Delta^{-1}\|E(\hat{U}H - U)\|_F \leq \Delta^{-1}\|E\|_{op}\|\hat{U}H - U\|_F. \quad (B.4.4)
\]
Second, the definition of \( f \) forces \( f(\mathbf{UM}) = 0 \) for all \( \mathbf{M} \in \mathbb{R}^{K \times K} \). Hence the decomposition \( \hat{U}(\hat{\Lambda} - \Lambda)\mathbf{H} = \mathbf{U}\mathbf{H}^{-1}(\hat{\Lambda} - \Lambda)\mathbf{H} + (\hat{\mathbf{U}}\mathbf{H} - \mathbf{U})\mathbf{H}^{-1}(\hat{\Lambda} - \Lambda)\mathbf{H} \) yields

\[
\|f[\hat{U}(\hat{\Lambda} - \Lambda)\mathbf{H}]\|_F = \|f[(\hat{\mathbf{U}}\mathbf{H} - \mathbf{U})\mathbf{H}^{-1}(\hat{\Lambda} - \Lambda)\mathbf{H}]\|_F \\
\leq \Delta^{-1}\|[(\hat{\mathbf{U}}\mathbf{H} - \mathbf{U})\mathbf{H}^{-1}(\hat{\Lambda} - \Lambda)\mathbf{H}]\|_F \leq \Delta^{-1}\|\hat{\mathbf{U}}\mathbf{H} - \mathbf{U}\|_F\|\mathbf{H}^{-1}(\hat{\Lambda} - \Lambda)\mathbf{H}\|_{op} \\
\leq \Delta^{-1}\|\hat{\mathbf{U}}\mathbf{H} - \mathbf{U}\|_F\|\mathbf{H}^{-1}\|_{op}\|\mathbf{H}\|_{op} \\
\leq \Delta^{-1}\|\mathbf{E}\|_{op}\|\hat{\mathbf{U}}\mathbf{H} - \mathbf{U}\|_F\|\mathbf{H}^{-1}\|_{op}.
\]

(B.4.5)  

Here we applied Weyl's inequality \( \|\hat{\Lambda} - \Lambda\|_{op} \leq \|\mathbf{E}\|_{op} \) (Corollary IV.4.9 in [109]) and used the fact that \( \|\mathbf{H}\|_{op} = \|\hat{\mathbf{V}}\mathbf{V}^\top\|_{op} \leq 1 \). Third, by similar tricks we decompose the third term

\[
\hat{U}(\Lambda\mathbf{H} - \mathbf{H}\Lambda) = \mathbf{U}\mathbf{H}^{-1}(\Lambda\mathbf{H} - \mathbf{H}\Lambda) + (\hat{\mathbf{U}}\mathbf{H} - \mathbf{U})\mathbf{H}^{-1}(\Lambda\mathbf{H} - \mathbf{H}\Lambda)
\]

(B.4.6)

and write

\[
\|f[\hat{U}(\Lambda\mathbf{H} - \mathbf{H}\Lambda)]\|_F = \|f[(\hat{\mathbf{U}}\mathbf{H} - \mathbf{U})\mathbf{H}^{-1}(\Lambda\mathbf{H} - \mathbf{H}\Lambda)]\|_F \\
\leq \Delta^{-1}\|[(\hat{\mathbf{U}}\mathbf{H} - \mathbf{U})\mathbf{H}^{-1}(\Lambda\mathbf{H} - \mathbf{H}\Lambda)]\|_F \leq \Delta^{-1}\|\hat{\mathbf{U}}\mathbf{H} - \mathbf{U}\|_F\|\mathbf{H}^{-1}(\Lambda\mathbf{H} - \mathbf{H}\Lambda)\|_{op} \\
\leq \Delta^{-1}\|\hat{\mathbf{U}}\mathbf{H} - \mathbf{U}\|_F\|\mathbf{H}^{-1}\|_{op}\|\Lambda\mathbf{H} - \mathbf{H}\Lambda\|_{op}.
\]

(B.4.7)

As an intermediate step, we are going to show that \( \|\Lambda\mathbf{H} - \mathbf{H}\Lambda\|_{op} \leq 2\|\mathbf{E}\|_{op} \). On the one hand, \( \hat{\Lambda}\hat{\mathbf{U}} = \hat{\mathbf{U}}\hat{\Lambda} \) yields

\[
L(\hat{\mathbf{U}}) = (\mathbf{A} - \hat{\Lambda})\hat{\mathbf{U}} + (\hat{\Lambda}\hat{\mathbf{U}} - \hat{\mathbf{U}}\hat{\Lambda}) + \hat{\mathbf{U}}(\hat{\Lambda} - \Lambda) = -\mathbf{E}\hat{\mathbf{U}} + \hat{\mathbf{U}}(\hat{\Lambda} - \Lambda).
\]

(B.4.8)
On the other hand, let \( \mathbf{U}_1 = (\mathbf{u}_1, \ldots, \mathbf{u}_s, \mathbf{u}_{s+K+1}, \ldots, \mathbf{u}_d) \), \( \hat{\mathbf{U}}_1 = (\hat{\mathbf{u}}_1, \ldots, \hat{\mathbf{u}}_s, \hat{\mathbf{u}}_{s+K+1}, \ldots, \hat{\mathbf{u}}_d) \), \( \mathbf{H} = \hat{\mathbf{U}}^\text{T} \mathbf{U} \) and

\[
\Lambda_1 = \text{diag}(\lambda_1, \ldots, \lambda_s, \lambda_{s+K+1}, \ldots, \lambda_d),
\]

\[
\hat{\Lambda}_1 = \text{diag}(\hat{\lambda}_1, \ldots, \hat{\lambda}_s, \hat{\lambda}_{s+K+1}, \ldots, \hat{\lambda}_d).
\]

We have

\[
\mathbf{A} \hat{\mathbf{U}} = \begin{pmatrix} \mathbf{U} & \mathbf{U}_1 \end{pmatrix} \begin{pmatrix} \Lambda & \mathbf{0} \\ \mathbf{0} & \Lambda_1 \end{pmatrix} \begin{pmatrix} \mathbf{U}^\text{T} \\ \mathbf{U}_1^\text{T} \end{pmatrix} \hat{\mathbf{U}} = \begin{pmatrix} \mathbf{U} & \mathbf{U}_1 \end{pmatrix} \begin{pmatrix} \Lambda \mathbf{U} \\ \Lambda_1 \mathbf{U}_1^\text{T} \hat{\mathbf{U}} \end{pmatrix},
\]

\[
\hat{\mathbf{U}} \Lambda = \begin{pmatrix} \mathbf{U} & \mathbf{U}_1 \end{pmatrix} \begin{pmatrix} \mathbf{U}^\text{T} \\ \mathbf{U}_1^\text{T} \end{pmatrix} \hat{\mathbf{U}} \Lambda = \begin{pmatrix} \mathbf{U} & \mathbf{U}_1 \end{pmatrix} \begin{pmatrix} \mathbf{H} \Lambda \\ \Lambda_1 \mathbf{U}_1^\text{T} \hat{\mathbf{U}} \end{pmatrix},
\]

and as a result,

\[
\| \Lambda \mathbf{H} - \mathbf{H} \Lambda \|_{\text{op}} \leq \| L(\hat{\mathbf{U}}) \|_{\text{op}} = -\mathbf{E} \hat{\mathbf{U}} + \hat{\mathbf{U}}(\hat{\Lambda} - \Lambda) \|_{\text{op}} \leq 2\| \mathbf{E} \|_{\text{op}}.
\]

By combining (B.4.4), (B.4.5), (B.4.7) and (B.4.11), we obtain that

\[
\| \mathbf{P}_\perp \hat{\mathbf{U}} \mathbf{H} - f(\mathbf{E} \mathbf{U}) \|_F = \| -f[L(\hat{\mathbf{U}} \mathbf{H})] - f(\mathbf{E} \mathbf{U}) \|_F
\]

\[
\leq \Delta^{-1}\| \mathbf{E} \|_{\text{op}} \| \hat{\mathbf{U}} \mathbf{H} - \mathbf{U} \|_F (1 + 3\| \mathbf{H}^{-1} \|_{\text{op}}).
\]

On the one hand, when \( \varepsilon \leq 1/\sqrt{2} \), Lemma B.11.1 forces \( \| \hat{\mathbf{H}} - \mathbf{H} \|_F \leq 1/2 \) and thus \( \| \mathbf{H}^{-1} \|_{\text{op}} \leq 2 \). On the other hand, the fact \( \mathbf{U} \in \mathcal{O}_{d\times K} \) implies that \( \mathbf{U}^\top \mathbf{U} = \mathbf{I}_K \) and

\[
\| \hat{\mathbf{U}} \mathbf{H} - \mathbf{U} \|_F = \| \hat{\mathbf{U}} \mathbf{U}^\top \mathbf{U} - \mathbf{U} \|_F = \| \hat{\mathbf{U}} \mathbf{U}^\top \mathbf{U} - \mathbf{U} \|_F
\]

\[
= \| \hat{\mathbf{U}} \mathbf{U}^\top - \mathbf{U} \|_F \leq \| \hat{\mathbf{U}} \mathbf{U}^\top - \mathbf{U} \|_F \leq \varepsilon.
\]
Hence \( \|P \perp \hat{U}H - f(EU)\|_F \leq 7\varepsilon^2/\sqrt{8K} \leq 3\varepsilon^2 \) and we have proved the second claim in Lemma 5.2.2. Finally we come to the last one. Lemma B.11.2 implies that

\[
\|\hat{U}\hat{U}^T - (P + P\hat{U}\hat{U}^T P \perp + P \perp \hat{U}\hat{U}^T P)\|_F \leq \sqrt{2}\|P \perp \hat{U}\|^2 \leq \sqrt{2}\varepsilon^2. \tag{B.4.14}
\]

Observe that

\[
P \perp \hat{U}\hat{U}^T P = P \perp \hat{U}H[U^T P + (\hat{U}H - U)^T P] = f(EU)U^T + [P \perp \hat{U}H - f(EU)]U^T + P \perp \hat{U}H(\hat{U}H - U)^T P. \tag{B.4.15}
\]

Also,

\[
\|P \perp \hat{U}H - f(EU)\|_F \leq \|P \perp \hat{U}H - f(EU)\|_F \|U^T\|_{op} \leq 4\varepsilon^2,
\]

\[
\|P \perp \hat{U}H(\hat{U}H - U)^T P\|_F = \|P \perp (\hat{U}H - U)(\hat{U}H - U)^T P\|_F \leq \|\hat{U}H - U\|^2 \leq \varepsilon^2. \tag{B.4.16}
\]

Hence \( \|P \perp \hat{U}\hat{U}^T P - f(EU)U^T\|_F \leq 5\varepsilon^2 \) and similarly, \( \|P\hat{U}\hat{U}^TP \perp - Uf(EU)^T\|_F \leq 5\varepsilon^2 \). To sum up, we have

\[
\|\hat{U}\hat{U}^T - (UU^T + f(EU)U^T + Uf(EU)^T)\|_F \leq \sqrt{2}\varepsilon^2 + 10\varepsilon^2 \leq 12\varepsilon^2. \tag{B.4.17}
\]

\( \square \)
B.5 Proof of Theorem 3.3.3

Proof. Define \( E = \hat{\Sigma}^{(1)} - \Sigma, P = V_K V_K^\top, \hat{P} = \hat{V}^{(1)}_K \hat{V}^{(1)}_K^\top, Q = f(EV_K)V_K^\top + V_K f(EV_K)^\top, W = \hat{P} - P - Q \) and \( \varepsilon = \|E\|_2/\Delta \). From \( EQ = 0 \) and

\[
\hat{P} - P - Q = W = W_{1_{\{\varepsilon \leq 1/10\}}} + (W + Q)_{1_{\{\varepsilon > 1/10\}}} - Q_{1_{\{\varepsilon > 1/10\}}},
\]

we derive that

\[
\mathbb{E} \hat{P} - P = \mathbb{E}(W_{1_{\{\varepsilon \leq 1/10\}}} + (W + Q)_{1_{\{\varepsilon > 1/10\}}}) - \mathbb{E}(Q_{1_{\{\varepsilon > 1/10\}}}),
\]

\[
\|\mathbb{E} \hat{P} - P\|_F \leq \mathbb{E}(\|W\|_F1_{\{\varepsilon \leq 1/10\}}) + \mathbb{E}(\|\hat{P} - P\|_F1_{\{\varepsilon > 1/10\}}) + \mathbb{E}(\|Q\|_F1_{\{\varepsilon > 1/10\}}).
\]

(B.5.1)

We are going to bound the three terms separately. On the one hand, Lemma 3.3.2 implies that \( \|W\|_F \leq 24\sqrt{K}\varepsilon^2 \) when \( \varepsilon \leq 1/10 \). Hence

\[
\mathbb{E}(\|W\|_F1_{\{\varepsilon \leq 1/10\}}) \leq \mathbb{E}(24\sqrt{K}\varepsilon^21_{\{\varepsilon \leq 1/10\}}) \lesssim \sqrt{K}\mathbb{E}\varepsilon^2.
\]

(B.5.2)

On the other hand, the Davis-Kahan theorem shows that \( \|\hat{P} - P\|_F \lesssim \sqrt{K}\varepsilon \). Besides, it is easily seen that \( \|Q\|_F \lesssim \|f(EV_K)\|_F \leq \sqrt{K}\|E\|_2/\Delta = \sqrt{K}\varepsilon \). Hence

\[
\mathbb{E}(\|\hat{P} - P\|_F1_{\{\varepsilon > 1/10\}}) + \mathbb{E}(\|Q\|_F1_{\{\varepsilon > 1/10\}}) \lesssim \sqrt{K}\mathbb{E}(\varepsilon 1_{\{\varepsilon > 1/10\}})
\]

\[
\leq 10\sqrt{K}\mathbb{E}(\varepsilon^21_{\{\varepsilon > 1/10\}}) \lesssim \sqrt{K}\mathbb{E}\varepsilon^2.
\]

(B.5.3)

By (B.5.1), (B.5.2), (B.5.3) and Lemma B.10.1 we have

\[
\|\mathbb{E} \hat{P} - P\|_F \lesssim \sqrt{K}\mathbb{E}\varepsilon^2 = \sqrt{K}\Delta^{-2}\mathbb{E}\|E\|_2^2 \lesssim \sqrt{K}\Delta^{-2}\|\|E\|_2\|_\psi_1^2 \lesssim \frac{k^2\sqrt{K}}{n}.
\]

(B.5.4)
B.6 Proof of Theorem 3.3.4

Proof. According to Theorem 3.3.3, there exists a constant $C$ such that $\|\Sigma^* - V_K V_K^\top\|_2 \leq 1/4$ as long as $n \geq C \kappa^2 \sqrt{K} r \geq r$. Then Theorem 3.3.1 implies that $\|\rho(\tilde{V}_K, V_K^*)\|_{\psi_1} \leq C_1 \kappa \sqrt{\frac{Kr}{N}}$ for some constant $C_1$.

When random samples have symmetric innovation, we have $\rho(V_K^*, V_K) = 0$ and

$$\|\rho(\tilde{V}_K, V_K)\|_{\psi_1} = \|\rho(\tilde{V}_K, V_K^*)\|_{\psi_1} \leq C_1 \kappa \sqrt{\frac{Kr}{N}}.$$

For general distribution, Theorem 3.3.3 implies that $\rho(V_K^*, V_K) \leq C_2 \kappa^2 \sqrt{K} r/n$ for some constant $C_2$ and

$$\|\rho(\tilde{V}_K, V_K)\|_{\psi_1} \leq \|\rho(\tilde{V}_K, V_K^*)\|_{\psi_1} + \rho(V_K^*, V_K) \leq C_1 \kappa \sqrt{\frac{Kr}{N}} + C_2 \kappa^2 \sqrt{\frac{Kr}{n}}.$$  \hfill (B.6.1)

When $m \leq C_3 n/(\kappa^2 r)$ for some constant $C_3$, we have

$$\kappa \sqrt{\frac{Kr}{N}} = \sqrt{\frac{\kappa^2 Kr}{nm}} \geq \sqrt{\frac{\kappa^2 Kr}{n \cdot C_3 n/(\kappa^2 r)}} = \frac{1}{\sqrt{C_3}} \cdot \frac{\kappa^2 \sqrt{K} r}{n},$$

and (B.6.1) forces

$$\|\rho(\tilde{V}_K, V_K)\|_{\psi_1} \leq (C_1 + C_2 \sqrt{C_3}) \kappa \sqrt{\frac{Kr}{N}}.$$
B.7 Proof of Theorem 3.3.5

Proof. We first focus on the first subsample \( \{X_i^{(1)}\}_{i=1}^n \) and the associated top eigenvector \( \hat{V}_1^{(1)} \). For ease of notation, we temporarily drop the superscript. Let \( S = \sum_{i=1}^n W_i \) and \( \hat{\Sigma}_Z = \frac{d-1}{n} \sum_{i=1}^n 1_{\{W_i=1\}} Z_i Z_i^\top \). From \( \hat{\Sigma} = \begin{pmatrix} \frac{\lambda}{n} (n-S) & 0_{1 \times (d-1)} \\ 0_{(d-1) \times 1} & 2 \hat{\Sigma}_Z \end{pmatrix} \) we know that \( \|\hat{\Sigma}_Z\|_2 > (\lambda/n)(n-S) \) and \( \|\hat{\Sigma}_Z\|_2 < (\lambda/n)(n-S) \) lead to \( \hat{V}_1 \perp V_1 \) and \( \hat{V}_1 \parallel V_1 \) (i.e. \( \hat{V}_1 = \pm V_1 \)), respectively. Besides, \( \|\hat{\Sigma}_Z\|_2 \) is a continuous random variable. Hence \( \mathbb{P}(\hat{V}_1 \perp V_1) + \mathbb{P}(\hat{V}_1 \parallel V_1) = 1 \). Note that

\[
\text{Tr}(\hat{\Sigma}_Z) = \frac{d-1}{n} \sum_{i=1}^n 1_{\{W_i=1\}} = \frac{(d-1)S}{n},
\]

\[
\|\hat{\Sigma}_Z\|_2 \geq \frac{\text{Tr}(\hat{\Sigma}_Z)}{\text{rank}(\hat{\Sigma}_Z)} \geq \frac{\text{Tr}(\hat{\Sigma}_Z)}{\min\{n, d-1\}} \geq \frac{(d-1)S}{n^2}.
\]

Then

\[
\mathbb{P}(\hat{V}_1 \parallel V_1) \leq \mathbb{P}\left(\|\hat{\Sigma}_Z\|_2 \leq \frac{\lambda}{n} (n-S)\right) \leq \mathbb{P}\left(\frac{(d-1)S}{n^2} \leq \frac{\lambda}{n} (n-S)\right)
\]

\[
= \mathbb{P}\left(\frac{S}{n} \leq 1 + \frac{d-1}{n\lambda}\right) \leq \mathbb{P}\left(\frac{S}{n} \leq \frac{1}{4}\right) = \mathbb{P}\left(\frac{S}{n} - \frac{1}{2} \leq -\frac{1}{4}\right) \leq e^{-n/8}.
\]

Above we used the assumption \( d \geq 3n\lambda + 1 \) and Hoeffding’s inequality. Now we finish the analysis of \( \hat{V}_1^{(1)} \) and collect back the superscript.

From now on we define \( S = \sum_{\ell=1}^m 1_{\{\hat{V}_1^{(\ell)} \not\parallel V_1\}} \). For \( \hat{V}_1^{(\ell)} \), let \( a_\ell \) be its first entry and \( b_\ell \) be the vector of its last \( (d-1) \) entries. The dichotomy \( \mathbb{P}(\hat{V}_1^{(\ell)} \parallel V_1) + \mathbb{P}(\hat{V}_1^{(\ell)} \perp V_1) = 1 \) mentioned above forces \( |a_\ell| = 1_{\{\hat{V}_1^{(\ell)} \parallel V_1\}}, \quad \|b_\ell\|_2 = 1_{\{\hat{V}_1^{(\ell)} \not\parallel V_1\}}, \quad \hat{V}_1^{(\ell)} \hat{V}_1^{(\ell)\top} = \begin{pmatrix} 1_{\{\hat{V}_1^{(\ell)} \not\parallel V_1\}} & 0_{1 \times (d-1)} \\ 0_{(d-1) \times 1} & b_\ell b_\ell^\top \end{pmatrix}, \quad \text{and}
\]

\[
\tilde{\Sigma} = \frac{1}{m} \sum_{\ell=1}^m \hat{V}_1^{(\ell)} \hat{V}_1^{(\ell)\top} = \begin{pmatrix} \frac{1}{m} S & 0_{1 \times (d-1)} \\ 0_{(d-1) \times 1} & \frac{1}{m} \sum_{\ell=1}^m b_\ell b_\ell^\top \end{pmatrix}.
\]

186
Note that \( n \geq 32 \log d \) forces \( \mathbb{P}(\tilde{V}_1^\perp / V_1) \leq e^{-n/8} \leq d^{-4} \).

**Case 1: \( m \leq d^3 \)**

In this case, \( \mathbb{P}(S = 0) = [1 - \mathbb{P}(\tilde{V}_1^\perp / V_1)]^m \geq 1 - m\mathbb{P}(\tilde{V}_1^\perp / V_1) \geq 1 - d^{-1} \).

When \( S = 0 \), we have \( \|b_\ell\|_2 = 1 \) for all \( \ell \in [m] \) and \( \|(1/m)\sum_{\ell=1}^m b_\ell b_\ell^\top\|_2 > 0 \), leading to \( \tilde{V}_1 \perp V_1 \).

**Case 2: \( m > d^3 \)**

On the one hand, by Hoeffding’s inequality we obtain

\[
\mathbb{P}\left(\frac{S}{m} \geq \frac{1}{d}\right) \leq \mathbb{P}\left(\frac{1}{m}(S - \mathbb{E}S) \geq \frac{1}{2d}\right) \leq e^{-2m(\frac{1}{2d})^2} = e^{-\frac{m}{2d^2}} < e^{-d/2}.
\]

On the other hand, note that

\[
\left\|\frac{1}{m} \sum_{k=1}^m b_\ell b_\ell^\top\right\|_2 \geq \frac{\text{Tr}(\frac{1}{m} \sum_{k=1}^m b_\ell b_\ell^\top)}{d - 1} = \frac{1}{m} \sum_{k=1}^m \|b_\ell\|_2^2 = \frac{1}{d - 1} \left(1 - \frac{S}{m}\right).
\]

Hence

\[
\mathbb{P}(\tilde{V}_1 \perp V_1) \geq \mathbb{P}\left(\left\|\frac{1}{m} \sum_{k=1}^m b_\ell b_\ell^\top\right\|_2 > \frac{S}{m}\right) \geq \mathbb{P}\left[\frac{1}{d - 1} \left(1 - \frac{S}{m}\right) > \frac{S}{m}\right] = \mathbb{P}\left(\frac{S}{m} < \frac{1}{d}\right)
\]

\[\geq 1 - e^{-d/2}.\]

\(\square\)

**B.8 Proof of Theorem 3.4.1**

Proof. With slight abuse of notations, here we define \( \Sigma_\ell^* = \mathbb{E}(\tilde{V}_K^\ell \tilde{V}_K^{\ell\top}) \), \( \Sigma^* = \frac{1}{m} \sum_{\ell=1}^m \Sigma_\ell^* \), and \( V_K^* \in \mathbb{R}^{d \times K} \) to be the top \( K \) eigenvectors of \( \Sigma^* \).
First we consider the general case. Note that $\lambda_K(\mathbf{V}_K \mathbf{V}_K^\top) = 1$ and $\lambda_K(\mathbf{V}_K \mathbf{V}_K^\top) = 0$. By the Davis-Kahan theorem, we have

$$
\rho(\tilde{\mathbf{V}}_K, \mathbf{V}_K) \lesssim \|\tilde{\Sigma} - \mathbf{V}_K \mathbf{V}_K^\top\|_F \leq \|\tilde{\Sigma} - \Sigma^*\|_F + \|\Sigma^* - \mathbf{V}_K \mathbf{V}_K^\top\|_F. \quad (B.8.1)
$$

Note that $\Sigma^* = \frac{1}{m} \sum_{\ell=1}^m \Sigma^*_\ell$. The first term in (B.8.1) is the norm of independent sums

$$
\|\tilde{\Sigma} - \Sigma^*\|_F = \left\| \frac{1}{m} \sum_{\ell=1}^m \left( \mathbf{V}^{(\ell)}_K \mathbf{V}^{(\ell)^\top}_K - \Sigma^*_\ell \right) \right\|_F.
$$

It follows from Lemma 3.3.1 that $\left\| \left( \mathbf{V}^{(\ell)}_K \mathbf{V}^{(\ell)^\top}_K - \Sigma^*_\ell \right) \right\|_F \lesssim \kappa_\ell \sqrt{\frac{Kr_\ell}{n}} = \sqrt{mS_\ell}$, from which Lemma B.10.2 leads to

$$
\left\| \tilde{\Sigma} - \Sigma^* \right\|_F \lesssim \frac{1}{m} \left( \sum_{\ell=1}^m (\sqrt{mS_\ell})^2 \right) = \sqrt{\frac{1}{m} \sum_{\ell=1}^m S^2_\ell}. \quad (B.8.2)
$$

The second term in (B.8.1) is bounded by

$$
\|\Sigma^* - \mathbf{V}_K \mathbf{V}_K^\top\|_F = \left\| \frac{1}{m} \sum_{\ell=1}^m (\Sigma^*_\ell - \mathbf{V}_K \mathbf{V}_K^\top) \right\|_F \leq \frac{1}{m} \sum_{\ell=1}^m \|\Sigma^*_\ell - \mathbf{V}_K \mathbf{V}_K^\top\|_F.
$$

Theorem 3.3.3 implies that when $n \geq r_\ell$,

$$
\|\Sigma^*_\ell - \mathbf{V}_K \mathbf{V}_K^\top\|_F \lesssim \kappa^2_\ell \sqrt{K}r_\ell/n = B_\ell. \quad (B.8.3)
$$

Hence

$$
\|\Sigma^* - \mathbf{V}_K \mathbf{V}_K^\top\|_F \lesssim \frac{1}{m} \sum_{\ell=1}^m B_\ell. \quad (B.8.4)
$$

The claim under general case follows from (B.8.1), (B.8.2) and (B.8.4).
Now we come to the symmetric case. If \( \| \Sigma^*_\ell - V_K V_K^T \|_2 < 1/2 \) for all \( \ell \in [m] \), then Theorem 3.3.2 implies that the top \( K \) eigenspace of \( \Sigma^*_\ell \) is \( \text{Col}(V_K) \). Therefore, the top \( K \) eigenspace of \( \Sigma^* \) is still \( \text{Col}(V_K) \) and \( \rho(V_K, V_K^*) = 0 \).

When \( n \geq C \sqrt{K} \max\ell \in [m] (\kappa_\ell^2 r_\ell) \) for large \( C \), (B.8.3) ensures \( \max\ell \in [m] \| \Sigma^*_\ell - V_K V_K^T \|_2 \leq 1/4 \), \( \| \Sigma^* - V_K V_K^T \|_2 \leq 1/4 \) and \( \rho(V_K, V_K^*) = 0 \). Weyl’s inequality forces \( \lambda_K(\Sigma^*) \geq 3/4 \) and \( \lambda_{K+1}(\Sigma^*) \leq 1/4 \). By the Davis-Kahan theorem and (B.8.2),

\[
\left\| \rho(\tilde{V}_K, V_K) \right\|_{\psi_1} = \left\| \rho(\tilde{V}_K, V_K^*) \right\|_{\psi_1} \lesssim \left\| \tilde{\Sigma} - \Sigma^* \right\|_F \lesssim \sqrt{\frac{1}{m} \sum_{\ell=1}^{m} S_\ell^2}.
\]

\[\Box\]

### B.9 Proof of Theorem 3.4.2

**Proof.** We define \( \Sigma^*_\ell = \mathbb{E} \tilde{\Sigma}^\ell \tilde{\Sigma}^\ell^\top \) and \( \Sigma^* = \frac{1}{m} \sum_{\ell=1}^{m} \Sigma^*_\ell \). Let \( V_K, \tilde{V}_K \in \mathcal{O}_{d \times K} \) be the top \( K \) eigenvectors of \( \Sigma^* \) and \( \Sigma^\ell \), respectively. By the Davis-Kahan theorem,

\[
\rho(\tilde{V}_K, V_K) \lesssim \| \tilde{\Sigma} - V_K V_K^T \|_F \leq \| \tilde{\Sigma} - \Sigma^* \|_F + \| \Sigma^* - V_K V_K^T \|_F. \tag{B.9.1}
\]

The first term in (B.9.1) is controlled in exactly the same way as (B.8.2). The second term is further decomposed as

\[
\| \Sigma^* - V_K V_K^T \|_F = \left\| \frac{1}{m} \sum_{\ell=1}^{m} (\Sigma^*_\ell - V_K V_K^T) \right\|_F \\
\leq \left\| \frac{1}{m} \sum_{\ell=1}^{m} (\Sigma^*_\ell - \tilde{\Sigma}^\ell \tilde{\Sigma}^\ell^\top) \right\|_F + \left\| \frac{1}{m} \sum_{\ell=1}^{m} (\tilde{\Sigma}^\ell \tilde{\Sigma}^\ell^\top - V_K V_K^T) \right\|_F. \tag{B.9.2}
\]

Similar to (B.8.3) and (B.8.4), with \( n \geq r_\ell \) we have \( \| \Sigma^*_\ell - \tilde{\Sigma}^\ell \tilde{\Sigma}^\ell^\top \|_F \lesssim B_\ell \) and

\[
\left\| \frac{1}{m} \sum_{\ell=1}^{m} (\Sigma^*_\ell - \tilde{\Sigma}^\ell \tilde{\Sigma}^\ell^\top) \right\|_F \leq \left\| \frac{1}{m} \sum_{\ell=1}^{m} (\Sigma^*_\ell - \tilde{\Sigma}^\ell \tilde{\Sigma}^\ell^\top) \right\|_F \lesssim \frac{1}{m} \sum_{\ell=1}^{m} B_\ell. \tag{B.9.3}
\]

189
For the last part in (B.9.2), note that $\bar{V}_K^{(\ell)}$ and $V_K$ contain eigenvectors of $\Sigma^{(\ell)}$ and $B^{(\ell)}B^{(\ell)\top}$. Hence the Davis-Kahan theorem forces

$$
\| \bar{V}_K^{(\ell)} \bar{V}_K^{(\ell)\top} - V_K V_K\|_F \lesssim \frac{\sqrt{K} \| \Sigma^{(\ell)}_u \|_{op}}{\lambda_K(\Lambda_K^{(\ell)})}.
$$

and

$$
\left\| \frac{1}{m} \sum_{\ell=1}^m (\bar{V}_K^{(\ell)} \bar{V}_K^{(\ell)\top} - V_K V_K) \right\|_F \lesssim \frac{\sqrt{K}}{m} \sum_{\ell=1}^m \| \Sigma^{(\ell)}_u \|_{op} \lambda_K(\Lambda_K^{(\ell)}) \text{.} \tag{B.9.4}
$$

The proof is completed by collecting (B.9.1), (B.9.2), (B.9.3) and (B.9.4).

\[\square\]

### B.9.1 Technical lemmas

### B.10 Tail bounds

**Lemma B.10.1.** Suppose $X$ and $\{X_i\}_{i=1}^n$ are i.i.d. sub-Gaussian random vectors in $\mathbb{R}^d$ with zero mean and covariance matrix $\Sigma \succeq 0$. Let $\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^n X_iX_i^\top$ be the sample covariance matrix, $\{\lambda_j\}_{j=1}^d$ be the eigenvalues of $\Sigma$ sorted in descending order, and $r = \text{Tr}(\Sigma)\|\Sigma\|_2$. There exist constants $c \geq 1$ and $C \geq 0$ such that when $n \geq r$, we have

$$
\mathbb{P}\left(\| \hat{\Sigma} - \Sigma \|_{op} \geq s \right) \leq \exp\left(-\frac{s}{c\lambda_1\sqrt{r/n}}\right), \quad \forall s \geq 0,
$$

and

$$
\left\| \| \hat{\Sigma} - \Sigma \|_{op} \right\|_{\psi_1} \leq C\lambda_1 \sqrt{r/n}.
$$

**Proof.** By the Theorem 9 in [70] and the simple fact

$$
(\mathbb{E}\|X\|_{op})^2/\|\Sigma\|_{op} \leq \mathbb{E}\|X\|_{op}^2/\|\Sigma\|_{op} = \text{Tr}(\Sigma)/\|\Sigma\|_{op} = r(\Sigma),
$$
we know the existence of a constant $c \geq 1$ such that

$$
\mathbb{P}\left( \| \hat{\Sigma} - \Sigma \|_{op} \geq c \lambda_1 \max \left\{ \sqrt{\frac{r}{n}}, \frac{r}{n}, \sqrt{\frac{t}{n}}, \frac{t}{n} \right\} \right) \leq e^{-t}, \quad \forall t \geq 1. \quad (B.10.1)
$$

Since $1 \leq r \leq n$, \((B.10.1)\) yields

$$
\mathbb{P}\left( \| \hat{\Sigma} - \Sigma \|_{op} \geq c \lambda_1 \sqrt{\frac{t}{n}} \right) \leq e^{-t}, \quad r \leq t \leq n, \quad (B.10.2)
$$

$$
\mathbb{P}\left( \| \hat{\Sigma} - \Sigma \|_{op} \geq c \lambda_1 \frac{t}{n} \right) \leq e^{-t}, \quad t \geq n. \quad (B.10.3)
$$

When $r \leq t \leq n$, we have $\sqrt{\frac{t}{n}} \leq \frac{t}{n} \sqrt{\frac{n}{r}}$. By letting $s = c \lambda_1 \frac{t}{n} \sqrt{\frac{n}{r}}$ we derive from \((B.10.2)\) that for $c \lambda_1 \frac{t}{n} \sqrt{\frac{n}{r}} \leq s \leq c \lambda_1 \sqrt{\frac{n}{r}}$,

$$
\mathbb{P}\left( \| \hat{\Sigma} - \Sigma \|_{op} \geq s \right) \leq \mathbb{P}\left( \| \hat{\Sigma} - \Sigma \|_{op} \geq c \lambda_1 \sqrt{\frac{t}{n}} \right) \leq e^{-t} = \exp \left( -\frac{s \sqrt{nr}}{c \lambda_1} \right). \quad (B.10.4)
$$

When $t \geq n$, we let $s = c \lambda_1 \frac{t}{n}$ and derive from \((B.10.3)\) that for $s \geq c \lambda_1$,

$$
\mathbb{P}\left( \| \hat{\Sigma} - \Sigma \|_{op} \geq s \right) = \mathbb{P}\left( \| \hat{\Sigma} - \Sigma \|_{op} \geq c \lambda_1 \frac{t}{n} \right) \leq e^{-t} = \exp \left( -\frac{ns}{c \lambda_1} \right). \quad (B.10.5)
$$

\((B.10.4), (B.10.5)\) and $n \geq r$ lead to

$$
\mathbb{P}\left( \| \hat{\Sigma} - \Sigma \|_{op} \geq s \right) \leq \exp \left( -\frac{s \sqrt{nr}}{c \lambda_1} \right), \quad \forall s \geq c \lambda_1 \sqrt{r/n}.
$$

and thus

$$
\mathbb{P}\left( \| \hat{\Sigma} - \Sigma \|_{op} \geq s \right) \leq \exp \left( 1 - \frac{s}{c \lambda_1 \sqrt{r/n}} \right), \quad \forall s \geq 0.
$$

According to the Definition 5.13 in [123], we get $\| \hat{\Sigma} - \Sigma \|_{op, \psi_1} \leq C \lambda_1 \sqrt{r/n}$ for some constant $C$. 

\[\square\]
The next lemma investigates the sum of independent random vectors in a Hilbert space whose norms are sub-exponential, which directly follows from Theorem 2.5 in [10].

**Lemma B.10.2.** If \( \{X_i\}_{i=1}^n \) are independent random vectors in a separable Hilbert space (where the norm is denoted by \( \| \cdot \| \)) with \( \mathbb{E}X_i = 0 \) and \( \|X_i\|_{\psi_1} \leq L_i < \infty \).

We have

\[
\left\| \sum_{i=1}^n X_i \right\|_{\psi_1} \lesssim \sqrt{\sum_{i=1}^n L_i^2}.
\]

**Proof.** We are going to apply Theorem 2.5 in [10]. By definition \( k^{-1} \mathbb{E}^{1/k} \|X_i\|^k \leq \|X_i\|_{\psi_1} \leq L_i \) for all \( k \geq 1 \), and

\[
\mathbb{E}\|X_i\|^k \leq (kL_i)^k \leq \sqrt{2\pi k} (k/e)^k (eL_i)^k \lesssim k! (eL_i)^k.
\]

Hence there exists some constant \( c \) such that \( \mathbb{E}\|X_i\|^k \leq \frac{k!}{2} (cL_i)^k \) for \( k \geq 2 \). Let \( \ell = \sqrt{c^2 \sum_{i=1}^n L_i^2} \) and \( b = c \cdot \max_{i \in [n]} L_i \). We have

\[
\sum_{i=1}^n \mathbb{E}\|X_i\|^k \leq \frac{k!}{2} \sum_{i=1}^n (cL_i)^k \leq \frac{k!}{2} \left( \sum_{i=1}^n c^2 L_i^2 \right) \left( c \cdot \max_{i \in [n]} L_i \right)^{k-2} = \frac{k!}{2} \ell^2 b^{k-2}, \quad \forall k \geq 2.
\]

Let \( S_n = \sum_{i=1}^n X_i \). Theorem 2.5 in [10] implies that

\[
\mathbb{P}(\|S_n\| \geq t) \leq 2 \exp \left( -\frac{t^2}{2\ell^2 + 2bt} \right), \quad \forall t > 0.
\]

When \( 4\ell \leq t \leq \ell^2/b \) (this cannot happen if \( 4b > \ell \)), we have \( 2\ell^2 \geq 2bt \) and

\[
\mathbb{P}(\|S_n\| \geq t) \leq 2 \exp \left( -\frac{t^2}{2\ell^2 + 2\ell^2} \right) \leq 2 \exp \left( -\frac{4\ell \cdot t}{4\ell^2} \right) = 2 \exp \left( -\frac{t}{\ell} \right) \leq \exp \left( 1 - \frac{t}{4\ell} \right).
\]
When \( t \geq \ell^2/b \), we have \( 2bt \geq 2\ell^2 \) and

\[
\mathbb{P}(\|S_n\| \geq t) \leq 2 \exp\left(-\frac{t^2}{2bt + 2bt}\right) = 2 \exp\left(-\frac{t}{4b}\right) \leq \exp\left(1 - \frac{t}{4\ell}\right),
\]

where the last inequality follows from \( 2 \leq e \) and \( b \leq \ell \). It is then easily seen that

\[
\mathbb{P}(\|S_n\| \geq t) \leq \exp\left(1 - \frac{t}{4\ell}\right), \quad \forall t \geq 0.
\]

With the help of Definition 5.13 in [123], we can conclude that

\[
\|\|S_n\|\|_{\psi_1} \lesssim \ell \lesssim \sqrt{\sum_{i=1}^{n} L_i^2}.
\]

\( \square \)

**B.11 Matrix analysis**

**Lemma B.11.1.** Let \( V, \hat{V} \in O_{d \times K} \), \( H = \hat{V}^T V \), \( H = XDY^T \) be its SVD, and \( \hat{H} = XY^T \). We have

\[
\left\| \hat{H} - H \right\|_{F}^{1/2} \leq \left\| (I -VV^T) \hat{V} \right\|_F \leq \min_{O \in O_{K,K}} \left\| \hat{V} O - V \right\|_F \leq \left\| \hat{V} \right\|_F \left\| \hat{V} \right\|_F \leq \sqrt{2} \sin \Theta(\hat{V}, V) \right\|_F = \left\| \hat{V} V^T - VV^T \right\|_F.
\]

*Proof.* Obviously, \( 0 \leq D_{ii} \leq 1 \) for all \( i \in [K] \), and \( \| \hat{H} - H \|_F = \| X(D - I)Y^T \|_F = \| D - I \|_F \). Let \( P = VV^T \). By Lemma B.11.3, we obtain that

\[
\| (I - P) \hat{V} \|_F \geq \| P \hat{V} \|_F = \| VH^T HV^T - VV^T \|_F = \| VYD^2 Y^T V^T - VV^T \|_F = \| (VY) (D^2 - I)(VY)^T \|_F = \| D^2 - I \|_F \geq \| D - I \|_F.
\]

(B.11.2)
Hence \( \|\hat{\mathbf{H}} - \mathbf{H}\|_F^{1/2} \leq \|(\mathbf{I} - \mathbf{V}\mathbf{V}^\top)\hat{\mathbf{V}}\|_F \). For any \( \mathbf{O} \in \mathcal{O}_{K,K} \), we have

\[
\|(\mathbf{I} - \mathbf{V}\mathbf{V}^\top)\hat{\mathbf{V}}\|_F = \|(\mathbf{I} - \mathbf{V}\mathbf{V}^\top)\hat{\mathbf{V}}\mathbf{O}\|_F = \|(\mathbf{I} - \mathbf{V}\mathbf{V}^\top)(\hat{\mathbf{V}}\mathbf{O} - \mathbf{V})\|_F \leq \|\hat{\mathbf{V}}\mathbf{O} - \mathbf{V}\|_F.
\]  
(B.11.3)

\[
\|(\mathbf{I} - \mathbf{V}\mathbf{V}^\top)\hat{\mathbf{V}}\|_F \leq \min_{\mathbf{O} \in \mathcal{O}_{K,K}} \|\hat{\mathbf{V}}\mathbf{O} - \mathbf{V}\|_F
\] is derived by taking infimum. Again, for any \( \mathbf{O} \in \mathcal{O}_{K,K} \) we have

\[
\|\hat{\mathbf{V}}\mathbf{O} - \mathbf{V}\|_F^2 = \|\hat{\mathbf{V}}\mathbf{O}\|_F^2 + \|\mathbf{V}\|_F^2 - 2\text{Tr}(\mathbf{O}^\top\hat{\mathbf{V}}\mathbf{V}) = 2K - 2\text{Tr}(\mathbf{O}^\top\mathbf{H}) \geq 2K - 2\|\mathbf{O}\|_{\text{op}}\|\mathbf{H}\|_* = 2K - 2\text{Tr}(\hat{\mathbf{H}}^\top\mathbf{H}) = \|\hat{\mathbf{V}}\mathbf{H} - \mathbf{V}\|_F^2.
\]  
(B.11.4)

This proves \( \min_{\mathbf{O} \in \mathcal{O}_{K,K}} \|\hat{\mathbf{V}}\mathbf{O} - \mathbf{V}\|_F = \|\hat{\mathbf{V}}\mathbf{H} - \mathbf{V}\|_F \). From above we also see that

\[
\|\hat{\mathbf{V}}\mathbf{H} - \mathbf{V}\|_F^2 = 2\sum_{j=1}^K (1 - D_{jj}) \leq 2\sum_{j=1}^K (1 - D_{jj}^2) = 2\|\sin \Theta(\hat{\mathbf{V}}, \mathbf{V})\|_F^2.
\]  
(B.11.5)

Besides,

\[
\|\hat{\mathbf{V}}\mathbf{V}^\top - \mathbf{V}\mathbf{V}^\top\|_F^2 = \|\hat{\mathbf{V}}\mathbf{V}^\top\|_F^2 + \|\mathbf{V}\mathbf{V}^\top\|_F^2 - 2\text{Tr}(\hat{\mathbf{V}}\mathbf{V}^\top\mathbf{V}\mathbf{V}^\top) = 2K - 2\|\mathbf{H}\|_F^2 = 2\sum_{j=1}^K (1 - D_{jj}^2) = 2\|\sin \Theta(\hat{\mathbf{V}}, \mathbf{V})\|_F^2.
\]  
(B.11.6)

Then the proof is completed by combining the estimates above. \( \square \)

**Lemma B.11.2.** Suppose \( \mathbf{V} \in \mathcal{O}_{d \times K} \) and \( \mathbf{P} \in \mathbb{R}^{d \times d} \) is a projection matrix with rank \( K \). Then

\[
\|(\mathbf{V}\mathbf{V}^\top - \mathbf{P}) - (\mathbf{P}\mathbf{V}\mathbf{V}^\top\mathbf{P} + \mathbf{P}_\perp\mathbf{V}\mathbf{V}^\top\mathbf{P})\|_F \leq \sqrt{2}\|\mathbf{I}_d - \mathbf{P}\|F^2.
\]  
(B.11.7)
Proof. Let $P_\perp = I_d - P$. We have

$$VV^T - P = (P + P_\perp)VV^T(P + P_\perp) - P = (PVV^T P - P) + (PVV^T P_\perp + P_\perp VV^T P) + P_\perp VV^T P_\perp.$$  \hfill (B.11.8)

Therefore,

$$\|(VV^T - P) - (PVV^T P_\perp + P_\perp VV^T P)\|_F^2 = \|(PVV^T P - P) + P_\perp VV^T P_\perp\|_F^2$$

$$= \|PVV^T P - P\|_F^2 + \|P_\perp VV^T P_\perp\|_F^2$$

$$\leq 2\|(I_d - P)V\|_F^4.$$  \hfill (B.11.9)

The inequality follows from Lemma [B.11.3] and submultiplicativity of $\|\cdot\|_F$. \hfill \(\square\)

Lemma B.11.3. Suppose $V \in O_{d \times K}$ and $P \in \mathbb{R}^{d \times d}$ is a projection matrix with rank $K$. Then $\|PVV^T P - P\|_F \leq \|(I_d - P)V\|_F^4$.

Proof. Since $VV^T$ is a projection matrix with rank $K$, Lemma [B.11.4] tells us that

$$\|PVV^T P - P\|_F^2 = \sum_{j=1}^K |1 - \sigma_j(PVV^T P)|^2 \leq \left( \sum_{j=1}^K |1 - \sigma_j(PVV^T P)| \right)^2$$

$$= \left( \sum_{j=1}^K |1 - \sigma_j(PVV^T P)| \right)^2 = \left( K - \sum_{j=1}^K \sigma_j(PVV^T P) \right)^2$$  \hfill (B.11.10)

$$= [K - \text{Tr}(PVV^T P)]^2 = (K - \|PV\|_F^2)^2$$

$$= \|(I_d - P)V\|_F^4.$$  Here we used the fact that $K = \|V\|_F^2 = \|PV\|_F^2 + \|(I_d - P)V\|_F^2$. \hfill \(\square\)

Lemma B.11.4. Suppose $P$ and $\hat{P}$ are $d \times d$ projection matrices with rank $K \leq d$. Then $\|P\hat{P}P - P\|_F^2 = \sum_{j=1}^K [1 - \sigma_j(P\hat{P}P)]^2$. 

195
Proof. There exists $Q \in \mathcal{O}_{d \times d}$ such that $QPQ^\top = D = \begin{pmatrix} I_K & 0_{K \times (d-K)} \\ 0_{(d-K) \times K} & 0_{(d-K) \times (d-K)} \end{pmatrix}$.

$$\|P\hat{P}P - P\|_F^2 = \|Q(P\hat{P}P - P)Q^\top\|_F^2 = \|QP\hat{P}PQ^\top - D\|_F^2. \quad (B.11.11)$$

We claim that $QP\hat{P}PQ^\top = \begin{pmatrix} (QP\hat{P}PQ^\top)_{[K][K]} & 0_{K \times (d-K)} \\ 0_{(d-K) \times K} & 0_{(d-K) \times (d-K)} \end{pmatrix}$, where $(\cdot)_{[K][K]}$ denotes the $K \times K$ submatrix in the upper left corner. This is because

$$QP\hat{P}PQ^\top = (QPQ^\top)(\hat{P}Q^\top)(QPQ^\top) = D(\hat{P}Q^\top)D. \quad (B.11.12)$$

Lemma B.11.5 yields

$$\|P\hat{P}P - P\|_F^2 = \|(QP\hat{P}PQ^\top)_{[K][K]} - I_K\|_F^2 = \sum_{j=1}^{K} (1 - \sigma_j[(QP\hat{P}PQ^\top)_{[K][K]}])^2$$

$$= \sum_{j=1}^{K} [1 - \sigma_j(QP\hat{P}PQ^\top)]^2 = \sum_{j=1}^{K} [1 - \sigma_j(\hat{P}P)]^2. \quad (B.11.13)$$

Lemma B.11.5. If $A \in \mathbb{I}_{r \times r}$ is symmetric, then $\|A - I_r\|_F^2 = \sum_{j=1}^{r} [1 - \sigma_j(A)]^2$. 

196
Appendix C

Proofs for Chapter 4

C.1 Proof of Theorem 4.2.1

Proof. This Lemma is just a simple application of the theoretical framework established in [93], but for completeness and clarity, we present the whole proof here. We start from the optimality of \( \hat{\Theta} \):

\[
-\langle \hat{\Sigma}_{YX}, \hat{\Theta} \rangle + \frac{1}{2} \text{vec}(\hat{\Theta})^T \hat{\Sigma}_{XX} \text{vec}(\hat{\Theta}) + \lambda_N \| \hat{\Theta} \|_* \\
\leq -\langle \hat{\Sigma}_{YX}, \Theta^* \rangle + \frac{1}{2} \text{vec}(\Theta^*)^T \hat{\Sigma}_{XX} \text{vec}(\Theta^*) + \lambda_N \| \Theta^* \|_* .
\]

Note that \( \hat{\Delta} = \hat{\Theta} - \Theta^* \). Simple algebra delivers that

\[
\frac{1}{2} \text{vec}(\hat{\Delta})^T \hat{\Sigma}_{XX} \text{vec}(\hat{\Delta}) \leq \langle \hat{\Sigma}_{YX} - \text{mat}(\hat{\Sigma}_{XX} \text{vec}(\Theta^*)), \hat{\Delta} \rangle + \lambda_N \| \hat{\Delta} \|_* \\
\leq \| \hat{\Sigma}_{YX} - \text{mat}(\hat{\Sigma}_{XX} \text{vec}(\Theta^*)) \|_{op} \cdot \| \hat{\Delta} \|_* + \lambda_N \| \hat{\Delta} \|_* \quad (\text{C.1.1})
\]

if \( \lambda_N \geq 2 \| \hat{\Sigma}_{YX} - \text{mat}(\hat{\Sigma}_{XX} \text{vec}(\Theta^*)) \|_{op} \). To bound the RHS of (C.1.1), we need to decompose \( \hat{\Delta} \) as [90] did. Let \( \Theta^* = UDV^T \) be the SVD of \( \Theta^* \), where the diagonals of \( D \) are in the decreasing order. Denote the first \( r \) columns of \( U \) and \( V \) by \( U^r \) and ...
\(V^r \) respectively, and define

\[
\mathcal{M} := \{ \Theta \in \mathbb{R}^{d_1 \times d_2} | \text{row}(\Theta) \subseteq \text{col}(V^r), \text{col}(\Theta) \subseteq \text{col}(U^r) \}, \\
\mathcal{M}^\perp := \{ \Theta \in \mathbb{R}^{d_1 \times d_2} | \text{row}(\Theta) \perp \text{col}(V^r), \text{col}(\Theta) \perp \text{col}(U^r) \},
\]

where \(\text{col}(\cdot)\) and \(\text{row}(\cdot)\) denote the column space and row space respectively. For any \(\Delta \in \mathbb{R}^{d_1 \times d_2}\) and Hilbert space \(W \subseteq \mathbb{R}^{d_1 \times d_2}\), let \(\Delta_W\) be the projection of \(\Delta\) onto \(W\).

We first clarify here what \(\Delta_M, \Delta_M^\perp\) and \(\Delta_{M^\perp}\) are. Write \(\Delta\) as

\[
\Delta = [U^r, U^r^\perp] \begin{bmatrix} \Gamma_{11} & \Gamma_{12} \\ \Gamma_{21} & \Gamma_{22} \end{bmatrix} \begin{bmatrix} V^r, V^r^\perp \end{bmatrix}^T,
\]

then the following equalities hold:

\[
\Delta_M = U^r \Gamma_{11}(V^r)^T, \quad \Delta_M^\perp = U^r^\perp \Gamma_{22}(V^r^\perp)^T, \\
\Delta_{M^\perp} = [U^r, U^r^\perp] \begin{bmatrix} \Gamma_{11} & \Gamma_{12} \\ \Gamma_{21} & 0 \end{bmatrix} [V^r, V^r^\perp]^T.
\]

Applying Lemma 1 in [93] to our new loss function implies that if \(\lambda_N \geq 2\|\hat{\Sigma}_{XY} - \text{mat}(\hat{\Sigma}_{XX} \text{vec}(\Theta^*))\|_{op}\), it holds that

\[
\|\hat{\Delta}_{M^\perp}\|_* \leq 3\|\hat{\Delta}_{M^\perp}\|_* + 4 \sum_{j \geq r+1} \sigma_j(\Theta^*). \tag{C.1.4}
\]

Note that \(\text{rank}(\hat{\Delta}_{M^\perp}) \leq 2r\); we thus have

\[
\|\hat{\Delta}\|_* \leq \|\hat{\Delta}_{M^\perp}\|_* + \|\hat{\Delta}_{M^\perp}\|_* \leq 4\|\hat{\Delta}_{M^\perp}\|_* + 4 \sum_{j \geq r+1} \sigma_j(\Theta^*) \leq 4\sqrt{2r}\|\hat{\Delta}\|_F + 4 \sum_{j \geq r+1} \sigma_j(\Theta^*). \tag{C.1.5}
\]

198
Following the proof of Corollary 2 in [90], we determine the value of $r$ here. For a threshold $\tau > 0$, we choose

$$r = \#\{j \in \{1, 2, \ldots, (d_1 \wedge d_2)\} | \sigma_j(\Theta^*) \geq \tau\}.$$ 

Then it follows that

$$\sum_{j \geq r+1} \sigma_j(\Theta^*) \leq \tau \sum_{j \geq r+1} \frac{\sigma_j(\Theta^*)}{\tau} \leq \tau \sum_{j \geq r+1} (\frac{\sigma_j(\Theta^*)}{\tau})^q \leq \tau^{1-q} \sum_{j \geq r+1} \sigma_j(\Theta^*)^q \leq \tau^{1-q}\rho. \tag{C.1.6}$$

On the other hand, $\rho \geq \sum_{j \leq r} \sigma_j(\Theta^*)^q \geq r\tau^q$, so $r \leq \rho\tau^{-q}$. Combining (C.1.1), (C.1.5) and $\text{vec}(\hat{\Delta})^T \hat{\Sigma}_{XX} \text{vec}(\hat{\Delta}) \geq \kappa_L \|\hat{\Delta}\|_F^2$, we have

$$\frac{1}{2} \kappa_L \|\hat{\Delta}\|_F^2 \leq 2\lambda_N(4\sqrt{2r}\|\hat{\Delta}\|_F + 4\tau^{1-q}\rho),$$

which implies that

$$\|\hat{\Delta}\|_F \leq 4\sqrt{\frac{\lambda_N\rho}{\kappa_L}} \left(\sqrt{\frac{32\lambda_N\tau^{-q}}{\kappa_L}} + \sqrt{\tau^{1-q}}\right).$$

Choosing $\tau = \lambda_N/\kappa_L$, we have for some constant $C_1$,

$$\|\hat{\Delta}\|_F \leq C_1 \sqrt{\rho} \left(\frac{\lambda_N}{\kappa_L}\right)^{\frac{1-q}{2}}.$$ 

Combining this result with (C.1.5) and (C.1.6), we can further derive the statistical error rate in terms of the nuclear norm as follows.

$$\|\hat{\Delta}\|_* \leq C_2 \rho \left(\frac{\lambda_N}{\kappa_L}\right)^{1-q},$$

where $C_2$ is certain positive constant.
Proof of Lemma 4.2.1

Proof. (a) We first prove for the case of the sub-Gaussian design. Recall that we use \( \tilde{Y}_i \) to denote \( \text{sgn}(Y_i)(|Y_i| \wedge \tau) \) and \( \tilde{x}_i = x_i \) in this case. Let \( \hat{\sigma}_{x_j\tilde{Y}}(\tau) = \frac{1}{N} \sum_{i=1}^{N} \tilde{Y}_i x_{ij} \).

Note that

\[
\text{Var}(\tilde{Y}_i x_{ij}) \leq \mathbb{E}(\tilde{Y}_i^2 x_{ij}^2) \leq (\mathbb{E}Y_i^2)^{\frac{1}{2}}(\mathbb{E}x_{ij}^{2k})^{\frac{k-1}{k}} \leq 2M^{\frac{k}{k}}k_0^2/(k - 1) < \infty, \tag{C.2.1}
\]

which is a constant that we denote by \( v_1 \). In addition, for \( p > 2 \),

\[
\mathbb{E}|\tilde{Y}_i x_{ij}|^p \leq \tau^{p-2} \mathbb{E}(Y_i^2 |x_{ij}|^p) \leq \tau^{p-2} M^{\frac{k}{k}} \left( \mathbb{E}|x_{ij}|^{kp} \right)^{\frac{k-1}{k}} \leq \tau^{p-2} M^{\frac{k}{k}} \left( \kappa_0 \sqrt{\frac{kp}{k - 1}} \right)^p.
\]

By the Jensen’s inequality and then Stirling approximation, it follows that for some constants \( c_1 \) and \( c_2 \),

\[
\mathbb{E}|\tilde{Y}_i x_{ij} - \mathbb{E}\tilde{Y}_i x_{ij}|^p \leq 2^{p-1}(\mathbb{E}|\tilde{Y}_i x_{ij}|^p + \mathbb{E}|\tilde{Y}_i x_{ij}|^p) \leq 2^{p-1}(\mathbb{E}|Y_i x_{ij}|^p + \mathbb{E}|Y_i x_{ij}|^p) \leq \frac{c_1 p! (c_2 \tau)^{p-2}}{2}.
\]

Define \( v := c_1 \lor v_1 \). According to Bernstein’s Inequality (Theorem 2.10 in [11]), we have for \( j = 1, \ldots, d \),

\[
P\left( |\hat{\sigma}_{x_j\tilde{Y}}(\tau) - \mathbb{E}(\tilde{Y}_i x_{ij})| \geq \sqrt{\frac{2vt}{N} + \frac{c_2 \tau t}{N}} \right) \leq 2 \exp(-t).
\]

Thus by the union bound,

\[
P\left( \|\hat{\sigma}_{x_j\tilde{Y}}(\tau) - \mathbb{E}(\tilde{Y}_i x_{ij})\|_{\max} \geq \sqrt{\frac{2vt}{N} + \frac{c_2 \tau t}{N}} \right) \leq 2d \exp(-t). \tag{C.2.2}
\]
Also note that by Markov's inequality,

\[
\mathbb{E}((\tilde{Y}_i - Y_i)x_{ij}) \leq \mathbb{E}(|Y_i x_{ij}| \cdot 1_{|Y_i| > \tau}) \leq \sqrt{\mathbb{E}(Y_i^2 x_{ij}^2) P(|Y_i| > \tau)} \leq \sqrt{vY_i^2 \tau^2} \leq \frac{\sqrt{vM^2 \tau}}{\tau}.
\]  

(C.2.3)

Choose \(\tau \approx \sqrt{N/\log d}\). Then we have for any \(\delta > 1\),

\[
P\left(\|\hat{\sigma}_{x_j}\tilde{Y}(\tau) - \mathbb{E}(\tilde{Y}_i x_{ij})\|_{\max} \geq C_1 \sqrt{\frac{\delta \log d}{N}}\right) \leq 2d^{1-\delta},
\]  

(C.2.4)

where \(C_1\) is a constant depending on \(v\) and \(M\). Besides, since \(\text{Var}(Y_i) = \beta^T \Sigma_x \beta^* + \mathbb{E}e^2 \leq M^2\frac{1}{\kappa} \) and \(\lambda_{\text{min}}(\Sigma_{xx}) \geq \kappa_{\ell}, \|\beta^*\|_2 \leq M^{\frac{1}{2}} \kappa_0 / \kappa_{\ell}\). Therefore, \(\|x_i^T \beta^*\|_{\psi_2} \leq M^{\frac{1}{2}} \kappa_0 / \kappa_{\ell}\) and \(\|x_i x_i^T \beta^*\|_{\psi_1} \leq 2M^{\frac{1}{2}} \kappa_0^2 / \kappa_{\ell}\). By Proposition 5.16 (Bernstein-type inequality) in [123], we have for sufficiently small \(t\),

\[
P\left(\|(\hat{\Sigma}_{x_j \tilde{Y}}(\tau))^T \beta^* - \mathbb{E}(\tilde{Y}_i x_{ij})\| \geq C_2 \sqrt{\frac{t}{N}}\right) \leq \exp(-t)
\]  

(C.2.5)

for some constant \(C_2\). An application of the triangle inequality and union bound yields that as long as \(\log d/N < \gamma_1\) for certain \(\gamma_1\), we have for some constant \(\nu_1 > 0\),

\[
P\left(\left\|\hat{\Sigma}_{YX}(\tau) - \frac{1}{N} \sum_{i=1}^{N} (x_i^T \theta^*) x_i\right\|_{\infty} \geq \nu_1 \sqrt{\frac{\delta \log d}{N}}\right) \leq 2d^{1-\delta}.
\]

(b) Now we switch to the case where both the noise and the design have only bounded moments. Note that

\[
\left\|\hat{\Sigma}_{YX} - \hat{\Sigma}_{XX} \theta^*\right\|_{\infty} \leq \left\|\hat{\Sigma}_{YX} - \Sigma_{YX}\right\|_{\infty} + \left\|\Sigma_{YX} - \Sigma_{XX}\right\|_{\infty} + \left\|\hat{\Sigma}_{XX} - \Sigma_{XX}\right\|_{\infty} = T_1 + T_2 + T_3 + T_4.
\]
We bound the four terms one by one. For $1 \leq j \leq d$, analogous to (C.2.1),
\[
\text{Var}(\tilde{Y}_i \tilde{x}_{ij}) \leq \mathbb{E}(\tilde{Y}_i \tilde{x}_{ij})^2 \leq \mathbb{E}(Y_i x_{ij})^2 \leq \sqrt{\mathbb{E}Y_i^4\mathbb{E}x_{ij}^4} =: v_1 < \infty.
\]

In addition, $\mathbb{E}|\tilde{Y}_i \tilde{x}_{ij}|^p \leq (\tau_1 \tau_2)^{p-2}v_1$. Therefore according to Bernstein’s Inequality (Theorem 2.10 in [11]), we have
\[
P\left(|\tilde{\sigma}_{\tilde{Y}_i \tilde{x}_j} - \sigma_{\tilde{Y}_i \tilde{x}_j}| \geq \sqrt{\frac{2v_1 t}{N} + \frac{c\tau_1 \tau_2 t}{N}}\right) \leq \exp(-t),
\]
where $\tilde{\sigma}_{\tilde{Y}_i \tilde{x}_j} = \frac{1}{N} \sum_{i=1}^N \tilde{Y}_i \tilde{x}_{ij}$, $\sigma_{\tilde{Y}_i \tilde{x}_j} = \mathbb{E}\tilde{Y}_i \tilde{x}_{ij}$ and $c$ is certain constant. Then by the union bound, we have
\[
P\left(|T_1| > \sqrt{\frac{2v_1 t}{N} + \frac{c\tau_1 \tau_2 t}{N}}\right) \leq d \exp(-t).
\]

Next we bound $T_2$. Note that for $1 \leq j \leq d$,
\[
\mathbb{E}\tilde{Y}_i \tilde{x}_{ij} - \mathbb{E}Y_i x_{ij} = \mathbb{E}\tilde{Y}_i \tilde{x}_{ij} - \mathbb{E}\tilde{Y}_i x_{ij} + \mathbb{E}\tilde{Y}_i x_{ij} - \mathbb{E}Y_i x_{ij}
\]
\[
= \mathbb{E}\tilde{Y}_i (\tilde{x}_{ij} - x_{ij}) + \mathbb{E}(\tilde{Y}_i - Y_i) x_{ij}
\]
\[
\leq \sqrt{\mathbb{E}(Y_i^2(\tilde{x}_{ij} - x_{ij})^2)} P(|x_{ij}| \geq \tau_2) + \sqrt{\mathbb{E}((\tilde{Y}_i - Y_i)^2 x_{ij}^2)} P(|Y_i| \geq \tau_1)
\]
\[
\leq \sqrt{M v_1 \left(\frac{1}{\tau_2^2} + \frac{1}{\tau_1^2}\right)},
\]

202
which delivers that $T_2 \leq \sqrt{Mv_1(1/\tau_1^2 + 1/\tau_2^2)}$. Then we bound $T_3$. For $1 \leq j \leq d$,

$$
\mathbb{E}((\mathbf{x}_i^T \theta^*) \tilde{x}_{ij}) - \mathbb{E}((\mathbf{x}_i^T \theta^*) x_{ij}) = \sum_{k=1}^{d} \mathbb{E}(\theta_k^* (\tilde{x}_{ik} \tilde{x}_{ij} - x_{ik} x_{ij})) 
\leq \sum_{k=1}^{d} |\theta_k^*| \mathbb{E}(|\tilde{x}_{ij} \tilde{x}_{ik} - x_{ij} x_{ik}|)
\leq \sum_{k=1}^{d} |\theta_k^*| (\mathbb{E}|x_{ij}(\tilde{x}_{ik} - x_{ik})| + \mathbb{E}|(\tilde{x}_{ij} - x_{ij}) x_{ik}|)
\leq \sum_{k=1}^{d} |\theta_k^*| \left( \mathbb{E}|x_{ij}(\tilde{x}_{ik} - x_{ik})| \mathbb{I}\{|x_{ik}| > \tau_2\} + \mathbb{E}|(\tilde{x}_{ij} - x_{ij}) x_{ik}| \mathbb{I}\{|x_{ij}| > \tau_2\} \right)
\leq CR \tau_2^2.
$$

Finally we bound $T_4$. For $1 \leq j, k \leq d$, we have $|\tilde{x}_{ik} \tilde{x}_{ij}| \leq \tau_2^2$ and $\text{Var}(\tilde{x}_{ij} \tilde{x}_{ik}) \leq M =: v_2$. Therefore according to Bernstein’s inequality (Theorem 2.10 in [11]),

$$
P\left(|\tilde{\sigma}_{\tilde{x}_j, \tilde{x}_k} - \sigma_{\tilde{x}_j, \tilde{x}_k}| \geq \sqrt{\frac{2v_2 t}{N} + \frac{c\tau_2^2 t}{N}} \right) \leq \exp(-t), \quad (C.2.6)
$$

where $\tilde{\sigma}_{\tilde{x}_j, \tilde{x}_k} = (1/N) \sum_{i=1}^{N} \tilde{x}_{ij} \tilde{x}_{ik}$, $\sigma_{\tilde{x}_j, \tilde{x}_k} = \mathbb{E}(\tilde{x}_{ij} \tilde{x}_{ik})$ and $c$ is certain constant. We therefore have

$$
P\left(|T_4| \geq R\left(\sqrt{\frac{2v_2 t}{N} + \frac{c\tau_2^2 t}{N}} \right) \right) \leq d^2 \exp(-t).
$$

Now we choose $\tau_1, \tau_2 \asymp (N / \log d)^{1/4}$, then it follows that for some constant $\nu_2 > 0$,

$$
P\left(\|\hat{\Sigma}_{\tilde{x}} - \tilde{\Sigma}_{\tilde{x}x}\theta^*\|_{\max} > \nu_2 \sqrt{\frac{\delta \log d}{N}} \right) \leq 2d^{1-d}.
$$

\[\square\]
C.3 Proof of Lemma 4.2.2

Proof. (a) We first consider the sub-Gaussian design. Again, since \( \tilde{x}_i = x_i \) in this case, we do not add any tilde above \( x_i \) in this proof. Define \( \hat{D}_{xx} \) to be the diagonal of \( \hat{\Sigma}_{xx} \). Note that given \( \|x_i\|_{\psi_2} \leq \kappa_0 \) and \( \lambda_{\min}(\Sigma_{xx}) \geq \kappa_\mathcal{L} > 0 \), it follows that

\[
\forall v \in S^{d-1}, \sqrt{\mathbb{E}(v^T x_i)^4} \leq 4\kappa_0^2 \leq (4\kappa_0^2/\kappa_\mathcal{L}) v^T \Sigma_{xx} v.
\]

In addition, since \( \|x_{ij}\|_{\psi_2} \leq \kappa_0 \), \( (\mathbb{E}x_{ij}^8)^{\frac{1}{4}} \leq 8\kappa_0^3 \leq (8\kappa_0^2/\kappa_\mathcal{L}) \mathbb{E}x_{ij}^2 \). According to Lemma 5.2 in [97], for any \( \eta_1 > 1 \),

\[
P\left( \forall v \in \mathbb{R}^d : v^T \hat{\Sigma}_{xx} v \geq \frac{1}{2} v^T \Sigma_{xx} v - \frac{c_1(1 + 2\eta_1 \log d)}{N} \|\hat{D}_{xx}^{1/2} v\|_1^2 \right) \geq 1 - \frac{d^{1-\eta_1}}{3},
\]

where \( c_1 \) is some universal constant. For any \( 1 \leq j \leq d \), we know that \( \|x_{ij}\|_{\psi_1} \leq 2\|x_{ij}\|_{\psi_2} = 2\kappa_0^2 \), therefore by the Bernstein-type inequality we have for sufficiently small \( t \),

\[
P\left( \left| \frac{1}{N} \sum_{i=1}^n x_{ij}^2 - \mathbb{E}x_{ij}^2 \right| \geq t \right) \leq 2 \exp(-cN t^2),
\]

where \( c \) depends on \( \kappa_0 \). Note that \( \mathbb{E}x_{ij}^2 \leq 2\kappa_0^2 \). An application of the union bound delivers that when \( t \) is sufficiently small,

\[
P(\|\hat{D}_{xx}\|_{op} \geq 2\kappa_0^2 + t) \leq 2d \exp(-cN t^2).
\]

Let \( t < \kappa_0^2 \), then the inequality above yields that for a new constant \( C_2 > 0 \),

\[
P(\|\hat{D}_{xx}\|_{op} \geq 3\kappa_0^2) \leq 2d \exp(-C_2 N).
\]
Combining the inequality above with (C.3.1), it follows for a new constant $C_1 > 0$ that

$$P\left( \forall \mathbf{v} \in \mathbb{R}^d : \mathbf{v}^T \hat{\Sigma}_{xx} \mathbf{v} \geq \frac{1}{2} \mathbf{v}^T \Sigma_{xx} \mathbf{v} - \frac{C_1 (1 + 2\eta_1 \log d)}{N} \| \mathbf{v} \|_1^2 \right)$$

$$\geq 1 - \frac{d^{1-\eta}}{3} - 2d \exp(-C_2 N).$$

**(b)** Now we switch to the case of designs with only bounded moments. We will show that the sample covariance of the truncated design has the restricted eigenvalue property. Recall that for $i = 1, ..., N, j = 1, ..., d$, $\tilde{x}_{ij} = \text{sgn}(x_{ij})(|x_{ij}| \wedge \tau_2)$. Note that

$$\mathbf{v}^T \hat{\Sigma}_{xx} \mathbf{v} = \mathbf{v}^T (\hat{\Sigma}_{xx} - \Sigma_{xx}) \mathbf{v} + \mathbf{v}^T (\Sigma_{xx} - \Sigma_{xx}) \mathbf{v} + \mathbf{v}^T \Sigma_{xx} \mathbf{v}.$$ 

According to (C.2.6), we have for some $c_1 > 0$,

$$P\left( \| \hat{\Sigma}_{xx} - \Sigma_{xx} \|_{\infty} \geq \sqrt{\frac{2Mt}{N}} + \frac{c_1 \tau^2}{N} \right) \leq d^2 \exp(-t).$$

Given $\tau_2 \asymp (N/\log d)^{1/4}$, we have for some constant $c_2 > 0$ and any $\eta_2 > 2$,

$$P\left( \| \hat{\Sigma}_{xx} - \Sigma_{xx} \|_{\infty} \geq c_2 \sqrt{\frac{\eta_2 \log d}{N}} \right) \leq d^{2-\eta_2}.$$

In addition, for any $1 \leq j_1, j_2 \leq d$, we have

$$\mathbb{E}x_{ij_1}x_{ij_2} - \mathbb{E}\tilde{x}_{ij_1}\tilde{x}_{ij_2} \leq \mathbb{E}\left( |x_{ij_1}x_{ij_2}| (\mathbb{1}_{|x_{ij_1}| \geq \tau_2} + \mathbb{1}_{|x_{ij_2}| \geq \tau_2}) \right)$$

$$\leq \sqrt{M} \left( \sqrt{P(|x_{ij_1}| \geq \tau_2)} + \sqrt{P(|x_{ij_2}| \geq \tau_2)} \right)$$

$$\leq \sqrt{M} \left( \sqrt{\frac{\mathbb{E}x_{ij_1}^2}{\tau^2_2}} + \sqrt{\frac{\mathbb{E}x_{ij_2}^2}{\tau^2_2}} \right) \leq c_3 \sqrt{\frac{\log d}{N}},$$

for some $c_3 > 0$, which implies that $\| \Sigma_{xx} - \Sigma_{xx} \|_{\infty} = \mathcal{O}_P(\sqrt{\log d/N})$. Therefore we have

$$P\left( \forall \mathbf{v} \in \mathbb{R}^d, \mathbf{v}^T \hat{\Sigma}_{xx} \mathbf{v} \geq \mathbf{v}^T \Sigma_{xx} \mathbf{v} - (c_2 \eta_2 + c_3) \sqrt{\frac{\log d}{N}} \| \mathbf{v} \|_1 \right) \leq d^{2-\eta_2}.$$
C.4 Proof of Theorem 4.2.2

Proof. Suppose \( \| \hat{\Delta} \|_2 \geq c_1 \sqrt{\rho \left( \frac{\lambda_N}{\kappa_L} \right)^{1-\frac{q}{2}}} \) for some \( c_1 > 0 \), then by (C.1.5) and (C.1.6),

\[
\| \hat{\Delta} \|_1 \leq 4\sqrt{2r} \| \hat{\Delta} \|_2 + 4r^{1-q} \rho \leq 4\sqrt{2\kappa_L^2 \rho \lambda_N^{\frac{q}{2}}} \| \hat{\Delta} \|_2 + 4\kappa_L^{-1} \rho \lambda_N^{1-q}
\]

(C.4.1)

For any \( \delta > 1 \), since \( \lambda_N = 2\nu \sqrt{\delta \log d/N} \), it is easy to verify that as long as \( \rho (\log d/N)^{1-\frac{q}{2}} \leq C_1 \) for some constant \( C_1 \), we will have by (C.4.1) and (4.2.3) in Lemma 4.2.2

\[
P \left( \hat{\Delta}^T \hat{\Sigma}_{xx} \hat{\Delta} \geq \frac{\kappa_L}{4} \| \hat{\Delta} \|_2^2 \right) \geq 1 - d^{1-\delta}.
\]

An application of Theorem 4.2.1 delivers that for constants \( c_2, c_3 > 0 \),

\[
\| \hat{\Delta} \|_2 \leq c_2 \sqrt{\rho \left( \frac{\lambda_N}{\kappa_L} \right)^{1-\frac{q}{2}}} \quad \text{and} \quad \| \hat{\Delta} \|_1 \leq c_3 \rho \left( \frac{\lambda_N}{\kappa_L} \right)^{1-\frac{q}{2}}.
\]

When \( \| \hat{\Delta} \|_2 \leq c_4 \sqrt{\rho \left( \frac{\lambda_N}{\kappa_L} \right)^{1-\frac{q}{2}}} \), we can still obtain the \( \ell_1 \) norm bound that \( \| \hat{\Delta} \|_1 \leq c_4 \rho \left( \frac{\lambda_N}{\kappa_L} \right)^{1-\frac{q}{2}} \) for some constant \( c_4 \) through (C.1.5) and (C.1.6). Overall, we can achieve the conclusion for some constants \( C_2 \) and \( C_3 \) that with probability at least \( 1 - 3d^{1-\delta} \),

\[
\| \hat{\Delta} \|_2 \leq C_2 \rho \left( \frac{\delta \log d}{N} \right)^{1-\frac{q}{2}} \quad \text{and} \quad \| \hat{\Delta} \|_1 \leq C_3 \rho \left( \frac{\delta \log d}{N} \right)^{1-\frac{q}{2}}.
\]

When the design only satisfies bounded moment conditions, again we first assume that \( \| \hat{\Delta} \|_2 \geq c_5 \sqrt{\rho (\lambda_N/\kappa_L)^{1-\frac{q}{2}}} \) for some constant \( c_5 \). Analogous to the case of the sub-Gaussian design, it is easy to verify that for any \( \delta > 1 \), as long as \( \rho (\log d/N)^{1-\frac{q}{2}} \leq C_4 \)
for some constant $C_4$, we will have by (C.4.1) and (4.2.3) in Lemma 4.2.2

$$P\left(\hat{\Delta}^T \hat{\Sigma}_{xx} \hat{\Delta} \geq \frac{\kappa L}{2} \|\hat{\Delta}\|_2^2\right) \geq 1 - d^{1-\delta},$$

An application of Theorem [4.2.1] delivers that for some constants $c_6, c_7 > 0$,

$$\|\hat{\Delta}\|_2 \leq c_7 \sqrt{\rho} \left(\frac{\lambda N}{\kappa L}\right)^{1-\frac{q}{2}}$$

and

$$\|\hat{\Delta}\|_1 \leq c_8 \sqrt{\rho} \left(\frac{\lambda N}{\kappa L}\right)^{1-\frac{q}{2}}.$$

When $\|\hat{\Delta}\|_2 \leq c_5 \sqrt{\rho} (\lambda N / \kappa L)^{1-\frac{q}{2}}$, we have exactly the same steps as those in the case of the sub-Gaussian design. \qed

### C.5 Proof of Lemma 4.2.3

**Proof.** Recall that $\tilde{Y}_i = \text{sgn}(Y_i)(|Y_i| \wedge \tau)$, then

$$\|\hat{\Sigma}_{YX}(\tau) - \frac{1}{N} \sum_{i=1}^{N} \langle X_i, \Theta^* \rangle X_i \|_{op}$$

$$\leq \|\hat{\Sigma}_{YX}(\tau) - \mathbb{E}(\tilde{Y}X)\|_{op} + \|\mathbb{E}((\tilde{Y} - Y)X)\|_{op} + \|\frac{1}{n} \sum_{i=1}^{n} \langle X_i, \Theta^* \rangle X_i - \mathbb{E}YX\|_{op}. \quad (C.5.1)$$

Next we use the covering argument to bound each term of the RHS. Let $S^{d-1} = \{u \in \mathbb{R}^d : \|u\|_2 = 1\}$, $\mathcal{N}^{d-1}$ be the $1/4$-net on $S^{d-1}$ and $\Phi(A) = \sup_{u \in \mathcal{N}^{d-1}, v \in \mathcal{N}^{d-2}} u^T Av$ for any matrix $A \in \mathbb{R}^{d_1 \times d_2}$, then we claim

$$\|A\|_{op} \leq \frac{16}{7} \Phi(A). \quad (C.5.2)$$

To establish the claim above, note that by the definition of the $1/4$-net, for any $u \in S^{d_1-1}$ and $v \in S^{d_2-1}$, there exist $u_1 \in \mathcal{N}^{d_1-1}$ and $v_1 \in \mathcal{N}^{d_2-1}$ such that $\|u -
\( \|u_1\|_2 \leq 1/4 \) and \( \|v - v_1\|_2 \leq 1/4 \). Then it follows that

\[
\begin{align*}
  u^T Av &= u_1^T Av_1 + (u - u_1)^T Av_1 + u_1^T A(v - v_1) + (u - u_1)^T A(v - v_1) \\
  &\leq \Phi(A) + \left( \frac{1}{4} + \frac{1}{4} + \frac{1}{16} \right) \sup_{u \in S^{d_1-1}, v \in S^{d_2-1}} u^T Av.
\end{align*}
\]

Taking the superlative over \( u \in S^{d_1-1} \) and \( v \in S^{d_2-1} \) on the LHS yields \( (C.5.2) \).

Now fix \( u \in S^{d_1-1} \) and \( v \in S^{d_2-1} \). Later on we always write \( u^T X_i v \) as \( Z_i \) and \( u^T X_i v \) as \( Z \) for convenience. Consider

\[
  u^T (\Sigma_{X_i}(\tau) - \mathbb{E}(\tilde{Y}X))v = \frac{1}{n} \sum_{i=1}^{n} \tilde{Y}_i Z_i - \mathbb{E}(\tilde{Y}Z).
\]

Note that

\[
\begin{align*}
  \mathbb{E}(\tilde{Y}_i Z_i - \mathbb{E}\tilde{Y}_i Z_i)^2 &\leq \mathbb{E}(\tilde{Y}_i Z_i)^2 \leq \mathbb{E}(Y_i Z_i)^2 = \mathbb{E}(Z_i^2 \mathbb{E}(Y_i^2 | X_i)) \\
  &= \mathbb{E}\left(Z_i^2 \left( (X_i, \Theta^*) \right)^2 \right) + \mathbb{E}\left(Z_i^2 \mathbb{E}(\tilde{e}_i^2 | X_i) \right) \\
  &\leq 16R^2 \kappa_0^4 + (2k\kappa_0^2/(k-1)) \left( \mathbb{E}(\mathbb{E}(\tilde{e}_i^2 | X_i)^k) \right)^{\frac{1}{k}} \\
  &\leq 16R^2 \kappa_0^4 + 2k\kappa_0^2 M^\frac{1}{k} / (k-1) < \infty,
\end{align*}
\]

which we denote by \( v_1 \) for convenience. Also we have

\[
\begin{align*}
  \mathbb{E}|\tilde{Y}_i Z_i|^p &\leq \tau^{p-2} \mathbb{E}(Y_i^2 | Z_i |^p) = \tau^{p-2} \mathbb{E}\left( (X_i, \Theta^*)^2 | Z_i |^p + \tilde{e}_i^2 | Z_i |^p \right) \\
  &\leq \tau^{p-2} \left( \sqrt{\mathbb{E}(X_i, \Theta^*)^4 \mathbb{E}|Z_i |^2p} + \left( \mathbb{E}(\mathbb{E}(\tilde{e}_i^2 | X_i)^k) \right)^{\frac{1}{k}} \left( \mathbb{E}|x_{ij}^p\right)^{1-\frac{1}{k}} \right) \\
  &\leq \tau^{p-2} \left( 4R^2 \kappa_0^2 (\kappa_0 \sqrt{2p})^p + M^\frac{1}{k} (\kappa_0 \sqrt{\frac{pk}{k-1}})^p \right)
\end{align*}
\]

and it holds for constants \( c_1 \) and \( c_2 \) that

\[
\begin{align*}
  \mathbb{E}|\tilde{Y}_i Z_i - \mathbb{E}\tilde{Y}_i Z_i|^p &\leq 2^{p-1}(\mathbb{E}|\tilde{Y}_i Z_i|^p + \mathbb{E}|\tilde{Y}_i Z_i|^p) \leq 2^{p-1}(\mathbb{E}|\tilde{Y}_i Z_i|^p + (\mathbb{E}|Y_i Z_i|)^p) \\
  &\leq c_1 p! (c_2 \tau)^{p-2}.
\end{align*}
\]
where the last inequality uses the Stirling approximation of $p!$. Define $v := c_1 \lor v_1$. Then an application of Bernstein’s Inequality (Theorem 2.10 in [11]) to \( \{ \tilde{Y}_i Z_i \}_{i=1}^N \) delivers
\[
P\left( \left| \frac{1}{N} \sum_{i=1}^N \tilde{Y}_i Z_i - \mathbb{E} \tilde{Y}_i Z_i \right| > \sqrt{\frac{2vt}{N} + \frac{c_2 \tau t}{N}} \right) \leq \exp(-t).
\]
By taking the union bound over all \((u, v) \in \mathcal{N}^{d_1-1} \times \mathcal{N}^{d_2-1}\) and (C.5.2) it follows that
\[
P\left( \left\| \tilde{Y} \mathbf{X} \right\|_{\infty} \geq \frac{16}{t} \left( \sqrt{\frac{2vt}{N} + \frac{c_2 \tau t}{N}} \right) \right) \leq \exp \left( (d_1 + d_2) \log 8 - t \right), \tag{C.5.5}
\]
where $c_1$ is a constant.

Next we aim to bound \( \left\| \mathbb{E}((\tilde{Y} - Y) \mathbf{X}) \right\|_{\infty} \). For any $u \in \mathcal{S}^{d_1-1}$ and $v \in \mathcal{S}^{d_2-1}$, by the Cauchy-Schwartz inequality and the Markov inequality,
\[
\mathbb{E}((\tilde{Y} - Y) Z) \leq \mathbb{E}(|YZ| \cdot 1_{|Y| > \tau}) \leq \sqrt{\mathbb{E}(Y^2 Z^2) P(|Y| > \tau)} \leq \sqrt{\frac{v \mathbb{E}Y^2}{\tau^2}} \leq \sqrt{\frac{v(2R^2 \kappa_0^2 + M_1)}{\tau}}.
\]
Note that the inequality above holds for any \((u, v)\). Therefore
\[
\left\| \mathbb{E}((\tilde{Y} - Y) \mathbf{X}) \right\|_{\infty} \leq \frac{\sqrt{v(2R^2 \kappa_0^2 + M_1)}}{\tau}. \tag{C.5.6}
\]

Now we give an upper bound of the third term on the RHS of (C.5.1). For any $u \in \mathcal{S}^{d_1-1}$ and $v \in \mathcal{S}^{d_2-1}$, \( \left\| (\mathbf{X}_i, \Theta^*) Z_i \right\|_{\psi_1} \leq R \kappa_0^2 \), so by Proposition 5.16 (Bernstein-type inequality) in [123] it follows that for sufficiently small $t$,
\[
P\left( \left| \frac{1}{N} \sum_{i=1}^N (\mathbf{X}_i, \Theta^*) Z_i - \mathbb{E} Y Z \right| > t \right) \leq 2 \exp \left( -\frac{c_3 N t^2}{R^2 \kappa_0^4} \right),
\]
where $c_3$ is a constant. Then an combination of the union bound over all points on $\mathcal{N}^{d_1-1} \times \mathcal{N}^{d_2-1}$ and (C.5.2) delivers
\[
P\left(\left\| \frac{1}{N} \sum_{i=1}^{N} \langle X_i, \Theta^* \rangle X_i - \mathbb{E}YX \right\|_{op} \geq \frac{16}{t} \right) \leq 2 \exp\left( (d_1 + d_2) \log 8 - \frac{c_3 N t^2}{R^2 \kappa_0^4} \right), \tag{C.5.7}
\]

Finally we choose $\tau \approx \sqrt{N/(d_1 + d_2)}$. Combining (C.5.5), (C.5.6) and (C.5.7), we can find a constant $\gamma > 0$ such that as long as $(d_1 + d_2)/N < \gamma$, it holds that
\[
P\left(\left\| \hat{\Sigma} Y X(\tau) - \frac{1}{N} \sum_{i=1}^{N} \langle X_i, \Theta^* \rangle X_i \right\|_{op} > \nu \sqrt{\frac{d_1 + d_2}{N}} \right) \leq \eta \exp(-(d_1 + d_2)),
\]
where $\nu$ and $\eta$ are constants.

\[
\square
\]

\section{C.6 Proof of Theorem 4.2.3}

\textit{Proof.} We first verify the RSC property. According to Proposition 1 in [90], the following inequality holds for all $\Delta \in \mathbb{R}^{d_1 \times d_2}$ with probability at least $1 - 2 \exp(-N/32)$,
\[
\sqrt{\text{vec}(\Delta)^T \Sigma_{XX} \text{vec}(\Delta)} \geq \frac{1}{4} \left( \sqrt{\Sigma_{XX} \text{vec}(\Delta)} \right)_2 - c \left( \sqrt{\frac{d_1}{N}} + \sqrt{\frac{d_2}{N}} \right) \| \Delta \|_* \tag{C.6.1}
\]

Let $\kappa_L = (1/32) \lambda_{\min}(\Sigma_{XX}) > 0$. For ease of notations, write $\tilde{\Theta} - \Theta^*$ as $\hat{\Delta}$. Suppose $\| \hat{\Delta} \|_F \geq c_1 \sqrt{\rho \left( \frac{\lambda_{\min}}{\kappa_L} \right)^{1/2}}$ for some $c_1 > 0$, then by (C.1.5) and (C.1.6),
\[
\| \hat{\Delta} \|_* \leq 4 \sqrt{2} \| \hat{\Delta} \|_F + 4 \tau_1^{-q} \rho \leq 4 \sqrt{2} \rho \kappa_L^{1/2} \lambda_{\min}^{-1/2} \| \hat{\Delta} \|_F + 4 \kappa_L^{-1} \rho \lambda_{\min}^{1/2} \tag{C.6.2}
\]

\[
\leq (4 \sqrt{2} + 4 c_1^{-1}) \sqrt{\rho \kappa_L^{1/2} \lambda_{\min}^{-1/2}} \| \hat{\Delta} \|_F.
\]

210
Since we choose $\lambda_N = 2\nu\sqrt{(d_1 + d_2)/N}$, there exists a constant $C_1$ such that as long as $\rho((d_1 + d_2)/N)^{1-\frac{q}{2}} \leq C_1$, we have by (C.6.1) and (C.6.2)

$$\text{vec}(\hat{\Delta}^T \hat{\Sigma}_{XX} \text{vec}(\hat{\Delta}) \geq \kappa \|\hat{\Delta}\|^2_F$$

with probability at least $1 - 2\exp(-N/32)$. An application of Theorem 4.2.1 delivers that for some constants $c_2, c_3 > 0$, it holds with high probability that

$$\|\hat{\Delta}\|_F \leq c_2 \sqrt{\rho \left( \frac{\lambda_N}{\kappa L} \right)^{1-\frac{q}{2}}} \quad \text{and} \quad \|\hat{\Delta}\|_* \leq c_3 \rho \left( \frac{\lambda_N}{\kappa L} \right)^{\frac{1-q}{2}}.$$ 

When $\|\hat{\Delta}\|_F \leq c_1 \sqrt{\rho \left( \frac{\lambda_N}{\kappa L} \right)^{1-\frac{q}{2}}}$, we can still obtain the $\ell_1$ norm bound that $\|\hat{\Delta}\|_1 \leq c_4 \rho \left( \frac{\lambda_N}{\kappa L} \right)^{\frac{1-q}{2}}$ through (C.1.5) and (C.1.6), where $c_4$ is some constant. Overall, we can achieve the conclusion that,

$$\|\hat{\Delta}\|^2_F \leq C_2 \rho \left( \frac{d_1 + d_2}{N} \right)^{1-\frac{q}{2}} \quad \text{and} \quad \|\hat{\Delta}\|_* \leq C_3 \rho \left( \frac{d_1 + d_2}{N} \right)^{\frac{1-q}{2}}$$ 

with probability at least $1 - C_4 \exp(-(d_1 + d_2))$ for constants $C_2, C_3$ and $C_4$.

\[\square\]

\section*{C.7 Proof of Lemma 4.2.4}

\textit{Proof.} We follow essentially the same strategies as in proof of Lemma 4.2.3. The only difference is that we do not use the covering argument to bound the first term in (C.5.1). Instead we apply the Matrix Bernstein inequality (Theorem 6.1.1 in [116]) to take advantage of the singleton design under the matrix completion setting.
For any fixed \( u \in S^{d_1-1} \) and \( v \in S^{d_2-1} \), write \( u^T X_i v \) as \( Z_i \). Then we have

\[
E(\tilde{Y}_i Z_i) = \sqrt{d_1 d_2} E(\tilde{Y}_i u_{j(i)} v_{k(i)})
\]

\[
= \frac{1}{\sqrt{d_1 d_2}} \sum_{j_0=1}^{d_1} \sum_{k_0=1}^{d_2} E(|\tilde{Y}_i| | j(i) = j_0, k(i) = k_0 \cdot |u_{j_0} v_{k_0}|
\]

\[
\leq \frac{1}{\sqrt{d_1 d_2}} \sum_{j_0=1}^{d_1} \sum_{k_0=1}^{d_2} E(|Y_i| | j(i) = j_0, k(i) = k_0 \cdot |u_{j_0} v_{k_0}|
\]

\[
\leq \sqrt{R^2 + M^\frac{1}{2}} \cdot \frac{1}{\sqrt{d_1 d_2}} \sum_{j_0=1}^{d_1} \sum_{k_0=1}^{d_2} |u_{j_0} v_{k_0}| \leq \sqrt{R^2 + M^\frac{1}{2}}.
\]

Since the above argument holds for all \( u \in S^{d_1-1} \) and \( v \in S^{d_2-1} \), we have

\[
\|E(\tilde{Y}_i X_i)\|_{op} \leq \sqrt{R^2 + M^\frac{1}{2}}.
\]

In addition,

\[
\|E\tilde{Y}_i^2 X_i^T X_i\|_{op} = d_1 d_2 \|E\tilde{Y}_i^2 e_{k(i)} e_{j(i)}^T e_{k(i)}^T e_{k(i)}\|_{op} = d_1 d_2 \|E\tilde{Y}_i^2 e_{k(i)} e_{k(i)}^T\|_{op}
\]

\[
= d_1 d_2 \left\|E \left( \tilde{Y}_i^2 | X_i \right) e_{k(i)} e_{k(i)}^T \right\|_{op}
\]

\[
= \left\| \sum_{k_0=1}^{d_1} \sum_{j_0=1}^{d_2} E \left( \tilde{Y}_i^2 | k(i) = k_0, j(i) = j_0 \right) e_{k_0} e_{k_0}^T \right\|_{op}
\]

\[
\leq d_1 \|E\tilde{Y}_i^2 \leq d_1 (R^2 + M^\frac{1}{2})
\]

Similarly we can get \( \|E\tilde{Y}_i^2 X_i X_i^T\|_{op} \leq d_2 (R^2 + M^\frac{1}{2}) \). Write \( \tilde{Y}_i X_i - E\tilde{Y}_i X_i \) as \( A_i \). Therefore by the triangle inequality, \( \max(\|EA_i^T A_i\|_{op}, \|EA_i A_i^T\|_{op}) \leq (d_1 \vee d_2)(R^2 + M^\frac{1}{2}) \), which we denote by \( \nu \) for convenience. Since \( \|A_i\|_{op} \leq 2\sqrt{d_1 d_2} \tau \), an application of the Matrix Bernstein inequality delivers,

\[
P \left( \|\tilde{\Sigma}_i Y_i \|_{op} \geq t \right) \leq (d_1 + d_2) \exp \left( \frac{-N t^2 / 2}{\nu + \sqrt{d_1 d_2} \tau t / 3} \right), \quad (C.7.1)
\]

212
Next we aim to bound \( \|E((\tilde{Y}_i - Y_i)X_i)\|_\text{op} \). Fix \( u \in S^{d_1-1} \) and \( v \in S^{d_2-1} \). By the Cauchy-Schwartz inequality and the Markov inequality,

\[
E((\tilde{Y}_i - Y_i)Z_i) = \sqrt{d_1 d_2} E((\tilde{Y}_i - Y_i)u_j(i)v_k(i)) = \frac{1}{\sqrt{d_1 d_2}} \sum_{j_0=1}^{d_1} \sum_{k_0=1}^{d_2} E(|\tilde{Y}_i - Y_i| | j(i) = j_0, k(i) = k_0|) \leq \frac{R^2 + M^1 \tau}{\tau} \cdot \frac{1}{\sqrt{d_1 d_2}} \sum_{j_0=1}^{d_1} \sum_{k_0=1}^{d_2} |u_{j_0} v_{k_0}| \leq \frac{R^2 + M^1 \tau}{\tau}.
\]

Note that the inequality above holds for any \((u, v)\). Therefore

\[
\| E((\tilde{Y}_i - Y_i)X_i) \|_\text{op} \leq \frac{R^2 + M^1 \tau}{\tau}. \tag{C.7.2}
\]

Now we give an upper bound of the third term on the RHS of (C.5.1). Denote \( \langle X_i, \Theta^* \rangle X_i - E(\langle X_i, \Theta^* \rangle X_i) \) by \( B_i \). It is not hard to verify that \( \|B_i\|_\text{op} \leq 2R\sqrt{d_1 d_2} \) and \( \max(\|EB_i^T B_i\|_\text{op}, \|EB_i B_i^T\|_\text{op}) \leq (d_1 \lor d_2) R^2 \), it follows that for any \( t \in \mathbb{R} \),

\[
P\left( \frac{1}{N} \sum_{i=1}^{N} \langle X_i, \Theta^* \rangle X_i - EY_iX_i \right) \geq t \left< \frac{-Nt^2/2}{(d_1 \lor d_2) R^2 + 2R\sqrt{d_1 d_2} t/3} \right. \tag{C.7.3}
\]

Finally, choose \( \tau = \sqrt{N/((d_1 \lor d_2) \log(d_1 + d_2))} \). Combining (C.7.1), (C.7.2) and (C.7.3), it is easy to verify that for any \( \delta > 0 \), there exist constants \( \nu \) and \( \gamma \) such that the conclusion holds as long as \( (d_1 \lor d_2) \log(d_1 + d_2)/N < \gamma \).
**C.8 Proof of Theorem 4.2.4**

*Proof.* This proof essentially follows the proof of Lemma 1 in [91]. Write \((d_1 + d_2)/2\) as \(d\). Define a constraint set

\[
C(N; c_0) = \left\{ \Delta \in \mathbb{R}^{d_1 \times d_2}, \Delta \neq 0 \middle| \sqrt{d_1 d_2} \frac{\|\Delta\|_{\max} \cdot \|\Delta\|_*}{\|\Delta\|_F} \leq \frac{1}{c_0 L} \sqrt{\frac{N}{(d_1 + d_2) \log(d_1 + d_2)}} \right\}.
\]

According to Case 1 in the proof of Lemma 1 in [91], if \(\hat{\Delta} \notin C(N; c_0)\), we have

\[
\|\hat{\Delta}\|_F^2 \leq 2c_0 R \sqrt{\frac{d \log d}{N}} \left\{ 8\sqrt{\tau} \|\hat{\Delta}\|_F + 4 \sum_{j=r+1}^{d_1 \wedge d_2} \sigma_j(\Theta^*) \right\},
\]

Following the same strategies in the proof of Theorem 4.2.1 in our work, we have

\[
\|\hat{\Delta}\|_F \leq c_1 \sqrt{\rho} \left(2c_0 R \sqrt{\frac{d \log d}{N}}\right)^{1-\frac{q}{2}}
\]

for some constant \(c_1\). If \(\hat{\Delta} \in C(N; c_0)\), according to Case 2 in proof of Lemma 1 in [91], with probability at least \(1 - C_1 \exp(-C_2 d \log d)\), either \(\|\hat{\Delta}\|_F \leq 512 R / \sqrt{N}\), or \(\text{vec}(\hat{\Delta})^T \hat{\Sigma}_{XX} \text{vec}(\hat{\Delta}) \geq (1/256) \|\hat{\Delta}\|_F^2\), where \(C_1\) and \(C_2\) are certain constants. For the case where \(\|\hat{\Delta}\|_F \leq 512 R / \sqrt{N}\), combining this fact with (C.1.5) and (C.1.6) delivers that

\[
\|\hat{\Delta}\|_* \leq 4 \sqrt{2\rho \tau^{-\frac{q}{2}}} \frac{512 R}{\sqrt{N}} + 4\tau^{1-q} \rho.
\]

We minimize the RHS of the inequality above by plugging in \(\tau = \left(\frac{R^2}{\rho N}\right)^{\frac{1}{2}}\). Then we have for some constant \(c_3 > 0\),

\[
\|\hat{\Delta}\|_* \leq c_3 \left(\rho \left(\frac{R^2}{N}\right)^{1-q}\right)^{\frac{1}{2}}.
\]
For the case where \( \text{vec}(\hat{\Delta})^T \hat{\Sigma}_{XX} \text{vec}(\hat{\Delta}) \geq (1/256)\|\hat{\Delta}\|_F^2 \), it is implied by Theorem 4.2.1 in our work that
\[
\|\hat{\Delta}\|_F \leq c_4 \sqrt{\rho \lambda_N^{1-\frac{q}{2}}} \quad \text{and} \quad \|\hat{\Delta}\|_* \leq c_5 \rho \lambda_N^{1-\frac{q}{2}}
\]
for some constants \( c_4 \) and \( c_5 \). Since \( \lambda_N = \nu \sqrt{\delta (d_1 \vee d_2) \log(d_1 + d_2)/N} \) and \( R \geq 1 \), by Lemma 4.2.4, it holds with probability at least \( 1 - C_1 \exp\left(-C_2(d_1 + d_2)\right) - 2(d_1 + d_2)^{1-\delta} \) that
\[
\|\hat{\Delta}\|_F^2 \leq C_3 \max\left\{ \rho \left(\frac{\delta R^2(d_1 + d_2) \log(d_1 + d_2)}{N}\right)^{1-\frac{q}{2}}, \frac{R^2}{N}\right\}
\]
and
\[
\|\hat{\Delta}\|_* \leq C_4 \max\left\{ \rho \left(\frac{\delta R^2(d_1 + d_2) \log(d_1 + d_2)}{N}\right)^{1-\frac{q}{2}}, \left(\frac{\rho R^{2-2q}}{N^{1-q}}\right)^{\frac{1}{2-q}}\right\}
\]
for constants \( C_3 \) and \( C_4 \).

\[\Box\]

**C.9 Proof of Lemma 4.2.5**

*Proof.* (a) First we study the sub-Gaussian design. Since \( \tilde{x}_i = x_i \), we do not add tildes above \( x, x_i \) or \( x_{ij} \) in this proof. Denote \( (\|y_i\|_4 \wedge \tau) y_i / \|y_i\|_4 \) by \( \tilde{y}_i \), then we have
\[
\|\tilde{\Sigma}_{x\tilde{y}}(\tau) - \frac{1}{n} \sum_{j=1}^n x_j \tilde{x}_j^T \Theta^*\|_\text{op} = \|\tilde{\Sigma}_{y\tilde{y}}(\tau) - \frac{1}{n} \sum_{j=1}^n \Theta^T x_j \tilde{x}_j^T\|_\text{op}
\leq \|\tilde{\Sigma}_{y\tilde{y}}(\tau) - \mathbb{E}(\tilde{y}x^T)\|_\text{op} + \|\mathbb{E}(\tilde{y}x^T - yx^T)\|_\text{op} + \|\tilde{\Sigma}_{xx} \Theta^* - \mathbb{E}(yx^T)\|_\text{op}
\]
(C.9.1)

Let
\[
S_i = \begin{bmatrix}
0 & \tilde{y}_i x_i^T - \mathbb{E}\tilde{y}_i x_i^T \\
x_i \tilde{y}_i^T - \mathbb{E}x_i \tilde{y}_i^T & 0
\end{bmatrix}.
\]
Now we bound $\|\mathbf{E}\mathbf{S}_p^l\|_{op}$ for $p > 2$. When $p$ is even, i.e., $p = 2l$, we have
\[
\mathbf{S}_i^{2l} = \begin{bmatrix}
((\bar{y}_i \mathbf{x}_i^T - \mathbf{E}\bar{y}_i \mathbf{x}_i^T)(\mathbf{x}_i \bar{y}_i - \mathbf{E}\mathbf{x}_i \bar{y}_i))^l & 0 \\
0 & ((\mathbf{x}_i \bar{y}_i - \mathbf{E}\mathbf{x}_i \bar{y}_i)(\bar{y}_i \mathbf{x}_i^T - \mathbf{E}\bar{y}_i \mathbf{x}_i^T))^l
\end{bmatrix}.
\]

For any $\mathbf{v} \in \mathcal{S}^{d_2-1}$,
\[
\mathbf{v}^T(\mathbf{E}(\bar{y}_i \mathbf{x}_i^T \mathbf{x}_i \bar{y}_i^T)) \mathbf{v} = \mathbf{E}((\mathbf{x}_i^T \mathbf{x}_i)^l(\bar{y}_i \mathbf{y}_i)^{l-1}(\mathbf{v}^T \bar{y}_i)^2) \leq \tau^{2l-2} \mathbf{E}((\mathbf{x}_i^T \mathbf{x}_i)^l(\mathbf{v}^T \bar{y}_i)^2) \\
\leq \tau^{2l-2} \mathbf{E}((\mathbf{x}_i^T \mathbf{x}_i)^l(\mathbf{v}^T \mathbf{y}_i)^2) = \tau^{2l-2} \mathbf{E}((\mathbf{v}^T \mathbf{y}_i)^2 | \mathbf{x}_i)(\mathbf{x}_i^T \mathbf{x}_i)^l \\
= \tau^{2l-2} \mathbf{E} ((\mathbf{v}^T \Theta \mathbf{x}_i)^2(\mathbf{x}_i^T \mathbf{x}_i)^l + \mathbf{E}((\mathbf{v}^T \mathbf{e}_i)^2 | \mathbf{x}_i)(\mathbf{x}_i^T \mathbf{x}_i)^l).
\]

(C.9.2)

Also note that for any $\mathbf{v} \in \mathcal{S}^{d_2-1}$, we have
\[
R\|\mathbf{v}\|^2 \geq \mathbf{v}^T \mathbf{E}\mathbf{y}_i \mathbf{y}_i^T \mathbf{v} \geq (\Theta^* \mathbf{v})^T \Sigma_{xx}(\Theta^* \mathbf{v}) \geq \kappa_{\mathcal{E}}\|\Theta^* \mathbf{v}\|^2.
\]

(C.9.3)

Therefore it follows that $\|\Theta^*\|_{op} \leq \sqrt{R/k_{\mathcal{E}}}$. Combining this with the fact that $\|\mathbf{x}_i^T \mathbf{x}_i\|_{\psi_1} \leq 2d_1\kappa_0^2$, it follows that
\[
\mathbf{E}((\mathbf{v}^T \Theta \mathbf{x}_i)^2(\mathbf{x}_i^T \mathbf{x}_i)^l) \leq \sqrt{\mathbf{E}(\mathbf{v}^T \Theta \mathbf{x}_i)^4} \sqrt{\mathbf{E}(\mathbf{x}_i^T \mathbf{x}_i)^{2l}} \leq (2\sqrt{R/k_{\mathcal{E}}}\kappa_0)^2(4d_1\kappa_0^2l)^l.
\]

Also by Hölder Inequality, we have
\[
\mathbf{E}((\mathbf{v}^T \mathbf{e}_i)^2 | \mathbf{x}_i)(\mathbf{x}_i^T \mathbf{x}_i) \leq \mathbf{E}((\mathbf{v}^T \mathbf{e}_i)^2)^{\frac{1}{k}}(\mathbf{E}(\mathbf{x}_i^T \mathbf{x}_i)^{\frac{k}{k-1}})^{1-\frac{1}{k}} \\
\leq M^{\frac{1}{k}}(2d_1\kappa_0^2)^{\frac{lk}{k-1}}(4d_1\kappa_0^2l)^l.
\]

Therefore we have
\[
\|\mathbf{E}(\bar{y}_i \mathbf{x}_i^T \mathbf{x}_i \bar{y}_i^T)\|_{op} \leq c_1 \tau^{2l-2}(d_1l)^l
\]

(C.9.4)

for $l \geq 1$, where $c_1$ and $c_1'$ are some constants. Letting $l = 1$ in the equation above implies that $\|\mathbf{E}(\bar{y}_i \mathbf{x}_i^T \mathbf{x}_i \bar{y}_i^T)\|_{op} \leq \|\mathbf{E}\bar{y}_i \mathbf{x}_i^T \mathbf{x}_i \bar{y}_i^T\|_{op} \leq c_1' d_1$. Therefore it holds that
(\mathbf{y}_i^T - \mathbb{E}\mathbf{y}_i^T)(\mathbf{x}_i\mathbf{y}_i^T - \mathbb{E}\mathbf{x}_i\mathbf{y}_i^T) \leq 2\mathbf{y}_i^T \mathbf{x}_i \mathbf{y}_i^T + 2\mathbb{E}\mathbf{y}_i^T \mathbb{E}\mathbf{x}_i \mathbf{y}_i^T \leq 2\mathbf{y}_i^T \mathbf{x}_i \mathbf{y}_i^T + 2c'_l d_1 I.

In addition, for any commutative positive semi-definite (PSD) matrices \( \mathbf{A} \) and \( \mathbf{B} \), it is true that \( (\mathbf{A} + \mathbf{B})^l \leq 2^{l-1}(\mathbf{A}^l + \mathbf{B}^l) \) for \( l > 1 \). Then it follows that

\[
((\mathbf{y}_i^T - \mathbb{E}\mathbf{y}_i^T)(\mathbf{x}_i\mathbf{y}_i^T - \mathbb{E}\mathbf{x}_i\mathbf{y}_i^T))^l \leq (2\mathbf{y}_i^T \mathbf{x}_i \mathbf{y}_i^T + 2c'_l d_1 I)^l \leq 2^{l-1}((\mathbf{y}_i^T \mathbf{x}_i \mathbf{y}_i^T)^l + (c'_l d_1)^l I).
\]

Therefore we have \( \|\mathbb{E}((\mathbf{y}_i^T - \mathbb{E}\mathbf{y}_i^T)(\mathbf{x}_i\mathbf{y}_i^T - \mathbb{E}\mathbf{x}_i\mathbf{y}_i^T))^l\|_{op} \leq c'_2 (c_2 \tau)^{2l-2}(d_1 l)^l \) for some constants \( c_2 \) and \( c'_2 \). Using similar methods, we can derive \( \|\mathbb{E}((\mathbf{x}_i\mathbf{y}_i^T - \mathbb{E}\mathbf{x}_i\mathbf{y}_i^T)(\mathbf{y}_i^T - \mathbb{E}\mathbf{y}_i^T))^l\|_{op} \leq c'_3 (c_3 \tau)^{2l-2}(d_2 l)^l \) for some constants \( c_3 \) and \( c'_3 \). So we can achieve for some constants \( c_4 \) and \( c'_4 \),

\[
\|\mathbb{E}S_i^{2l}\|_{op} \leq c'_4 (c_4 \tau \sqrt{d_1 \vee d_2})^{2l-2} l^l (d_1 \vee d_2)
\]

When \( p \) is an odd number, i.e., \( p = 2l + 1 \), by \( (C.9.5) \) we have

\[
S_i^{2l+1} = \sqrt{S_i^{4l+2}} \leq 2^{l+1/2} \text{diag} \left( \sqrt{(\mathbf{y}_i^T \mathbf{x}_i \mathbf{y}_i^T)^{2l+1} + (\mathbb{E}\mathbf{y}_i^T \mathbb{E}\mathbf{x}_i \mathbf{y}_i^T)^{2l+1}}, \sqrt{\mathbf{x}_i \mathbf{y}_i^T \mathbf{x}_i \mathbf{y}_i^T)^{2l+1} + (\mathbb{E}\mathbf{x}_i \mathbf{y}_i^T \mathbb{E}\mathbf{y}_i \mathbf{y}_i^T)^{2l+1}} \right)
\]

\[
\leq 2^{l+1/2} \text{diag} \left( (\mathbf{y}_i^T \mathbf{x}_i \mathbf{y}_i^T)^{l+1/2} + (\mathbb{E}\mathbf{y}_i^T \mathbb{E}\mathbf{x}_i \mathbf{y}_i^T)^{l+1/2}, (\mathbf{x}_i \mathbf{y}_i^T \mathbf{x}_i \mathbf{y}_i^T)^{l+1/2} + (\mathbb{E}\mathbf{x}_i \mathbf{y}_i^T \mathbb{E}\mathbf{y}_i \mathbf{y}_i^T)^{l+1/2} \right).
\]

Note that \( (\mathbf{x}_i \mathbf{y}_i^T \mathbf{x}_i)^{1/2} = \mathbf{x}_i \mathbf{y}_i^T \mathbf{y}_i^T / (\|\mathbf{x}_i\|_2 \|\mathbf{y}_i\|_2) \) and \( (\mathbf{y}_i^T \mathbf{x}_i \mathbf{y}_i^T)^{1/2} = \mathbf{y}_i^T \mathbf{x}_i \mathbf{y}_i^T / (\|\mathbf{x}_i\|_2 \|\mathbf{y}_i\|_2) \), so for any \( \mathbf{v} \in S^{d-1} \), we have

\[
\mathbf{v}^T \mathbb{E}(\mathbf{y}_i^T \mathbf{x}_i \mathbf{y}_i^T)^{l+1/2} \mathbf{v} = \mathbb{E}((\mathbf{x}_i^T \mathbf{x}_i)^{l+1/2} (\mathbf{y}_i^T \mathbf{y}_i)^{l-1/2} (\mathbf{v}^T \mathbf{y}_i)^2)
\]

\[
\leq \tau^{2l-1} \mathbb{E}((\mathbf{x}_i^T \mathbf{x}_i)^{l+1/2} (\mathbf{v}^T \mathbf{y}_i)^2).
\]
Following the same steps as in the case of \( l \) being even, we can derive \( \|E(\tilde{y}_i x_i^T \tilde{y}_i^T \bar{y}_i^T)^{l+1/2}\|_{op} \leq c'_5 (c_5 \tau)^{2l-1} (d_1 (l + 1/2))^{l+1/2} \), \( \|E(x_i \tilde{y}_i^T \tilde{y}_i x_i^T)^{l+1/2}\|_{op} \leq c'_6 (c_6 \tau)^{2l-1} (d_2 (l + 1/2))^{l+1/2} \) and finally

\[
\|E S_i^{2l+1}\|_{op} \leq c'_7 (c_7 \tau \sqrt{d_1 \vee d_2})^{2l-1} (l + 1/2)^{l+1/2} (d_1 \vee d_2). \tag{C.9.7}
\]

Define \( \sigma^2 = (c_7 \vee c'_7) (d_1 \vee d_2) \). Then by (C.9.6), we have \( \|E S_i^2\|_{op} \leq \sigma^2 \). Also by combining (C.9.6) and (C.9.7), we can verify the moment constraints of Lemma [C.9.1] for \( p > 2 \), we have

\[
\|E S_i^p\|_{op} \leq p!((c_4 \vee c_7) \tau \sqrt{d_1 \vee d_2})^{p-2} \sigma^2. \tag{C.9.8}
\]

Therefore by Lemma [C.9.1] we have

\[
P\left( \|\tilde{\Sigma}_{\tilde{y}x}(\tau) - E(\tilde{y}x^T)\|_{op} \geq t \right) \leq P\left( \frac{1}{n} \sum_{i=1}^{n} S_i \|_{op} \geq t \right)
\]

\[
\leq (d_1 + d_2) \cdot \exp \left( -c \min \left( \frac{nt^2}{\sigma^2}, \frac{nt}{(c_4 \vee c_7) \tau \sqrt{d_1 \vee d_2}} \right) \right). \tag{C.9.9}
\]

Next we aim to bound \( \|E(\tilde{y}_i x_i^T - y_i x_i^T)\|_{op} \). Note that for any \( u \in S^{d_1-1} \) and \( v \in S^{d_2-1} \),

\[
E((v^T y_i)^2 (u^T x_i)^2) = E((E(v^T y_i)^2 | x_i)(u^T x_i)^2)
\]

\[
= \mathbb{E} \left( (v^T \Theta^* x_i)^2 (u^T x_i)^2 + E((v^T \varepsilon_i)^2 | x_i)(u^T x_i)^2 \right)
\]

\[
:= v^2 < \infty.
\]
For any $u \in S^{d_1-1}$ and $v \in S^{d_2-1}$, by the Cauchy-Schwartz inequality and the Markov inequality,

$$E((v^T \tilde{y_i})(u^T x_i) - (v^T y_i)(u^T x_i)) \leq E(|(v^T y_i)(u^T x_i)| \cdot 1_{\|y\|_2 > \tau})$$

$$\leq \sqrt{E((v^T y_i)(u^T x_i))^2} P(\|y\|_2 > \tau)$$

$$\leq \sqrt{\frac{v^2 E\|y\|^2_2}{\tau^2}} = \frac{v}{\tau} \sqrt{R^2 E\|x\|^2_2 + E\|\epsilon\|^2_2} \leq \frac{v \sqrt{d_2}}{\tau} \sqrt{R^2 \kappa_0^2 + M \delta}.$$  

Since $u$ and $v$ are arbitrary, we have

$$\|E(\tilde{y}_i x_i^T - y_i x_i^T)\|_{op} \leq \frac{v \sqrt{d_2}}{\tau} \sqrt{R^2 \kappa_0^2 + M \delta}. \quad (C.9.10)$$

Now we give an upper bound of the third term on the RHS of (C.9.1). For this term, we still follow the covering argument. Let $N^{d-1}$ be the $1/4$-net on $S^{d-1}$. For any $u \in S^{d_1-1}$ and $v \in S^{d_2-1}$, $\|u^T x_i x_i^T \Theta^* v\|_{\psi_1} \leq \sqrt{R/\kappa_L \kappa_0^2}$, so by Proposition 5.16 (Bernstein-type inequality) in [123] it follows that for sufficiently small $t$,

$$P\left(\left|u^T \tilde{\Sigma}_{xx} \Theta^* v - E(u^T x_i)(v^T y_i)\right| > t\right) \leq 2 \exp \left(-c_8 n t^2\right),$$

where $c_8$ is some positive constant. Then an combination of the union bound over all points on $N^{d_1-1} \times N^{d_2-1}$ and (C.5.2) delivers

$$P\left(\|\tilde{\Sigma}_{xx} \Theta^* - E y x^T\|_2 \geq \frac{16}{t} t\right) \leq 2 \exp \left((d_1 + d_2) \log 8 - c_8 n t^2\right), \quad (C.9.11)$$

where $c_8$ is a constant.

Finally we choose $\tau = O(\sqrt{n/\log(d_1 + d_2)})$, and combining (C.9.9), (C.9.10) and (C.9.11) delivers that for any $\delta > 0$, as long as $(d_1 + d_2) \log(d_1 + d_2)/n < \gamma$ for some
constant $\gamma > 0$,

$$P\left(\|\tilde{\Sigma}_{xy}(\tau_1) - \frac{1}{n} \sum_{j=1}^{n} \Theta^{*T} x_j x_j^T\|_{op} \geq \sqrt{\left(\frac{(\nu + \delta)(d_1 + d_2) \log(d_1 + d_2)}{n}\right)}\right) \leq 2(d_1 + d_2)^{1-\eta \delta},$$

where $\nu$ and $\eta$ are universal constants.

(b) Now we switch to the case of designs with only bounded moments. We first show that for two constants $c_3$ and $c_4$,

$$P\left(\|\tilde{\Sigma}_{xy} - \Sigma_{xy}\|_{op} \geq \sqrt{\left(\frac{(c_3 + \delta)(d_1 + d_2) \log(d_1 + d_2)}{n}\right)}\right) \leq (d_1 + d_2)^{1-c_4 \delta}. \quad (C.9.12)$$

Note that

$$\mathbb{E}(v^T y_i)^4 = \mathbb{E}(v^T \Theta^{*T} x_i + v^T \epsilon_i)^4 \leq 16(\mathbb{E}(v^T \Theta^{*T} x_i)^4 + \mathbb{E}(v^T \epsilon_i)^4).$$

By Rosenthal-type inequality in [58], we have

$$\mathbb{E}(v^T y_i)^4 \leq 16(16R^4 \kappa_0^4 + M) =: V_1.$$

In addition, $\mathbb{E}(v^T x_i)^4 \leq 16 \kappa_0^4 =: V_2$. Let $V := \max(V_1, V_2)$, $Z_i := x_i y_i^T$ and $\tilde{X}_i := \tilde{x}_i \tilde{y}_i^T$. We have

$$\|\tilde{Z}_i - \mathbb{E}\tilde{Z}_i\|_{op} \leq \|Z_i\|_{op} + \|\mathbb{E}Z_i\|_{op} = \|\tilde{x}_i\|_2\|\tilde{y}_i\|_2 + \sqrt{R} \leq (d_1 d_2)^{\frac{3}{2}} \tau^2 + \sqrt{V}.$$  

Also for any $v \in S^{d-1}$, we have

$$\mathbb{E}(v^T \tilde{Z}_i^T \tilde{Z}_i v) = \mathbb{E}(\|\tilde{x}_i\|_2^2(v^T \tilde{y}_i)^2) \leq \mathbb{E}(\|x_i\|_2^2(v^T y_i)^2)$$

$$= \sum_{j=1}^{d} \mathbb{E}(x_{ij}^4(v^T y_i)^2) \leq \sum_{j=1}^{d} \sqrt{\mathbb{E}(x_{ij}^4) \mathbb{E}(v^T y_i)^4} \leq V d_1.$$
Then it follows that \( \|\mathbb{E}\tilde{Z}_i^T\tilde{Z}_i\|_{op} \leq Vd_1 \). Similarly we can obtain \( \|\mathbb{E}\tilde{Z}_i\tilde{Z}_i^T\|_{op} \leq Vd_2 \).

Denote \( d_1 + d_2 \) by \( d \). Since \( \max(\|\mathbb{E}\tilde{X}_i\|_{op}, \|\mathbb{E}\tilde{Z}_i\|_{op}) \leq \|\mathbb{E}Z_i\|_{op} \leq V \),

\[
\max(\|\mathbb{E}(\tilde{Z}_i - \mathbb{E}\tilde{Z}_i)^T(\tilde{Z}_i - \mathbb{E}\tilde{Z}_i)\|_{op}, \|\mathbb{E}(\tilde{Z}_i - \mathbb{E}\tilde{Z}_i)(\tilde{Z}_i - \mathbb{E}\tilde{Z}_i)^T\|_{op}) \leq V(d + 1).
\]

By Corollary 6.2.1 in [115], we have for some constant \( c_1 \),

\[
P\left( \frac{1}{n} \sum_{i=1}^{n} \tilde{Z}_i - \mathbb{E}\tilde{Z}_i \|_{op} > t \right) \leq d \exp\left( -c_1 \left( \frac{nt^2}{V(d + 1)} \wedge \frac{nt}{\sqrt{d}T_1T_2 + \sqrt{V}} \right) \right). \tag{C.9.13}
\]

Now we bound the bias \( \mathbb{E}\|\tilde{Z}_i - \tilde{Z}_i\|_{op} \). For any \( v \in S^{d-1} \), it holds that

\[
\mathbb{E}(v^T(Z_i - \tilde{Z}_i)v) = \mathbb{E}\left( \|v^T x_i \|_4^4 \right) - (v^T \tilde{x}_i)(v^T \tilde{y}_i) \mathbb{I}_{\{\|x_i\|_4 \geq \tau_1 \text{ or } \|y_i\|_4 \geq \tau_2 \}}
\]

\[
\leq \mathbb{E}(\|v^T x_i \|_4^4) \mathbb{I}_{\{\|x_i\|_4 \geq \tau_1 \text{ or } \|y_i\|_4 \geq \tau_2 \}} + \mathbb{E}(\|v^T \tilde{x}_i \|_4^4) \mathbb{I}_{\{\|y_i\|_4 \geq \tau_2 \}}
\]

\[
\leq \sqrt{\mathbb{E}(\|v^T x_i \|_4^4)^2} \mathbb{E}(\|v^T \tilde{x}_i \|_4^4) \mathbb{I}_{\{\|x_i\|_4 \geq \tau_1 \}} + \sqrt{\mathbb{E}(\|v^T y_i \|_4^4)^2} \mathbb{E}(\|v^T \tilde{y}_i \|_4^4) \mathbb{I}_{\{\|y_i\|_4 \geq \tau_2 \}}
\]

\[
\leq \frac{V\sqrt{d_1}}{\tau_1^2} + \frac{V\sqrt{d_2}}{\tau_2^2}.
\]

Therefore we have \( \|\mathbb{E}(Z_i - \tilde{Z}_i)\|_{op} \leq V\sqrt{d(1/\tau_1^2 + 1/\tau_2^2)} \). Choose \( \tau_1 \asymp (n/\log d_1)^{\frac{1}{4}} \), \( \tau_2 \asymp (n/\log d_2)^{\frac{1}{4}} \) and substitute \( t \) with \( \sqrt{\delta d \log d/n} \). Then we obtain \( \text{(C.9.12)} \) by combining the bound of bias and \( \text{(C.9.13)} \).

Finally, note that

\[
\|\hat{\Sigma}_{\tilde{X}\tilde{Y}} - \hat{\Sigma}_{XX}\Theta^*\|_{op} \leq \|\hat{\Sigma}_{\tilde{X}\tilde{Y}} - \Sigma_{XY}\|_{op} + \|\hat{\Sigma}_{XX} - \Sigma_{XX}\|_{op} + \|\Sigma_{XY} - \Sigma_{XX}\Theta^*\|_{op}.
\]

Combining consistency of \( \hat{\Sigma}_{XX} \) established in Theorem 4.3.1 Condition (C1) and \( \text{(C.9.12)} \), we can reach the conclusion of the lemma.

\[\Box\]

**Lemma C.9.1** (Matrix Bernstein Inequality with Moment Constraint). Consider a finite sequence \( \{S_i\}_{i=1}^{n} \) of independent, random, Hermitian matrices with dimensions \( d \times d \). Assume that \( \mathbb{E}S_i = 0 \) and \( \|\mathbb{E}S_i^2\|_{op} \leq \sigma^2 \). Also the following moment conditions...
hold for all $1 \leq i \leq n$ and $p \geq 2$:

$$\|\mathbb{E}S_i^p\|_{op} \leq p!L^{p-2}\sigma^2,$$

where $L$ is a constant. Then for every $t \geq 0$ we have

$$P\left(\left\|\frac{1}{n}\sum_{i=1}^{n} S_i\right\|_{op} \geq t\right) \leq d \exp\left(-\frac{nt^2}{4\sigma^2 + 2Lt}\right) \leq d \cdot \exp\left(-c \min\left(\frac{nt^2}{\sigma^2}, \frac{nt}{L}\right)\right).$$

**Proof of Lemma** [C.9.1] Given the moment constraints, we have for $0 < \theta < 1/L$,

$$\mathbb{E}e^{\theta S_i} = I + \sum_{p=2}^{\infty} \frac{\theta^p \mathbb{E}S_i^p}{p!} \leq I + \sum_{p=2}^{\infty} \sigma^2 L^{p-2}\theta^p I = I + \frac{\theta^2 \sigma^2}{1 - \theta L} I \leq \exp\left(\frac{\theta^2 \sigma^2}{1 - \theta L}\right) I.$$

Let $g(\theta) = \theta^2/(1 - \theta L)$. Owing to the master tail inequality (Theorem 3.6.1 in [116]), we have

$$P\left(\left\|\sum_{i=1}^{n} S_i\right\|_{op} \geq t\right) \leq \inf_{\theta > 0} e^{-\theta t} \text{Tr} \exp\left(\sum_{i} \log \mathbb{E}e^{\theta S_i}\right)$$

$$\leq \inf_{0 < \theta < 1/L} e^{-\theta t} \text{Tr} \exp\left(n\sigma^2 g(\theta)I\right) \leq \inf_{0 < \theta < 1/L} de^{-\theta t} \exp\left(n\sigma^2 g(\theta)\right).$$

Choosing $\theta = t/(2n\sigma^2 + Lt)$, we can reach the conclusion. \qed

**C.10 Proof of Theorem 4.2.5**

**Proof.** (a) First consider the case of the sub-Gaussian design. Denote $\frac{1}{n}\sum_{j=1}^{n} x_j x_j^T$ by $\Sigma_{xx}$. Since $x_j$ is a sub-Gaussian vector, an application of Theorem 5.39 in [123] implies that with probability at least $1 - \exp(-c_1d_1)$, $\|\Sigma_{xx} - \Sigma_{xx}\|_{op} \leq c_2\sqrt{d_1/n}$ and furthermore $\lambda_{\min}(\Sigma_{xx}) \geq \frac{1}{2}\lambda_{\min}(\Sigma_{xx}) > 0$ as long as $d_1/n$ is sufficiently small.
Therefore
\[
\text{vec}(\hat{\Delta})^T \hat{\Sigma}_{XX} \text{vec}(\hat{\Delta}) = \frac{1}{N} \sum_{i=1}^{N} (\hat{\Delta}, X_i)^2 = \frac{1}{N} \sum_{j=1}^{n} \sum_{k=1}^{d_2} (\hat{\Delta}, x_j e_k^T)^2
\]
\[
= \frac{1}{N} \sum_{j=1}^{n} \sum_{k=1}^{d_2} \text{Tr}(x_j^T \hat{\Delta} e_k)^2 = \frac{1}{N} \sum_{k=1}^{d_2} \sum_{j=1}^{n} (x_j^T \hat{\Delta} e_k)^2 \quad (C.10.1)
\]
\[
= \frac{1}{d_2} \sum_{k=1}^{d_2} (\Delta e_k)^T \Sigma_{xx} (\Delta e_k) \geq \frac{1}{2d_2} \lambda_{\min}(\Sigma_{xx}) \| \hat{\Delta} \|^2_F,
\]
with probability at least \(1 - \exp(-c_1 d_1)\). In addition, by Lemma 4.2.5, as long as we choose
\[
\lambda_N = \frac{1}{d_2} \sqrt{(\nu_1 + \delta)(d_1 + d_2) \log(d_1 + d_2)/n}, \quad \lambda_N \geq 2 \| \hat{\Sigma}_{YY} X - \frac{1}{N} \sum_{i=1}^{N} Y_i X_i \|_{op}
\]
with probability at least \(1 - 2(d_1 + d_2)^{1-n\delta}\). Finally, by Theorem 4.2.1 we establish the statistical error rate: there exist \(\gamma_1, \gamma_2 > 0\) that as long as \((d_1 + d_2) \log(d_1 + d_2)/n < \gamma_1\) and \(d_1 + d_2 > \gamma_2\), we have
\[
\| \hat{\Theta}(\lambda_N, \tau) - \Theta^* \|_F \leq C_1 \rho \left( \frac{(\nu_1 + \delta)(d_1 + d_2) \log (d_1 + d_2)}{n} \right)^{1-q/2}
\]
and
\[
\| \hat{\Theta}(\lambda_N, \tau) - \Theta^* \|_* \leq C_2 \rho \left( \frac{(\nu_1 + \delta)(d_1 + d_2) \log (d_1 + d_2)}{n} \right)^{1-q/2}
\]
for some constants \(C_1\) and \(C_2\).

(b) Now we switch to the designs with bounded moments. According to Theorem 4.3.1 with probability at least \(1 - d_1^{1-C\delta}\),
\[
\| \hat{\Sigma}_{XX} - \Sigma_{XX} \|_{op} \leq \sqrt{\frac{\delta d_1 \log d_1}{n}},
\]
which furthermore implies that \(\lambda_{\min}(\hat{\Sigma}_{XX}) \geq \frac{1}{2} \lambda_{\min}(\Sigma_{xx}) > 0\) as long as \(d_1/n\) is sufficiently small. Analogous to \(C.10.1\), we therefore have
\[
\text{vec}(\hat{\Delta})^T \hat{\Sigma}_{XX} \text{vec}(\hat{\Delta}) \geq \frac{1}{2d_2} \lambda_{\min}(\Sigma_{xx}) \| \hat{\Delta} \|^2_F.
\]
223
Combining Theorem 4.2.1 with the inequality above, we can reach the final conclusion.

□

C.11 Proof of Theorem 4.3.1

Proof. We denote \( x_i x_i^T \) by \( \mathbf{X}_i \) and \( \tilde{x}_i \tilde{x}_i^T \) by \( \mathbf{\tilde{X}}_i \) for ease of notations. Note that

\[
\|\mathbf{\tilde{X}}_i - \mathbb{E}\mathbf{\tilde{X}}_i\|_{op} \leq \|\mathbf{\tilde{X}}_i\|_{op} + \|\mathbb{E}\mathbf{\tilde{X}}_i\|_{op} = \|\tilde{x}_i\|_2^2 + \sqrt{R} \leq \sqrt{d\tau^2 + \sqrt{R}} \tag{C.11.1}
\]

Also for any \( v \in S^{d-1} \), we have

\[
\mathbb{E}(v^T \mathbf{\tilde{X}}_i v) = \mathbb{E}((v^T \tilde{x}_i)^2) \leq \mathbb{E}((v^T x_i)^2)
= \sum_{j=1}^{d} \mathbb{E}(x_{ij}^2(v^T x_i)^2) \leq \sum_{j=1}^{d} \sqrt{\mathbb{E}(x_{ij}^4)} \mathbb{E}(v^T x_i)^4 \leq Rd
\]

Then it follows that \( \|\mathbb{E}\mathbf{\tilde{X}}_i^T \mathbf{\tilde{X}}_i\|_{op} \leq Rd \). Since \( \|(\mathbb{E}\mathbf{\tilde{X}}_i)^T \mathbb{E}\mathbf{\tilde{X}}_i\|_{op} \leq \|\mathbb{E}\mathbf{X}_i\|_{op}^2 \leq R \), \( \|\mathbb{E}((\tilde{X}_i - \mathbb{E}\tilde{X}_i)^T(\tilde{X}_i - \mathbb{E}\tilde{X}_i))\|_{op} \leq R(d + 1) \). By Theorem 5.29 (Non-commutative Bernstein-type inequality) in [123], we have for some constant \( c \),

\[
P\left(\left\|\frac{1}{n} \sum_{i=1}^{n} \tilde{X}_i - \mathbb{E}\tilde{X}_i\right\|_{op} > t\right) \leq 2d \exp\left(-c\left(\frac{nt^2}{R(d + 1)} \wedge \frac{nt}{\sqrt{d\tau^2 + \sqrt{R}}}\right)\right). \tag{C.11.2}
\]

Now we bound the bias \( \mathbb{E}\|\mathbf{X}_i - \mathbf{\tilde{X}}_i\|_{op} \). For any \( v \in S^{d-1} \), it holds that

\[
\mathbb{E}(v^T (\mathbf{X}_i - \mathbf{\tilde{X}}_i)v) = \mathbb{E}(((v^T x_i)^2 - (v^T \tilde{x}_i)^2)1_{\{|x_i|_4 > \tau\}})
\leq \mathbb{E}((v^T x_i)^2 1_{\{|x_i|_4 > \tau\}}) \leq \sqrt{\mathbb{E}(v^T x_i)^4 P(|x_i|_4 > \tau)}
\leq \sqrt{\frac{R^2 d}{\tau^4}} = \frac{R \sqrt{d}}{\tau^2}.
\]
Therefore we have $\|\mathbb{E}(X_i - \tilde{X}_i)\|_{op} \leq R\sqrt{d/\tau^2}$. Choose $\tau \asymp (nR/\delta \log d)^{1/4}$ and substitute $t$ with $\sqrt{Rd\log d/n}$. Then we reach the final conclusion by combining the bound of bias and (C.11.2).

Lemma C.11.1 (Gilbert–Varshamov [84], Lemma 4.7). There exist binary vectors $w_1, \ldots, w_T \in \{0, 1\}^d$ such that

1. $d_H(w_j, w_k) \geq d/4$ for all $j \neq k$;
2. $T \geq \exp(d/8)$.

Lemma C.11.2. Suppose that $\beta, \eta \in \mathbb{R}^d$ and $\|\eta\|_2 = \|\beta\|_2$. Let $\Sigma_1 := I + \beta\beta^\top$ and $\Sigma_2 := I + \eta\eta^\top$. Then

$$\text{KL}(N(0, \Sigma_1), N(0, \Sigma_2)) = \frac{\|\eta\|_2^4 - (\eta^\top\beta)^2}{2(1 + \|\eta\|_2^2)}.$$

Proof of Lemma C.11.2. Since $\|\eta\|_2 = \|\beta\|_2$, the matrices $\Sigma_1$ and $\Sigma_2$ share the same set of eigenvalues. Hence $|\Sigma_1| = |\Sigma_2|$ and we have

$$\text{KL}(N(0, \Sigma_1), N(0, \Sigma_2)) = \frac{1}{2} \left\{ \log \left| \frac{\Sigma_2}{\Sigma_1} \right| - d + \text{Tr}\left( (\Sigma_2)^{-1}\Sigma_1 \right) \right\}$$

$$= \frac{1}{2} \left\{ \text{Tr}\left( (\Sigma_2)^{-1}\Sigma_1 \right) - d \right\}$$

$$= \frac{1}{2} \left\{ \text{Tr}\left( (I + \eta\eta^\top)^{-1}(I + \beta\beta^\top) \right) - d \right\}.$$ 

Now, by the Sherman–Morrison Formula,

$$(I + \eta\eta^\top)^{-1} = I - \frac{\eta\eta^\top}{1 + \|\eta\|_2^2}$$

and thus we have

$$\text{KL}(N(0, \Sigma_1), N(0, \Sigma_2)) = \frac{1}{2} \left[ \text{Tr}\left( \left( I - \frac{\eta\eta^\top}{1 + \|\eta\|_2^2} \right)(I + \beta\beta^\top) \right) - d \right]$$

$$= \frac{1}{2} \left( \|\beta\|_2^2 - \frac{\|\eta\|_2^4}{1 + \|\eta\|_2^2} - \frac{(\eta^\top\beta)^2}{1 + \|\eta\|_2^2} \right) = \frac{\|\eta\|_2^4 - (\eta^\top\beta)^2}{2(1 + \|\eta\|_2^2)}.$$
C.12 Proof of Theorem 4.3.2

Proof. Without loss of generality, we assume the dimension $d$ is a multiple of 2, i.e., $d = 2h$, where $h$ is a positive integer. Note that for any $v \in \mathbb{R}^d$, $\mathbb{E}x_v = 0$ and

$$\Sigma_v := \text{cov}(x_v) = I + vv^\top = I + \begin{pmatrix} B_{11} & \cdots & B_{1h} \\ \vdots & \ddots & \vdots \\ B_{h1} & \cdots & B_{hh} \end{pmatrix},$$

where $B_{st} \in \mathbb{R}^{2 \times 2}$ for $s, t \in [h]$. In other words, we partition $vv^\top$ into $h^2$ 2-by-2 blocks.

According to Lemma C.11.1, there exist binary vectors $w_1, \ldots, w_T \in \{0, 1\}^h$ such that

1. $d_H(w_j, w_k) \geq h/4$ for all $j \neq k$;
2. $T \geq \exp(h/8)$.

Let $\theta \leq \pi/4$. Define $b = (\cos \theta, \sin \theta)^\top$ and $c = (\cos \theta, -\sin \theta)^\top$. For any $j \in [T]$, construct $v^{(j)} \in \mathbb{R}^d$ such that

$$v^{(j)} = \frac{1}{\sqrt{h}}(w_{j1}b^\top + (1 - w_{j1})c^\top, w_{j2}b^\top + (1 - w_{j2})c^\top, \ldots, w_{jh}b^\top + (1 - w_{jh})c^\top)^\top,$$

where $w_{jk}$ denotes the $k$th component of $w_j$. Correspondingly, write $I + v^{(j)}v^{(j)^\top}$ as $\Sigma^{(j)}$. Note that $w_{jk}b^\top + (1 - w_{jk})c^\top = b^\top$ if $w_{jk} = 1$ and $w_{jk}b^\top + (1 - w_{jk})c^\top = c^\top$ if $w_{jk} = 0$. Therefore, $v^{(j)}$ is just assembled by several $\frac{1}{\sqrt{h}}b$'s and $\frac{1}{\sqrt{h}}c$'s. We calculate the KL-divergence from $P_j$ to $P_k$ as

$$\text{KL}(P_j, P_k) = \frac{\|v^{(k)}\|_2^4 - (v^{(k)^\top}v^{(j)})^2}{2(1 + \|v^{(k)}\|_2^2)} \leq \frac{1 - \cos^2 2\theta}{4} = \frac{\sin^2 2\theta}{4}. \quad (C.12.1)$$

as required. □

226
For $j \in [T]$, suppose we have $n$ i.i.d. random samples $\{x_i^{(j)}\}_{i=1}^n$ from $P_j$ and we use $P^n_j$ to denote the joint distribution of these $n$ samples. Then it follows that

$$\text{KL}(P^n_j, P^n_k) \leq \frac{n \sin^2 \frac{2\theta}{4}}{4} \leq n \sin^2 \theta.$$ 

Besides, recall that $d_H(w_j, w_k) \geq h/4$ for all $j \neq k$. We have

$$\|\Sigma^{(j)} - \Sigma^{(k)}\|_{\text{op}}^2 \geq \frac{1}{2} \|\mathbf{v}^{(j)} \mathbf{v}^{(j)\top} - \mathbf{v}^{(k)} \mathbf{v}^{(k)\top}\|_F^2 = \frac{1}{2} \sum_{1 \leq s, t \leq h} \|B^{(j)}_{st} - B^{(k)}_{st}\|_F^2$$

$$\geq \frac{1}{2} \sum_{1 \leq s, t \leq h} \|B^{(j)}_{st} - B^{(k)}_{st}\|_F^2 \cdot 1_{\{w_s^{(j)} \neq w_s^{(k)}, w_t^{(j)} \neq w_t^{(k)}\}}$$

$$= \frac{h^{-2} \sin^2 \theta \cos^2 \theta h^2}{4} \geq \frac{\sin^2 \theta}{8}.$$ 

By Fano’s inequality ([103]),

$$\inf \max_{\hat{\Sigma}} P\left(\|\hat{\Sigma} - \Sigma^{(j)}\|_{\text{op}} \geq \frac{\sqrt{2}}{4} \sin \theta\right) \geq 1 - \frac{16(n \sin^2 \theta + \log 2)}{d}.$$ 

Choose $\theta \in \mathbb{R}$ such that $\sin^2 \theta = d/(48n)$. When $d \geq 34 > 48 \log 2$, we have

$$\inf \max_{\hat{\Sigma}} P\left(\|\hat{\Sigma} - \Sigma^{(j)}\|_{\text{op}} \geq \frac{1}{48} \cdot \sqrt{\frac{6d}{n}}\right) \geq \frac{1}{3}.$$ 

Then we have

$$\inf \max_{\Sigma, \mathbf{v}, \|\mathbf{v}\|_2=1} P\left(\|\hat{\Sigma} - \Sigma_v\|_{\text{op}} \geq \frac{1}{48} \sqrt{\frac{6d}{n}}\right) \geq \frac{1}{3}.$$ 

$\Box$
Appendix D

Technical lemmas and proofs for Chapter 5

Lemma D.0.1. Suppose \( E(v^T x_i)^4 \leq R \) for any \( v \in S^{d-1} \). Define the \( \ell_4 \)-norm shrunk samples

\[
\tilde{x}_i := \min(\|x_i\|_4, \tau) \cdot x_i,
\]

where \( \tau \) is a threshold value. Then we have the following:

1. \( \|\tilde{x}_i \tilde{x}_i^T - \mathbb{E}\tilde{x}_i \tilde{x}_i^T\|_{op} \leq \|\tilde{x}_i\|_2^2 + \sqrt{R} \leq \sqrt{d\tau^2} + \sqrt{R} \);

2. \( \|E((\tilde{x}_i \tilde{x}_i^T - \mathbb{E}\tilde{x}_i \tilde{x}_i^T)(\tilde{x}_i \tilde{x}_i^T - \mathbb{E}\tilde{x}_i \tilde{x}_i^T))\|_{op} \leq R(d + 1) \);

3. For all \( \xi > 0 \), \( \mathbb{P}(\|\tilde{\Sigma}_n(\tau) - \Sigma\|_{op} \geq \xi \sqrt{\frac{Rd\log n}{n}}) \leq n^{1-C\xi} \), where \( \tau \propto (nR/(\log n))^{1/4} \) and \( C \) is a universal constant.

Proof. This result is from [46]. For convenience of adapting the lemma to other settings, we present its proof here. Notice that

\[
\|\tilde{x}_i \tilde{x}_i^T - \mathbb{E}\tilde{x}_i \tilde{x}_i^T\|_{op} \leq \|\tilde{x}_i \tilde{x}_i^T\|_{op} + \|\mathbb{E}\tilde{x}_i \tilde{x}_i^T\|_{op} = \|\tilde{x}_i\|_2^2 + \sqrt{R} \leq \sqrt{d\tau^2} + \sqrt{R}. \quad (D.0.1)
\]
Also for any \( v \in S^{d-1} \), we have

\[
\mathbb{E}(v^T \tilde{x}_i^T \tilde{x}_i^T \tilde{x}_i^T v) = \mathbb{E}(\| \tilde{x}_i \|_2^2 (v^T \tilde{x}_i)^2) \leq \mathbb{E}(\| x_i \|_2^2 (v^T x_i)^2) \\
= \sum_{j=1}^{d} \mathbb{E}(x_{ij}^2 (v^T x_i)^2) \leq \sum_{j=1}^{d} \sqrt{\mathbb{E}(x_{ij}^4) \mathbb{E}(v^T x_i)^4} \leq Rd
\]

Then it follows that \( \| \mathbb{E} \tilde{x}_i^T \tilde{x}_i^T \tilde{x}_i^T \|_{op} \leq Rd \). Since \( \| (\mathbb{E} \tilde{x}_i^T \mathbb{E} \tilde{x}_i^T) \|_{op} \leq R \),

\[
\| \mathbb{E}((\tilde{x}_i^T \tilde{x}_i^T - \mathbb{E} \tilde{x}_i^T \mathbb{E} \tilde{x}_i^T)(\tilde{x}_i^T \mathbb{E} \tilde{x}_i^T - \mathbb{E} \tilde{x}_i^T \mathbb{E} \tilde{x}_i^T)) \|_{op} \leq R(d + 1). \tag{D.0.2}
\]

By the matrix Bernstein’s inequality (Theorem 5.29 in [123]), we have for some constant \( c_1 \),

\[
\mathbb{P}\left( \left\| \frac{1}{n} \sum_{i=1}^{n} \tilde{x}_i \tilde{x}_i^T - \mathbb{E} \tilde{x}_i^T \mathbb{E} \tilde{x}_i^T \right\|_{op} > t \right) \leq 2d \exp\left( -c_1 \left( \frac{nt^2}{R(d + 1)} \wedge \frac{nt}{\sqrt{d} \tau^2 + \sqrt{R}} \right) \right). \tag{D.0.3}
\]

For any \( v \in S^{d-1} \), it holds that

\[
\mathbb{E}(v^T (x_i^T x_i^T) v \cdot 1_{\{\| x_i \|_4 \geq \tau \}}) \leq \sqrt{\mathbb{E}(v^T x_i)^4 P(\| x_i \|_4 > \tau)} \leq \sqrt{\frac{R^2 d}{\tau^4}} = \frac{R \sqrt{d}}{\tau^2}. \tag{D.0.4}
\]

Therefore we have

\[
\| \mathbb{E}(x_i x_i^T - \tilde{x}_i \tilde{x}_i^T) \|_{op} \leq R \sqrt{d} / \tau^2. \tag{D.0.5}
\]

Choose \( \tau \asymp (nR / \log d)^{\frac{1}{4}} \) and substitute \( t \) with \( \xi \sqrt{Rd \log n/n} \). Then we reach the final conclusion by combining the concentration bound and bias bound.
D.1 Proof of Lemma 5.2.1

Proof. Define a contraction function

$$\phi(x; \theta) = x^2 \cdot \mathbb{1}_{\{x \leq \theta\}} + (x - 2\theta)^2 \cdot \mathbb{1}_{\{\theta < x \leq 2\theta\}} + (x + 2\theta)^2 \cdot \mathbb{1}_{\{-2\theta \leq x < -\theta\}}.$$ 

One can verify that $\phi(x; \theta) \leq x^2$ for any $\theta$. This contraction function was used in a preliminary version of [93] to establish the RSC of negative log-likelihood. Given any $\Delta \in B_2(0, r)$, by the Taylor expansion, we can find $v \in (0, 1)$ such that

$$\delta \ell_n(\beta^* + \Delta; \beta^*) = \ell_n(\beta^* + \Delta) - \ell_n(\beta^*) - \nabla \ell_n(\beta^*)^\top \Delta = \frac{1}{2} \Delta^\top \mathbf{H}_n(\beta^* + v\Delta) \Delta$$

$$= \frac{1}{2n} \sum_{i=1}^n b''(\tilde{x}_i^\top (\beta^* + v\Delta))((\Delta^\top \tilde{x}_i)^2 \geq \frac{1}{2n} \sum_{i=1}^n b''(\tilde{x}_i^\top (\beta^* + v\Delta))\phi(\Delta^\top \tilde{x}_i; \alpha_1 r) \cdot \mathbb{1}_{\{\beta^* \tilde{x}_i \leq \alpha_2\}}$$

$$\geq \frac{m(\omega)}{2n} \sum_{i=1}^n \phi(\Delta^\top \tilde{x}_i; \alpha_1 r) \cdot \mathbb{1}_{\{\beta^* \tilde{x}_i \leq \alpha_2\}},$$

(D.1.1)

where we choose $\omega = \alpha_1 + \alpha_2 > \alpha_1 r + \alpha_2$ so that the last inequality holds by Condition (1). For ease of notation, let $A_i := \{\|\Delta^\top \tilde{x}_i\| \leq \alpha_1 r\}$ and $B_i := \{\|\beta^* \tilde{x}_i\| \leq \alpha_2\}$. We have

$$\mathbb{E}[\phi(\Delta^\top \tilde{x}_i; \alpha_1 r) \cdot \mathbb{1}_{B_i}] \geq \mathbb{E}[(\Delta^\top \tilde{x}_i)^2 \cdot \mathbb{1}_{A_i \cap B_i}]$$

$$\geq \Delta^\top \mathbb{E}[\mathbb{1}_{A_i \cap B_i}] \Delta - \Delta^\top \mathbb{E}[\mathbb{1}_{A_i \cap B_i}] \Delta$$

$$\geq \Delta^\top \mathbb{E}[\mathbb{1}_{A_i \cap B_i}] \Delta - \Delta^\top \mathbb{E}[\mathbb{1}_{A_i \cap B_i}] \Delta$$

$$\geq \Delta^\top \mathbb{E}[\mathbb{1}_{A_i \cap B_i}] \Delta - \Delta^\top \mathbb{E}[\mathbb{1}_{A_i \cap B_i}] \Delta$$

$$\geq \kappa_0 \|\Delta\|_2^2 - \sqrt{\mathbb{E}(\Delta^\top \tilde{x}_i)^4} \cdot (\mathbb{P}(A_i) + \mathbb{P}(B_i)) - \Delta^\top \mathbb{E}[\mathbb{1}_{A_i}] \Delta.$$

By the Markov Inequality,

$$\mathbb{P}(A_i) \leq \frac{\mathbb{E}(\Delta^\top \tilde{x}_i)^4}{\alpha_1^4 r^4} \leq \frac{R}{\alpha_1^4}$$

and

$$\mathbb{P}(B_i) \leq \frac{\mathbb{E}(\beta^* \tilde{x}_i)^4}{\alpha_2^4 r^4} \leq \frac{R \|\beta^*\|_2^4}{\alpha_2^4} \leq \frac{R L^4}{\alpha_2^4}.$$
Besides, according to (D.0.5),
\[
\Delta^\top \mathbb{E}[x_i x_i^\top - \tilde{x}_i \tilde{x}_i^\top] \Delta \leq \frac{R \sqrt{d} \|\Delta\|_2^2}{\tau_1^2} \leq C_1 R \|\Delta\|_2^2 \sqrt{\frac{d \log d}{n}},
\]
where \(C_1\) is certain constant. Therefore, for sufficiently large \(\alpha_1, \alpha_2, n\) and \(d\),
\[
\mathbb{E}[\phi(\Delta^\top \tilde{x}_i; \alpha_1 r) \cdot 1_{B_i}] \geq \frac{\kappa_0}{2} \|\Delta\|_2^2.
\]  
(D.1.2)

For notational convenience, define \(Z_i := \phi(\Delta^\top \tilde{x}_i; \alpha_1 r) \cdot 1_{B_i} = \phi(\Delta^T \tilde{x}_i \cdot 1_{B_i}; \alpha_1 r)\) and \(\Gamma_r := \sup_{\|\Delta\|_2 \leq r} n^{-1} \sum_{i=1}^n Z_i - \mathbb{E} Z_i\). Then an application of Massart’s inequality (85) delivers that
\[
\mathbb{P}(|\Gamma_r - \mathbb{E} \Gamma_r| \geq \alpha_1^2 r^2 \sqrt{\frac{t}{n}}) \leq 2 \exp\left(-\frac{t}{8}\right).
\]  
(D.1.3)

The remaining job is to derive the order of \(\mathbb{E} \Gamma_r\). Note that \(|\phi(x_1; \theta) - \phi(x_2; \theta)| \leq 2\theta |x_1 - x_2|\) for any \(x_1, x_2 \in \mathbb{R}\). By the symmetrization argument and then Ledoux-Talagrand contraction inequality (see [73], p. 112), for a sequence of i.i.d. Rademacher variables \(\{\gamma_i\}_{i=1}^n\),
\[
\mathbb{E} \Gamma_r \leq 2 \mathbb{E} \sup_{\|\Delta\|_2 \leq r} \left| \frac{1}{n} \sum_{i=1}^n \gamma_i Z_i \right| \leq 8 \alpha_1 r \cdot \mathbb{E} \sup_{\|\Delta\|_2 \leq r} \left| \frac{1}{n} \sum_{i=1}^n \gamma_i \tilde{x}_i \cdot 1_{\{|\beta^* \top \tilde{x}_i| \leq \alpha_2\}, \Delta}\right|
\leq 8 \alpha_1 r^2 \cdot \mathbb{E} \left\| \frac{1}{n} \sum_{i=1}^n \gamma_i \tilde{x}_i \cdot 1_{\{|\beta^* \top \tilde{x}_i| \leq \alpha_2\}} \right\|_2 \leq 8 \alpha_1 r^2 \cdot \sqrt{\mathbb{E} \left\| \frac{1}{n} \sum_{i=1}^n \gamma_i \tilde{x}_i \cdot 1_{\{|\beta^* \top \tilde{x}_i| \leq \alpha_2\}} \right\|_2^2}
\leq 8 \alpha_1 r^2 \cdot \sqrt{\frac{1}{n^2} \sum_{i=1}^n \mathbb{E} \|\tilde{x}_i\|_2^2} \leq 8 \alpha_1 r^2 R^4 \cdot \sqrt{\frac{d}{n}}.
\]

Combining the above inequality with (D.1.1), (D.1.2) and (D.1.3) yields that for any \(t > 0\), with probability at least \(1 - 2 \exp(-t)\), for all \(\Delta \in \mathbb{R}^d\) such that \(\|\Delta\|_2 \leq r\),
\[
\delta \tilde{\ell}_n(\beta; \beta^*) \geq \frac{m \kappa_0}{4} \|\Delta\|_2^2 - \alpha_1^2 \sqrt{\frac{8t}{n}} r^2 - 8 \alpha_1 R^4 \cdot \sqrt{\frac{d}{n}} r^2.
\]
D.2 Proof of Theorem 5.2.1

Proof. Construct an intermediate estimator \( \tilde{\beta}_\eta \) between \( \tilde{\beta} \) and \( \beta^* \):

\[
\tilde{\beta}_\eta = \beta^* + \eta(\tilde{\beta} - \beta^*),
\]

where \( \eta \) = 1 if \( \|\tilde{\beta} - \beta^*\|_2 \leq r \) and \( \eta = r/\|\tilde{\beta} - \beta^*\|_2 \) if \( \|\tilde{\beta} - \beta^*\|_2 > r \). Write \( \tilde{\beta}_\eta - \beta^* \) as \( \tilde{\Delta}_\eta \). By Lemma 5.2.1, it holds with probability at least 1 - \( 2 \exp(-t) \) that

\[
\kappa \|	ilde{\Delta}_\eta\|_2^2 \leq C r^2 (\sqrt{\frac{t}{n}} + \sqrt{\frac{d}{n}}) \leq -\nabla \tilde{\ell}_n(\beta^*)^\top \tilde{\Delta}_\eta \leq \|
abla \tilde{\ell}_n(\beta^*)\|_2 \cdot \|	ilde{\Delta}_\eta\|_2,
\]

which further implies that

\[
\|	ilde{\Delta}_\eta\|_2 \leq \frac{3\|
abla \tilde{\ell}_n(\beta^*)\|_2}{\kappa} + \sqrt{\frac{3c_1 r^2}{\kappa}} \cdot \left(\frac{t}{n}\right)^{\frac{1}{4}} + \sqrt{\frac{3c_2 r^2}{\kappa}} \cdot \left(\frac{d}{n}\right)^{\frac{1}{4}}. \tag{D.2.1}
\]

Now we derive the rate of \( \|
abla \tilde{\ell}_n(\beta^*)\|_2 \).

\[
\nabla \tilde{\ell}_n(\beta^*) = \frac{1}{n} \sum_{i=1}^{n} (\tilde{z}_i - b'(\tilde{x}_i^T \beta^*)) \tilde{x}_i
\]

\[
= \frac{1}{n} \sum_{i=1}^{n} \tilde{z}_i \tilde{x}_i - \mathbb{E}\tilde{z}_i \tilde{x}_i + \mathbb{E}(\tilde{z}_i - b'(\tilde{x}_i^T \beta^*)) \tilde{x}_i + \frac{1}{n} \sum_{i=1}^{n} b'(\tilde{x}_i^T \beta^*) \tilde{x}_i - \mathbb{E}(b'(\tilde{x}_i^T \beta^*) \tilde{x}_i).
\]

(D.2.2)

where \( \tilde{x}_i \) is between \( x_i \) and \( \bar{x}_i \) by the mean value theorem. In the following we will bound \( T_1, T_2 \) and \( T_3 \) respectively.
**Bound for $T_1$:** Define the Hermitian dilation matrix

$$
\tilde{Z}_i := \tilde{z}_i \cdot \begin{pmatrix} 0 & \bar{x}_i^T \\ \bar{x}_i & 0 \end{pmatrix}
$$

Note that

$$
\|\mathbb{E}\tilde{Z}_i^2\|_{op} = \|\mathbb{E}[\tilde{z}_i^2 \cdot \begin{pmatrix} \bar{x}_i^T \bar{x}_i & 0^T \\ 0 & \bar{x}_i \bar{x}_i^T \end{pmatrix}]\|_{op} = \max(\mathbb{E}(\tilde{z}_i^2 \bar{x}_i^T \bar{x}_i), \|\mathbb{E}(\tilde{z}_i^2 \bar{x}_i \bar{x}_i^T)\|_{op})
$$

For any $j \in [d]$,

$$
\mathbb{E}(\tilde{z}_i^2 \cdot \bar{x}_{ij}^2) \leq \sqrt{\mathbb{E}z_i^4 \cdot \mathbb{E}x_{ij}^4} \leq \sqrt{M_1 R},
$$

so $\mathbb{E}[\tilde{z}_i^2 \cdot \bar{x}_i^T \bar{x}_i] \leq d \sqrt{M_1 R}$. In addition, for any $v \in \mathbb{R}^d$ such that $\|v\|_2 = 1$,

$$
\mathbb{E}(\tilde{z}_i^2 (v^T \bar{x}_i)^2) \leq \sqrt{M_1 R}.
$$

We thus have $\|\mathbb{E}\tilde{Z}_i^2\|_{op} \leq d \sqrt{M_1 R}$. In addition, $\|\mathbb{E}\tilde{Z}_i\|_{op} = \mathbb{E}(\tilde{z}_i \cdot \|\bar{x}_i\|_2) \leq \sqrt{\mathbb{E}z_i^2 \mathbb{E}\|\bar{x}_i\|_2^2} \leq \sqrt{d(M_1 R)^{\frac{1}{2}}}$, which further implies that $\|\mathbb{E}(\tilde{Z}_i - \mathbb{E}\tilde{Z}_i)^2\|_{op} \leq (d + 1) \sqrt{M_1 R}$. Also notice that since $\|\bar{x}_i\|_4 \leq \tau_1$ and $\tilde{z}_i \leq \tau_2$, $\|\tilde{Z}_i\|_{op} \leq \frac{1}{2} d^{\frac{1}{2}} \cdot \tau_1 \tau_2$. By the matrix Bernstein’s inequality,

$$
P\left(\|\frac{1}{n} \sum_{i=1}^n \tilde{Z}_i - \mathbb{E}\tilde{Z}_i\|_{op} \geq t\right) \leq d \cdot \exp\left(-c_1 \min\left(\frac{nt^2}{(d + 1)\sqrt{M_1 R}}, \frac{2nt}{d^{\frac{1}{2}} \tau_1 \tau_2}\right)\right).
$$

Given that $\|T_1\|_2 = 2\|n^{-1} \sum_{i=1}^n \tilde{Z}_i - \mathbb{E}\tilde{Z}_i\|_{op}$, it thus holds that

$$
\mathbb{P}\left(\|T_1\|_2 \geq 2t\right) \leq d \cdot \exp\left(-c_1 \min\left(\frac{nt^2}{(d + 1)\sqrt{M_1 R}}, \frac{2nt}{d^{\frac{1}{2}} \tau_1 \tau_2}\right)\right), \quad (D.2.3)
$$

233
To summarize here, we have

\[ \|T_2\|_2 \leq \left| \mathbb{E}((\tilde{z}_i - z_i)\bar{x}_i) \right|_2 + \left| \mathbb{E}(z_i - y_i)\bar{x}_i) \right|_2 + \left| \mathbb{E}(y_i - b'(x_i^T \beta^*)\bar{x}_i) \right|_2 + \left| b'(x_i^T \beta^*) - b'(\bar{x}_i^T \beta^*) \right|_2. \]

Now we work on \( \{T_{2i}\}_{i=1}^{4} \) one by one. For any \( v \in \mathbb{R}^d \) such that \( \|v\|_2 = 1 \),

\[ \left| \mathbb{E}(\tilde{z}_i - z_i)(v^T \bar{x}_i) \right| \leq \mathbb{E}(\|z_i\|^2(v^T x_i) \cdot 1_{\{|z_i| > \tau_2\}}) \leq \sqrt{\mathbb{E}(z_i^2(v^T x_i)^2) \cdot P(|z_i| > \tau_2)}. \]

thus we have \( \|T_{21}\|_2 \leq M_1^3 R^2/\tau_2^2 \). Again, for any \( v \in \mathbb{R}^d \) such that \( \|v\|_2 = 1 \), since \( \|\mathbb{E}\epsilon_i x_i\|_2 \leq M_2 \sqrt{d/n} \),

\[ \mathbb{E}[\epsilon_i(x_i^T v)] = \mathbb{E}[(\bar{x}_i - x_i)^T v) + \mathbb{E}[\epsilon_i(x_i^T v)] \leq \mathbb{E}[\epsilon_i x_i^T v) \cdot 1_{\{|x_i| \geq \tau_1\}}] + M_2 \sqrt{\frac{d}{n}} \leq \sqrt{\mathbb{E}(\epsilon_i x_i^T v)^2} \cdot P(|x_i| \geq \tau_1) + M_2 \sqrt{\frac{d}{n}} \leq (M_1 R)^2 \cdot \frac{\sqrt{dR}}{\tau_1^2} + M_2 \sqrt{\frac{d}{n}}. \]

Therefore \( \|T_{22}\|_2 \leq (M_1 R)^2 \sqrt{dR}/\tau_1^2 + M_2 \sqrt{d/n} \). For \( T_{23} \), since \( \mathbb{E}[y_i - b'(x_i^T \beta^*)|x_i] = 0 \),

\( T_{23} = 0 \). Finally we bound \( T_{24} \). For any \( v \in \mathbb{R}^d \) such that \( \|v\|_2 = 1 \),

\[ \|T_{24}\|_2 \leq M \mathbb{E}((\beta^{T} (x_i - \bar{x}_i)) (v^T \bar{x}_i) \leq M \mathbb{E}[(\beta^{T} x_i) (v^T x_i) \cdot 1_{\{|x_i| \geq \tau_1\}}] \leq M \sqrt{\mathbb{E}(\beta^{T} x_i)^2 (v^T x_i)^2} \cdot P(|x_i| \geq \tau_1) \leq M L \sqrt{dR}/\tau_1^2. \]

To summarize here, we have

\[ \|T_2\|_2 \leq (M_1 R)^2 \left( \frac{\sqrt{M_1}}{\tau_2^2} + \frac{\sqrt{dR}}{\tau_1^2} \right) + M L \sqrt{dR}/\tau_1^2 + M_2 \sqrt{\frac{d}{n}}. \quad (D.2.4) \]
Bound for $T_3$: We apply a similar proof strategy as in the bound for $T_1$. Define the following Hermitian dilation matrix:

$$\tilde{X}_i := b'(\tilde{x}_i^T \beta^*) \cdot \begin{pmatrix} 0 & \tilde{x}_i^T \\ \tilde{x}_i & 0 \end{pmatrix}.$$ 

First, 

$$\|\mathbb{E}\tilde{X}_i^2\|_{op} = \max(\mathbb{E}(b'(\tilde{x}_i^T \beta^*)\tilde{X}_i^2), \|\mathbb{E}b'(\tilde{x}_i^T \beta^*)^2\tilde{x}_i\|_{op}).$$

Write $|b'(1)|$ as $b_1$. For any $j \in [d],$

$$\mathbb{E}(b'(\tilde{x}_i^T \beta^*)^2 \cdot \tilde{x}_{ij}^2) \leq \mathbb{E}[(b_1 + M|\tilde{x}_i^T \beta^* - 1)|^2 \tilde{x}_{ij}^2] \leq 2\mathbb{E}[(b_1 + M)^2 + M^2(\tilde{x}_i^T \beta^*)^2]\tilde{x}_{ij}^2 \leq 2M^2R\|\beta^*\|_2^2 + 2(b_1 + M)^2\sqrt{R} =: V,$$

so $\mathbb{E}[b'(\tilde{x}_i^T \beta^*)^2 \cdot \tilde{x}_i^T \tilde{x}_i] \leq dV.$ In addition, for any $v \in \mathbb{R}^d$ such that $\|v\|_2 = 1$,

$$\mathbb{E}(b'(\tilde{x}_i^T \beta^*)^2(v^T \tilde{x}_i)^2) \leq \mathbb{E}((b_1 + M|\tilde{x}_i^T \beta^* - 1)|^2(v^T \tilde{x}_i)^2) \leq V.$$ 

We thus have $\|\mathbb{E}\tilde{X}_i^2\|_{op} \leq dV$. In addition, $\|\mathbb{E}\tilde{X}_i\|_{op} = \mathbb{E}(b'(\tilde{x}_i^T \beta^*) \cdot \tilde{x}_i) \leq \sqrt{\mathbb{E}b'(\tilde{x}_i^T \beta^*)^2\mathbb{E}\tilde{x}_i^2} \leq dV$, which further implies that $\|\mathbb{E}(\tilde{X}_i - \mathbb{E}\tilde{X}_i)^2\|_{op} \leq (d + \sqrt{d})V$. Also notice that $\|\tilde{X}_i\|_{op} \leq ((b_1 + M) + M\|\beta^*\|_2 \cdot d^{\frac{1}{4}}\tau_1)d^{\frac{1}{4}}\tau_1$. By the matrix Bernstein’s inequality,

$$\mathbb{P}\left(\left\|\frac{1}{n} \sum_{i=1}^{n} \tilde{X}_i - \mathbb{E}\tilde{X}_i\right\|_{op} \geq t\right) \leq d \cdot \exp\left(-c_1 \min\left(\frac{nt^2}{(d + \sqrt{d})V}, \frac{nt}{(b_1 + M + M\|\beta^*\|_2 \cdot d^{\frac{1}{4}}\tau_1)d^{\frac{1}{4}}\tau_1}\right)\right).$$

Given that $\|T_3\|_2 = 2\|n^{-1} \sum_{i=1}^{n} \tilde{X}_i - \mathbb{E}\tilde{X}_i\|_{op}$, it thus holds that

$$\mathbb{P}\left(\|T_3\|_2 \geq 2t\right) \leq d \cdot \exp\left(-c_1 \min\left(\frac{nt^2}{(d + \sqrt{d})V}, \frac{nt}{(b_1 + M + M\|\beta^*\|_2 \cdot d^{\frac{1}{4}}\tau_1)d^{\frac{1}{4}}\tau_1}\right)\right).$$

(D.2.5)
Finally, choose $\tau_1, \tau_2 \asymp (n/\log n)^{\frac{1}{4}}$. Combining \eqref{eq:D.2.3}, \eqref{eq:D.2.4} and \eqref{eq:D.2.5} delivers that for some constant $C_1$ any $\xi > 1$,

$$
P(\|\nabla \tilde{\ell}_n(\beta^*)\|_2 \geq C_1 \xi \sqrt{\frac{d \log n}{n}}) \leq n^{1-\xi}.
\tag{D.2.6}
$$

Choose $t = \xi \log n$ and let $r$ be larger than the RHS of \eqref{eq:D.2.1}. When $d/n$ is sufficiently small and $n$ is sufficiently large, we can obtain that

$$
r \geq C_2 \xi \sqrt{\frac{d \log n}{n}} =: r_0,
$$

where $C_2$ is a constant. Choose $r = r_0$. Then by \eqref{eq:D.2.1}, $\|\Delta_\eta\|_2 \leq r_0$ and thus $\tilde{\Delta} = \tilde{\Delta}_\eta$. Finally, we reach the conclusion that

$$
P(\|\tilde{\Delta}\|_2 \geq C_2 \xi \sqrt{\frac{d \log n}{n}}) \leq n^{1-\xi} + 2n^{-\xi} \leq 3n^{1-\xi}.
$$

\hfill \Box

\section{Proof of Corollary \ref{cor:5.2.1}}

\textbf{Proof.} The proof strategy is nearly the same as that for deriving Theorem \ref{thm:5.2.1}, so we provide a roadmap here and do not dive into great details. For ease of notation, write $n^{-1} \sum_{i=1}^n \ell^u(\tilde{x}_i, z_i; \beta)$ as $\tilde{\ell}^u(\beta)$ and denote the hessian matrix of $\tilde{\ell}^u(\beta)$ by $\tilde{H}^u(\beta)$. Since $\tilde{H}^u(\beta) = \nabla^2 \tilde{\ell}_n(\beta) = \tilde{H}_n(\beta)$, we can directly obtain the uniform strong convexity of
$\tilde{H}_n^w(\beta)$ from Lemma 5.2.1. In addition,
\[
\nabla_{\beta} \tilde{\ell}_n^w(\beta^*) = \frac{1-p}{1-2p} \cdot \frac{1}{n} \sum_{i=1}^{n} (b'(\tilde{x}_i^T \beta^*) - z_i) \tilde{x}_i - \frac{p}{1-2p} \cdot \frac{1}{n} \sum_{i=1}^{n} (b'(\tilde{x}_i^T \beta^*) - (1-z_i)) \tilde{x}_i
\]
\[
= \frac{1-p}{1-2p} (T_1 - E T_1) - \frac{p}{1-2p} (T_2 - E T_2) + \frac{1-p}{1-2p} E T_1 - \frac{p}{1-2p} E T_2
\]
\[
= \frac{1-p}{1-2p} (T_1 - E T_1) - \frac{p}{1-2p} (T_2 - E T_2) + E (b'(\tilde{x}_i^T \beta^*) - y_i) \tilde{x}_i.
\]

Since $|b'(\tilde{x}_i^T \beta^*) - z_i| \leq 1$ and $|b'(\tilde{x}_i^T \beta^*) - (1-z_i)| \leq 1$, following the bound for $T_1$ in Theorem 5.2.1, we will obtain
\[
P(\| \frac{1-p}{1-2p} (T_1 - E T_1) - \frac{p}{1-2p} (T_2 - E T_2) \|_2 \geq c_1 \xi \sqrt{\frac{d \log n}{n}} \n) \leq n^{1-\xi},
\]
where $c_1 > 0$ depends on $R$ and $p$ and $\xi > 1$. In addition, following the bound for $T_{23}$ and $T_{24}$ in Theorem 1, we shall obtain
\[
\| E (b'(\tilde{x}_i^T \beta^*) - y_i) \tilde{x}_i \|_2 \leq M_2 L \frac{\sqrt{d R}}{\tau_1^2} \leq c_2 M_2 \sqrt{\frac{d R \log n}{n}}.
\]
where $c_2 > 0$ is a constant. Therefore, for some constant $c_3$ depending on $R, p, M_2, R$, we have
\[
P(\| \nabla_{\beta} \tilde{\ell}_n^w(\beta^*) \|_2 \geq c_3 \xi \sqrt{\frac{d \log n}{n}} \) \leq n^{1-\xi}.
\]
Combining this with the uniform strong convexity of $\tilde{H}_n^w(\beta)$ delivers the final conclusion.
D.4 Proof of Lemma 5.2.2

Proof. According to (5.1.3), \( \nabla_\beta \tilde{\ell}(\beta^*) \big|_j = (b'(\tilde{x}_i^T \beta^*) - \tilde{z}_i)\tilde{x}_{ij} \). Then we have

\[
\left| \frac{1}{n} \sum_{i=1}^{n} (b'(\tilde{x}_i^T \beta^*) - \tilde{z}_i)\tilde{x}_{ij} \right| \leq \frac{1}{n} \sum_{i=1}^{n} b'(\tilde{x}_i^T \beta^*)\tilde{x}_{ij} - \mathbb{E}b'(\tilde{x}_i^T \beta^*)\tilde{x}_{ij} \right| + \frac{1}{n} \sum_{i=1}^{n} (b'(\tilde{x}_i^T \beta^*) - \tilde{z}_i)\tilde{x}_{ij} + \mathbb{E}(b'(\tilde{x}_i^T \beta^*) - \tilde{z}_i)\tilde{x}_{ij} \right|
\]

\[
+ \left| \frac{1}{n} \sum_{i=1}^{n} \tilde{z}_i\tilde{x}_{ij} - \mathbb{E}\tilde{z}_i\tilde{x}_{ij} \right|
\]

We start with the upper bound of \( T_1 \). By the Mean Value Theorem, for any \( i \in [n] \), there exists \( \xi_i \) between 1 and \( \tilde{x}_i^T \beta^* \) such that \( b'(\tilde{x}_i^T \beta^*) = b'(1) + b''(\xi_i) \cdot (\tilde{x}_i^T \beta^* - 1) \). Therefore we have

\[
T_1 \leq \left| \frac{1}{n} \sum_{i=1}^{n} b'(1)\tilde{x}_{ij} - \mathbb{E}(b'(1)\tilde{x}_{ij}) \right| + \left| \frac{1}{n} \sum_{i=1}^{n} b''(\xi_i)\tilde{x}_{ij} - \mathbb{E}(b''(\xi_i)\tilde{x}_{ij})(\tilde{x}_i^T \beta^* - 1) \right|
\]

\[
\leq \left| \frac{1}{n} \sum_{i=1}^{n} b'(1)\tilde{x}_{ij} - \mathbb{E}(b'(1)\tilde{x}_{ij}) \right| + \sum_{k=1}^{d} |\beta_k| \cdot \left| \frac{1}{n} \sum_{i=1}^{n} b''(\xi_i)\tilde{x}_{ij}\tilde{x}_{ik} - \mathbb{E}b''(\xi_i)\tilde{x}_{ij}\tilde{x}_{ik} \right|
\]

\[
+ \left| \frac{1}{n} \sum_{i=1}^{n} b''(\xi_i)\tilde{x}_{ij} - \mathbb{E}(b''(\xi_i)\tilde{x}_{ij}) \right|
\]

Since \( \text{var}(\tilde{x}_{ij}) \leq \sqrt{R} \) and \( |\tilde{x}_{ij}| \leq \tau_1 \), an application of Bernstein’s inequality (Theorem 2.10 in [1]) yields that

\[
\mathbb{P}\left( \frac{1}{n} \sum_{i=1}^{n} b'(1)\tilde{x}_{ij} - \mathbb{E}(b'(1)\tilde{x}_{ij}) \right) \geq |b'(1)| \left( \frac{\sqrt{R} \cdot 2t}{n} + \frac{c_1 \tau_1 t}{n} \right) \leq 2 \exp(-t),
\]

where \( c_1 > 0 \) is some universal constant. In addition, \( b''(\xi_i)\tilde{x}_{ij}\tilde{x}_{ik} \leq M\tau_1^2 \) and \( \text{var}(b''(\xi_i)\tilde{x}_{ij}\tilde{x}_{ik}) \leq \mathbb{E}(b''(\xi_i)\tilde{x}_{ij}\tilde{x}_{ik})^2 \leq M^2 R \). Again by Bernstein’s inequality,

\[
\mathbb{P}\left( \frac{1}{n} \sum_{i=1}^{n} b''(\xi_i)\tilde{x}_{ij}\tilde{x}_{ik} - \mathbb{E}(b''(\xi_i)\tilde{x}_{ij}\tilde{x}_{ik}) \right) \geq \frac{2M^2 Rt}{n} + \frac{c_1 M\tau_1^2 t}{n} \leq 2 \exp(-t).
\]

238
Similarly,
\[
\mathbb{P}\left( \frac{1}{n} \sum_{i=1}^{n} b''(\xi_i) \tilde{x}_{ij} - \mathbb{E}(b''(\xi_i) \tilde{x}_{ij}) \right) \geq \sqrt{\frac{M^2 \sqrt{R t}}{n}} + \frac{M \tau_1 t}{n} \leq 2 \exp(-t).
\]

Combining the above three inequalities delivers that
\[
\mathbb{P}\left( T_1 \geq b'(1) \left( \sqrt{\frac{R \cdot 2t}{n}} + \frac{c_1 \tau_1 t}{n} + \frac{2M^2 R t}{n} + \frac{c_1 M \tau_1^2 t}{n} + \frac{M^2 \sqrt{R t}}{n} + \frac{M \tau_1 t}{n} \right) \right) \leq 6 \exp(-t).
\]

Now we bound \( T_2 \).
\[
T_2 = \mathbb{E}((z_i - \tilde{z}_i) \tilde{x}_{ij}) + \mathbb{E} \epsilon_i \tilde{x}_{ij} + \mathbb{E}[(b'(\mathbf{x}_i^T \beta^*) - b'(\tilde{\mathbf{x}}_i^T \beta^*)) \tilde{x}_{ij}]
\leq \mathbb{E}[(z_i \tilde{x}_{ij}) \cdot 1_{|z_i| \geq \tau_2}] + \mathbb{E} \epsilon_i \tilde{x}_{ij} + \mathbb{E} \epsilon_i (x_{ij} - \tilde{x}_{ij}) + M \cdot \sum_{k=1}^{d} |\beta^*_k| \cdot \mathbb{E}|\tilde{x}_{ik}(\tilde{x}_{ij} - x_{ij})|
\leq (M_1 R)^{\frac{1}{4}} \cdot \frac{\sqrt{M_1}}{\tau_2} + \frac{M_3}{\sqrt{n}} + \frac{(M_1 R)^{\frac{1}{2}}}{\tau_1} + M M_2 \cdot \frac{\sqrt{R}}{\tau_1}.
\]

Finally we bound \( T_3 \). Note that \(|z_i \tilde{x}_{ij}| \leq \tau_1 \tau_2\), \(\text{var}(\tilde{x}_{ij} \tilde{z}_i) \leq \mathbb{E}|\tilde{z}_i \tilde{x}_{ij}|^2 \leq \sqrt{M_1 R}\). According to the Bernstein’s inequality,
\[
\mathbb{P}\left( |T_3| \geq \sqrt{\frac{2t \sqrt{M_1 R}}{n}} + \frac{c_1 \tau_1 \tau_2 t}{n} \right) \leq 2 \exp(-t).
\]

Choose \( \tau_1, \tau_2 \approx (n/ \log d)^{\frac{1}{4}} \). Combining (D.4.1), (D.4.2) and (D.4.3) delivers that for some constant \( C_1 > 0 \) that depends on \( M, R, \{M_i\}_{i=1}^{3}, b'(1) \) and any \( \xi > 1 \),
\[
\mathbb{P}( |\nabla_\beta \tilde{\ell}(\beta^*)| \geq C_1 \xi \sqrt{\frac{\log d}{n}} ) \leq 2d^{-\xi}.
\]
Then by the union bound for all \( j \in [d] \), it holds that
\[
P(\max_{j \in [d]} |\nabla_{\beta} \tilde{\ell}(\beta^*)|_j | \geq C_1 \xi \sqrt{\frac{\log d}{n}}) \leq 2d^{1-\xi}.
\]

\[\square\]

D.5 Proof of Lemma 5.2.3

Proof. The proof strategy is quite similar to that for Lemma 5.2.1, except that we need to take advantage of the restricted cone \( \mathcal{C}(S) \) that \( \Delta \) lies in. First of all, for any \( 1 \leq j, k \leq d \),
\[
|\mathbb{E}(\tilde{x}_{ij} \tilde{x}_{ik} - x_{ij} x_{ik})| \leq \sqrt{\mathbb{E}(x_{ij} x_{ik})^2 \cdot (\mathbb{P}(|x_{ij}| \geq \tau_1) + \mathbb{P}(|x_{ik}| \geq \tau_1))} \leq \frac{\sqrt{2} R}{\tau^2}.
\]
We thus have
\[
\|\mathbb{E}[x_i x_i^\top - \tilde{x}_i \tilde{x}_i^\top]\|_{\text{max}} \leq \frac{\sqrt{2} R}{\tau^2} \leq CR \sqrt{\frac{2 \log d}{n}}, \tag{D.5.1}
\]
where \( C > 0 \) is some constant. Again, define a contraction function
\[
\phi(x; \theta) = x^2 \cdot 1_{\{x < \theta\}} + (x - 2\theta)^2 \cdot 1_{\{\theta < x < 2\theta\}} + (x + 2\theta)^2 \cdot 1_{\{-2\theta < x < -\theta\}}.
\]
Given any \( \Delta \in B_2(0, r) \cap \mathcal{C}(S) \), by the Taylor expansion, we can find \( v \in (0, 1) \) such that
\[
\delta \tilde{\ell}_n(\beta^* + \Delta, \beta^*) = \tilde{\ell}_n(\beta^* + \Delta) - \tilde{\ell}_n(\beta^*) - \nabla \tilde{\ell}_n(\beta^*)^\top \Delta = \frac{1}{2} \Delta^\top \tilde{H}_n(\beta^* + v\Delta) \Delta
\]
\[
= \frac{1}{2n} \sum_{i=1}^n b''(\tilde{x}_i^\top (\beta^* + v\Delta)) (\Delta^\top \tilde{x}_i)^2 \geq \frac{1}{2n} \sum_{i=1}^n b''(\tilde{x}_i^\top (\beta^* + v\Delta)) \phi(\Delta^\top \tilde{x}_i; \alpha_1 r) \cdot 1_{\{\beta^*^\top \tilde{x}_i \leq \alpha_2\}}
\]
\[
\geq \frac{m(\omega)}{2n} \sum_{i=1}^n \phi(\Delta^\top \tilde{x}_i; \alpha_1 r) \cdot 1_{\{\beta^*^\top \tilde{x}_i \leq \alpha_2\}}, \tag{D.5.2}
\]
where we choose $\omega = \alpha_1 + \alpha_2 > \alpha_1 r + \alpha_2$ so that the last inequality holds by Condition (1). For ease of notation, let $A_i := \{|\Delta^T \tilde{x}_i| \leq \alpha_1 r\}$ and $B_i := \{|\beta^T \tilde{x}_i| \leq \alpha_2\}$. We have

$$
\mathbb{E}[\phi(\Delta^T \tilde{x}_i; \alpha_1 r) \cdot 1_{B_i}] \geq \mathbb{E}[(\Delta^T \tilde{x}_i)^2 \cdot 1_{A_i \cap B_i}]
$$

$$
\geq \Delta^T \mathbb{E}[x_i x_i^T \cdot 1_{A_i \cap B_i}] \Delta - \Delta^T \mathbb{E}[(x_i x_i^T - \tilde{x}_i \tilde{x}_i^T) \cdot 1_{A_i \cap B_i}] \Delta
$$

$$
\geq \Delta^T \mathbb{E}(x_i x_i^T) \Delta - \Delta^T \mathbb{E}(x_i x_i^T \cdot 1_{A_i \cup B_i^c}) \Delta - \Delta^T \mathbb{E}[x_i x_i^T - \tilde{x}_i \tilde{x}_i^T] \Delta
$$

$$
\geq \kappa_0 \|\Delta\|^2 - \sqrt{\mathbb{E}(\Delta^T x_i)^4 \cdot (\mathbb{P}(A_i^c) + \mathbb{P}(B_i^c)) - \Delta^T \mathbb{E}[x_i x_i^T - \tilde{x}_i \tilde{x}_i^T] \Delta}
$$

$$
\geq \kappa_0 \|\Delta\|^2 - \sqrt{R \mathbb{P}(A_i^c) + \mathbb{P}(B_i^c)) \cdot \|\Delta\|^2 - \|\mathbb{E}[x_i x_i^T - \tilde{x}_i \tilde{x}_i^T]\|_{\max} \cdot \|\Delta\|^2}
$$

By the Markov Inequality and (D.5.1),

$$
\mathbb{P}(A_i^c) \leq \frac{\mathbb{E}(\Delta^T \tilde{x}_i)^2}{\alpha_1^2 r^2} \leq \frac{\mathbb{E}(\Delta^T x_i)^2 + \Delta^T \mathbb{E}(\tilde{x}_i \tilde{x}_i^T - x_i x_i^T) \Delta}{\alpha_1^2 r^2}
$$

$$
\leq \frac{\sqrt{R} \|\Delta\|^2 + C R s \|\Delta\|^2 \sqrt{2 \log d/n}}{\alpha_1^2 r^2} \leq \frac{\sqrt{R} + C R s \sqrt{\log d/n}}{\alpha_1^2 r^2}
$$

and

$$
\mathbb{P}(B_i^c) \leq \frac{\mathbb{E}(\beta^T \tilde{x}_i)^2}{\alpha_2^2} \leq \frac{\mathbb{E}(\beta^T x_i)^2 + \beta^T \mathbb{E}(\tilde{x}_i \tilde{x}_i^T - x_i x_i^T) \beta}{\alpha_2^2}
$$

$$
\leq \frac{\sqrt{R} \|\beta\|^2 + C R s \|\beta\|^2 \sqrt{2 \log d/n}}{\alpha_2^2} \leq \frac{\sqrt{R} L^2 + C R L^2 s \sqrt{2 \log d/n}}{\alpha_2^2}.
$$

Overall, as long as $\alpha_1, \alpha_2$ are sufficiently large and $s \sqrt{\log d/n}$ is not large,

$$
\mathbb{E}[\phi(\Delta^T \tilde{x}_i; \alpha_1 r) \cdot 1_{B_i}] \geq \frac{\kappa_0}{2} \|\Delta\|^2. \quad (D.5.3)
$$

For notational convenience, define $Z_i := \phi(\Delta^T \tilde{x}_i; \alpha_1 r) \cdot 1_{B_i} = \phi(\Delta^T \tilde{x}_i \cdot 1_{B_i}; \alpha_1 r)$ and $\Gamma_r := \sup_{\|\Delta\| \leq r, \Delta \in \mathcal{C}(S)} \{n^{-1} \sum_{i=1}^n Z_i - \mathbb{E} Z_i\}$. Then an application of Massart’s inequality
(\text{D.5.4}) delivers that

\[
P\left( |\Gamma_r - \mathbb{E}\Gamma_r| \geq \alpha_1^2 r^2 \sqrt{\frac{t}{n}} \right) \leq 2 \exp\left(-\frac{t}{8}\right). \tag{D.5.4}
\]

The remaining job is to derive the order of $\mathbb{E}\Gamma_r$. By the symmetrization argument and Ledoux-Talagrand contraction inequality, for a sequence of i.i.d. Rademacher variables $\{\gamma_i\}_{i=1}^n$,

\[
\mathbb{E}\Gamma_r \leq 2 \mathbb{E} \sup_{\|\Delta\|_2 \leq r, \Delta \in \mathcal{C}(S)} \left| \frac{1}{n} \sum_{i=1}^n \gamma_i Z_i \right| \leq 8 \alpha_1 r \cdot \mathbb{E} \sup_{\|\Delta\|_2 \leq r, \Delta \in \mathcal{C}(S)} \left| \left\langle \frac{1}{n} \sum_{i=1}^n \gamma_i \tilde{x}_i \cdot 1_{\{|\beta^* \top \tilde{x}_i| \leq \alpha_2\}}, \Delta \right\rangle \right|
\]

\[
\leq 8 \alpha_1 \sqrt{s} r^2 \cdot \mathbb{E} \left\| \frac{1}{n} \sum_{i=1}^n \gamma_i \tilde{x}_i \cdot 1_{\{|\beta^* \top \tilde{x}_i| \leq \alpha_2\}} \right\|_\infty.
\]

For any $1 \leq j \leq d$, by Bernstein inequality,

\[
P\left( \left| \frac{1}{n} \sum_{i=1}^n \gamma_i \tilde{x}_{ij} \cdot 1_{\{|\beta^* \top \tilde{x}_i| \leq \alpha_2\}} \right| \geq \sqrt{\frac{2\sqrt{R} t}{n}} + \frac{C_1 \tau t}{n} \right) \leq 2 \exp(-t),
\]

where $C_1$ is some constant. By the union bound, we can deduce that for some constant $C_2$,

\[
P\left( \left\| \frac{1}{n} \sum_{i=1}^n \gamma_i \tilde{x}_i \cdot 1_{\{|\beta^* \top \tilde{x}_i| \leq \alpha_2\}} \right\|_\infty \geq C_2 \sqrt{\frac{t \log d}{n}} \right) \leq 2d^{1-t},
\]

which further implies that

\[
\mathbb{E}\Gamma_r \leq 8 \alpha_1 \sqrt{s} r^2 \cdot \mathbb{E} \left\| \frac{1}{n} \sum_{i=1}^n \gamma_i \tilde{x}_i \cdot 1_{\{|\beta^* \top \tilde{x}_i| \leq \alpha_2\}} \right\|_\infty \leq 8C_3 \alpha_1 r^2 \sqrt{s \log d \over n}.
\]

for some constant $C_3$. Combining the above inequality with (D.5.2), (D.5.3) and (D.5.4) yields that for any $t > 0$, with probability at least $1 - 2 \exp(-t)$,

\[
\delta_{\tilde{\ell}_n}(\beta; \beta^*) \geq \frac{m \kappa_0}{4} \|\Delta\|_2^2 - \alpha_1^2 r^2 \sqrt{8t \over n} - 8C_3 \alpha_1 r^2 \sqrt{s \log d \over n}.
\]
D.6 Proof of Theorem 5.2.2

Proof. According to Lemma 1 in [93], as long as \( \lambda \geq 2 \left\| \nabla \tilde{\ell}_n(\beta) \right\|_\infty, \tilde{\Delta} \in C(S) \). We construct an intermediate estimator \( \tilde{\beta}_\eta \) between \( \tilde{\beta} \) and \( \beta^* \):

\[
\tilde{\beta}_\eta = \beta^* + \eta (\tilde{\beta} - \beta^*),
\]

where \( \eta = 1 \) if \( \| \tilde{\beta} - \beta^* \|_2 \leq r \) and \( \eta = r / \| \tilde{\beta} - \beta^* \|_2 \) if \( \| \tilde{\beta} - \beta^* \|_2 > r \). Choose \( \lambda = 2C\xi \sqrt{\log d/n} \), where \( C \) and \( \xi \) are the same as in Lemma 5.2.2. By Lemmas 5.2.2 and 5.2.3, it holds with probability at least 1 – 2 exp(-t),

\[
\kappa \| \tilde{\Delta}_\eta \|_2^2 - C_0 r^2 \left( \sqrt{\frac{t}{n} + \frac{s \log d}{n}} \right) \leq \delta \tilde{\ell}_n(\tilde{\beta}; \beta^*) \leq -\nabla \tilde{\ell}_n(\beta^*)^\top \tilde{\Delta}_\eta \leq \left\| \nabla \tilde{\ell}_n(\beta^*) \right\|_\infty \cdot \| \tilde{\Delta}_\eta \|_1 \leq 4 \left\| \nabla \tilde{\ell}_n(\beta^*) \right\|_\infty \cdot \| [\tilde{\Delta}_\eta]_S \|_1 \leq 4 \sqrt{s} \left\| \nabla \tilde{\ell}_n(\beta^*) \right\|_\infty \cdot \| \tilde{\Delta}_\eta \|_2.
\]

(D.6.1)

Some algebra delivers that

\[
\| \tilde{\Delta}_\eta \|_2 \leq \frac{4\sqrt{s} \left\| \nabla \tilde{\ell}_n(\beta^*) \right\|_\infty}{\kappa} + r \sqrt{\frac{C_0}{\kappa} \left( \sqrt{\frac{t}{n} + \frac{s \log d}{n}} \right)}.
\]

(D.6.2)

Choose \( t = \xi \log d \) above. Let \( r \) be greater than the RHS of the inequality above. For sufficiently sufficiently small \( s \log d/n \), we have \( r \geq 5\sqrt{s} \left\| \nabla \tilde{\ell}_n(\beta^*) \right\|_\infty / \kappa \). Define \( r_0 := 5\sqrt{s} \left\| \nabla \tilde{\ell}_n(\beta^*) \right\|_\infty / \kappa \) and choose \( r = r_0 \). Therefore, \( \| \tilde{\Delta}_\eta \|_2 \leq r \) and \( \tilde{\Delta}_\eta = \tilde{\Delta} \).

By Lemma 5.2.2, we reach the conclusion.
Bibliography


Yongming Qu, George Ostrouchov, Nagiza Samatova, and Al Geist. Principal component analysis for dimension reduction in massive distributed data sets. In IEEE International Conference on Data Mining (ICDM), 2002.


253


