Sliced Inverse Regression with Large Structural Dimensions

Dongming Huang[*], Songtao Tian [†‡], Qian Lin[§]

Abstract

The central space of a joint distribution $(\boldsymbol{X}, Y)$ is the minimal subspace $\mathcal{S}$ such that $Y \perp\!\!\!\perp \boldsymbol{X} \mid P_{\mathcal{S}}\boldsymbol{X}$ where $P_{\mathcal{S}}$ is the projection onto $\mathcal{S}$. Sliced inverse regression (SIR), one of the most popular methods for estimating the central space, often performs poorly when the structural dimension $d = \dim(\mathcal{S})$ is large (e.g., $\geqslant 5$). In this paper, we demonstrate that the generalized signal-noise-ratio (gSNR) tends to be extremely small for a general multiple-index model when d is large. Then we determine the minimax rate for estimating the central space over a large class of high dimensional distributions with a large structural dimension d (i.e., there is no constant upper bound on d) in the low gSNR regime. This result not only extends the existing minimax rate results for estimating the central space of distributions with fixed d to that with a large d, but also clarifies that the degradation in SIR performance is caused by the decay of signal strength. The technical tools developed here might be of independent interest for studying other central space estimation methods.

***Keywords*:** Central space, sufficient dimension reduction, sliced inverse regression, structural dimension, minimax rates, multiple-index model

1 Introduction

A subspace $\mathcal{S} \subset \mathbb{R}^p$ is a dimension reduction subspace of a joint distribution of $(\boldsymbol{X}, Y) \in \mathbb{R}^p \times \mathbb{R}$ if $Y \perp\!\!\!\perp \boldsymbol{X} | P_{\mathcal{S}}\boldsymbol{X}$, where $P_{\mathcal{S}}$ denotes the projection operator from $\mathbb{R}^p$ to the subspace $\mathcal{S}$. The central space, denoted by $\mathcal{S}_{Y|\boldsymbol{X}}$, is the intersection of all dimension reduction subspaces. The objective of sufficient dimension reduction (SDR) is to estimate the central space $\mathcal{S}_{Y|\boldsymbol{X}}$. In the case where $(\boldsymbol{X}, Y)$ follows a multiple-index model

$$Y = f(\boldsymbol{\beta}_1^\top \boldsymbol{X}, \dots, \boldsymbol{\beta}_d^\top \boldsymbol{X}, \epsilon), \quad d \ll p \tag{1}$$

($f : \mathbb{R}^{d+1} \to \mathbb{R}$ is an unknown non-parametric link function, $\boldsymbol{\beta}_i \in \mathbb{R}^p$ are the index vectors and ϵ is a random noise independent of $\boldsymbol{X}$), sliced inverse regression(SIR), the first SDR method proposed by Li [1991], aims to estimate the central space span$\{\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_d\}$. Besides SIR, various SDR algorithms have been proposed in the literature: *sliced average variance estimation* (SAVE, Cook and Weisberg [1991]), *principal hessian directions* (PHD, Li [1992]), *directional regression* (DR, Li and Wang [2007]), *minimum average variance estimation* (MAVE, Xia et al. [2009]) and many others. All these methods are widely applied and served as an intermediate step in modelling the relation between Y and $\boldsymbol{X}$ in various fields. Though these SDR methods gained successes widely, the increasing dimension of modern data raises the new challenges to them:

[*] Department of Statistics and Data Science, National University of Singapore. Email: stahd@nus.edu.sg

[†] Department of Mathematical Sciences, Tsinghua University. Email: tst20@mails.tsinghua.edu.cn

[‡] Co-first author

[§] Center for Statistical Science & Department of Industrial Engineering, Tsinghua University; Beijing Academy of Artificial Intelligence. Email: qianlin@tsinghua.edu.cn

we have to develop the high dimensional SDR methods. Understanding the theoretical limitation of these SDR algorithms might be the first step to propose new high dimensional SDR methods.

The asymptotic properties of SIR are of particular interest in recent decades, as SIR is considered one of the most popular SDR methods due to its simplicity and computational efficiency. When the dimension p is either fixed or growing at a slower rate than the sample size n, researchers have extensively studied SIR's asymptotic properties in various settings. For example, Hsing and Carroll [1992] established the root-n consistency and asymptotic normality of SIR when the sample size in each slice (denoted by c) equals 2; Zhu and Ng [1995] provided asymptotic results of SIR for an arbitrary fixed constant c or a growing c with an increasing sample size n; Zhu et al. [2006] discussed the condition for SIR to give a consistent estimator when $p = o(\sqrt{n})$; Wu and Li [2011] determined the convergence rate of SIR for sparse multiple-index models with $p = o(n/\log(n))$.

It becomes more challenging to study the properties of SIR in the high-dimensional regime, where the dimension p could be comparable to or even larger than n. Lin et al. [2018a] studied the situation where $\delta := \lim p/n$ is a constant and proved that the SIR estimator for the central space is consistent if and only if $\delta = 0$. They proposed a modification, DT-SIR, which is consistent in the high-dimensional setting under the sparsity assumption that the central space only depends on a small proportion of the predictors. We hereafter call by *sparse SIR* any method that modifies the original SIR so as to estimate the central space under the sparsity assumption. Lin et al. [2021] established the minimax rate optimality of sparse SIR over a large class of high dimensional multiple-index models. Lin et al. [2019] proposed the Lasso-SIR algorithm, which is computationally efficient and achieves the minimax rate. In a different setting, Tan et al. [2020] studied the minimax rates under various loss functions and proposed a computationally tractable adaptive estimation scheme for sparse SIR.

There are still some limitations in the theory of high-dimensional SIR despite the mentioned progress. First of all, existing minimax theories are restricted because they assume that *the structural dimension* d (i.e., the dimension of the central space $\mathcal{S}$) is either bounded or fixed [Lin et al., 2021, Tan et al., 2020]. It is of interest to relax this assumption and establish the minimax rate for estimating the central space when d is large (i.e., there is no constant upper bound on d). Determining this minimax rate can help explain the observed poor performance of SIR for large d (e.g., $d \geqslant 5$). While this problem has been observed in previous literature (e.g., Ferré [1998], Lin et al. [2021]), no theoretical explanation has been provided so far.

Secondly, a crucial technical condition used in previous studies remains unclear. Lin et al. [2018a] introduced the sliced stable condition (SSC, see Definition 1) to obtain the "key lemma", which is a concentration inequality for the SIR estimator of the conditional covariance matrix $\boldsymbol{\Lambda} := \mathrm{Cov}(\mathbb{E}[\boldsymbol{X}|Y])$. This lemma serves as the main technical tool for developing the asymptotics of SIR in Lin et al. [2018a, 2019, 2021]. However, the main drawback of SSC is its lack of clarity. This issue is partially addressed in Lin et al. [2018a, 2021], which showed that SSC can be derived from a slight modification of the smoothness and tail conditions proposed by Hsing and Carroll [1992]. Although SSC and the modified smoothness and tail conditions are considered as relatively mild, they are defined in terms of the central curve $\boldsymbol{m}(y) = \mathbb{E}[\boldsymbol{X}|Y = y]$ and thus it remains unclear how to verify whether a specific index model satisfies these conditions. This lack of clarity in the conditions poses challenges for further theoretical development of high-dimensional SDR.

1.1 Major contributions

In this article, we address the aforementioned limitations in the theory of the high dimensional SIR, specifically the vagueness of SSC and the boundedness condition on the structural dimension d in the minimax theory. We relax the requirement in SSC and study the fundamental limits of estimating the central space when d is large, using a decision-theoretic approach.

Our first major contribution is the introduction of a relatively mild condition called weak SSC (see Definition 2) to overcome the vagueness of SSC. Our key finding is that all the results established under SSC in previous studies (such as Lin et al. [2018a, 2019, 2021]), including the "key lemma," still hold if SSC is replaced by weak SSC. We prove that weak SSC holds under some mild conditions that are readily interpretable: $\sup_{\|\boldsymbol{\beta}\|=1} \mathbb{E}[|\langle \boldsymbol{X}, \boldsymbol{\beta} \rangle|^{\ell}] < \infty$ for some $\ell > 2$, Y is a continuous random variable, and

$\mathbb{E}[\boldsymbol{X} \mid Y = y]$ is a continuous function. This relaxation allows us to investigate the asymptotics of high-dimensional SIR, such as consistency and minimax rate optimality, under the least restrictive conditions found in the SIR literature. We anticipate that these conditions will simplify the theoretical investigation of other slicing-based SDR algorithms in high-dimensional settings.

Our second major contribution is to establish the minimax optimal rate for SIR estimation when the structural dimension d is large. We first demonstrate a phenomenon that the generalized signal-to-noise ratio (gSNR), defined as the d-th largest eigenvalue of $\mathrm{Cov}(\mathbb{E}[\boldsymbol{X}|Y])$, decays as the structural dimension d increases: we prove a tight upper bound on the decay rate of the gSNR and perform extensive simulations that showcase a faster decay for various link functions, including random functions drawn from a Gaussian process (GP). This phenomenon suggests that we should focus on scenarios with small gSNRs. We then prove a lower bound on the minimax risk in low gSNR scenarios. In our theory, λ is a parameter that governs the range of gSNR such that $\lambda \leqslant$ gSNR. Our lower bound depends on the quadruple (n, p, d, λ) and matches with an upper bound on the risk of SIR. Consequently, we conclude that the minimax rate is $\frac{dp}{n\lambda}$ and SIR is minimax rate optimal (see Theorem 6). We also extend the minimax results to high dimensional settings: the optimal rate is $\frac{ds+s\log(p/s)}{n\lambda}$, where s is a sparsity parameter (see Theorem 9). Our proof of the lower bound involves an application of Fano's method and a novel construction of nonlinear dependent distributions. This novel construction is based on the observation that SSC can be weakened. The minimax rates, together with the gSNR decay phenomenon, provides a preliminary explanation for the poor performance of SIR in practice when the structural dimension d is large (e.g., $d \geqslant 5$): the gSNR generally diminishes to a level that is insufficient for SIR to provide a good estimate of the central space. To the best of our knowledge, our work is the first effort to establish the optimal rate for estimating the central space for high-dimensional multiple-index models with large structural dimensions. It is also the first attempt to provide a theoretical explanation of the poor performance of SIR when the structural dimension is large. We believe that our findings contribute significantly to a deeper understanding of SIR.

1.2 Organization of the paper

The paper is organized as follows. In Section 2.1, we briefly review the SIR procedure for estimating the central space. Sections 2.2 and 2.3 introduce the new mild condition WSSC for analyzing SIR estimators. In Section 3, we illustrate the phenomenon that the gSNR decays as the structural dimension increases. Section 4 establishes the minimax rates of convergence for estimating the central space with a large structural dimension by matching an upper bound on the risk of SIR with a minimax lower bound. Related problems and future directions are discussed in Section 5.

1.3 Notation

For a matrix $\boldsymbol{V} \in \mathbb{R}^{l\times m}$, we denote its column space by $\mathrm{col}(\boldsymbol{V})$, its i-th row, j-th column and k-th singular value by $\boldsymbol{V}_{i,*}$, $\boldsymbol{V}_{*,j}$ and $\sigma_k(\boldsymbol{V})$ respectively. We use $P_{\boldsymbol{V}}$ to denote the orthogonal projection w.r.t. the standard inner product $\langle\cdot,\cdot\rangle$ in Euclidean space onto $\mathrm{col}(\boldsymbol{V})$. For a square matrix $\boldsymbol{A} \in \mathbb{R}^{m\times m}$, we denote by $\lambda_i(\boldsymbol{A})$ and $\mathrm{Tr}(\boldsymbol{A}) := \sum_{i=1}^{m} \boldsymbol{A}_{i,i}$ the i-th largest eigenvalue and the trace of $\boldsymbol{A}$ respectively. The Frobenius norm and the operator norm (2-norm) of the matrix $\boldsymbol{V} \in \mathbb{R}^{l\times m}$ are defined as $\|\boldsymbol{V}\|_F := \sqrt{\mathrm{Tr}(\boldsymbol{V}^\top \boldsymbol{V})}$ and $\|\boldsymbol{V}\| := \sqrt{\lambda_1(\boldsymbol{V}^\top \boldsymbol{V})}$ respectively. For a vector $\boldsymbol{X}$, denote by X_k the k-th entry of $\boldsymbol{X}$ and $\boldsymbol{X}^{\otimes} = \boldsymbol{X}\boldsymbol{X}^\top$. For two numbers a and b, we use $a \vee b$ and $a \wedge b$ to denote $\max\{a, b\}$ and $\min\{a, b\}$ respectively. For a positive integer c, denote by $[c]$ the index set $\{1, 2, ..., c\}$. For any positive integers p and d such that $p \geqslant d$, denote by $\mathbb{O}(p, d)$ the set of all $p \times d$ orthogonal matrices (i.e., those $\boldsymbol{B}$ with $\boldsymbol{B}^\top \boldsymbol{B} = \boldsymbol{I}_d$). We use $\mathbb{S}^{p-1}$ to denote the $(p-1)$-sphere, i.e., $\mathbb{S}^{p-1} := \{\boldsymbol{x} \in \mathbb{R}^p : \|\boldsymbol{x}\| = 1\}$.

We use C, C', C_1, and C_2 to denote generic absolute constants, though their actual values may vary from case to case. For two sequences a_n and b_n, we write $a_n \gtrsim b_n$ (resp. $a_n \lesssim b_n$) when there exists a positive constant C such that $a_n \geqslant Cb_n$ (resp. $a_n \leqslant C'b_n$). If both $a_n \gtrsim b_n$ and $a_n \lesssim b_n$ hold, we write $a_n \asymp b_n$. We write $a_n = o(b_n)$ if $\lim_{n\to\infty} a_n/b_n = 0$.

2 Asymptotics of SIR under Weak Sliced Stable Condition

Recently, there has been a series of work [Lin et al., 2018a, 2021, 2019] investigating the behavior of SIR on high dimensional data. These studies showed that SIR can produce a consistent estimate of the central space if and only if $\lim \frac{p}{n} = 0$ and SIR attains the minimax rate of estimating the central space in various settings. One of the key technical tools developed in these studies is the 'key lemma': a 'deviation property' of the quantity $\boldsymbol{\beta}^{\top} \widehat{\boldsymbol{\Lambda}}_H \boldsymbol{\beta}$ (see the definition in Equation (4)) for any unit vector $\boldsymbol{\beta} \in \mathbb{R}^p$. The 'key lemma' heavily relies on the sliced stability condition (SSC), a scarcely seen condition. In this section, we propose the weak sliced stable condition (WSSC), a mild condition that is easy to verify, and we investigate the asymptotic behavior of SIR under WSSC.

2.1 A brief review of SIR

Suppose that we observed n i.i.d. samples $\{(\boldsymbol{X}_i, Y_i)\}_{i \in [n]}$ drawn from the joint distribution of $(\boldsymbol{X}, Y)$ given by the following multiple-index model:

$$Y = f(\boldsymbol{B}^{\top} \boldsymbol{X}, \epsilon), \quad \boldsymbol{B}^{\top} \boldsymbol{\Sigma} \boldsymbol{B} = \mathbf{I}_d, \tag{2}$$

where f is an unknown link function, $\boldsymbol{B} = [\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_d] \in \mathbb{R}^{p \times d}$ is the indices matrix, $\epsilon \sim N(0,1)$ is independent of $\boldsymbol{X}$, and $\boldsymbol{\Sigma}$ is the covariance matrix of $\boldsymbol{X}$. Throughout the paper, we assume that $\mathbb{E}[\boldsymbol{X}] = 0$ without loss of generality. Though $\boldsymbol{B}$ is not identifiable because of the unknown link function f, the central space $\mathcal{S}_{Y|\boldsymbol{X}} := \operatorname{col}(\boldsymbol{B})$ can be estimated.

The SIR procedure for estimating $\operatorname{col}(\boldsymbol{B})$ can be briefly described as follows. First of all, the samples $\{(\boldsymbol{X}_i, Y_i)\}_{i \in [n]}$ are divided into H equal-sized slices according to the order statistics $Y_{(i)}$; for simplicity, we assume that $n = cH$, where c is a positive integer. Next, the data can be re-expressed as $\boldsymbol{X}_{h,j}$ and $Y_{h,j}$, with (h,j) as the double subscript, where h denotes the order of the slice and j the order of the sample in the h-th slice, i.e.,

$$\boldsymbol{X}_{h,j} = \boldsymbol{X}_{(c(h-1)+j)}, \qquad Y_{h,j} = Y_{(c(h-1)+j)}. \tag{3}$$

Here $\boldsymbol{X}_{(k)}$ is the concomitant of $Y_{(k)}$ [Yang, 1977]. Let $\mathcal{S}_h$ be the h-th interval $(Y_{(h-1,c)}, Y_{(h,c)}]$ for $h = 2, \ldots, H-1$, $\mathcal{S}_1 = \{y \mid y \leqslant Y_{(1,c)}\}$, and $\mathcal{S}_H = \{y \mid y > Y_{(H-1,c)}\}$. Consequently, $\mathfrak{S}_H(n) := \{\mathcal{S}_h, h = 1, .., H\}$ is a partition of $\mathbb{R}$ and is referred to as the *sliced partition*. Denote the sample mean of $\boldsymbol{X}$ in the h-th slice by $\overline{\boldsymbol{X}}_{h,\cdot}$. The SIR algorithm estimates the candidate matrix $\boldsymbol{\Lambda} := \operatorname{Cov}(\mathbb{E}[\boldsymbol{X}|Y])$ by

$$\widehat{\boldsymbol{\Lambda}}_H = \frac{1}{H} \sum_{h=1}^{H} \overline{\boldsymbol{X}}_{h,\cdot} \overline{\boldsymbol{X}}_{h,\cdot}^{\top}. \tag{4}$$

In Lin et al. [2018a], it is shown that when the ratio $p/n \to 0$, a consistent estimator of the central space $\operatorname{col}(\boldsymbol{B})$ is given by $\widehat{\boldsymbol{\Sigma}}^{-1} \operatorname{col}(\widehat{\eta}_H)$, where $\widehat{\boldsymbol{\Sigma}}$ is the sample covariance matrix of $\boldsymbol{X}$ and $\widehat{\eta}_H$ is the matrix formed by the top d eigenvectors of $\widehat{\boldsymbol{\Lambda}}_H$. Alternatively, the central space $\operatorname{col}(\boldsymbol{B})$ could be estimated by $\operatorname{col}(\widehat{\boldsymbol{B}}_H)$, where $\widehat{\boldsymbol{B}}_H$ is defined by

$$\widehat{\boldsymbol{B}}_H := \arg\max_{\boldsymbol{B}} \quad \operatorname{Tr}(\boldsymbol{B}^{\top} \widehat{\boldsymbol{\Lambda}}_H \boldsymbol{B}) \quad \text{s.t. } \boldsymbol{B}^{\top} \widehat{\boldsymbol{\Sigma}} \boldsymbol{B} = \mathbf{I}_d, \tag{5}$$

since $\operatorname{col}(\widehat{\eta}_H) = \widehat{\boldsymbol{\Sigma}} \operatorname{col}(\widehat{\boldsymbol{B}}_H)$ (see, e.g., Proposition 4 in Li [2007]).

To ensure that SIR provides a consistent estimator of the central space, the following conditions have been suggested to show the relation $\operatorname{col}(\boldsymbol{\Lambda}) = \boldsymbol{\Sigma} \mathcal{S}_{Y|\boldsymbol{X}}$ [Li, 1991, Hsing and Carroll, 1992, Zhu et al., 2006, Lin et al., 2018a, 2021, 2019].

Assumption 1. *The joint distribution of $(\boldsymbol{X}, Y) \in \mathbb{R}^p \times \mathbb{R}$ satisfies the following conditions:*

i) *Linearity condition: For any $\boldsymbol{a} \in \mathbb{R}^p$, the conditional expectation $\mathbb{E}\left[\langle \boldsymbol{a}, \boldsymbol{X} \rangle \mid \boldsymbol{B}^{\top} \boldsymbol{X}\right]$ is linear in $\boldsymbol{B}^{\top} \boldsymbol{X}$.*

ii) *Coverage condition:* $\lambda \leqslant \lambda_d(\mathrm{Cov}(\mathbb{E}[\boldsymbol{X} \mid Y])) \leqslant \lambda_1(\mathrm{Cov}(\mathbb{E}[\boldsymbol{X} \mid Y])) \leqslant \kappa\lambda \leqslant \lambda_{\max}(\boldsymbol{\Sigma})$ *where* $\kappa > 1$ *is a positive constant.*

The condition (**ii**) is a refinement of the coverage condition in the SIR literature, as explained in Condition A2 of Lin et al. [2019]. Inspired by a similar assumption in Cai et al. [2013], we introduce the regularity parameter κ to control the condition number of $\mathrm{Cov}(\mathbb{E}[\boldsymbol{X} \mid Y])$. We want to emphasize that the coverage condition is critical. As demonstrated by the minimax lower bounds (see Theorems 4 and 7), if the eigenvalues of $\mathrm{Cov}(\mathbb{E}[\boldsymbol{X} \mid Y])$ are too small, then any estimation method will fail to accurately estimate the central space.

We consider the following loss function for our minimax theory. Let $\widehat{\boldsymbol{B}}$ be an estimate of $\boldsymbol{B}$, whose columns form a basis of the central space $\mathcal{S}_{Y|\boldsymbol{X}}$. Note that the parameter $\boldsymbol{B}$ itself is not identifiable while $\boldsymbol{B}\boldsymbol{B}^\top$ is identifiable. To evaluate the estimated central space $\mathrm{col}(\widehat{\boldsymbol{B}})$, we consider the loss function $\|\widehat{\boldsymbol{B}}\widehat{\boldsymbol{B}}^\top - \boldsymbol{B}\boldsymbol{B}^\top\|_{\mathrm{F}}^2$, which we will refer to as the *general loss*. This loss function is commonly used in the sufficient dimension reduction literature; for more details, see Section 1.1 of Tan et al. [2020].

Throughout the paper, we assume that the structural dimension d is known but we allow d to be arbitrarily large (i.e., there is no constant upper bound on d).

2.2 Key lemma under weak sliced stable condition

We begin by revisiting the sliced stable condition (SSC) used in the works of Lin et al. [2018a, 2021, 2019] and then introduce an alternative condition. Throughout the paper, γ is a fixed small positive constant.

Definition 1 (Sliced Stable Condition). Let $Y \in \mathbb{R}$ be a random variable, K a positive integer and $\vartheta > 0$ a constant. A continuous curve $\boldsymbol{\kappa}(y) : \mathbb{R} \to \mathbb{R}^p$ is said to be (K, ϑ)*-sliced stable* w.r.t. Y, if for any $H \geqslant K$ and any partition $\mathcal{B}_H := \{-\infty = a_0 < a_1 < \cdots < a_{H-1} < a_H = \infty\}$ of $\mathbb{R}$ such that

$$\frac{1-\gamma}{H} \leqslant \mathbb{P}(a_h \leqslant Y \leqslant a_{h+1}) \leqslant \frac{1+\gamma}{H}, \qquad \forall h = 0, 1, \ldots, H-1, \tag{6}$$

it holds that

$$\frac{1}{H}\sum_{h=0}^{H-1} \mathrm{var}\left(\boldsymbol{\beta}^\top \boldsymbol{\kappa}(Y) \middle| a_h \leqslant Y \leqslant a_{h+1}\right) \leqslant \frac{1}{H^\vartheta}\mathrm{var}\left(\boldsymbol{\beta}^\top\boldsymbol{\kappa}(Y)\right) \quad (\forall \boldsymbol{\beta} \in \mathbb{S}^{p-1}).$$

Lin et al. [2018a] utilized this condition to establish the deviation properties of the eigenvalues, eigenvectors, and entries of $\widehat{\boldsymbol{\Lambda}}_H$. Although they showed that the SSC is a mild condition, it is hard to verify whether the central curve $\boldsymbol{m}(y) := \mathbb{E}[\boldsymbol{X} \mid Y = y]$ of a given joint distribution $(\boldsymbol{X}, Y)$ satisfies SSC.

To motivate a more manageable condition than SSC, we revisit an intuitive explanation of SSC: for any well-behaved continuous curve $\boldsymbol{\kappa}(y)$, if it is divided into K pieces with roughly equal probability mass of the distribution of Y, then the average of the variances of $\boldsymbol{\kappa}(Y)$ in each piece tends to 0 as the slice number H tends to ∞. The (K, ϑ)-sliced stable condition requires the average of the variances to tend to 0 at certain rate (e.g., $H^{-\vartheta}$). This geometric explanation leads us to introduce the following definition of weak sliced stable condition (WSSC), which only requires the average of the variances to be sufficiently small.

Definition 2 (Weak Sliced Stable Condition). Let $Y \in \mathbb{R}$ be a random variable, K a positive integer and $\tau > 1$ a constant. A continuous curve $\boldsymbol{\kappa}(y) : \mathbb{R} \to \mathbb{R}^p$ is said to be *weak* (K, τ)*-sliced stable* w.r.t. Y, if for any $H \geqslant K$ and any partition $\mathcal{B}_H$ of $\mathbb{R}$ such that $\frac{1-\gamma}{H} \leqslant \mathbb{P}(a_h \leqslant Y \leqslant a_{h+1}) \leqslant \frac{1+\gamma}{H}, \forall h = 0, 1, \ldots, H-1$, it holds that

$$\frac{1}{H}\sum_{h=0}^{H-1} \mathrm{var}\left(\boldsymbol{\beta}^\top \boldsymbol{\kappa}(Y) \middle| a_h \leqslant Y \leqslant a_{h+1}\right) \leqslant \frac{1}{\tau}\mathrm{var}\left(\boldsymbol{\beta}^\top\boldsymbol{\kappa}(Y)\right) \quad (\forall \boldsymbol{\beta} \in \mathbb{S}^{p-1}).$$

In the following, we show that Lemma 1 in Lin et al. [2018a], referred to as the ‘key lemma’ therein, still holds if we replace SSC with WSSC. Specifically, we can establish the deviation properties of $\boldsymbol{\beta}^\top\widehat{\boldsymbol{\Lambda}}_H\boldsymbol{\beta}$ under WSSC if H is sufficiently large. Before proceeding, we first recall the following normality assumption.

Assumption 2 (Normality)**.** *The random vector $\boldsymbol{X}$ satisfies $\boldsymbol{X} \sim N(0, \boldsymbol{\Sigma})$ and $\|\boldsymbol{\Sigma}\| \vee \|\boldsymbol{\Sigma}^{-1}\| \leqslant M$, where M is a positive constant.*

This assumption is frequently used in SIR literature, such as Lin et al. [2018a, 2019, 2021]. In fact, it implies the linearity condition in Assumption 1 since $\mathbb{E}[\boldsymbol{X} \mid \boldsymbol{B}^\top \boldsymbol{X}] = \boldsymbol{\Sigma}\boldsymbol{B}\boldsymbol{B}^\top \boldsymbol{X}$. The next important result is the 'key lemma' under WSSC.

Lemma 1. *Suppose that the central curve $\boldsymbol{m}(y) := \mathbb{E}[\boldsymbol{X} \mid Y]$ satisfies weak (K, τ)-SSC w.r.t. Y with $\tau > 16$ and Assumptions 1 (ii) and 2 hold. Then there exist positive absolute constants C, C_1, C_2, and C_3 such that if $H \geqslant K \vee Cd$, then for any $\nu \in (1, \tau/16]$, any unit vector $\boldsymbol{\beta} \in \operatorname{col}(\boldsymbol{\Lambda})$ and any sufficiently large $n > 1 + 1H/\gamma$, it holds that*

$$\mathbb{P}\left(\left|\boldsymbol{\beta}^\top\left(\widehat{\boldsymbol{\Lambda}}_H - \boldsymbol{\Lambda}\right)\boldsymbol{\beta}\right| \geqslant \frac{1}{2\nu}\boldsymbol{\beta}^\top\boldsymbol{\Lambda}\boldsymbol{\beta}\right) \leqslant C_1 \exp\left(-C_2\frac{n\boldsymbol{\beta}^\top\boldsymbol{\Lambda}\boldsymbol{\beta}}{H^2\nu^2} + C_3\log(nH)\right).$$

Compared with the 'key lemma' in Lin et al. [2018a], the range of ν in the current lemma is narrower, i.e., ν can not go to infinity. The reason behind this difference is that, in the current lemma, we have relaxed the condition from SSC to WSSC. However, it is worth noting that the narrower range of ν makes no essential difference of theory developed for sparse SIR in high dimensional settings. Specifically, the minimax optimal rate for sparse SIR still holds under WSSC, as will be shown in Section 4.

2.3 Weak sliced stable condition is very mild

Though it is hard to verify whether the central curve $\boldsymbol{m}(y)$ satisfies SSC, we can show that $\boldsymbol{m}(y)$ satisfies WSSC under plausible assumptions.

Theorem 1. *Suppose that the joint distribution of $(\boldsymbol{X}, Y) \in \mathbb{R}^p \times \mathbb{R}$ satisfies the following conditions:*

- **i)** *for any $\boldsymbol{\beta} \in \mathbb{S}^{p-1}$, $\mathbb{E}\left[|\langle \boldsymbol{\beta}, \boldsymbol{X}\rangle|^\ell\right] \leqslant c_1$ holds for absolute constants $\ell > 2$ and $c_1 > 0$;*
- **ii)** *Y is a continuous random variable;*
- **iii)** *the central curve $\boldsymbol{m}(y) := \mathbb{E}[\boldsymbol{X}|Y = y]$ is continuous.*

Then for any $\tau > 1$, there exists an integer $K = K(\tau, d) \geqslant d$ such that $\boldsymbol{m}(y)$ is weak (K, τ)-sliced stable w.r.t. Y.

A result in Lin et al. [2018a] guarantees that with high probability, the probability mass of the distribution of Y in each slice $\mathcal{S}_h$ is roughly the same, i.e., (6) holds for the sliced partition $\mathfrak{S}_H(n) := \{\mathcal{S}_h, h = 1, .., H\}$. This leads to the following corollary.

Corollary 1. Suppose that the conditions in Theorem 1 hold. For any sufficiently large $H \geqslant K$, if $n > 1 + 4H/\gamma$ is sufficiently large, then for the sliced partition $\mathfrak{S}_H(n) = \{\mathcal{S}_h, h = 1, .., H\}$, the inequality

$$\frac{1}{H}\sum_{h=1}^{H} \operatorname{var}\left(\boldsymbol{\beta}^\top \boldsymbol{m}(Y)\middle| Y \in \mathcal{S}_h\right) \leqslant \frac{1}{\tau}\operatorname{var}\left(\boldsymbol{\beta}^\top \boldsymbol{m}(Y)\right) \quad (\forall \boldsymbol{\beta} \in \mathbb{S}^{p-1})$$

holds with probability at least $1 - CH^2\sqrt{n+1}\exp\left(-\gamma^2(n+1)/32H^2\right)$ for some absolute constant $C > 0$.

Remark 1. If we further assume that $\|\boldsymbol{m}(y) - \boldsymbol{m}(y')\| \leqslant Cd|y - y'|$ for any y, y' defined on a compact set of $\mathbb{R}$ and $C > 0$ a constant, then the WSSC coefficient K equals to $K_0 d$ for some integer $K_0 \geqslant 1$. This assumption is mild since one can always turn the last $p - d$ entries of $\boldsymbol{m}(y)$ into zero through orthogonal transformation by noting that $\dim\{\operatorname{span}\{\boldsymbol{m}(y) : y \in \mathbb{R}\}\} = \operatorname{rank}(\operatorname{Cov}(\boldsymbol{m}(Y))) = d$.

3 Small gSNR with a large structural dimension

Lin et al. [2021] has made a conjecture that the minimax optimal rate for estimating the central space under multiple-index models should be inverse proportional to the gSNR. This indicates that a small gSNR would result in a large estimation error, so it is important to carefully examine how the gSNR depends on the structural dimension d.

In this section, we demonstrate that for a general multiple-index model, the gSNR always decreases as the structural dimension d grows and often becomes extremely small.

3.1 An upper bound on the gSNR

We show here that as the structural dimension d increases, the gSNR must decrease at least at the rate of $\frac{\log(d)}{d}$.

Theorem 2. *Assume $\boldsymbol{X} \sim N(\mathbf{0}, \boldsymbol{I}_p)$ and Y is a random variable. If $\boldsymbol{m}(y)$ satisfies the weak (K,τ)-SSC w.r.t. Y with $\tau > 1+\gamma$ and $K = O(d)$, then the smallest positive eigenvalue of* $\operatorname{Cov}\left[\mathbb{E}\left(\boldsymbol{X} \mid Y\right)\right]$ *(i.e., the gSNR) is no greater than $C_1 \frac{\log(d)}{d}$, where C_1 only depends on τ and γ.*

It is worth mentioning that the rate of such an upper bound is tight, as there exists a joint distribution of $(\boldsymbol{X}, Y)$ for which the gSNR is asymptotically $O(\frac{\log(d)}{d})$. To illustrate this, consider a function $\psi^0(\boldsymbol{z} = (z_1, \ldots, z_d))$ from $\mathbb{R}^d$ to $\{-d, \ldots, 0, 1, \ldots, d\}$ such that if $|z_i|$ is uniquely the largest among $|z_i|$'s, then $\psi^0(\boldsymbol{z}) = \operatorname{sgn}(z_i) i$; otherwise, $\psi^0(\boldsymbol{z}) = 0$. Let $\boldsymbol{B} = [\boldsymbol{e}_1, \ldots, \boldsymbol{e}_d]$ whose columns are the first d standard basis vectors in $\mathbb{R}^p$. We construct the following joint distribution of $(\boldsymbol{X}, Y)$:

$$\begin{aligned} \boldsymbol{X} &\sim N(0, \boldsymbol{I}_p), \\ Y &= \psi^0(\boldsymbol{B}^\top \boldsymbol{X}) + \eta, \quad \eta \sim \operatorname{Unif}(-1/2, 1/2), \end{aligned} \tag{7}$$

where $\boldsymbol{X}$ and η are independent. This distribution satisfies the WSSC and its gSNR decays to 0 at the rate of $\frac{\log d}{d}$ as d tends to ∞.

Theorem 2 relies on the following Theorem 3, which may be of independent interest. It provides an upper bound on the smallest eigenvalue of the covariance matrix of the conditional expectation of a normal random vector given any discrete random variable.

Theorem 3. *Suppose $\boldsymbol{Z} \sim N(0, \boldsymbol{I}_d)$ and W is a discrete random variable whose probability mass function is smaller than $1/2$. The entropy of W is defined as $\sum_w \mathbb{P}(W = w) \log \frac{1}{\mathbb{P}(W=w)}$ and is denote by $\operatorname{Ent}(W)$. It holds that*

$$\lambda_{\min}\left\{\operatorname{Cov}\left[\mathbb{E}\left(\boldsymbol{Z} \mid W\right)\right]\right\} \leqslant 37\ d^{-1} \operatorname{Ent}(W).$$

In particular, if the support of W has K elements, we have $\lambda_{\min}\left\{\operatorname{Cov}\left[\mathbb{E}\left(\boldsymbol{Z} \mid W\right)\right]\right\} \leqslant 37\ d^{-1} \log K$.

Theorem 2 states that the gSNR must decay at least at the rate of $\frac{\log(d)}{d}$ as the structural dimension d grows. However, in practice, the gSNR is often observed to decay much more rapidly. We will elaborate on this phenomenon in the next subsection.

3.2 Small gSNR, a realistic scenario

In this subsection, we examine synthetic and random GP models to demonstrate that for a general multiple-index model, the gSNR tends to be extremely small when d is large. This observation suggests that we should focus on scenarios with small gSNRs.

We first present two exemplary five-index models to illustrate the notorious performance of SIR when $d = 5$ [Ferré, 1998, Lin et al., 2021] and point out that the gSNR is very small. Then we investigate a broader scenario using a random link function sampled from a GP and find that the estimated gSNR decays rapidly as d increases and becomes very small at $d = 5$.

Synthetic experiments

Our following numerical results on synthetic data demonstrate that when $d = 5$, SIR behaves poorly and the gSNR is very close to 0.

Let us consider the following two models:

$$\begin{aligned}
&\boldsymbol{X} = (X_1, ..., X_{15})^\top \sim N(0, \mathbf{I}_{15}), \quad \epsilon \sim N(0,1);\\
&\mathcal{M}_1 : Y = X_1 + \exp(X_2) + \log(|X_3 + 1| + 1) + \sin(X_4) + \arctan(X_5) + 0.01 * \epsilon;\\
&\mathcal{M}_2 : Y = X_1^3 + \frac{X_2}{(1 + X_3)^2} + \mathrm{sgn}(X_4)\log(|X_5 + 0.02| + 5) + 0.01 * \epsilon.
\end{aligned}$$

In both $\mathcal{M}_1$ and $\mathcal{M}_2$, the columns of $\boldsymbol{B}$ are $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_5$, the first five standard basis vectors in $\mathbb{R}^{15}$. Table 1 shows the general loss $\|\widehat{\boldsymbol{B}}\widehat{\boldsymbol{B}}^\top - \boldsymbol{B}\boldsymbol{B}^\top\|_F^2$ of SIR under models $\mathcal{M}_1$ and $\mathcal{M}_2$ over 100 replications. The value of n ranges in $\{10^3, 10^4, 10^5, 10^6, 10^7, 8 \cdot 10^7\}$ and H in $\{2, 5, 10, 20, 50, 100, 200, 500\}$. Each entry is the average of the general loss of SIR estimates for a given pair of n and H. The results in Table 1 show that whatever H is chosen, the estimation error decays very slowly with increasing n and remains large even when $n = 8 \cdot 10^7$.

	n	$H = 2$	$H = 5$	$H = 10$	$H = 20$	$H = 50$	$H = 100$	$H = 200$	$H = 500$
$\mathcal{M}_1$	10^3	6.350	4.303	4.821	4.850	4.668	4.644	4.641	4.602
		(0.2122)	(0.0699)	(0.0604)	(0.0507)	(0.0492)	(0.0482)	(0.0534)	(0.0511)
	10^4	5.762	3.505	4.468	4.479	4.438	4.511	4.434	4.564
		(0.2012)	(0.0510)	(0.0510)	(0.0520)	(0.0505)	(0.0571)	(0.0530)	(0.0468)
	10^5	5.578	3.247	3.726	3.548	3.620	3.681	3.644	3.958
		(0.2242)	(0.0703)	(0.0497)	(0.0459)	(0.0437)	(0.0375)	(0.0432)	(0.0566)
	10^6	5.735	2.492	2.806	2.712	2.907	2.914	3.112	3.216
		(0.2314)	(0.0665)	(0.0422)	(0.0445)	(0.0435)	(0.0468)	(0.0452)	(0.0451)
	10^7	5.924	1.646	1.929	1.877	1.915	1.912	1.956	2.068
		(0.2509)	(0.0683)	(0.0293)	(0.0260)	(0.0263)	(0.0271)	(0.0207)	(0.0240)
	$8 \cdot 10^7$	6.004	**1.062**	1.832	1.851	1.829	1.835	**1.802**	1.819
		(0.2301)	(0.0818)	(0.0220)	(0.0202)	(0.0243)	(0.0235)	(0.0233)	(0.0251)
$\mathcal{M}_2$	10^3	8.027	3.749	3.709	3.601	3.549	3.514	3.628	3.972
		(0.2795)	(0.0681)	(0.0426)	(0.0468)	(0.0456)	(0.0376)	(0.0404)	(0.0493)
	10^4	7.274	2.795	2.780	2.866	3.024	3.123	3.192	3.243
		(0.2617)	(0.0683)	(0.0411)	(0.0458)	(0.0483)	(0.0453)	(0.0399)	(0.0414)
	10^5	7.214	2.324	1.984	1.963	1.962	1.984	2.096	2.375
		(0.2909)	(0.0604)	(0.0218)	(0.0248)	(0.0240)	(0.0274)	(0.0213)	(0.0374)
	10^6	6.553	1.859	1.844	1.840	1.865	1.857	1.863	1.828
		(0.3084)	(0.0312)	(0.0221)	(0.0194)	(0.0171)	(0.0182)	(0.0216)	(0.0249)
	10^7	5.445	1.659	1.747	1.722	1.759	1.837	1.792	1.824
		(0.2813)	(0.0413)	(0.0258)	(0.0299)	(0.0246)	(0.0177)	(0.0226)	(0.0220)
	$8 \cdot 10^7$	5.152	1.239	**0.688**	**0.821**	1.025	1.298	1.397	1.652
		(0.2724)	(0.0627)	(0.0355)	(0.0444)	(0.0506)	(0.0492)	(0.0480)	(0.0407)

Table 1: The general loss of SIR under models $\mathcal{M}_1$ and $\mathcal{M}_2$ with each (n, H) combination. The two best combinations are highlighted in bold. The average and the standard error (in parenthesis) are based on 100 replications.

To show the poor performance more clearly, we check for model $\mathcal{M}_1$ whether $\widehat{\boldsymbol{\beta}}_i$ $(i = 1, \ldots, 5)$ the estimated directions of SIR lie in the central space, i.e., the space spanned by $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_5$. Specifically, for model $\mathcal{M}_1$, we compute the mean squared value of the last 10 entries of $\widehat{\boldsymbol{\beta}}_i$ $(i = 1, \ldots, 5)$, based on 100 replications, with $n = 8 \cdot 10^7$ and $H \in \{5, 200\}$ (the first two winners with the smallest general loss among all choices of n and H).

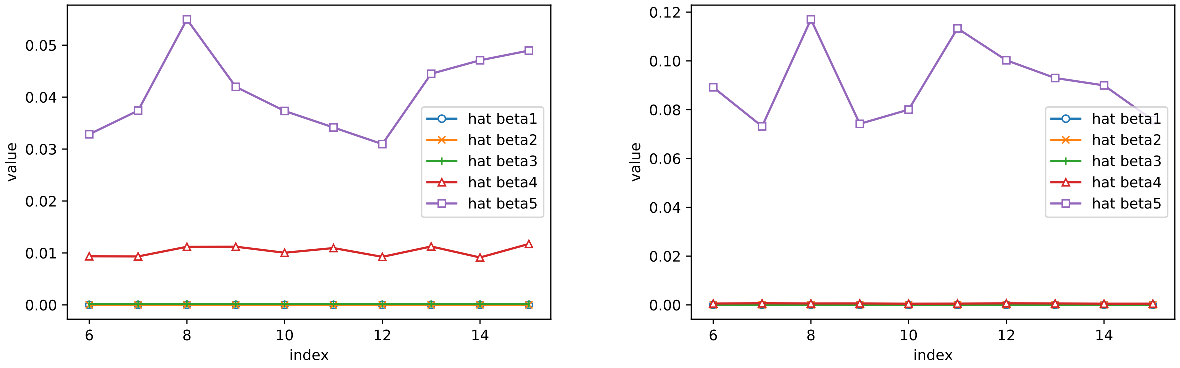


Figure 1: Mean squared value of the last 10 entries of estimated directions of SIR for model $\mathcal{M}_1$ with $n = 8 \cdot 10^7$, $H = 5$ (left) and $H = 200$ (right).

Figure 1 shows that the last two directions ($\widehat{\boldsymbol{\beta}}_4$ and $\widehat{\boldsymbol{\beta}}_5$) of the SIR estimate often lie outside the central space of $\mathcal{M}_1$. This implies that SIR can not provide a good estimate of the central space when $d = 5$ even if the sample size is as large as $8 \cdot 10^7$ and H enumerates all reasonable choices. This observation is consistent with the high loss presented in Table 1.

We proceed to examine the gSNR in these models by calculating the average of the logarithm of $\lambda_i(\widehat{\boldsymbol{\Lambda}}_H)$, the eigenvalues of the SIR estimate of $\mathrm{Cov}(\mathbb{E}(\boldsymbol{X} \mid Y))$ with $n = 8 \cdot 10^7$ (the largest sample size in this experiment). As seen in Table 2, both models $\mathcal{M}_1$ and $\mathcal{M}_2$ exhibit rapid decay of $\lambda_i(\widehat{\boldsymbol{\Lambda}}_H)$ as the index increases and their estimated gSNRs $\lambda_5(\widehat{\boldsymbol{\Lambda}}_H)$ are very close to 0.

$n = 8 \cdot 10^7$	index	$H = 2$	$H = 5$	$H = 10$	$H = 20$	$H = 50$	$H = 100$	$H = 200$	$H = 500$
$\mathcal{M}_1$	1	-0.54	-0.22	-0.16	-0.15	-0.14	-0.14	-0.14	-0.14
	2	-39.19	-2.83	-2.33	-2.16	-2.09	-2.08	-2.07	-2.07
	3	-45.88	-9.45	-8.21	-7.60	-7.24	-7.12	-7.05	-7.00
	4	-51.60	-13.61	-11.37	-10.88	-10.53	-10.70	-10.62	-10.48
	gSNR	**-54.56**	**-41.31**	**-15.07**	**-14.45**	**-13.71**	**-13.15**	**-12.56**	**-11.75**
$\mathcal{M}_2$	1	-0.78	-0.58	-0.54	-0.53	-0.53	-0.53	-0.53	-0.53
	2	-40.11	-1.72	-1.43	-1.39	-1.38	-1.37	-1.37	-1.37
	3	-47.09	-3.32	-2.30	-2.25	-2.22	-2.22	-2.22	-2.21
	4	-51.97	-9.50	-6.78	-6.68	-6.62	-6.61	-6.61	-6.60
	gSNR	**-53.97**	**-47.14**	**-14.52**	**-14.10**	**-13.57**	**-13.09**	**-12.54**	**-11.75**

Table 2: Logarithm of the eigenvalues of $\widehat{\boldsymbol{\Lambda}}_H$ the SIR estimate of $\mathrm{Cov}(\mathbb{E}(\boldsymbol{X} \mid Y))$ (averaged based on 100 replications) under models $\mathcal{M}_1$ and $\mathcal{M}_2$. For each model, the i-th row corresponds to the i-th eigenvalue.

Gaussian process To further explore the decay of gSNR as d increases, we study a general setting where the link function is a random continuous function sampled from a GP. Our findings reveal that the gSNR decays rapidly with increasing d and is close to 0 when $d = 5$.

For each $d \in [5] := \{1, 2, 3, 4, 5\}$, let $\boldsymbol{B} = [\boldsymbol{e}_1, \ldots, \boldsymbol{e}_d]$, where $\boldsymbol{e}_i$ is the i-th standard basis vector of $\mathbb{R}^{15}$ and consider the following joint distribution of $(\boldsymbol{X}, Y)$:

$$\begin{aligned} \boldsymbol{X} &= (X_1, ..., X_{15})^\top \sim N(0, \mathbf{I}_{15}); \\ \mathcal{M}_3 : Y &= f(\boldsymbol{B}^\top \boldsymbol{X}) + 0.01 * \epsilon, \quad \epsilon \sim N(0, 1), \end{aligned} \tag{8}$$

where f is a random function generated from the GP with mean function $\mu(\boldsymbol{x}) = \mathbf{0}$ and covariance function $\Sigma(\boldsymbol{x}, \boldsymbol{x}') = e^{-\frac{\|\boldsymbol{x}-\boldsymbol{x}'\|^2}{2}}$ (see Seeger [2004] for an introduction to GPs).

In the following, we explore the decay of the gSNR for model $\mathcal{M}_3$ in (8) with increasing d. Specifically, for each $d \in [5]$, we sample f for $1,000$ times. For each sampled f, we draw a sample of $(\boldsymbol{X}, Y)$ of size n from model $\mathcal{M}_3$ and compute the estimated gSNR of $\mathcal{M}_3$ by $\lambda_d(\widehat{\boldsymbol{\Lambda}}_H)$, the d-th eigenvalue of the SIR estimate of $\mathrm{Cov}(\mathbb{E}(\boldsymbol{X} \mid Y))$. Due to computational limitations, we set the maximum sample size at $50,000$. Here we present the result for $H = 15$ (the results are not sensitive to the choice of H). Detailed sampling procedures and results for other values of H can be found in Appendix I.3.

Figure 2 plots the average logarithm of the estimated gSNR over $1,000$ replications. We show the average as a function of n for various values of d in the left subfigure and show it as a function of d for various values of n in right subfigure. All of the associated standard error are less than 0.005. Histograms of the estimated gSNR can be found in Appendix I.3. Based on the left subfigure, the estimated gSNR keeps decreasing as n grows, indicating that it overestimates the true gSNR. The right subfigure shows that for a sufficiently large n, the estimated gSNR appears to decay exponentially with respect to d and becomes extremely small when $d = 5$.

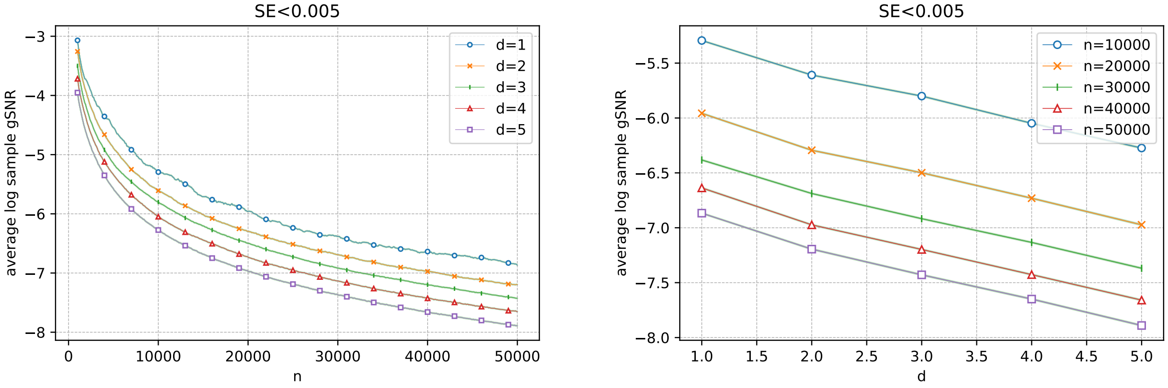


Figure 2: Average logarithm of gSNR with increasing n (left) and increasing d (right).

These experiments suggest that in realistic situations, the gSNR tends to decay rapidly as the structural dimension d increases. Consequently, when assessing the SIR method when d is allowed to grow, our focus should be on the low gSNR regime where the gSNR can be arbitrarily small and no greater than some constant ϖ_d that depends only on d. We will establish the minimax rate in this regime in the next section.

4 Minimax rate optimality of SIR

In this section, we first establish the minimax rate optimality of SIR for a broad range of distributions in the low gSNR regime. Our minimax results precisely capture the impact of gSNR on the estimation risk of the central space and clarify that it is the weakened signal strength that causes the decline in performance of SIR when d is large. We then extend the results to sparse SIR for high-dimensional problems.

4.1 A large class of distributions

To determine the minimax rate for estimating the central space, we first introduce a large class of functions.

Definition 3. Suppose that $\boldsymbol{Z} \sim N(0, \mathbf{I}_d)$ and $\epsilon \sim N(0,1)$. A function $f : \mathbb{R}^{d+1} \to \mathbb{R}$ is said to be in the class $\mathcal{F}_d(\lambda, \kappa, K)$ if the joint distribution for $(\boldsymbol{Z}, Y = f(\boldsymbol{Z}, \epsilon))$ satisfies the following two properties:

(i) the eigenvalues of the conditional covariance matrix $\boldsymbol{\Lambda}_z := \mathrm{Cov}(\mathbb{E}[\boldsymbol{Z}|Y])$ satisfy that

$$0 < \lambda \leqslant \lambda_d(\boldsymbol{\Lambda}_z) \leqslant \cdots \leqslant \lambda_1(\boldsymbol{\Lambda}_z) \leqslant \kappa\lambda \leqslant 1;$$

(ii) the central curve $\boldsymbol{m}_z(y) = \mathbb{E}[\boldsymbol{Z}|y]$ is weak $(K, 32\kappa)$-sliced stable w.r.t. Y.

It is clear from Theorem 1 that if f and $\boldsymbol{m}_z(y)$ are continuous functions, then the joint distribution for $(\boldsymbol{Z}, Y = f(\boldsymbol{Z}, \epsilon))$ satisfies property **(ii)** for some K immediately. Thus $\mathcal{F}_d(\lambda, \kappa, K)$ is a fairly large class of functions.

We proceed to define $\mathfrak{M}\,(p, d, \lambda)$ the class of distributions for $(\boldsymbol{X}, Y)$ where the minimax rate of estimation risk is determined. This class is given by:

$$\mathfrak{M}\,(p,d,\lambda) := \left\{ \begin{array}{c|l} & \boldsymbol{X} \sim N(0,\boldsymbol{\Sigma}), \epsilon \sim N(0,1) \text{ is independent of } \boldsymbol{X}, \\ \text{distribution of} & \boldsymbol{\Sigma} \text{ is a } p \times p \text{ matrix}, \|\boldsymbol{\Sigma}\| \vee \|\boldsymbol{\Sigma}^{-1}\| \leqslant M, \\ \big(\boldsymbol{X}, Y = f(\boldsymbol{B}^\top \boldsymbol{X}, \epsilon)\big) & \boldsymbol{B} \text{ is a } p \times d \text{ matrix}, \boldsymbol{B}^\top \boldsymbol{\Sigma} \boldsymbol{B} = \mathbf{I}_d, \\ & f \in \mathcal{F}_d(\lambda,\kappa,K), K = K_0 d, \lambda \leqslant \varpi_d. \end{array} \right\}, \tag{9}$$

where K_0, κ, and M are constants throughout this section. We have also imposed the constraint that $\lambda \leqslant \varpi_d$, where $\{\varpi_d\}_{d=1}^{\infty}$ is a sequence of constants. Whenever the constants in (9) are determined, the class $\mathfrak{M}(p, d, \lambda)$ is well-defined. In the following, we will use the phrase *a specification of* $\mathfrak{M}(p, d, \lambda)$ to refer to a set of values for these constants.

It is easy to see from Remark 1 that $K = K_0 d$ is a natural condition. As evidenced by the examples in Section 3.2, our primary focus should be on the low gSNR regime. Consequently, we allow ϖ_d to decay as d increases, provided that $\lambda \leqslant \varpi_d$ covers the majority of relevant situations. Besides, the other conditions in the definition of $\mathfrak{M}(p, d, \lambda)$ are mild regularity conditions. Thus, $\mathfrak{M}(p, d, \lambda)$ is a fairly large class of distributions.

We now briefly discuss the consequences of the conditions imposed in $\mathfrak{M}\,(p, d, \lambda)$. If the distribution of $(\boldsymbol{X}, Y)$ lies in $\mathfrak{M}\,(p, d, \lambda)$, the candidate matrix $\boldsymbol{\Lambda} = \mathrm{Cov}(\mathbb{E}[\boldsymbol{X} \mid Y])$ satisfies coverage condition and the central curve $\mathbb{E}[\boldsymbol{X} \mid Y = y]$ satisfies WSSC. More precisely, let $\boldsymbol{Z} = \boldsymbol{B}^\top \boldsymbol{X}$. Then $\boldsymbol{Z} \sim N(0, \mathbf{I}_d)$. By the law of total expectation, $\mathbb{E}[\boldsymbol{X}|Y] = \mathbb{E}[\mathbb{E}[\boldsymbol{X}|\boldsymbol{B}^\top\boldsymbol{X}]|Y] = \mathbb{E}[\boldsymbol{\Sigma}\boldsymbol{B}\boldsymbol{B}^\top\boldsymbol{X}|Y] = \boldsymbol{\Sigma}\boldsymbol{B}\mathbb{E}[\boldsymbol{Z}|Y]$, thus $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}\boldsymbol{B}\boldsymbol{\Lambda}_z\boldsymbol{B}^\top\boldsymbol{\Sigma}$. On the one hand, $\lambda_i(\boldsymbol{\Lambda}) = \lambda_i(\boldsymbol{\Sigma}\boldsymbol{B}\boldsymbol{\Lambda}_z\boldsymbol{B}^\top\boldsymbol{\Sigma}) \leqslant \lambda_i(\boldsymbol{\Sigma}^{1/2}\boldsymbol{B}\boldsymbol{\Lambda}_z\boldsymbol{B}^\top\boldsymbol{\Sigma}^{1/2})\lambda_{\max}(\boldsymbol{\Sigma}) \leqslant M\lambda_i(\boldsymbol{\Lambda}_z)$. Similarly, $\lambda_i(\boldsymbol{\Lambda}) \geqslant M^{-1}\lambda_i(\boldsymbol{\Lambda}_z)$. Since the coverage condition holds for $\boldsymbol{\Lambda}_z$, it holds for $\boldsymbol{\Lambda}$ as well. On the other hand, it is obviously that WSSC for $\mathbb{E}[\boldsymbol{Z}|Y = y]$ implies WSSC for $\mathbb{E}[\boldsymbol{X}|Y = y]$.

4.2 Minimax rate optimality of SIR

We establish the minimax rate optimality of SIR over the model class $\mathfrak{M}\,(p, d, \lambda)$ under the general loss.

Our analysis in this section does not require the population parameters (p, d, λ) of the model class $\mathfrak{M}(p, d, \lambda)$ to be fixed. Instead, they are allowed to depend on the sample size n, i.e., p and d may grow and λ may decay as n increases. Throughout this section, the infimum $\inf_{\widehat{\boldsymbol{B}}}$ is taken over all estimators that depend on the sample $\{(\boldsymbol{X}_i, Y_i)\}_{i=1}^{n}$, which consists of n i.i.d. draws from $\mathcal{M} \in \mathfrak{M}(p, d, \lambda)$. In addition, the expectation $\mathbb{E}_{\mathcal{M}}$ is taken with respect to the randomness of the sample.

4.2.1 Minimax lower bound

In this subsection, we prove a minimax lower bound for estimating the central space over the model class $\mathfrak{M}(p,d,\lambda)$, stated in terms of the triplet (n,p,d) as well as λ.

Theorem 4 (Lower bound)**.** *There exists a specification of $\mathfrak{M}(p,d,\lambda)$ and a universal constant C_1 such that if $2d<p$ and $dp<C_1 n\lambda$, then*

$$\inf_{\widehat{\boldsymbol{B}}} \sup_{\mathcal{M}\in\mathfrak{M}(p,d,\lambda)} \mathbb{E}_{\mathcal{M}} \left\| \widehat{\boldsymbol{B}}\widehat{\boldsymbol{B}}^\top - \boldsymbol{B}\boldsymbol{B}^\top \right\|_{\mathrm{F}}^2 \gtrsim \frac{d(p-d)}{n\lambda}.$$

Theorem 4 states a sharp lower bound for the minimax rate, which matches with the upper bound in Section 4.2.2. It is very challenging to obtain a lower bound that depends optimally on all parameters, in particular λ and d.

The proof of our minimax lower bound follows the standard Fano method framework (see e.g., [Yu, 1997]). This framework requires a family of distributions that are separated from each other in the parameter space but close to each other in terms of the Kullback–Leibler divergence (KL-divergence). One of the main technical contributions of this paper is the explicit construction of such a family of distributions in $\mathfrak{M}(p,d,\lambda)$. The construction is novel in the literature. Furthermore, it is highly nontrivial to obtain a sharp upper bound on the pairwise KL-divergence. The difficulty arises from the nonlinear relationship between Y and $\boldsymbol{X}$ encoded by the multiple-index model and the semiparametric nature of the SDR problem.

We will first describe how to construct the family of distributions of $(\boldsymbol{X},Y)$ and then sketch the proof of the minimax lower bound based on this construction. As a building block for the construction, we introduce a piecewise constant function as follows.

Definition 4. Let m be the median of χ^2_d distribution. $\psi(s_1,\dots,s_d)$ is a function from $\mathbb{R}^d$ to $\{-d,\dots,0,1,\dots,d\}$.

1. If $\sum_i s_i^2 \leqslant m$, suppose $|s_i|$ is uniquely the largest among $|s_1|,\dots,|s_d|$, then $\psi(s_1,\dots,s_d)=\mathrm{sgn}(s_i)i$;
2. if $\sum_i s_i^2 > m$ or if the largest number is not unique, then $\psi(s_1,\dots,s_d)=0$.

For each $\boldsymbol{B}\in\mathbb{O}(p,d)$, we construct the following joint distribution $\mathbb{P}_{\boldsymbol{B}}$ of $(\boldsymbol{X},Y)$:

$$\begin{aligned} \boldsymbol{X} &\sim N(0,\boldsymbol{I}_p), \\ \boldsymbol{Z} &= \rho\boldsymbol{B}^\top\boldsymbol{X} + \sqrt{1-\rho^2}\xi, \quad \xi\sim N(0,\boldsymbol{I}_d), \\ W &= \psi(\boldsymbol{Z}), \\ Y &= W+\eta, \quad \eta\sim \mathrm{Unif}(-\sigma,\sigma), \end{aligned} \tag{10}$$

where $\boldsymbol{X},\xi,\eta$ are independent of each other, and $\sigma\in(0,1/2]$ and $\rho\in(0,1)$ are fixed constants.

We are now ready to present the sketch of the proof for the minimax lower bound. For any $\lambda\leqslant\varpi_d$, we can choose ρ such that $\mathbb{P}_{\boldsymbol{B}}$ constructed in (10) satisfies the following two properties:

(i) for any $\boldsymbol{B}\in\mathbb{O}(p,d)$, Y can be represented as $f(\boldsymbol{B}^\top\boldsymbol{X},\epsilon)$ where $\epsilon\sim N(0,1)$ and $\mathbb{P}_{\boldsymbol{B}}$ belongs to $\mathfrak{M}(p,d,\lambda)$;

(ii) for any $\boldsymbol{B}_1,\boldsymbol{B}_2\in\mathbb{O}(p,d)$, $\mathrm{KL}(\mathbb{P}_{\boldsymbol{B}_1},\mathbb{P}_{\boldsymbol{B}_2})\leqslant C\lambda\|\boldsymbol{B}_1-\boldsymbol{B}_2\|_F^2$ for some absolute constant C.

Recall that for any sufficiently small $\varepsilon>0$ and any $\alpha\in(0,1)$, there is a subset $\Theta\subset\mathbb{O}(p,d)$ such that

$$|\Theta| \geqslant \left(\frac{C_0}{\alpha}\right)^{d(p-d)} \quad \text{and}$$
$$\|\boldsymbol{B}-\widetilde{\boldsymbol{B}}\|_{\mathrm{F}}\leqslant 2\varepsilon, \qquad \|\boldsymbol{B}\boldsymbol{B}^\top-\widetilde{\boldsymbol{B}}\widetilde{\boldsymbol{B}}^\top\|_F\geqslant\alpha\varepsilon$$

for any $\boldsymbol{B},\widetilde{\boldsymbol{B}}\in\Theta$ and some absolute constant C_0. Therefore, the class of distributions $\{\mathbb{P}_{\boldsymbol{B}}:\boldsymbol{B}\in\Theta\}$ are separated from each other in terms of $\boldsymbol{B}\boldsymbol{B}^\top$ and are close to each other in terms of KL-divergence. We can apply Fano's inequality to obtain the lower bound on the minimax risk. The details of the proof can be found in Appendix E. The main technical difficulty lies in bounding the KL-divergence between any two joint distributions, $\mathbb{P}_{\boldsymbol{B}_1}$ and $\mathbb{P}_{\boldsymbol{B}_2}$, sharply by $\lambda\|\boldsymbol{B}_1-\boldsymbol{B}_2\|_F^2$.

Remark 2. The gSNR of the distribution constructed by (10) can be explicitly computed as

$$2d^{-1}\rho^2\left(\mathbb{E}[\max_{i\in[d]}(|Z_i|)\,1_{\|\boldsymbol{Z}\|^2\leqslant m}]\right)^2,$$

where $\boldsymbol{Z}$ is a d-variate standard normal vector. This will allow researchers to conduct numerical experiments to examine dependence of the estimation error on the gSNR and the structural dimension via simulations.

The following result provides a characterization of the specification of $\mathfrak{M}(p,d,\lambda)$ in Theorem 4.

Proposition 1. *There exist two universal constants C_2 and c_2, such that a specification of $\mathfrak{M}(p,d,\lambda)$ will satisfy Theorem 4 whenever $K_0\geqslant C_2$, $\varpi_d\leqslant c_2 d^{-8.1}$, $M\geqslant 1$, and $\kappa\geqslant 1$.*

As demonstrated in Section 3.2, the gSNR often decay rapidly or even appears to decay exponentially as the structural dimension d increases. The choice of ϖ_d in Proposition 1 allows the model class $\mathfrak{M}(p,d,\lambda)$ to encompass the most interesting central space estimation problems in the low gSNR regime.

4.2.2 Minimax rate

In this subsection, we provide an upper bound on the minimax risk of estimating the central space over $\mathfrak{M}(p,d,\lambda)$, which is achieved by the SIR method.

Theorem 5 (Upper bound)**.** *Consider the specification of $\mathfrak{M}(p,d,\lambda)$ in Theorem 4. There is a universal constant C_1 such that*

$$\inf_{\widehat{\boldsymbol{B}}}\sup_{\mathcal{M}\in\mathfrak{M}(p,d,\lambda)}\mathbb{E}_{\mathcal{M}}\left\|\widehat{\boldsymbol{B}}\widehat{\boldsymbol{B}}^\top-\boldsymbol{B}\boldsymbol{B}^\top\right\|_{\mathrm{F}}^2\lesssim\frac{dp}{n\lambda} \tag{11}$$

holds whenever $dp+d^2(\log(nd)+d)<C_1 n\lambda<C_1e^p$.

We prove this upper bound through a careful analysis of the SIR method (see Appendices C.2 and D.1). Since this upper bound matches the lower bound in Theorem 4, we conclude the following minimax optimal rate for estimating the central space over $\mathfrak{M}(p,d,\lambda)$ when the structural dimension d is allowed to grow.

Theorem 6 (Minimax optimal rate)**.** *Consider the specification of $\mathfrak{M}(p,d,\lambda)$ in Theorem 4. There exist a universal constant C_1 such that the following holds*

$$\inf_{\widehat{\boldsymbol{B}}}\sup_{\mathcal{M}\in\mathfrak{M}(p,d,\lambda)}\mathbb{E}_{\mathcal{M}}\left\|\widehat{\boldsymbol{B}}\widehat{\boldsymbol{B}}^\top-\boldsymbol{B}\boldsymbol{B}^\top\right\|_{\mathrm{F}}^2\asymp\frac{dp}{n\lambda} \tag{12}$$

whenever $dp+d^2(\log(nd)+d)<C_1n\lambda<C_1e^p$.

Theorem 6 captures the dependence of the minimax rate for estimating the central space on the parameters p, n, λ, and the structural dimension d. In particular, the minimax rate is linear w.r.t. the structural dimension d and is inverse proportional to λ. This provides a precise characterization of the challenges in estimating the central space across a wide range of models.

We conclude that the estimation problem becomes difficult not only when the structural dimension d is large but also when the gSNR is small. For instance, consider a case where $d=10$, $p=10^3$, and $n=10^6$. The ratio of the number of parameters in $\boldsymbol{B}$ to the sample size, $\frac{dp}{n}$, is as small as 0.001. However, if the gSNR is on the scale of $n^{-1/3}$, Theorem 6 suggests that the estimation error remains significant. Nonetheless, under the conditions of Theorem 6, the regular SIR method is minimax rate optimal for the model $\mathfrak{M}(p,d,\lambda)$. Therefore, in the cases where the structural dimension d is large and the gSNR is small, the poor performance of SIR should be attributed to the intrinsic difficulty of the estimation problem rather than flaws in the method itself.

Dependence of error w.r.t. d and λ In the following examples, we illustrate the dependence of the optimal rate on the structural dimension d and the lower bound of the gSNR (i.e., λ). We observe that for various values of d and λ, the average loss of SIR exhibits a linear relationship with d and an inverse proportion relationship with λ, consistent with the theoretical result in Theorem 6.

In this experiment, we construct $(\boldsymbol{X}, Y)$ as per Equation (10), with $\boldsymbol{B}$ chosen as $\left[\boldsymbol{I}_d, \mathbf{0}_{d\times(p-d)}\right]^\top$, $\sigma = 0.5$. We set $\rho = \theta \cdot \sqrt{d}\left(\mathbb{E}[\max|Z_i|1_{\|\boldsymbol{Z}\|^2\leqslant m}]\right)^{-1}$, where θ is a scaling factor ensuring that ρ lies in $(0,1)$, $\boldsymbol{Z}$ is a d-variate standard normal random vector, and the expectation is computed numerically. Under such a specification of ρ, two properties are implied by Remark 2: 1) the gSNR remains constant if we fix the value of θ, and 2) gSNR $\propto \theta^2$ if we fixed the value of d. Therefore, we can illustrate the linear dependence of the estimation loss on d by fixing (n, p, θ) and varying the value of d. In addition, we can demonstrate that the estimation loss is inverse proportional to λ by examining the relationship between the loss and θ^2. We vary the value of d within $\{2, 4, 6, 8, 10\}$ and θ within $\{0.01, 0.02, \ldots, 0.07\}$, while fixing $n = 10^6$, $p = 200$, and $H = 1000$.

Figure 3 shows the relationship between the general loss and the varying factors d and θ based on 100 replications. In the left subplot, the solid line represents the average general loss as d increases (with a fixed $\theta = 0.05$), which aligns perfectly with the dotted straight line fitted by least squares regression ($1 - R^2 < 0.001$). The shaded areas represent the standard error associated with these estimates and all of them are less than 0.003. This plot indicates a linear dependence of the estimation loss on d. In the right subplot, the solid line plots the average of the logarithm of general loss against the logarithm of θ (with a fixed $d = 10$). The dotted line is the straight line fitted by least squares regression, featuring a slope of -2.021 and $1 - R^2 < 0.03$. This plot indicates that the estimation loss is inverse proportional to θ^2, and thus to λ. The observations in this experiment are consistent with the theoretical result in Theorem 6.

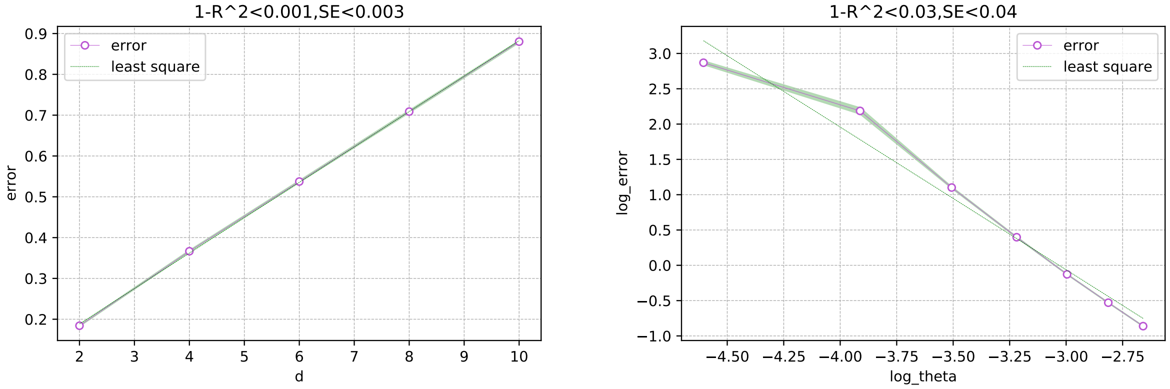


Figure 3: Left: error with increasing d; Right: logarithm of error with increasing $\log(\theta)$

4.3 Optimal rate for high-dimensional sparse SIR

In cases where p the dimension of the predictor is comparable to or larger than the sample size n, the SIR estimator for the central space is inconsistent [Lin et al., 2018a]. Therefore, we need to impose certain structural assumptions to ensure consistent estimation in high dimensional settings. In this subsection, we determine the minimax rate for estimating the central space under a sparsity assumption.

We impose the sparsity assumption on the indices matrix $\boldsymbol{B}$ as follows. Denote by $\mathrm{supp}(\boldsymbol{B})$ the support of $\boldsymbol{B}$:

$$\mathrm{supp}(\boldsymbol{B}) = \{j \in [p] : \|\boldsymbol{B}_{j,*}\| > 0\},$$

and by $\|\boldsymbol{B}\|_0$ the number of non-zero rows of $\boldsymbol{B}$, i.e., $\|\boldsymbol{B}\|_0 = |\mathrm{supp}(\boldsymbol{B})|$. We assume that $\boldsymbol{B}$ *is ℓ_0-sparse*, i.e., $\|\boldsymbol{B}\|_0 \leqslant s$ for some integer s.

We consider the estimation of the central space for the class of sparse models:

$$\mathfrak{M}_s(p,d,\lambda) = \mathfrak{M}(p,d,\lambda) \cap \left\{\text{distribution of } (\boldsymbol{X}, Y = f(\boldsymbol{B}^\top \boldsymbol{X}, \epsilon)) \,\middle|\, \|\boldsymbol{B}\|_0 \leqslant s\right\}. \tag{13}$$

As in the definition of the model class $\mathfrak{M}(p,d,\lambda)$ in (9), we are interested in the low gSNR regime and require that $\lambda \leqslant \varpi_d$ for some constant ϖ_d depending only on the structural dimension d. We allow the parameters (p,s,d) to grow and λ to decay as n increases. Particularly, in a high-dimensional setting, p might be much larger than n while d and s might grow at a slow rate in n. Similar to the definition of $\mathfrak{M}(p,d,\lambda)$, *a specification of* $\mathfrak{M}_s(p,d,\lambda)$ refers to a set of values of the constants in (13).

We begin with an extension of the lower bound for the minimax risk in Theorem 6 to the high dimensional sparse model (13). The proof can be found in Appendix F.

Theorem 7 (Lower bound for sparse models)**.** *There exists a specification of $\mathfrak{M}_s(p,d,\lambda)$ and a universal constant C_1 such that*

$$\inf_{\widehat{\boldsymbol{B}}} \sup_{\mathcal{M}\in\mathfrak{M}_s(p,d,\lambda)} \mathbb{E}_{\mathcal{M}} \left\|\widehat{\boldsymbol{B}}\widehat{\boldsymbol{B}}^\top - \boldsymbol{B}\boldsymbol{B}^\top\right\|_{\mathrm{F}}^2 \gtrsim \frac{ds + s\log(ep/s)}{n\lambda}. \tag{14}$$

holds whenever $2d < s$ and $ds + s\log(p/s) < C_1 n\lambda$.

An upper bound on the minimax risk of estimating the central space over $\mathfrak{M}_s(p,d,\lambda)$ is given in the following theorem.

Theorem 8 (Upper bound for sparse models)**.** *Consider the specification of $\mathfrak{M}_s(p,d,\lambda)$ in Theorem 7. There is a universal constant C_1 such that*

$$\inf_{\widehat{\boldsymbol{B}}} \sup_{\mathcal{M}\in\mathfrak{M}_s(p,d,\lambda)} \mathbb{E}_{\mathcal{M}} \left\|\widehat{\boldsymbol{B}}\widehat{\boldsymbol{B}}^\top - \boldsymbol{B}\boldsymbol{B}^\top\right\|_{\mathrm{F}}^2 \lesssim \frac{ds + s\log(ep/s)}{n\lambda} \tag{15}$$

holds wherever $d^2(\log(nd) + d) + s\log(p/s) + ds < C_1 n\lambda < C_1 e^s$.

In order to obtain the upper bound in Theorem 8, we introduce an aggregation estimator following the idea in Cai et al. [2013] and Lin et al. [2021]. This estimator is constructed by sample splitting and aggregation. For simplicity, assume that there are $n = 2Hc$ samples for some positive integer c and these samples are divided into two equal-sized sets. We denote by $\boldsymbol{\Lambda}_H^{(i)} (i = 1,2)$ the SIR estimates of $\boldsymbol{\Lambda} = \mathrm{Cov}(\mathbb{E}[\boldsymbol{X}|Y])$ using the i-th set of samples. Similarly, denote by $\widehat{\boldsymbol{\Sigma}}^{(1)}$ the sample covariance matrix based on the first set of samples. Let $\mathcal{L}(s)$ be the set of all subsets of $[p]$ with size s. The aggregation estimator $\widehat{\boldsymbol{B}}$ is constructed as follows:

(i) For each $L \in \mathcal{L}(s)$, let

$$\begin{aligned} \widehat{\boldsymbol{B}}_L := \arg\max_{\boldsymbol{B}} \quad & \mathrm{Tr}(\boldsymbol{B}^\top \boldsymbol{\Lambda}_H^{(1)} \boldsymbol{B}) \\ \text{s.t. } \boldsymbol{B}^\top \widehat{\boldsymbol{\Sigma}}^{(1)} \boldsymbol{B} = \mathbf{I}_d & \text{ and } \mathrm{supp}(\boldsymbol{B}) \subset L. \end{aligned} \tag{16}$$

(ii) Our aggregation estimator $\widehat{\boldsymbol{B}}$ is defined to be $\widehat{\boldsymbol{B}}_{L^*}$ where

$$L^* := \arg\max_{L\in\mathcal{L}(s)} \quad \mathrm{Tr}(\widehat{\boldsymbol{B}}_L^\top \boldsymbol{\Lambda}_H^{(2)} \widehat{\boldsymbol{B}}_L).$$

We show that $\widehat{\boldsymbol{B}}_{L^*}$ attains the rate in the right hand side of (15) (see Appendix C.3 and D.2). Furthermore, this rate is optimal because it matches the lower bound in Theorem 7, as summarized in Theorem 9.

Theorem 9 (Optimal rate of sparse models)**.** *Consider the specification of $\mathfrak{M}_s(p,d,\lambda)$ in Theorem 7. There exists a universal constant C_1 such that the following holds*

$$\inf_{\widehat{\boldsymbol{B}}} \sup_{\mathcal{M}\in\mathfrak{M}_s(p,d,\lambda)} \mathbb{E}_{\mathcal{M}} \left\|\widehat{\boldsymbol{B}}\widehat{\boldsymbol{B}}^\top - \boldsymbol{B}\boldsymbol{B}^\top\right\|_{\mathrm{F}}^2 \asymp \frac{ds + s\log(ep/s)}{n\lambda} \tag{17}$$

whenever $2d < s$ and $d^2(\log(nd) + d) + s\log(p/s) + ds < C_1 n\lambda < C_1 e^s$.

5 Discussions

We have provided a complete characterization of the minimax rate for sufficient dimension reduction (SDR) under multiple-index models. Our theory encompasses a broad scenario: the generalized signal-to-noise ratio (gSNR) may approach 0, the covariance matrix $\boldsymbol{\Sigma}$ is unknown, and, most importantly, the structural dimension d can be large. We established the matching upper bounds and lower bounds on the minimax risk for estimating the central space, addressing both the ordinary scaling and the high-dimensional scaling separately. To the best of our knowledge, this paper is the first attempt to investigate the minimax rate in SDR where the structural dimension d can grow along with the sample size n.

The results in the current paper are different from the ones in Lin et al. [2021]. First, the upper bound result in the earlier paper imposed more restrictive conditions on the covariance matrix $\boldsymbol{\Sigma}$: either it had to be the identity matrix or it had to satisfy a technical condition involving the index matrix $\boldsymbol{B}$, which is hard to verify (see Equation (28) therein). In our study, we only assume that the largest and the smallest eigenvalues of $\boldsymbol{\Sigma}$ are bounded. Second, the minimax rate results for multiple-index models in the earlier paper had certain limitations: the structural dimension d had a fixed upper bound, and the gSNR was required to be bounded away from 0 by a fixed constant. The authors speculated a more general result that relaxed the constraint on the gSNR, but it relied on a conjecture that had not yet been proven. In contrast, the minimax rate results in our study not only allow d to grow but also allow the gSNR to approach 0. Besides, our results are also different from the ones in Tan et al. [2020], who considered Gaussian mixture models and required d to be fixed. Our minimax rates have a clear advantage as they are applicable to a wider range of situations and they highlight the crucial roles of the structural dimension and the gSNR in SDR problems. In conjunction with the empirical observation that gSNR tends to decay rapidly as d grows, we provide a theoretical explanation for the underperformance of the SIR method when d is large.

Our introduction of the weak sliced stable condition (WSSC) also contributes to the theoretical development of SDR. This condition simplifies our proofs for both the upper bounds and the lower bounds of the minimax rates presented in this paper. Furthermore, we expect that deriving the WSSC from the moment condition in Theorem 1 would inspire future research on the high-dimensional behavior of other SDR methods, such as SAVE, without assuming higher-order sliced stability conditions. For instance, if one intends to study the phase transition phenomenon of SAVE in high dimensions, as was done for SIR in Lin et al. [2018a], it might initially seem that a higher-order SSC is indispensable. This speculation stems from the observation that in low-dimensional settings, the asymptotic theory of SAVE developed in Li and Zhu [2007] required several conditions similar to those proposed for SIR in Hsing and Carroll [1992], Zhu and Ng [1995], but with a higher order. However, our results shed light on another possibility: the higher-order SSC could be replaced by the existence of $\ell > 2$ moments of predictors.

Our findings raise several open questions. First, our current theory focuses on the SIR method and our notion of gSNR is related to $\operatorname{Cov}(\mathbb{E}[\boldsymbol{X} \mid Y])$. When studying other SDR methods, it might be possible to consider alternative definitions of gSNR in the corresponding contexts and establish the minimax rate for estimating the central space. Second, although Theorem 2 proved a tight upper bound on the decay rate of gSNR, this upper bound is rarely attained by a joint distribution that is likely to be encountered in real practice. Indeed, in our simulation studies in Section 3.2, the gSNR appears to decay at an exponential rate as d grows. It is interesting to develop a theory that explains the rapid decay of gSNR, as observed empirically. Third, we construct the aggregation estimator for the theoretical purpose of proving the upper bound on the minimax risk for high-dimensional sparse models. From a practical perspective, it is beneficial to develop computationally efficient algorithms that attain or nearly attain the optimal rate. Lastly, our theory presumes that the structural dimension d has been known. However, in real practice, the determination of d has to be inferred from the data. This issue, known as order determination, has been widely studied in the literature (see Li [2017, Chapter 9] and references therein). It is still unclear whether a minimax theory can be established when d is unknown a priori.

Acknowledgements

Huang's research is supported in part by NUS Start-up Grant A-0004824-00-0 and Singapore Ministry of Education AcRF Tier 1 Grant A-8000466-00-00. Lin's research was supported in part by the National Natural Science Foundation of China (Grant 11971257), Beijing Natural Science Foundation (Grant Z190001), National Key R&D Program of China (2020AAA0105200), and Beijing Academy of Artificial Intelligence.

References

Anderson, T. W. (1955). The integral of a symmetric unimodal function over a symmetric convex set and some probability inequalities. *Proceedings of the American Mathematical Society 6*(2), 170–176.

Berg, C. and H. L. Pedersen (2006). The Chen–Rubin conjecture in a continuous setting. *Methods and Applications of Analysis 13*(1), 63–88.

Boucheron, S., G. Lugosi, and P. Massart (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press.

Cai, T. T., Z. Ma, and Y. Wu (2013). Sparse PCA: Optimal rates and adaptive estimation. *The Annals of Statistics 41*(6), 3074–3110.

Cook, R. D. and S. Weisberg (1991). Sliced inverse regression for dimension reduction: Comment. *Journal of the American Statistical Association 86*(414), 328–332.

Ferré, L. (1998). Determining the dimension in sliced inverse regression and related methods. *Journal of the American Statistical Association 93*(441), 132–140.

Gao, C., Z. Ma, Z. Ren, H. H. Zhou, et al. (2015). Minimax estimation in sparse canonical correlation analysis. *The Annals of Statistics 43*(5), 2168–2197.

Hsing, T. and R. J. Carroll (1992). An asymptotic theory for sliced inverse regression. *The Annals of Statistics 20*(2), 1040–1061.

Laurent, B. and P. Massart (2000). Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics 28*(5), 1302–1338.

Ledoux, M. and M. Talagrand (1991). *Probability in Banach Spaces: Isoperimetry and processes*. Springer–Verlag Berlin Heide1berg.

Li, B. (2017). *Sufficient Dimension Reduction: Methods and Applications with R*. Chapman and Hall/CRC.

Li, B. and S. Wang (2007). On directional regression for dimension reduction. *Journal of the American Statistical Association 102*(479), 997–1008.

Li, K.-C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association 86*(414), 316–327.

Li, K.-C. (1992). On principal hessian directions for data visualization and dimension reduction: Another application of stein's lemma. *Journal of the American Statistical Association 87*(420), 1025–1039.

Li, L. (2007). Sparse sufficient dimension reduction. *Biometrika 94*(3), 603–613.

Li, Y. and L.-X. Zhu (2007). Asymptotics for sliced average variance estimation. *The Annals of Statistics 35*(1), 41–69.

Lin, Q., X. Li, D. Huang, and J. S. Liu (2021). On the optimality of sliced inverse regression in high dimensions. *The Annals of Statistics 49*(1), 1–20.

Lin, Q., Z. Zhao, and J. S. Liu (2018a). On consistency and sparsity for sliced inverse regression in high dimensions. *The Annals of Statistics 46*(2), 580–610.

Lin, Q., Z. Zhao, and J. S. Liu (2018b). Supplement to "On consistency and sparsity for sliced inverse regression in high dimensions".

Lin, Q., Z. Zhao, and J. S. Liu (2019). Sparse sliced inverse regression via lasso. *Journal of the American Statistical Association 114*(528), 1726–1739.

Ma, Z. and X. Li (2020). Subspace perspective on canonical correlation analysis: Dimension reduction and minimax rates. *Bernoulli 26*(1), 432–470.

Schmitt, B. A. (1992). Perturbation bounds for matrix square roots and pythagorean sums. *Linear algebra and its applications 174*, 215–227.

Seeger, M. (2004). Gaussian processes for machine learning. *International journal of neural systems 14*(02), 69–106.

Tan, K., L. Shi, and Z. Yu (2020). Sparse SIR: Optimal rates and adaptive estimation. *The Annals of Statistics 48*(1), 64–85.

Tao, T. (2012). *Topics in random matrix theory*, Volume 132 of *Graduate Studies in Mathematics*. American Mathematical Society.

Tsybakov, A. B. (2008). *Introduction to Nonparametric Estimation.* Springer Series in Statistics. Springer New York, NY.

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

Vu, V. and J. Lei (2012, 21–23 Apr). Minimax rates of estimation for sparse PCA in high dimensions. In N. D. Lawrence and M. Girolami (Eds.), *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, Volume 22 of *Proceedings of Machine Learning Research*, La Palma, Canary Islands, pp. 1278–1286. PMLR.

Watson, G. N. (1959). A note on gamma functions. *Edinburgh Mathematical Notes 42*, 7–9.

Williams, C. K. and C. E. Rasmussen (2006). *Gaussian processes for machine learning*, Volume 2. MIT press Cambridge, MA.

Wu, Y. and L. Li (2011). Asymptotic properties of sufficient dimension reduction with a diverging number of predictors. *Statistica Sinica 2011*(21), 707.

Xia, Y., H. Tong, W. K. Li, and L.-X. Zhu (2009). An adaptive estimation of dimension reduction space. In *Exploration of A Nonlinear World: An Appreciation of Howell Tong's Contributions to Statistics*, pp. 299–346. World Scientific.

Yang, S. S. (1977). General distribution theory of the concomitants of order statistics. *The Annals of Statistics 5*(5), 996–1002.

Yu, B. (1997). Assouad, Fano, and Le Cam. In *Festschrift for Lucien Le Cam*, pp. 423–435. Springer.

Zhang, A. R. and Y. Zhou (2020). On the non-asymptotic and sharp lower tail bounds of random variables. *Stat 9*(1), e314.

Zhu, L., B. Miao, and H. Peng (2006). On sliced inverse regression with high-dimensional covariates. *Journal of the American Statistical Association 101*(474), 630–643.

Zhu, L.-X. and K. W. Ng (1995). Asymptotics of sliced inverse regression. *Statistica Sinica 5*(2), 727–736.

A Proof of Lemma 1

Proof. It is a direct corollary of Lemma 1 (*'key lemma'*) in Lin et al. [2018a] by noticing that factor $\frac{\gamma_3}{H^\nu}$ in the proof of Lemma 2(i) therein corresponds to factor $\frac{1}{\tau}$ in Definition 2. □

B Proof of Theorem 1 and Corollary 1

We first introduce the following lemma, which shows that the WSSC is easy to satisfy for general curves.

Lemma 2. *Let $Y \in \mathbb{R}$ be a random variable and $\boldsymbol{\kappa}: \mathbb{R} \to \mathbb{R}^p$ be a nonzero continuous function satisfying*

(i) $\sup_{\|\boldsymbol{\beta}\|=1} \mathbb{E}[|\langle \boldsymbol{\kappa}(Y), \boldsymbol{\beta}\rangle|^\ell] \leqslant c_1$ holds for some $\ell > 2$ and $c_1 > 0$;

(ii) Y is a continuous random variable.

Then for any $\tau > 1$, there exists a constant $K = K(\tau, d') \geqslant d'$ such that $\boldsymbol{\kappa}(y)$ is weak (K, τ)-sliced stable w.r.t. Y where $d' := \dim\{\mathrm{span}\{\boldsymbol{\kappa}(y) : y \in \mathbb{R}\}\}$.

If we further assume that $\|\boldsymbol{\kappa}(y) - \boldsymbol{\kappa}(y')\|^2 \leqslant Cd'(y-y')^2$ for any y, y' defined on a compact set of $\mathbb{R}$ and $C > 0$ a constant, then the WSSC coefficient $K = K_0 d'$ for some integer $K_0 \geqslant 1$.

By density transformation formula of continuous random variable, it is easy to check that there exists a monotonic function f such that the (probability density function) p.d.f. of $f(Y)$ is continuous and positive everywhere. Since any monotonic transformation $f(Y)$ keeps SIR procedures unchanged, (ii) in Lemma 2 can be replaced by

(ii') The p.d.f. of Y is continuous and positive everywhere on $\mathbb{R}$.

B.1 Proof of Lemma 2

To prove Lemma 2, we need some properties of any partition $\mathcal{B}_H := \{-\infty = a_0 < a_1 < \cdots < a_{H-1} < a_H = \infty\}$ of $\mathbb{R}$. Before this, we introduce some important definitions. For simplicity, we let $U_h := (a_{h-1}, a_h]$.

Definition 5. Let Y be a real random variable and $\gamma \in (0, 1)$. A partition $\mathcal{B}_H := \{U_h\}_{h=1}^H$ of $\mathbb{R}$ is called a γ-partition w.r.t. Y if

$$\frac{1-\gamma}{H} \leqslant \mathbb{P}(Y \in U_h) \leqslant \frac{1+\gamma}{H}$$

for $h = 1, \ldots, H$.

Definition 6. Suppose $\delta > 0$, $D \subset \mathbb{R}$ is compact and Y is a real random variable. A γ-partition $\mathcal{B}_H$ w.r.t. Y is (δ, D)-admissible if there exists a compact set $D' \supset D$ with the following properties:

(1) For any $U_h \in \mathcal{B}_H$ such that $U_h \bigcap D \neq \emptyset$, then $U_h \subset D'$.

(2) For any $U_h \in \mathcal{B}_H$ such that $U_h \subset D'$, then $\mathrm{diam}(U_h) < \delta$, where $\mathrm{diam}(U_h)$ is the diameter of U_h.

Intuitively, when $\mathbb{R}$ is partitioned into sufficiently many intervals of the same probability mass of the distribution of Y, then the Euclidean lengths of these intervals within a specified compact set are all sufficiently small. Thus the partition $\mathcal{B}_H$ is (δ, D)-admissible. We shall illustrate this intuition by the following Lemma, which is an essential property of the partition.

Lemma 3. *Suppose that $Y \in \mathbb{R}$ is a real random variable and $p(y)$ is continuous and positive everywhere. Let $\{\mathcal{B}_H\}$ be a sequence of γ-partition. For any $\delta > 0$ and compact set D, there exists an $H_0 = H_0(\gamma, \delta, D)$ such that for any $H > H_0$, $\mathcal{B}_H$ is a (δ, D)-admissible γ-partition.*

Proof. Let D' be a compact interval containing the set $\{x \mid \mathrm{dist}(x, D) \leqslant 1/2\}$. Since $p(y)$ is positive and continuous over D', there exists a positive number μ, such that for any interval $[a, b] \subset D'$ satisfying $|b - a| \geqslant 1/2$, one has $\mathbb{P}(Y \in [a, b]) \geqslant \mu$. Choose a real number $H_1 > \frac{1+\gamma}{\mu}$. For any $H > H_1$, one has $\mathbb{P}(Y \in U_h) < \mu$. Hence, if $U_h \cap D \neq \emptyset$, we must have diameter$(U_h) \leqslant 1/2$. This implies that $U_h \subset D'$. Let $\mu' := \min_{Y \in D'} p(Y)$. As D' is compact, our assumption about $p(Y)$ implies that μ' is well-defined and $\mu' > 0$. Let $H_2 \in \mathbb{Z}_{\geqslant 1}$ satisfying $\frac{1+\gamma}{H_2 \mu'} < \delta$. For any $H > H_2$, let $a_h < b_h$ be two endpoints of $U_h \subset D'$, one has $\mu'(b_h - a_h) \leqslant \mathbb{P}(Y \in U_h) \leqslant \frac{1+\gamma}{H}$. Then we obtain $\mathrm{diam}(U_h) = b_h - a_h \leqslant \frac{1+\gamma}{H\mu'} < \delta$. Choose $H_0 = \max\{H_1, H_2\}$, then for any $H > H_0$, one has $\mathcal{B}_H$ is (δ, D)-admissible. □

Proof of Lemma 2: Now we are ready to prove Lemma 2 under condition (i) and (ii'). We first prove that for any $\boldsymbol{\beta} \in \mathbb{S}^{p-1}$,

$$\frac{1}{H} \sum_{h: U_h \in \mathcal{B}_H} \mathrm{var}(\boldsymbol{\kappa}(\boldsymbol{\beta}) \mid Y \in U_h) \leqslant \frac{1}{\tau} \lambda_{\min}^{+}(\mathrm{Cov}(\boldsymbol{\kappa}(Y))) \tag{18}$$

where $\boldsymbol{\kappa}(\boldsymbol{\beta}) := \boldsymbol{\beta}^\top \boldsymbol{\kappa}(Y)$.

For any $\tau > 0$, let us choose a compact set D such that $\mathbb{P}(Y \in D^c) < \epsilon_1^{\frac{\ell}{\ell-2}}$, where $\epsilon_1 = \frac{(1-\gamma)\lambda_{\min}^{+}(\mathrm{Cov}(\boldsymbol{\kappa}(Y)))}{\tau \sup_{\|\boldsymbol{\beta}\|=1} \mathbb{E}[|\langle \boldsymbol{\kappa}(Y), \boldsymbol{\beta} \rangle|^{\ell}]^{2/\ell}}$. Let δ be some small positive constant, say, 0.2. By Lemma 3, we know that there exists an H_0, such that for any $H > H_0$, the partition $\mathcal{B}_H$ is a (δ, D) admissible γ-partition. This means there exists a compact set D', such that if $U_h \cap D \neq \emptyset$, then $U_h \subset D'$ and $\mathrm{diam}(U_h) < \delta$. Since $\boldsymbol{\kappa}$ is uniformly continuous on D', for $\epsilon_2 = \sqrt{\frac{\lambda_{\min}^{+}(\mathrm{Cov}(\boldsymbol{\kappa}(Y)))}{\tau}}$, there exists an ϵ' such that $\|\boldsymbol{\kappa}(y_1) - \boldsymbol{\kappa}(y_2)\| \leqslant \epsilon_2$ for any $y_1, y_2 \in D'$ satisfying that $|y_1 - y_2| < \epsilon'$. Let $\delta' = \min\{\delta, \epsilon'\}$. By Lemma 3, there exists an H_0', such that if $\mathcal{B}_H$ is a γ partition, then it is (δ', D) admissible for any $H > H_0'$ (If we further assume that $\|\boldsymbol{\kappa}(y) - \boldsymbol{\kappa}(y')\| \leqslant Cd'|y - y'|$, then $\epsilon' \asymp \frac{1}{d'}$ and $H_0' \asymp d'$).

For such (δ', D) admissible partition $\mathcal{B}_H$, one has

1 For any $U_h \cap D \neq \emptyset$ and any $\boldsymbol{\beta} \in \mathbb{S}^{p-1}$, by intermediate value theorem, we know that there exists a $\xi \in U_h$, such that

$$\int_{U_h} \boldsymbol{\kappa}(\boldsymbol{\beta}) p(y \mid Y \in U_h) dy = \boldsymbol{\beta}^\top \boldsymbol{\kappa}(\xi).$$

Thus, one has

$$\mathrm{var}(\boldsymbol{\kappa}(\boldsymbol{\beta}) \mid Y \in U_h) = \int_{U_h} (\boldsymbol{\kappa}(\boldsymbol{\beta}) - \boldsymbol{\beta}^\top \boldsymbol{\kappa}(\xi))^2 p(y \mid Y \in U_h) dy \leqslant \epsilon_2^2 = \frac{\lambda_{\min}^{+}(\mathrm{Cov}(\boldsymbol{\kappa}(Y)))}{\tau}.$$

2 For any $U_h \subset D^c$ and any $\boldsymbol{\beta} \in \mathbb{S}^{p-1}$, one has

$$\mathrm{var}(\boldsymbol{\kappa}(\boldsymbol{\beta}) \mid Y \in U_h) \leqslant \int_{U_h} (\boldsymbol{\kappa}(\boldsymbol{\beta}))^2 p(y \mid Y \in U_h) dy \leqslant \frac{H}{1-\gamma} \int_{U_h} (\boldsymbol{\kappa}(\boldsymbol{\beta}))^2 p(y) dy$$

because $\gamma \leqslant \frac{1}{\tau}$. Thus,

$$\frac{1}{H} \sum_{h: U_h \subset D^c} \mathrm{var}(\boldsymbol{\kappa}(\boldsymbol{\beta}) \mid Y \in U_h) \leqslant \frac{1}{1-\gamma} \int_{D^c} (\boldsymbol{\kappa}(\boldsymbol{\beta}))^2 p(y) dy.$$

Let $f(y) = \kappa(\boldsymbol{\beta})^2$ and $q(y) = p(y)/\mathbb{P}(Y \in D^c)$. By Jensen's inequality, we have $\int_{D^c} f(y) q(y) dy \leqslant \left(\int_{D^c} f^{\ell/2}(y) q(y) dy\right)^{2/\ell}$. This can be written as

$$\int_{D^c} (\boldsymbol{\kappa}(\boldsymbol{\beta}))^2 p(y) dy \leqslant \left(\int_{D^c} (\kappa(\boldsymbol{\beta}))^{\ell} p(y) dy\right)^{2/\ell} \mathbb{P}(Y \in D^c)^{1-2/\ell},$$

which is bounded by

$$\sup_{\|\boldsymbol{\beta}\|=1} \mathbb{E}[|\langle \boldsymbol{\kappa}(Y), \boldsymbol{\beta}\rangle|^{\ell}]^{2/\ell}\epsilon_1 = \frac{1}{\tau}\lambda_{\min}^{+}(\mathrm{Cov}(\boldsymbol{\kappa}(Y))).$$

Then

$$\frac{1}{H}\sum_{h:U_h\in\mathcal{B}_H} \mathrm{var}(\boldsymbol{\kappa}(\boldsymbol{\beta}) \mid Y \in U_h) = \frac{1}{H}\sum_{h:U_h\cap D\neq\emptyset} \mathrm{var}(\boldsymbol{\kappa}(\boldsymbol{\beta}) \mid Y \in U_h) + \frac{1}{H}\sum_{h:U_h\subset D^c} \mathrm{var}(\boldsymbol{\kappa}(\boldsymbol{\beta}) \mid Y \in U_h)$$
$$\leqslant \frac{1}{H\tau}N_1\lambda_{\min}^{+}(\mathrm{Cov}(\boldsymbol{\kappa}(Y))) + \frac{1}{\tau}\lambda_{\min}^{+}(\mathrm{Cov}(\boldsymbol{\kappa}(Y))) \leqslant \frac{2}{\tau}\lambda_{\min}^{+}(\mathrm{Cov}(\boldsymbol{\kappa}(Y)))$$

where N_1 is the number of U_h such that $U_h \cap D \neq \emptyset$. This completes the proof of (18).

Note that for any $\boldsymbol{\beta} \in \mathrm{col}(\mathrm{Cov}(\boldsymbol{\kappa}(Y)))$, $\lambda_{\min}^{+}(\mathrm{Cov}(\boldsymbol{\kappa}(Y))) \leqslant \boldsymbol{\beta}^{\top}\mathrm{Cov}(\boldsymbol{\kappa}(Y))\boldsymbol{\beta}$. In the case when $\mathrm{var}(\boldsymbol{\beta}^{\top}\boldsymbol{\kappa}(Y)) = 0$, it holds that $\mathrm{var}(\boldsymbol{\beta}^{\top}\boldsymbol{\kappa}(Y)|Y \in U_h) = 0(\forall U_h)$. Thus the proof is completed.

□

B.2 Proof of Theorem 1

Proof. Note that $\langle \boldsymbol{m}(y), \boldsymbol{\beta}\rangle = \mathbb{E}[\langle \boldsymbol{X}, \boldsymbol{\beta}\rangle|Y]$. By Jensen's inequality for conditional expectation, one has

$$\mathbb{E}[|\mathbb{E}[\langle \boldsymbol{X}, \boldsymbol{\beta}\rangle|Y]|^{\ell}] \leqslant \mathbb{E}[\mathbb{E}[|\langle \boldsymbol{X}, \boldsymbol{\beta}\rangle|^{\ell}|Y]] = \mathbb{E}[|\langle \boldsymbol{X}, \boldsymbol{\beta}\rangle|^{\ell}] \leqslant c_1.$$

Then the proof is completed by Lemma 2. □

B.3 Proof of Corollary 1

It is a direct corollary of Theorem 1 by noticing that the following result.

Lemma 4 (Lemma 11 in Lin et al. [2018b]). *For any sufficiently large H, c and $n > \frac{4H}{\gamma} + 1$, $\mathfrak{S}_H(n)$ is a γ-partition with probability at least*

$$1 - CH^2\sqrt{n+1}\exp\left(-\frac{\gamma^2(n+1)}{32H^2}\right)$$

for some absolute constant C.

C Proofs of upper bounds with a known covariance matrix

Here we present the proofs of the upper bounds in Sections 4.2.2 and 4.3 with $\boldsymbol{\Sigma}$ known. These proofs are adapted from Lin et al. [2021]. Without loss of generality, we can assume $\boldsymbol{\Sigma} = \mathbf{I}_p$. The proof for general cases with unknown $\boldsymbol{\Sigma}$ is presented in Appendix D, which makes use of the results here.

C.1 Preliminary

Before we start proving the theorems, we need some preparations.

Notations: Suppose that we have $n = Hc$ samples $(\boldsymbol{X}_i, Y_i)$ from a distribution $\mathcal{M} \in \mathfrak{M}(p, d, \lambda)$. Throughout this section, H is taken to be an integer such that H satisfies the inequality $H > K \vee Cd$ in Lemma 1 and $H \leqslant H_0 d$ for some constant $H_0 > K_0 \vee C$. In this way, we can apply the result of Lemma 1 and we will implicitly use $H = O(d)$.

Since $\boldsymbol{\Sigma} = \mathbf{I}_p$, the SIR estimator in (5) is defined directly as $\widehat{\boldsymbol{B}} = \left[\widehat{\boldsymbol{B}}_1, ..., \widehat{\boldsymbol{B}}_d\right]$, where $\widehat{\boldsymbol{B}}_i$ is the i-th leading eigenvector of $\widehat{\boldsymbol{\Lambda}}_H$.

Let $\boldsymbol{B}_{\perp}$ be a $p \times (p-d)$ orthogonal matrix such that $\boldsymbol{B}^{\top}\boldsymbol{B}_{\perp} = 0$.

For any pair of $(\boldsymbol{X}, Y)$ sampled from $\mathcal{M}$, let $\boldsymbol{Z} = \boldsymbol{B}^\top \boldsymbol{X}$ and $\boldsymbol{E} = \boldsymbol{B}_\perp^\top \boldsymbol{X}$. Since $\boldsymbol{B}^\top \boldsymbol{B} = \mathbf{I}_d$, $\boldsymbol{B}_\perp^\top \boldsymbol{B}_\perp = \mathbf{I}_{p-d}$, one has $\boldsymbol{Z} \sim N(0, \boldsymbol{I}_d)$ and $\boldsymbol{E} \sim N(0, \boldsymbol{I}_{p-d})$. Furthermore, $\boldsymbol{Z} \perp\!\!\!\perp \boldsymbol{E}$ since $\mathrm{Cov}(\boldsymbol{Z}, \boldsymbol{E}) = \boldsymbol{B}^\top \boldsymbol{B}_\perp = 0$. Besides, we have

$$\boldsymbol{X} = P_{\mathcal{S}} \boldsymbol{X} + P_{\mathcal{S}^\perp} \boldsymbol{X} = \boldsymbol{B}\boldsymbol{Z} + \boldsymbol{B}_\perp \boldsymbol{E} \quad (\because \boldsymbol{B}\boldsymbol{B}^\top + \boldsymbol{B}_\perp \boldsymbol{B}_\perp^\top = \mathbf{I}_p)$$

where $\mathcal{S} = \mathrm{col}(\boldsymbol{B})$ is the central space.

Let $\boldsymbol{V} = \boldsymbol{B}\boldsymbol{Z}$ and $\boldsymbol{W} = \boldsymbol{B}_\perp \boldsymbol{E}$. Then $\boldsymbol{V}^\top \boldsymbol{W} = 0$. We introduce the notation $\overline{\boldsymbol{V}}_{h,\cdot}$, $\overline{\boldsymbol{Z}}_{h,\cdot}$, $\overline{\boldsymbol{W}}_{h,\cdot}$, and $\overline{\boldsymbol{E}}_{h,\cdot}$ similar to the definition of the sample mean in the h-th slice $\overline{\boldsymbol{X}}_{h,\cdot}$ near Equation (4).

Let $\mathcal{V} = \frac{1}{\sqrt{H}}\left[\overline{\boldsymbol{V}}_{1,\cdot}\ ,\ \overline{\boldsymbol{V}}_{2,\cdot}, ...,\ \overline{\boldsymbol{V}}_{H,\cdot}\right]$, $\mathcal{Z} = \frac{1}{\sqrt{H}}\left[\overline{\boldsymbol{Z}}_{1,\cdot}\ ,\ \overline{\boldsymbol{Z}}_{2,\cdot}, ...,\ \overline{\boldsymbol{Z}}_{H,\cdot}\right]$, $\mathcal{W} = \frac{1}{\sqrt{H}}\left[\overline{\boldsymbol{W}}_{1,\cdot}\ ,\ \overline{\boldsymbol{W}}_{2,\cdot}, ...,\ \overline{\boldsymbol{W}}_{H,\cdot}\right]$, and $\mathcal{E} = \frac{1}{\sqrt{H}}\left[\overline{\boldsymbol{E}}_{1,\cdot}\ ,\ \overline{\boldsymbol{E}}_{2,\cdot}, ...,\ \overline{\boldsymbol{E}}_{H,\cdot}\right]$ be four matrices formed by the p-dimensional vectors $\frac{1}{\sqrt{H}}\overline{\boldsymbol{V}}_{h,\cdot}$, $\frac{1}{\sqrt{H}}\overline{\boldsymbol{Z}}_{h,\cdot}$, $\frac{1}{\sqrt{H}}\overline{\boldsymbol{W}}_{h,\cdot}$, and $\frac{1}{\sqrt{H}}\overline{\boldsymbol{E}}_{h,\cdot}$. We have $\mathcal{V} = \boldsymbol{B}\mathcal{Z}$, $\mathcal{W} = \boldsymbol{B}_\perp \mathcal{E}$, and $\mathcal{V}^\top \mathcal{W} = 0$.

Define $\widehat{\boldsymbol{\Lambda}}_z = \mathcal{Z}\mathcal{Z}^\top$ and $\widehat{\boldsymbol{\Lambda}}_V = \mathcal{V}\mathcal{V}^\top = \boldsymbol{B}\widehat{\boldsymbol{\Lambda}}_z \boldsymbol{B}^\top$. Then we have the following decomposition

$$\begin{aligned}\widehat{\boldsymbol{\Lambda}}_H &= \mathcal{V}\mathcal{V}^\top + \mathcal{V}\mathcal{W}^\top + \mathcal{W}\mathcal{V}^\top + \mathcal{W}\mathcal{W}^\top \\ &= \widehat{\boldsymbol{\Lambda}}_V + \boldsymbol{B}\mathcal{Z}\mathcal{E}^\top \boldsymbol{B}_\perp^\top + \boldsymbol{B}_\perp \mathcal{E}\mathcal{Z}^\top \boldsymbol{B}^\top + \boldsymbol{B}_\perp \mathcal{E}\mathcal{E}^\top \boldsymbol{B}_\perp^\top.\end{aligned} \tag{19}$$

Since $\boldsymbol{E} \sim N(0, \mathbf{I}_{p-d})$ and is independent of Y, we know that the entries $\mathcal{E}_{i,j}$ of $\mathcal{E}$ are *i.i.d.* samples of $N(0, \frac{1}{n})$.

C.2 Proof of Theorem 5

First, we have the following lemma.

Lemma 5. *Assume that $f \in \mathcal{F}_d(\lambda, \kappa, K)$ in Definition 3 and $H \geqslant \max\{K, Cd\}$ for a sufficiently large constant C. We have the following statements.*

(a)

$$\mathbb{P}\left(\|\mathcal{W}\mathcal{W}^\top\| > 6\frac{p \vee H + t}{n}\right) \leqslant 2\exp\left(-t\right).$$

(b) For $\nu \in (\kappa, 2\kappa]$,

$$\mathbb{P}\left(\exists \boldsymbol{\beta} \in \mathbb{S}^{p-1},\ \textit{s.t.}\ \left|\boldsymbol{\beta}^\top \left(\widehat{\boldsymbol{\Lambda}}_V - \boldsymbol{\Lambda}\right)\boldsymbol{\beta}\right| > \frac{2}{3\nu}\boldsymbol{\beta}^\top \boldsymbol{\Lambda}\boldsymbol{\beta}\right) \leqslant C_1 \exp\left(-C_2 \frac{n\lambda}{H^2\nu^2} + C_3 \log(nH) + C_4 d\right).$$

Proof. (a): We apply Lemma 27 to $\sqrt{n}\cdot\mathcal{E}$ and note that

$$\left(\sqrt{p-d} + \sqrt{H} + \sqrt{2t}\right)^2 \leqslant 3\left(p - d + H + 2t\right) \leqslant 6(p \vee H + t).$$

(b): Since $\mathrm{col}(\widehat{\boldsymbol{\Lambda}}_V) = \mathrm{col}(\boldsymbol{\Lambda}) = \mathrm{col}(\boldsymbol{B})$, we only need to consider the vector $\boldsymbol{\beta}$ that lies in $\mathrm{col}(\boldsymbol{\Lambda})$. Let $\boldsymbol{\Lambda} = \boldsymbol{V}\boldsymbol{D}\boldsymbol{V}^\top$ be the eigen-decomposition of $\boldsymbol{\Lambda}$ where $\boldsymbol{V}$ is a $p \times d$ orthogonal matrix and $\boldsymbol{D}$ is a $d \times d$ invertible diagonal matrix. Let $\boldsymbol{\Omega} := \boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{V}^\top(\widehat{\boldsymbol{\Lambda}}_H - \boldsymbol{\Lambda})\boldsymbol{V}\boldsymbol{D}^{-\frac{1}{2}}$. For any unit vector $\boldsymbol{\beta} \in \mathrm{col}(\boldsymbol{\Lambda})$, consider the transformed vector $\boldsymbol{U} = \boldsymbol{D}^{1/2}\boldsymbol{V}^\top\boldsymbol{\beta}$. Since $\mathrm{col}(\boldsymbol{B}) = \mathrm{col}(\boldsymbol{\Lambda})$, one has $\boldsymbol{B}_\perp^\top \boldsymbol{\beta} = 0$. Then from (19) we turn to prove that

$$\begin{aligned}&\mathbb{P}\left(\exists \boldsymbol{\beta} \in \mathbb{S}^{p-1},\ \text{s.t.}\ \left|\boldsymbol{\beta}^\top \left(\widehat{\boldsymbol{\Lambda}}_H - \boldsymbol{\Lambda}\right)\boldsymbol{\beta}\right| > \frac{2}{3\nu}\boldsymbol{\beta}^\top \boldsymbol{\Lambda}\boldsymbol{\beta}\right) \\ =&\mathbb{P}\left(\exists \boldsymbol{U} \in \mathbb{R}^p,\ \text{s.t.}\ \left|\boldsymbol{U}^\top \boldsymbol{\Omega}\boldsymbol{U}\right| > \frac{2}{3\nu}\boldsymbol{U}^\top \boldsymbol{U}\right) \\ \leqslant& C_1 \exp\left(-C_2 \frac{n\lambda}{H^2\nu^2} + C_3 \log(nH) + C_4 d\right).\end{aligned}$$

Lemma 1 yields

$$\mathbb{P}\left(\left|\boldsymbol{U}^\top \boldsymbol{\Omega} \boldsymbol{U}\right| > \frac{1}{2\nu}\boldsymbol{U}^\top \boldsymbol{U}\right) \leqslant C_1 \exp\left(-C_2 \frac{n\lambda}{H^2\nu^2} + C_3 \log(nH)\right), \tag{20}$$

where we have used $\boldsymbol{U}^\top \boldsymbol{U} = \boldsymbol{\beta}^\top \boldsymbol{\Lambda} \boldsymbol{\beta} \geqslant \lambda$. We then use the standard ϵ-net argument (see, e.g., [Tao, 2012, Chapter 2.3.1]) to bound

$$\mathbb{P}\left(\exists \boldsymbol{U} \in \mathbb{R}^p, \text{ s.t. } \left|\boldsymbol{U}^\top \boldsymbol{\Omega} \boldsymbol{U}\right| > \frac{2}{3\nu}\boldsymbol{U}^\top \boldsymbol{U}\right) = \mathbb{P}\left(\|\boldsymbol{\Omega}\| > \frac{2}{3\nu}\right).$$

Let $\mathcal{N}$ be a $\frac{1}{8}$-net in $\mathbb{S}^{d-1}$: a minimal set of points in $\mathbb{S}^{d-1}$ such that for any $\boldsymbol{u} \in \mathbb{S}^{d-1}$, one can find find $\widetilde{\boldsymbol{u}} \in \mathcal{N}$ such that $\|\boldsymbol{u} - \widetilde{\boldsymbol{u}}\| \leqslant 1/8$. This implies that $\boldsymbol{u}^\top \boldsymbol{\Omega} \boldsymbol{u} = \widetilde{\boldsymbol{u}}^\top \boldsymbol{\Omega} \widetilde{\boldsymbol{u}} + (\boldsymbol{u} - \widetilde{\boldsymbol{u}})^\top \boldsymbol{\Omega} \widetilde{\boldsymbol{u}} + \boldsymbol{u}^\top \boldsymbol{\Omega}(\boldsymbol{u} - \widetilde{\boldsymbol{u}}) \leqslant \widetilde{\boldsymbol{u}}^\top \boldsymbol{\Omega} \widetilde{\boldsymbol{u}} + \|\boldsymbol{\Omega}\|/4$. Taking the maximum of $\boldsymbol{u} \in \mathbb{S}^{d-1}$, one has $\|\boldsymbol{\Omega}\| \leqslant 4/3 \cdot \max_{\widetilde{\boldsymbol{u}} \in \mathcal{N}} \widetilde{\boldsymbol{u}}^\top \boldsymbol{\Omega} \widetilde{\boldsymbol{u}}$. Therefore

$$\begin{aligned}\mathbb{P}\left(\|\boldsymbol{\Omega}\| > \frac{2}{3\nu}\right) &\leqslant \mathbb{P}\left(\max_{\widetilde{\boldsymbol{u}} \in \mathcal{N}} \widetilde{\boldsymbol{u}}^\top \boldsymbol{\Omega} \widetilde{\boldsymbol{u}} > \frac{1}{2\nu}\right) \leqslant \sum_{\widetilde{\boldsymbol{u}} \in \mathcal{N}} \mathbb{P}\left(\left|\widetilde{\boldsymbol{u}}^\top \boldsymbol{\Omega} \widetilde{\boldsymbol{u}}\right| > \frac{1}{2\nu}\right) \\ &\leqslant C_1 \exp\left(-C_2 \frac{n\lambda}{H^2\nu^2} + C_3 \log(nH) + C_4 d\right),\end{aligned}$$

where in the last inequality we use the the fact that $|\mathcal{N}| \leqslant 17^d$ (See e.g., Lemma 5.2 in Vershynin [2010]) and insert $\boldsymbol{\beta} = \boldsymbol{V}\boldsymbol{D}^{-1/2}\widetilde{\boldsymbol{u}}/\|\boldsymbol{V}\boldsymbol{D}^{-1/2}\widetilde{\boldsymbol{u}}\|$ and $\boldsymbol{U} = \boldsymbol{D}^{1/2}\boldsymbol{V}^\top \boldsymbol{\beta}$ into Equation (20). □

To proceed, we define some events: $\mathtt{E}_1 = \left\{ \|\mathcal{W}\mathcal{W}^\top\| \leqslant 6\frac{p \vee H + \log(n\lambda)}{n} \right\}$, $\mathtt{E}_2 = \left\{ \|\widehat{\boldsymbol{\Lambda}}_V - \boldsymbol{\Lambda}\| \leqslant \frac{2}{3\nu}\kappa\lambda \right\}$ and $\mathtt{E} = \mathtt{E}_1 \cap \mathtt{E}_2$.

Corollary 2. For any $\nu \in (\kappa, 2\kappa]$, we can find constants C and $\widetilde{C}$, such that $\mathbb{P}(\mathtt{E}^c) \leqslant \frac{\widetilde{C}}{n\lambda}$ holds if

$$\kappa^2 H^2 \left(\log(nH\kappa) + d\right) < Cn\lambda. \tag{21}$$

If further $\kappa\left(p \vee H + \log(n\lambda)\right) < 1800^{-1} n\lambda$, then on the event $\mathtt{E}$, the followings hold

a) $\frac{1}{3}\lambda \leqslant \lambda_d(\widehat{\boldsymbol{\Lambda}}_V) \leqslant \lambda_1(\widehat{\boldsymbol{\Lambda}}_V) \leqslant 2\kappa\lambda$.

b) $\|\widehat{\boldsymbol{\Lambda}}_H - \widehat{\boldsymbol{\Lambda}}_V\| \leqslant \lambda\sqrt{18\kappa\frac{p \vee H + \log(n\lambda)}{n\lambda}} < \frac{1}{4}\lambda$.

c) $\lambda_{d+1}(\widehat{\boldsymbol{\Lambda}}_H) < \frac{1}{4}\lambda$.

Proof. From Lemma 5, one has

$$\begin{aligned}\mathbb{P}(\mathtt{E}_1^c) &= \mathbb{P}\left(\|\mathcal{W}\mathcal{W}^\top\| > 6\frac{p \vee H + \log(n\lambda)}{n}\right) \leqslant 2\exp\left(-\log(n\lambda)\right) = \frac{2}{n\lambda} \\ \mathbb{P}(\mathtt{E}_2^c) &= \mathbb{P}\left(\exists \boldsymbol{\beta} \in \mathbb{S}^{p-1}, \text{ s.t. } \left|\boldsymbol{\beta}^\top \left(\widehat{\boldsymbol{\Lambda}}_V - \boldsymbol{\Lambda}\right)\boldsymbol{\beta}\right| > \frac{2}{3\nu}\kappa\lambda\right) \\ &\leqslant \mathbb{P}\left(\exists \boldsymbol{\beta}, \text{ s.t. } \left|\boldsymbol{\beta}^\top \left(\widehat{\boldsymbol{\Lambda}}_V - \boldsymbol{\Lambda}\right)\boldsymbol{\beta}\right| > \frac{2}{3\nu}\boldsymbol{\beta}^\top \boldsymbol{\Lambda}\boldsymbol{\beta}\right) \leqslant C_1 \exp\left(-C_2 \frac{n\lambda}{H^2\nu^2} + C_3 \log(nH) + C_4 d\right).\end{aligned}$$

Thus to show

$$\mathbb{P}(\mathtt{E}^c) = \mathbb{P}(\mathtt{E}_1^c \cup \mathtt{E}_2^c) \leqslant \mathbb{P}(\mathtt{E}_1^c) + \mathbb{P}(\mathtt{E}_2^c) \leqslant \frac{\widetilde{C}}{n\lambda}$$

for some $\widetilde{C}$, one only need $\exp\left(-C_2 \frac{n\lambda}{H^2\nu^2} + C_3 \log(nH) + C_4 d\right) \lesssim \frac{1}{n\lambda}$. This will be true if the followings are bounded from below by some positive constant

$$\frac{n\lambda}{H^2\nu^2}/\log(nH), \quad \frac{n\lambda}{H^2\nu^2}/d, \quad \frac{n\lambda}{H^2\nu^2}/\log(n\lambda).$$

Since $\nu \leqslant 2\kappa$, the first two are bounded by choosing a small C in Equation (21). Since $x/\log(x)$ is increasing for $x > e$, one has

$$n\lambda/\log(n\lambda) > C^{-1}H^2\kappa^2(\log(nH\kappa)+d)/\log\left[C^{-1}H^2\kappa^2(\log(nH\kappa)+d)\right] \gtrsim H^2\nu^2$$

and the last one is also bounded.

On the event $\mathtt{E}_2$, Weyl's inequality implies that $\lambda_d(\widehat{\boldsymbol{\Lambda}}_V) \geqslant \lambda_d(\boldsymbol{\Lambda}) - \frac{2}{3\nu}\kappa\lambda > \frac{1}{3}\lambda$ and $\lambda_1(\widehat{\boldsymbol{\Lambda}}_V) \leqslant \lambda_1(\boldsymbol{\Lambda}) + 2\kappa\lambda/(3\nu) \leqslant 2\kappa\lambda$.

From Equation (19),

$$\begin{aligned}\|\widehat{\boldsymbol{\Lambda}}_H - \widehat{\boldsymbol{\Lambda}}_V\| &\leqslant \|\mathcal{V}\mathcal{W}^\top\| + \|\mathcal{W}\mathcal{V}^\top\| + \|\mathcal{W}\mathcal{W}^\top\| \\ &\leqslant 2\sqrt{\|\widehat{\boldsymbol{\Lambda}}_V\|\|\mathcal{W}\mathcal{W}^\top\|} + \|\mathcal{W}\mathcal{W}^\top\|.\end{aligned} \tag{22}$$

On the event $\mathtt{E}$, if $6\frac{p\vee H+\log(n\lambda)}{n\lambda} \leqslant 2\kappa$, the last display is further bounded by $\sqrt{6\cdot 3^2\cdot 2\kappa\lambda\frac{p\vee H+\log(n\lambda)}{n}}$. If $\kappa\frac{p\vee H+\log(n\lambda)}{n\lambda} < 2^{-6}3^{-3}$, the bound is smaller than $\frac{1}{4}\lambda$. By Lemma 26 (Weyl's inequality) and $\lambda_{d+1}(\widehat{\boldsymbol{\Lambda}}_V) = 0$, one has $\lambda_{d+1}(\widehat{\boldsymbol{\Lambda}}_H) < \frac{1}{4}\lambda$. □

Now we start the proof of Theorem 5. Throughout the proof C is a constant independent of(p, d, H, n, λ) whose value may very from line to line. Note that

$$\begin{aligned}&\mathbb{E}\|\widehat{\boldsymbol{B}}\widehat{\boldsymbol{B}}^\top - \boldsymbol{B}\boldsymbol{B}^\top\|_F^2 \\ &\quad= \underbrace{\mathbb{E}\|\widehat{\boldsymbol{B}}\widehat{\boldsymbol{B}}^\top - \boldsymbol{B}\boldsymbol{B}^\top\|_F^2\mathbf{1}_{\mathtt{E}^c}}_{I} + \underbrace{\mathbb{E}\|\widehat{\boldsymbol{B}}\widehat{\boldsymbol{B}}^\top - \boldsymbol{B}\boldsymbol{B}^\top\|_F^2\mathbf{1}_{\mathtt{E}}}_{II}.\end{aligned}$$

For I: By the triangle inequality and the fact that the Frobenius norm of a projection matrix equals to its rank, we have

$$I \leqslant 2d\mathbb{P}(\mathtt{E}^c) \leqslant C\frac{d}{n\lambda}. \tag{23}$$

For II: Let $\widehat{\boldsymbol{\Lambda}}_V = \widetilde{\boldsymbol{B}}\boldsymbol{D}_H\widetilde{\boldsymbol{B}}^\top$ be the eigen-decomposition of $\widehat{\boldsymbol{\Lambda}}_V$, where $\widetilde{\boldsymbol{B}}$ is a $p\times d$ orthogonal matrix and $\boldsymbol{D}_H$ is a $d\times d$ diagonal matrix. Note that $\widetilde{\boldsymbol{B}}$ and $\boldsymbol{B}$ are sharing the same column space (i.e., $\widetilde{\boldsymbol{B}}\widetilde{\boldsymbol{B}}^\top = \boldsymbol{B}\boldsymbol{B}^\top$) since $\mathrm{col}(\boldsymbol{\Lambda}) = \mathrm{col}(\widehat{\boldsymbol{\Lambda}}_V)$.

Corollary 2 states that on $\mathtt{E}$, one has $\lambda_d(\widehat{\boldsymbol{\Lambda}}_V) = \lambda_d(\boldsymbol{D}_H) \geqslant \frac{\lambda}{3}$, $\|\widehat{\boldsymbol{\Lambda}}_V\| \leqslant 2\kappa\lambda$, and $\lambda_{d+1}(\widehat{\boldsymbol{\Lambda}}_H) \leqslant \frac{1}{4}\lambda$.

Let $\boldsymbol{Q} = \widehat{\boldsymbol{\Lambda}}_H - \widehat{\boldsymbol{\Lambda}}_V$ and let $\widehat{\boldsymbol{B}}_\perp$ be a $p\times(p-d)$ orthogonal matrix whose columns are the last $(p-d)$ eigenvectors of $\widehat{\boldsymbol{\Lambda}}_H$. Applying the Sin-Theta theorem (e.g., Lemma 28) to the pair of symmetric matrices $(\widehat{\boldsymbol{\Lambda}}_V, \widehat{\boldsymbol{\Lambda}}_H)$, one has

$$\begin{aligned}II =&\,\mathbb{E}\|\boldsymbol{B}\boldsymbol{B}^\top - \widehat{\boldsymbol{B}}\widehat{\boldsymbol{B}}^\top\|_F^2\mathbf{1}_{\mathtt{E}} = \mathbb{E}\|\widetilde{\boldsymbol{B}}\widetilde{\boldsymbol{B}}^\top - \widehat{\boldsymbol{B}}\widehat{\boldsymbol{B}}^\top\|_F^2\mathbf{1}_{\mathtt{E}} \\ \leqslant&\,\frac{288}{\lambda^2}\min\left(\mathbb{E}\|\widetilde{\boldsymbol{B}}_\perp^\top\boldsymbol{Q}\widehat{\boldsymbol{B}}\|_F^2\mathbf{1}_{\mathtt{E}}, \mathbb{E}\|\widetilde{\boldsymbol{B}}^\top\boldsymbol{Q}\widehat{\boldsymbol{B}}_\perp\|_F^2\mathbf{1}_{\mathtt{E}}\right) \\ \leqslant&\,\frac{288}{\lambda^2}\min\left(\mathbb{E}\|\widetilde{\boldsymbol{B}}_\perp^\top\boldsymbol{Q}\|_F^2\mathbf{1}_{\mathtt{E}}, \mathbb{E}\|\widetilde{\boldsymbol{B}}^\top\boldsymbol{Q}\|_F^2\mathbf{1}_{\mathtt{E}}\right) \quad (\text{Lemma 25 and } \|\widetilde{\boldsymbol{B}}\| \leqslant 1, \|\widehat{\boldsymbol{B}}_\perp\| \leqslant 1) \\ \leqslant&\,\frac{288}{\lambda^2}\mathbb{E}\|\widetilde{\boldsymbol{B}}^\top\boldsymbol{Q}\|_F^2\mathbf{1}_{\mathtt{E}}\end{aligned}$$

Since $\widetilde{\boldsymbol{B}}$ and $\boldsymbol{B}$ share the same column space, one has $\widetilde{\boldsymbol{B}}^\top\mathcal{W} = \mathbf{0}$. Thus, one has

$$\widetilde{\boldsymbol{B}}^\top\boldsymbol{Q} = \widetilde{\boldsymbol{B}}^\top\mathcal{V}\mathcal{W}^\top.$$

By Lemma 25 and note that $\|\widetilde{\boldsymbol{B}}^\top\mathcal{V}\|^2\mathbf{1}_{\mathtt{E}} \leqslant \|\widehat{\boldsymbol{\Lambda}}_V\|\mathbf{1}_{\mathtt{E}} \leqslant 2\kappa\lambda$, one has

$$\mathbb{E}\|\widetilde{\boldsymbol{B}}^\top\mathcal{V}\mathcal{W}^\top\|_F^2\mathbf{1}_{\mathtt{E}} \leqslant 2\kappa\lambda\mathbb{E}\|\mathcal{W}^\top\|_F^2 \leqslant \frac{2\kappa\lambda}{n}H(p-d) \tag{24}$$

where in the last inequality we apply Lemma 24 to $\sqrt{n}\cdot\mathcal{E}$. Since κ is assumed to be fixed, one has

$$\begin{aligned} II &\leqslant \frac{576\kappa}{n\lambda}H(p-d) \\ &\leqslant C''\frac{H(p-d)}{n\lambda}. \end{aligned} \tag{25}$$

Combining (23) and (25), we conclude that

$$\sup_{\mathcal{M}\in\mathfrak{M}(p,d,\kappa,\lambda)} \mathbb{E}\|\widehat{\boldsymbol{B}}\widehat{\boldsymbol{B}}^\top - \boldsymbol{B}\boldsymbol{B}^\top\|_F^2 \lesssim \frac{d+H(p-d)}{n\lambda} \lesssim \frac{dp}{n\lambda}.$$

C.3 Proof of Theorem 8

Preliminaries

Since we have assumed that $\boldsymbol{\Sigma}=\mathbf{I}_p$ in this section, the two-fold estimator $\widehat{\boldsymbol{B}}$ defined near Theorem 8 can be simplified. Specifically, we first divide the samples into two equal sets of samples and have the corresponding decomposition (19)

$$\begin{aligned} \widehat{\boldsymbol{\Lambda}}_H &= \mathcal{V}\mathcal{V}^\top + \mathcal{V}\mathcal{W}^\top + \mathcal{W}\mathcal{V}^\top + \mathcal{W}\mathcal{W}^\top \\ &= \widehat{\boldsymbol{\Lambda}}_V + \boldsymbol{B}\mathcal{Z}\mathcal{E}^\top\boldsymbol{B}_\perp^\top + \boldsymbol{B}_\perp\mathcal{E}\mathcal{Z}^\top\boldsymbol{B}^\top + \boldsymbol{B}_\perp\mathcal{E}\mathcal{E}^\top\boldsymbol{B}_\perp^\top. \end{aligned}$$

for these two sets of samples. That is, for $i=1,2$, we define $\boldsymbol{\Lambda}_H^{(i)}$, $\boldsymbol{\Lambda}_V^{(i)}$, $\mathcal{Z}^{(i)},\mathcal{W}^{(i)}$ $\mathcal{V}^{(i)},\widehat{\boldsymbol{\Lambda}}_z^{(i)}$ and $\mathcal{E}^{(i)}$ for the first and second set of samples respectively according to the decomposition (19). Then the two-fold aggregation estimator $\widehat{\boldsymbol{B}}$ can be defined as:

Two-fold Aggregation Estimator with identity covariance:

(i) For each $L\in\mathcal{L}(s)$ (the set of all subsets of $[p]$ with size s), let

$$\begin{aligned} &\widehat{\boldsymbol{B}}_L := \arg\max_{\boldsymbol{B}} \quad \operatorname{Tr}(\boldsymbol{B}^\top\boldsymbol{\Lambda}_H^{(1)}\boldsymbol{B}) \\ &\text{s.t. } \boldsymbol{B}^\top\boldsymbol{B}=\mathbf{I}_d \text{ and } \operatorname{supp}(\boldsymbol{B})\subset L \end{aligned} \tag{26}$$

(ii) Our aggregation estimator $\widehat{\boldsymbol{B}}$ is defined to be $\widehat{\boldsymbol{B}}_{L^*}$ where

$$L^* := \arg\max_{L\in\mathcal{L}(s)} \quad \operatorname{Tr}(\widehat{\boldsymbol{B}}_L^\top\boldsymbol{\Lambda}_H^{(2)}\widehat{\boldsymbol{B}}_L).$$

In addition, we introduce an ‘Oracle estimator’ $\widehat{\boldsymbol{B}}_O$ (as if we know the support of $\boldsymbol{B}$).

Oracle Estimator:

$$\begin{aligned} &\widehat{\boldsymbol{B}}_O := \arg\max_{\boldsymbol{B}}\langle\boldsymbol{\Lambda}_H^{(1)},\boldsymbol{B}\boldsymbol{B}^\top\rangle = \arg\max_{\boldsymbol{B}}\operatorname{Tr}(\boldsymbol{B}^\top\boldsymbol{\Lambda}_H^{(1)}\boldsymbol{B}) \\ &\text{s.t. } \boldsymbol{B}^\top\boldsymbol{B}=\mathbf{I}_d \text{ and } \mathit{supp}(\boldsymbol{B})=S. \end{aligned} \tag{27}$$

Let us first introduce some notations. For $i=1,2$, let $\boldsymbol{\Lambda}_V^{(i)}=\boldsymbol{B}^{(i)}\boldsymbol{D}^{(i)}\boldsymbol{B}^{(i),\top}$ where $\boldsymbol{B}^{(i)}$ is $p\times d$ orthogonal matrix and $\boldsymbol{D}^{(i)}:=\{\lambda_1^{(i)},\dots,\lambda_d^{(i)}\}$ is a diagonal matrix. For any subset S of $[p]$, let $\boldsymbol{J}_S$ be the diagonal matrix such that $\boldsymbol{J}_S(i,i)=1$ if $i\in[S]$ and $\boldsymbol{J}_S(i,i)=0$ otherwise.

For $i=1,2$, let $\mathtt{E}_2^{(i)}$ be the event defined similarly as $\mathtt{E}_2$ (which is introduced near Corollary 2). Let $\bar{\mathtt{E}}_2=\mathtt{E}_2^{(1)}\cap\mathtt{E}_2^{(2)}$ and $\boldsymbol{Q}_S=\boldsymbol{J}_S\left(\boldsymbol{\Lambda}_H^{(1)}-\boldsymbol{\Lambda}_V^{(1)}\right)\boldsymbol{J}_S$.

Let $\mathtt{F}$ consist of the events such that $\|\boldsymbol{J}_S\mathcal{W}^{(1)}\mathcal{W}^{(1),\top}\boldsymbol{J}_S\|\leqslant 6\frac{s\vee H+\log(n\lambda)}{n}$ and define $\mathtt{E}:=\bar{\mathtt{E}}_2\cap\mathtt{F}$. Following the reasoning of Corollary 2, if $\nu\in(\kappa,2\kappa]$, $\kappa^2H^2\left(\log(nH)+\log\kappa+d\right)/(n\lambda)$ and $\kappa\left(s\vee H+\log(n\lambda)\right)/(n\lambda)$ are sufficiently small, one has $\mathbb{P}\left(\mathtt{E}^c\right)\leqslant\frac{C}{n\lambda}$ and the followings hold on $\mathtt{E}$:

1.

$$\frac{\lambda}{3} \leqslant \lambda_d^{(i)} \leqslant \ldots \leqslant \lambda_1^{(i)} \leqslant 2\kappa\lambda \tag{28}$$

2. By Weyl's inequality, one has

$$\|\boldsymbol{Q}_S\| < \frac{1}{4}\lambda, \quad \lambda_{d+1}\left(\boldsymbol{J}_S \boldsymbol{\Lambda}_H^{(1)} \boldsymbol{J}_S\right) \leqslant \frac{\lambda}{4}. \tag{29}$$

Let $\widehat{\boldsymbol{B}}_O^\top \boldsymbol{B} = \boldsymbol{U}_1 \Delta \boldsymbol{U}_2^\top$ be the singular value decomposition of $\widehat{\boldsymbol{B}}_O^\top \boldsymbol{B}$ such that the entries of Δ are non-negative and $\boldsymbol{M} := \boldsymbol{U}_2^\top \widehat{\boldsymbol{\Lambda}}_z^{(2)} \boldsymbol{U}_2$.

Main part of the proof

Now, we start our proof of Theorem 8. It is easy to verify that

$$\|\widehat{\boldsymbol{B}}\widehat{\boldsymbol{B}}^\top - \boldsymbol{B}\boldsymbol{B}^\top\|_F^2 \leqslant C\left(\|\widehat{\boldsymbol{B}}\widehat{\boldsymbol{B}}^\top - \widehat{\boldsymbol{B}}_O\widehat{\boldsymbol{B}}_O^\top\|_F^2 + \|\widehat{\boldsymbol{B}}_O\widehat{\boldsymbol{B}}_O^\top - \boldsymbol{B}\boldsymbol{B}^\top\|_F^2\right).$$

For the first term $\|\widehat{\boldsymbol{B}}\widehat{\boldsymbol{B}}^\top - \widehat{\boldsymbol{B}}_O\widehat{\boldsymbol{B}}_O^\top\|_F^2$, conditioning on $\mathtt{E}$, we know

$$\begin{aligned}
\|\widehat{\boldsymbol{B}}\widehat{\boldsymbol{B}}^\top - \widehat{\boldsymbol{B}}_O\widehat{\boldsymbol{B}}_O^\top\|_F^2 \leqslant & \frac{2}{\lambda_d(\widehat{\boldsymbol{\Lambda}}_z^{(2)})} \langle \widehat{\boldsymbol{B}}_O \boldsymbol{U}_1 \boldsymbol{M} \boldsymbol{U}_1^\top \widehat{\boldsymbol{B}}_O^\top, \widehat{\boldsymbol{B}}_O\widehat{\boldsymbol{B}}_O^\top - \widehat{\boldsymbol{B}}\widehat{\boldsymbol{B}}^\top \rangle & (30)\\
\leqslant & \frac{C}{\lambda} \langle \widehat{\boldsymbol{B}}_O \boldsymbol{U}_1 \boldsymbol{M} \boldsymbol{U}_1^\top \widehat{\boldsymbol{B}}_O^\top - \boldsymbol{\Lambda}_H^{(2)}, \widehat{\boldsymbol{B}}_O\widehat{\boldsymbol{B}}_O^\top - \widehat{\boldsymbol{B}}\widehat{\boldsymbol{B}}^\top \rangle & (31)\\
:= & I + II.
\end{aligned}$$

where

$$\begin{aligned}
I = & \frac{C}{\lambda} \langle \widehat{\boldsymbol{B}}_O \boldsymbol{U}_1 \boldsymbol{M} \boldsymbol{U}_1^\top \widehat{\boldsymbol{B}}_O^\top - \boldsymbol{\Lambda}_V^{(2)}, \widehat{\boldsymbol{B}}_O\widehat{\boldsymbol{B}}_O^\top - \widehat{\boldsymbol{B}}\widehat{\boldsymbol{B}}^\top \rangle\\
II = & \frac{C}{\lambda} \langle \boldsymbol{\Lambda}_V^{(2)} - \boldsymbol{\Lambda}_H^{(2)}, \widehat{\boldsymbol{B}}_O\widehat{\boldsymbol{B}}_O^\top - \widehat{\boldsymbol{B}}\widehat{\boldsymbol{B}}^\top \rangle.
\end{aligned}$$

Inequality (30) follows from applying Lemma 32 with the positive definite matrix $\boldsymbol{U}_1 \boldsymbol{M} \boldsymbol{U}_1^\top$. The inequality (31) follows from the definition of $\widehat{\boldsymbol{B}}$ and the fact that the eigenvalues of $\widehat{\boldsymbol{\Lambda}}_z^{(2)}$ are in $(\lambda/3, 2\kappa\lambda)$ (See fact 1). To simplify the notation, we let

$$\delta = \|\widehat{\boldsymbol{B}}_O\widehat{\boldsymbol{B}}_O^\top - \boldsymbol{B}\boldsymbol{B}^\top\|_F.$$

For I: First, $\widehat{\boldsymbol{B}}_O \boldsymbol{U}_1$ and $\boldsymbol{B}\boldsymbol{U}_2$ satisfy the condition that $\boldsymbol{U}_1^\top \widehat{\boldsymbol{B}}_O^\top \boldsymbol{B} \boldsymbol{U}_2 = \Delta$ is a diagonal matrix with non-negative entries. Second, the eigenvalues of $\widehat{\boldsymbol{\Lambda}}_z^{(2)} \in (\frac{1}{3}\lambda, 2\kappa\lambda)$ thus $\boldsymbol{M} := \boldsymbol{U}_2^\top \widehat{\boldsymbol{\Lambda}}_z^{(2)} \boldsymbol{U}_2$ has eigenvalues in $(\lambda/3, 2\kappa\lambda)$. By Lemma 31, there exists a constant C such that

$$\|\widehat{\boldsymbol{B}}_O \boldsymbol{U}_1 \boldsymbol{M} \boldsymbol{U}_1^\top \widehat{\boldsymbol{B}}_O^\top - \boldsymbol{\Lambda}_V^{(2)}\|_F \leqslant C\lambda \|\widehat{\boldsymbol{B}}_O\widehat{\boldsymbol{B}}_O^\top - \boldsymbol{B}\boldsymbol{B}^\top\|_F.$$

Thus, conditioning on $\mathtt{E}$, one has

$$\begin{aligned}
|I| \leqslant & C\|\widehat{\boldsymbol{B}}_O\widehat{\boldsymbol{B}}_O^\top - \boldsymbol{B}\boldsymbol{B}^\top\|_F \|\widehat{\boldsymbol{B}}_O\widehat{\boldsymbol{B}}_O^\top - \widehat{\boldsymbol{B}}\widehat{\boldsymbol{B}}^\top\|_F\\
= & C\delta \|\widehat{\boldsymbol{B}}_O\widehat{\boldsymbol{B}}_O^\top - \widehat{\boldsymbol{B}}\widehat{\boldsymbol{B}}^\top\|_F.
\end{aligned} \tag{32}$$

For II: Define $\boldsymbol{K}_L = \|\widehat{\boldsymbol{B}}_O\widehat{\boldsymbol{B}}_O^\top - \widehat{\boldsymbol{B}}_L\widehat{\boldsymbol{B}}_L^\top\|_F^{-1}\left(\widehat{\boldsymbol{B}}_O\widehat{\boldsymbol{B}}_O^\top - \widehat{\boldsymbol{B}}_L\widehat{\boldsymbol{B}}_L^\top\right)$. (For any $L \in \mathcal{L}(s)$, $\widehat{\boldsymbol{B}}_L$ is introduced in (26)). Then

$$|II| = \frac{C}{\lambda} \langle \boldsymbol{\Lambda}_V^{(2)} - \boldsymbol{\Lambda}_H^{(2)}, (\widehat{\boldsymbol{B}}_O\widehat{\boldsymbol{B}}_O^\top - \widehat{\boldsymbol{B}}\widehat{\boldsymbol{B}}^\top)\|\widehat{\boldsymbol{B}}_O\widehat{\boldsymbol{B}}_O^\top - \widehat{\boldsymbol{B}}\widehat{\boldsymbol{B}}^\top\|_F^{-1}\rangle \|\widehat{\boldsymbol{B}}_O\widehat{\boldsymbol{B}}_O^\top - \widehat{\boldsymbol{B}}\widehat{\boldsymbol{B}}^\top\|_F$$

$$\leqslant \frac{C}{\lambda} \max_{L\in\mathcal{L}(s)} \left|\langle \boldsymbol{\Lambda}_V^{(2)} - \boldsymbol{\Lambda}_H^{(2)}, \boldsymbol{K}_L\rangle\right| \|\widehat{\boldsymbol{B}}_O\widehat{\boldsymbol{B}}_O^\top - \widehat{\boldsymbol{B}}\widehat{\boldsymbol{B}}^\top\|_F$$

From the equation (19), one has

$$\left|II\right| \leqslant \frac{C}{\lambda}\|\widehat{\boldsymbol{B}}_O\widehat{\boldsymbol{B}}_O^\top - \widehat{\boldsymbol{B}}\widehat{\boldsymbol{B}}^\top\|_F \,(2T_2 + T_1) \tag{33}$$

where $T_1 = \max_{L\in\mathcal{L}(s)} \left|\left\langle \mathcal{W}^{(2)}\mathcal{W}^{(2),\top}, \boldsymbol{K}_L\right\rangle\right|$, $T_2 = \max_{L\in\mathcal{L}(s)} \left|\left\langle \mathcal{V}^{(2)}\mathcal{W}^{(2),\top}, \boldsymbol{K}_L\right\rangle\right|$. To summarize, conditioning on $\mathtt{E}$, one has

$$\|\widehat{\boldsymbol{B}}\widehat{\boldsymbol{B}}^\top - \widehat{\boldsymbol{B}}_O\widehat{\boldsymbol{B}}_O^\top\|_F \leqslant C\left(\delta + \frac{1}{\lambda}(2T_2 + T_1)\right). \tag{34}$$

Thus, one has

$$\begin{aligned}\|\widehat{\boldsymbol{B}}\widehat{\boldsymbol{B}}^\top - \boldsymbol{B}\boldsymbol{B}^\top\|_F^2 \mathbf{1}_{\mathtt{E}} &\leqslant C\left(\delta^2 + \|\widehat{\boldsymbol{B}}_O\widehat{\boldsymbol{B}}_O^\top - \widehat{\boldsymbol{B}}\widehat{\boldsymbol{B}}^\top\|_F^2\right)\mathbf{1}_{\mathtt{E}} \\ &\leqslant C\left(\delta^2 + C\left(\delta + \frac{1}{\lambda}\left(2T_1 + T_2\right)\right)^2\right)\mathbf{1}_{\mathtt{E}} \\ &\leqslant C\left(\delta^2 + \left(\frac{1}{\lambda}\left(2T_1 + T_2\right)\right)^2\right)\mathbf{1}_{\mathtt{E}}.\end{aligned}$$

Note that $\mathbb{P}(\mathtt{E}^c) \leqslant \frac{C}{n\lambda}$. If we can prove

$$\mathbb{E}\delta^2\mathbf{1}_{\mathtt{E}} \leqslant C\epsilon_n^2 \quad \text{ and } \mathbb{E}(2T_1 + T_2)^2\mathbf{1}_{\mathtt{E}} \leqslant \lambda^2\epsilon_n^2, \tag{35}$$

then one has $\mathbb{E}\|\widehat{\boldsymbol{B}}\widehat{\boldsymbol{B}}^\top - \boldsymbol{B}\boldsymbol{B}^\top\|_F^2\mathbf{1}_{\mathtt{E}} \leqslant C\epsilon_n^2$ and then

$$\mathbb{E}\left[\|\widehat{\boldsymbol{B}}\widehat{\boldsymbol{B}}^\top - \boldsymbol{B}\boldsymbol{B}^\top\|_F^2 \ 1_{\mathtt{E}^c}\right] \leqslant 2d\frac{C}{n\lambda} \lesssim \epsilon_n^2.$$

Therefore, we conclude that $\mathbb{E}\left[\|\widehat{\boldsymbol{B}}\widehat{\boldsymbol{B}}^\top - \boldsymbol{B}\boldsymbol{B}^\top\|_F^2\right] \lesssim \epsilon_n^2$. Thus, it is suffice to prove the following two lemmas.

Lemma 6. *If $n\lambda \leqslant e^{s\vee H}$, then*

$$\mathbb{E}\delta^2\mathbf{1}_{\mathtt{E}} \leqslant C\epsilon_n^2. \tag{36}$$

Proof. By (28) and (29), we know that the eigenvalues of $\boldsymbol{J}_S\boldsymbol{\Lambda}_V^{(1)}\boldsymbol{J}_S = \boldsymbol{\Lambda}_V^{(1)}$ is in $(\frac{1}{3}\lambda, 2\kappa\lambda)$ and the $(d+1)$-th largest eigenvalues of $\boldsymbol{J}_S\boldsymbol{\Lambda}_H^{(1)}\boldsymbol{J}_S$ is less than $\frac{\lambda}{4}$. Let $\widehat{\boldsymbol{B}}_O^\perp$ be a $p\times(p-d)$ orthogonal matrix whose columns are the last $(p-d)$ eigenvectors of $\boldsymbol{J}_S\boldsymbol{\Lambda}_H^{(1)}\boldsymbol{J}_S$. After applying the Sin-Theta Theorem (Lemma 28) to the pair of symmetric matrices $(\boldsymbol{\Lambda}_V^{(1)} = \boldsymbol{J}_S\boldsymbol{\Lambda}_V^{(1)}\boldsymbol{J}_S, \boldsymbol{J}_S\boldsymbol{\Lambda}_H^{(1)}\boldsymbol{J}_S)$, one has

$$\delta^2 = \|\widehat{\boldsymbol{B}}_O\widehat{\boldsymbol{B}}_O^\top - \boldsymbol{B}^{(1)}\boldsymbol{B}^{(1),\top}\|_F \leqslant \frac{C}{\lambda^2}\|\widehat{\boldsymbol{B}}_O^{\perp,\top}\boldsymbol{Q}_S\boldsymbol{B}^{(1)}\|_F^2 \leqslant \frac{C}{\lambda^2}\|\boldsymbol{Q}_S\boldsymbol{B}^{(1)}\|_F^2.$$

Note that $\boldsymbol{B}^{(1),\top}\boldsymbol{J}_S\mathcal{W} = \boldsymbol{B}^{(1),\top}\boldsymbol{B}_\perp\mathcal{E} = 0$ because $\boldsymbol{J}_S\boldsymbol{B}^{(1)} = \boldsymbol{B}^{(1)}$ and $\boldsymbol{B}^{(1),\top}\boldsymbol{B}_\perp = 0$. Hence

$$\|\boldsymbol{Q}_S\boldsymbol{B}^{(1)}\|_F^2 = \|\boldsymbol{B}^\top\mathcal{V}^{(1)}\mathcal{W}^{(1),\top}\boldsymbol{J}_S\|_F^2 \leqslant \|\mathcal{V}^{(1)}\|^2\|\mathcal{W}^{(1),\top}\boldsymbol{J}_S\|_F^2.$$

On the event $\mathtt{E}$,$\|\mathcal{V}^{(1)}\|^2 \leqslant \|\boldsymbol{\Lambda}_V^{(1)}\| \leqslant 2\kappa\lambda$ and $\left\{\|\boldsymbol{J}_S\mathcal{W}^{(1)}\mathcal{W}^{(1),\top}\boldsymbol{J}_S\| \leqslant 6\frac{s\vee H + \log(n\lambda)}{n}\right\}$. Therefore,

$$\mathbb{E}[\delta^2\mathbf{1}_{\mathtt{E}}] \leqslant \frac{C}{\lambda}\mathbb{E}[\|\boldsymbol{J}_S\mathcal{W}^{(1)}\|_F^2\mathbf{1}_{\mathtt{E}}] \leqslant \frac{C(s\wedge H)}{\lambda}\mathbb{E}[\|\boldsymbol{J}_S\mathcal{W}^{(1)}\mathcal{W}^{(1),\top}\boldsymbol{J}_S\|\mathbf{1}_{\mathtt{E}}] \leqslant C(s\wedge H)\frac{s\vee H}{n\lambda} \leqslant C\epsilon_n^2$$

where the second inequality follows from the inequality $\mathrm{Tr}(\boldsymbol{A}) \leqslant \mathrm{rank}(\boldsymbol{A})\|\boldsymbol{A}\|$ and

$$\mathrm{rank}(\boldsymbol{J}_S\mathcal{W}^{(1)}\mathcal{W}^{(1),\top}\boldsymbol{J}_S) \leqslant s \wedge H$$

and the third inequality follows from $n\lambda < e^{s\vee H}$. □

Lemma 7. *There exists positive constant C such that*

$$\mathbb{E}(2T_1+T_2)^2\mathbf{1}_{\mathrm{E}} \leqslant C\lambda^2\epsilon_n^2$$

Proof. Since $(2T_1+T_2)^2 \leqslant C(T_1^2+T_2^2)$, we only need to bound $\mathbb{E}T_1^2$ and $\mathbb{E}T_2^2$ separately.
For T_1. Recall that $\mathcal{W}^{(2)} = \boldsymbol{B}_\perp\mathcal{E}^{(2)}$ (See notation near (19).) and for each fixed $L \in \mathcal{L}_s$, $\boldsymbol{K}_L \perp\!\!\!\perp \mathcal{W}^{(2)}$, hence

$$\langle\mathcal{W}^{(2)}\mathcal{W}^{(2),\top}, \boldsymbol{K}_L\rangle = \langle\mathcal{E}^{(2)}\mathcal{E}^{(2),\top}, \boldsymbol{B}_\perp^\top\boldsymbol{K}_L\boldsymbol{B}_\perp\rangle \tag{37}$$

and $\boldsymbol{B}_\perp^\top\boldsymbol{K}_L\boldsymbol{B}_\perp \perp\!\!\!\perp \mathcal{W}^{(2)}$.

By Lemma 25, $\|\boldsymbol{B}_\perp^\top\boldsymbol{K}_L\boldsymbol{B}_\perp\|_F \leqslant 1$. For any $m\times m$ symmetric matrix $\boldsymbol{A}$, $\|\boldsymbol{A} - \frac{\mathrm{Tr}(\boldsymbol{A})}{m}\boldsymbol{I}_m\|_F^2 = \mathrm{Tr}(\boldsymbol{A}^\top\boldsymbol{A}) - \frac{1}{m}\mathrm{Tr}(\boldsymbol{A})^2 \leqslant \|\boldsymbol{A}\|_F^2$. Therefore,

$$\|\boldsymbol{B}_\perp^\top\boldsymbol{K}_L\boldsymbol{B}_\perp - \frac{\mathrm{Tr}(\boldsymbol{B}_\perp^\top\boldsymbol{K}_L\boldsymbol{B}_\perp)}{p-d}\mathbf{I}_{p-d}\|_F \leqslant \|\boldsymbol{B}_\perp^\top\boldsymbol{K}_L\boldsymbol{B}_\perp\|_F \leqslant 1.$$

Note that $\mathcal{E}^{(2)}$ is a $(p-d)\times H$ matrix and $\sqrt{n}\mathcal{E}_{i,j}^{(2)} \sim N(0,1)$, we can apply Lemma 29 with $\boldsymbol{Z} = \sqrt{n}\mathcal{E}^{(2)}$ and $\boldsymbol{K} = \boldsymbol{B}_\perp^\top\boldsymbol{K}_L\boldsymbol{B}_\perp - \frac{\mathrm{Tr}(\boldsymbol{B}_\perp^\top\boldsymbol{K}_L\boldsymbol{B}_\perp)}{p-d}\mathbf{I}_{p-d}$ to derive that

$$\mathbb{P}\left(\left|\left\langle\mathcal{E}^{(2)}\mathcal{E}^{(2),\top}, \boldsymbol{B}_\perp^\top\boldsymbol{K}_L\boldsymbol{B}_\perp - \frac{\mathrm{Tr}(\boldsymbol{B}_\perp^\top\boldsymbol{K}_L\boldsymbol{B}_\perp)}{p-d}\mathbf{I}_{p-d}\right\rangle\right| \geqslant \frac{2\sqrt{H}}{n}t + \frac{2}{n}t^2\right) \leqslant 2\exp\left(-t^2\right). \tag{38}$$

After applying Lemma 30 with $N = |\mathcal{L}(s)| \leqslant \left(\frac{ep}{s}\right)^s$, $a = \frac{2\sqrt{H}}{n}$, $b = \frac{2}{n}$, $c = 2$ and $X_i = \left\langle\mathcal{E}^{(2)}\mathcal{E}^{(2),\top}, \boldsymbol{B}_\perp^\top\boldsymbol{K}_L\boldsymbol{B}_\perp - \frac{\mathrm{Tr}(\boldsymbol{B}_\perp^\top\boldsymbol{K}_L\boldsymbol{B}_\perp)}{p-d}\mathbf{I}_{p-d}\right\rangle$, one has

$$\begin{aligned}&\mathbb{E}\max_{L\in\mathcal{L}(s)}\left|\left\langle\mathcal{E}^{(2)}\mathcal{E}^{(2),\top}, \boldsymbol{B}_\perp^\top\boldsymbol{K}_L\boldsymbol{B}_\perp - \frac{\mathrm{Tr}(\boldsymbol{B}_\perp^\top\boldsymbol{K}_L\boldsymbol{B}_\perp)}{p-d}\mathbf{I}_{p-d}\right\rangle\right|^2\\ \leqslant&\frac{2H+32}{n^2}\log(2eN) + \frac{8}{n^2}\log^2(2N).\end{aligned}$$

Note that

$$\begin{aligned}\mathbb{E}\max_{L\in\mathcal{L}(s)}\left|\left\langle\mathcal{E}^{(2)}\mathcal{E}^{(2),\top}, \boldsymbol{B}_\perp^\top\boldsymbol{K}_L\boldsymbol{B}_\perp\right\rangle\right|^2 &\leqslant 2\mathbb{E}\max_{L\in\mathcal{L}(s)}\left|\left\langle\mathcal{E}^{(2)}\mathcal{E}^{(2),\top}, \frac{\mathrm{Tr}(\boldsymbol{B}_\perp^\top\boldsymbol{K}_L\boldsymbol{B}_\perp)}{p-d}\mathbf{I}_{p-d}\right\rangle\right|^2\\ +2\mathbb{E}\max_{L\in\mathcal{L}(s)}&\left|\left\langle\mathcal{E}^{(2)}\mathcal{E}^{(2),\top}, \boldsymbol{B}_\perp^\top\boldsymbol{K}_L\boldsymbol{B}_\perp - \frac{\mathrm{Tr}(\boldsymbol{B}_\perp^\top\boldsymbol{K}_L\boldsymbol{B}_\perp)}{p-d}\mathbf{I}_{p-d}\right\rangle\right|^2\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}\max_{L\in\mathcal{L}(s)}\left|\left\langle\mathcal{E}^{(2)}\mathcal{E}^{(2),\top}, \frac{\mathrm{Tr}(\boldsymbol{B}_\perp^\top\boldsymbol{K}_L\boldsymbol{B}_\perp)}{p-d}\mathbf{I}_{p-d}\right\rangle\right|^2 &= \mathbb{E}\max_{L\in\mathcal{L}(s)}\left(\frac{\mathrm{Tr}(\boldsymbol{B}_\perp^\top\boldsymbol{K}_L\boldsymbol{B}_\perp)}{p-d}\right)^2\|\mathcal{E}^{(2)}\|_F^4 \overset{(a)}{\leqslant} \frac{2s}{(p-d)^2}\mathbb{E}\|\mathcal{E}^{(2)}\|_F^4\\ &\overset{(b)}{=} \frac{2s}{(p-d)^2}\frac{(p-d)^2H^2+2H(p-d)}{n^2} \asymp \frac{H^2s}{n^2}.\end{aligned}$$

Here in (a) we used the inequality that

$|\mathrm{Tr}(\boldsymbol{A})| \leqslant \sqrt{\mathrm{rank}(\boldsymbol{A})}\|\boldsymbol{A}\|_F$ and the facts that $\mathrm{rank}(\boldsymbol{B}_\perp^\top \boldsymbol{K}_L \boldsymbol{B}_\perp) \leqslant 2s$ and $\|\boldsymbol{B}_\perp^\top \boldsymbol{K}_L \boldsymbol{B}_\perp\|_F \leqslant 1$, and in (b) we used Lemma 24. Then we have

$$\begin{aligned}
&\mathbb{E} \max_{L\in\mathcal{L}(s)} \left|\left\langle \mathcal{E}^{(2)}\mathcal{E}^{(2),\top}, \boldsymbol{B}_\perp^\top \boldsymbol{K}_L \boldsymbol{B}_\perp \right\rangle\right|^2 \\
\lesssim & \frac{4H+64}{n^2}\log(2eN) + \frac{16}{n^2}\log^2(2N) + \frac{H^2 s}{n^2} \\
\lesssim & \frac{H}{n^2}\log(2eN) + \frac{\log^2(2N)}{n^2} + \frac{H^2 s}{n^2} \lesssim \lambda^2\epsilon_n^4 \lesssim \lambda^2\epsilon_n^2.
\end{aligned}$$

For T_2. Fix $L \in \mathcal{L}(s)$. Since $\mathcal{V}^{(2)} \perp\!\!\!\perp \mathcal{W}^{(2)}$, $\boldsymbol{K}_L \perp\!\!\!\perp \mathcal{W}^{(2)}$ and $\boldsymbol{K}_L \perp\!\!\!\perp \mathcal{V}^{(2)}$, conditioned on the $\mathcal{V}^{(2)}$ and $\boldsymbol{K}_L$, we know that

$$\sqrt{n}\langle \mathcal{V}^{(2)}\mathcal{W}^{(2),\top}, \boldsymbol{K}_L\rangle = \langle \boldsymbol{B}_\perp^\top \boldsymbol{K}_L \mathcal{V}^{(2)}, \sqrt{n}\mathcal{E}^{(2)}\rangle$$

is distributed according to $N(0, \|\boldsymbol{B}_\perp^\top \boldsymbol{K}_L \mathcal{V}^{(2)}\|_F^2)$. Therefore

$$\sqrt{n}\langle \mathcal{V}^{(2)}\mathcal{W}^{(2),\top}, \boldsymbol{K}_L\rangle \overset{d}{=} \|\boldsymbol{B}_\perp^\top \boldsymbol{K}_L \mathcal{V}^{(2)}\|_F W$$

for some $W \sim N(0,1)$ independent of $\mathcal{V}^{(2)}$ and $\boldsymbol{K}_L$. For simplicity of notation, we denote $\sqrt{n}\langle \mathcal{V}^{(2)}\mathcal{W}^{(2),\top}, \boldsymbol{K}_L\rangle$ by F_L. Define the event $\tilde{\mathtt{E}} = \{\|\mathcal{V}^{(2)}\mathcal{V}^{(2),\top}\| \leqslant 2\kappa\lambda\}$. Then $\mathtt{E} \subset \tilde{\mathtt{E}}$ and $\tilde{\mathtt{E}}$ only depends on $\mathcal{V}^{(2)}$. Note that $\|\boldsymbol{B}_\perp^\top \boldsymbol{K}_L \mathcal{V}^{(2)}\|_F \leqslant \|\boldsymbol{B}_\perp^\top \boldsymbol{K}_L\|_F \|\mathcal{V}^{(2)}\| \leqslant \sqrt{\|\mathcal{V}^{(2)}\mathcal{V}^{(2),\top}\|}$. Consequently,

$$\mathbb{P}\left(|F_L| > t \mid \tilde{\mathtt{E}}\right) \leqslant \mathbb{P}\left(\sqrt{2\kappa\lambda}|W| > t\right) \leqslant 2\exp\left(-\frac{t^2}{2\kappa\lambda}\right). \tag{39}$$

In other words, conditioning on $\tilde{\mathtt{E}}$, each of F_L $(L \in \mathcal{L}(s))$ satisfies the premise in Lemma 30 with $(a,b,c) = (\sqrt{2\kappa\lambda}, 0, 2)$. Therefore,

$$\mathbb{E}\left(T_2^2 \mathbf{1}_{\mathtt{E}}\right) \leqslant \frac{1}{n}\mathbb{E}\max_{L\in\mathcal{L}(s)}\left(F_L^2 \mathbf{1}_{\tilde{\mathtt{E}}}\right) \leqslant \frac{4\kappa\lambda}{n}\log(2eN) \leqslant C\lambda^2\epsilon_n^2.$$

□

D Proofs of upper bounds with a unknown covariance matrix

In this section, we prove the upper bounds in Sections 4.2.2 and 4.3 for the general cases where $\boldsymbol{\Sigma}$ is unknown.

As in Appendix C, we take H to be an integer such that $H \leqslant H_0 d$ for some constant $H_0 > K_0 \vee C$ and the inequality in Lemma 1 holds.

D.1 Proof of Theorem 5

Let $\boldsymbol{\Sigma}^{1/2}$ be a square root of $\boldsymbol{\Sigma}$, $\widetilde{\boldsymbol{B}} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{B}$ and $\widetilde{\boldsymbol{X}_i} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{X}_i$. Then $\widetilde{\boldsymbol{X}_i} \sim N(0, \mathbf{I}_p)$ and $\boldsymbol{B}^\top \boldsymbol{X}_i = \widetilde{\boldsymbol{B}}^\top \widetilde{\boldsymbol{X}_i}$. We will use the results in Section C for the pairs $(\widetilde{\boldsymbol{X}_i}, Y_i)$'s by defining $\widetilde{\boldsymbol{B}}_\perp$, $\widetilde{\boldsymbol{E}}$, $\widetilde{\boldsymbol{V}}$, $\widetilde{\boldsymbol{W}}$ accordingly.

Let $\widetilde{\boldsymbol{B}}_\perp$ be a $p \times (p-d)$ orthogonal matrix such that $\widetilde{\boldsymbol{B}}^\top \widetilde{\boldsymbol{B}}_\perp = 0$.

For any pair of $(\boldsymbol{X}, Y)$ sampled from $\mathcal{M}$, let $\boldsymbol{Z} = \boldsymbol{B}^\top \boldsymbol{X} = \widetilde{\boldsymbol{B}}^\top \widetilde{\boldsymbol{X}}$ and $\widetilde{\boldsymbol{E}} = \widetilde{\boldsymbol{B}}_\perp^\top \widetilde{\boldsymbol{X}}$. We have the decomposition that

$$\widetilde{\boldsymbol{X}} = \widetilde{\boldsymbol{B}}\widetilde{\boldsymbol{B}}^\top \widetilde{\boldsymbol{X}} + \widetilde{\boldsymbol{B}}_\perp \widetilde{\boldsymbol{B}}_\perp^\top \widetilde{\boldsymbol{X}} = \widetilde{\boldsymbol{B}}\boldsymbol{Z} + \widetilde{\boldsymbol{B}}_\perp \widetilde{\boldsymbol{E}}.$$

Let $\widetilde{\boldsymbol{V}} = \widetilde{\boldsymbol{B}}\boldsymbol{Z}$ and $\widetilde{\boldsymbol{W}} = \widetilde{\boldsymbol{B}}_\perp \widetilde{\boldsymbol{E}}$. Then $\widetilde{\boldsymbol{V}}^\top \widetilde{\boldsymbol{W}} = 0$. Let $\widetilde{\boldsymbol{\Lambda}} = \mathrm{Cov}\left(\mathbb{E}[\widetilde{\boldsymbol{X}} \mid Y]\right)$. We introduce the notation $\overline{\widetilde{\boldsymbol{X}}}_{h,\cdot}$ similar to the definition of $\overline{\boldsymbol{X}}_{h,\cdot}$ in Equation 3 and let $\widetilde{\mathcal{X}} = \frac{1}{\sqrt{H}}\left[\overline{\widetilde{\boldsymbol{X}}}_{1,\cdot}, \overline{\widetilde{\boldsymbol{X}}}_{2,\cdot}, ..., \overline{\widetilde{\boldsymbol{X}}}_{H,\cdot}\right]$. Similarly we define $\overline{\widetilde{\boldsymbol{V}}}_{h,\cdot}$, $\overline{\boldsymbol{Z}}_{h,\cdot}$, $\overline{\widetilde{\boldsymbol{W}}}_{h,\cdot}$, $\overline{\widetilde{\boldsymbol{E}}}_{h,\cdot}$ and $\widetilde{\mathcal{V}}$, $\mathcal{Z}$, $\widetilde{\mathcal{W}}$, $\widetilde{\mathcal{E}}$. We have $\widetilde{\mathcal{V}} = \widetilde{\boldsymbol{B}}\mathcal{Z}$ and $\widetilde{\mathcal{W}} = \widetilde{\boldsymbol{B}}_\perp \widetilde{\mathcal{E}}$. Since $\widetilde{\boldsymbol{E}} \sim N(0, \mathbf{I}_{p-d})$ and is independent of Y, we know that the entries $\widetilde{\mathcal{E}}_{i,j}$ of $\widetilde{\mathcal{E}}$ are *i.i.d.* samples of $N(0, \frac{1}{n})$.

Define $\widehat{\widetilde{\boldsymbol{\Lambda}}}_H = \widetilde{\mathcal{X}}\widetilde{\mathcal{X}}^\top$, $\widehat{\boldsymbol{\Lambda}}_z = \mathcal{Z}\mathcal{Z}^\top$ and $\widehat{\widetilde{\boldsymbol{\Lambda}}}_V = \widetilde{\mathcal{V}}\widetilde{\mathcal{V}}^\top$. Then $\widehat{\widetilde{\boldsymbol{\Lambda}}}_V = \widetilde{\boldsymbol{B}}\widehat{\boldsymbol{\Lambda}}_z\widetilde{\boldsymbol{B}}^\top$, and $\widehat{\boldsymbol{\Lambda}}_H = \boldsymbol{\Sigma}^{1/2}\widehat{\widetilde{\boldsymbol{\Lambda}}}_H\boldsymbol{\Sigma}^{1/2}$. We have the following decomposition

$$\widehat{\widetilde{\boldsymbol{\Lambda}}}_H = \widetilde{\mathcal{V}}\widetilde{\mathcal{V}}^\top + \widetilde{\mathcal{V}}\widetilde{\mathcal{W}}^\top + \widetilde{\mathcal{W}}\widetilde{\mathcal{V}}^\top + \widetilde{\mathcal{W}}\widetilde{\mathcal{W}}^\top. \tag{40}$$

We define some events: $\widetilde{\mathtt{E}}_1 = \left\{ \|\widetilde{\mathcal{W}}\widetilde{\mathcal{W}}^\top\| \leqslant 6\frac{p\vee H+\log(n\lambda)}{n} \right\}$, $\widetilde{\mathtt{E}}_2 = \left\{ \|\widehat{\widetilde{\boldsymbol{\Lambda}}}_V - \widetilde{\boldsymbol{\Lambda}}\| \leqslant \frac{2}{3\nu}\kappa\lambda \right\}$ and $\widetilde{\mathtt{E}} = \widetilde{\mathtt{E}}_1 \cap \widetilde{\mathtt{E}}_2$. In view of Corollary 2, we have the following result.

Corollary 3. For $\nu \in (\kappa, 2\kappa]$, we can find constants C and $\widetilde{C}$, such that if

$$\kappa^2 H^2\left(\log(n\kappa H) + d\right) < Cn\lambda$$

and $\kappa\left(p \vee H + \log(n\lambda)\right) < Cn\lambda$, then $\mathbb{P}\left(\widetilde{\mathtt{E}}^c\right) \leqslant \frac{\widetilde{C}}{n\lambda}$ and on the event $\widetilde{\mathtt{E}}$, the followings hold

a) $\frac{1}{3}\lambda \leqslant \lambda_d(\widehat{\widetilde{\boldsymbol{\Lambda}}}_V) \leqslant \lambda_1(\widehat{\widetilde{\boldsymbol{\Lambda}}}_V) \leqslant 2\kappa\lambda$.

b) $\|\widehat{\widetilde{\boldsymbol{\Lambda}}}_H - \widehat{\widetilde{\boldsymbol{\Lambda}}}_V\| \leqslant \lambda\sqrt{18\kappa\frac{p\vee H+\log(n\lambda)}{n\lambda}} < \frac{1}{4}\lambda$.

c) $\lambda_{d+1}(\widehat{\widetilde{\boldsymbol{\Lambda}}}_H) < \frac{1}{4}\lambda$.

Recall the SIR estimator $\widehat{\boldsymbol{B}}$ in (5). Our goal is to bound the expectation of $\|\widehat{\boldsymbol{B}}^{\otimes} - \boldsymbol{B}^{\otimes}\|_F^2$. Under the assumption that $\|\boldsymbol{\Sigma}^{-1}\| < M$, one has

$$\begin{aligned}\|\widehat{\boldsymbol{B}}^{\otimes} - \boldsymbol{B}^{\otimes}\|_F &< M\|\boldsymbol{\Sigma}^{1/2}\left(\widehat{\boldsymbol{B}}^{\otimes} - \boldsymbol{B}^{\otimes}\right)\boldsymbol{\Sigma}^{1/2}\|_F \\ &\leqslant M\left(\|\boldsymbol{\Sigma}^{1/2}\widehat{\boldsymbol{B}}^{\otimes}\boldsymbol{\Sigma}^{1/2} - (\widehat{\boldsymbol{\Sigma}}^{1/2}\widehat{\boldsymbol{B}})^{\otimes}\|_F + \|(\widehat{\boldsymbol{\Sigma}}^{1/2}\widehat{\boldsymbol{B}})^{\otimes} - \boldsymbol{\Sigma}^{1/2}\boldsymbol{B}^{\otimes}\boldsymbol{\Sigma}^{1/2}\|_F\right).\end{aligned} \tag{41}$$

Step 1: Bounding the first term in (41).

For any matrices $\boldsymbol{A} \in \mathbb{R}^{p\times p}$ and $\boldsymbol{L} \in \mathbb{R}^{p\times p}$, one has the identity that $\boldsymbol{A}^\top\boldsymbol{L}\boldsymbol{A} - \boldsymbol{L} = (\boldsymbol{A} - \mathbf{I}_p)^\top\boldsymbol{L}\boldsymbol{A} + \boldsymbol{L}(\boldsymbol{A} - \mathbf{I}_p)$ and the inequality

$$\begin{aligned}\|\boldsymbol{A}^\top\boldsymbol{L}\boldsymbol{A} - \boldsymbol{L}\|_F &\leqslant \|\boldsymbol{A} - \mathbf{I}_p\|\|\boldsymbol{L}\|_F\|\boldsymbol{A}\| + \|\boldsymbol{L}\|_F\|\boldsymbol{A} - \mathbf{I}_p\| \\ &= (\|\boldsymbol{A}\| + 1)\|\boldsymbol{A} - \mathbf{I}_p\|\|\boldsymbol{L}\|_F.\end{aligned}$$

Substitute $\boldsymbol{A} = \widehat{\mathbf{I}} = \widehat{\boldsymbol{\Sigma}}^{-1/2}\boldsymbol{\Sigma}^{1/2}$ and $\boldsymbol{L} = (\widehat{\boldsymbol{\Sigma}}^{1/2}\widehat{\boldsymbol{B}})^{\otimes}$. Then

$$\|\boldsymbol{\Sigma}^{1/2}\widehat{\boldsymbol{B}}^{\otimes}\boldsymbol{\Sigma}^{1/2} - (\widehat{\boldsymbol{\Sigma}}^{1/2}\widehat{\boldsymbol{B}})^{\otimes}\|_F \leqslant (\|\widehat{\mathbf{I}}\| + 1)\|\widehat{\mathbf{I}} - \mathbf{I}_p\|\|(\widehat{\boldsymbol{\Sigma}}^{1/2}\widehat{\boldsymbol{B}})^{\otimes}\|_F \leqslant d(\|\widehat{\mathbf{I}}\| + 1)\|\widehat{\mathbf{I}} - \mathbf{I}_p\|$$

since $\widehat{\boldsymbol{\Sigma}}^{1/2}\widehat{\boldsymbol{B}}$ is a $p \times d$ orthogonal matrix. In order to obtain an upper bound for the right hand side in the last inequality, we present the following lemma, whose proof will be provided at the end of this subsection.

Lemma 8. *There exist constants C and $\widetilde{C}$, such that if $p{+}\log(n\lambda) < Cn$, then $\|\widehat{\mathbf{I}}\|^2 < 2M^2$, $\|\widehat{\boldsymbol{\Sigma}}^{-1/2}\|^2 < 2M$ and $\|\widehat{\mathbf{I}} - \mathbf{I}_p\|^2 < \widetilde{C}\frac{p+\log(n\lambda)}{n}$ hold with probability at least $1 - \widetilde{C}/n\lambda$.*

Let $\mathtt{E}$ be the intersection of the events in Corollary 3 and Lemma 8.

By Lemma 8, one has

$$1_{\mathtt{E}}\|\boldsymbol{\Sigma}^{1/2}\widehat{\boldsymbol{B}}^{\otimes}\boldsymbol{\Sigma}^{1/2} - \widehat{\boldsymbol{\Sigma}}^{1/2}\widehat{\boldsymbol{B}}^{\otimes}\widehat{\boldsymbol{\Sigma}}^{1/2}\|_F^2 \lesssim \frac{d[p + \log(n\lambda)]}{n}. \tag{42}$$

Step 2: Bounding the second term in (41).

We know that $\widehat{\boldsymbol{\Sigma}}^{1/2}\widehat{\boldsymbol{B}}$ is formed by the first d leading eigenvector of $\widehat{\widetilde{\boldsymbol{\Lambda}}}_H = \widehat{\boldsymbol{\Sigma}}^{-1/2}\widehat{\boldsymbol{\Lambda}}_H\widehat{\boldsymbol{\Sigma}}^{-1/2}$. Since $\widetilde{\boldsymbol{B}}$ is formed by the first d leading eigenvectors of $\widehat{\widetilde{\boldsymbol{\Lambda}}}_V$, by Lemma 28 and Corollary 3, on the event $\mathtt{E}$, it holds that

$$\|(\widehat{\boldsymbol{\Sigma}}^{1/2}\widehat{\boldsymbol{B}})^{\otimes} - \widetilde{\boldsymbol{B}}^{\otimes}\|_F^2 \lesssim \frac{1}{\lambda^2}\|\widehat{\boldsymbol{\Sigma}}^{-1/2}\widehat{\boldsymbol{\Lambda}}_H\widehat{\boldsymbol{\Sigma}}^{-1/2} - \widehat{\widetilde{\boldsymbol{\Lambda}}}_V\|_F^2. \tag{43}$$

Let $\Delta = \widehat{\boldsymbol{\Sigma}}^{-1/2}\widehat{\boldsymbol{\Lambda}}_H\widehat{\boldsymbol{\Sigma}}^{-1/2} - \widehat{\widetilde{\boldsymbol{\Lambda}}}_V$ and $\widehat{\mathbf{I}} = \widehat{\boldsymbol{\Sigma}}^{-1/2}\boldsymbol{\Sigma}^{1/2}$. Then $\Delta = \widehat{\mathbf{I}}\left(\widehat{\widetilde{\boldsymbol{\Lambda}}}_H - \widehat{\widetilde{\boldsymbol{\Lambda}}}_V\right)\widehat{\mathbf{I}}^\top + (\widehat{\mathbf{I}} - \mathbf{I}_p)\widehat{\widetilde{\boldsymbol{\Lambda}}}_V\widehat{\mathbf{I}}^\top + \widehat{\widetilde{\boldsymbol{\Lambda}}}_V\left(\widehat{\mathbf{I}} - \mathbf{I}_p\right)^\top$.

We then have

$$1_{\mathtt{E}}\|\Delta\|_F^2 \lesssim 1_{\mathtt{E}}\left(\|\widehat{\widetilde{\boldsymbol{\Lambda}}}_H - \widehat{\widetilde{\boldsymbol{\Lambda}}}_V\|_F^2 + \|\widehat{\widetilde{\boldsymbol{\Lambda}}}_V\|_F^2\frac{p+\log(n\lambda)}{n}\right). \tag{44}$$

Note that $\|\widehat{\widetilde{\boldsymbol{\Lambda}}}_H - \widehat{\widetilde{\boldsymbol{\Lambda}}}_V\|_F \leqslant 2\|\widetilde{\mathcal{V}}\widetilde{\mathcal{W}}^\top\|_F + \|\widetilde{\mathcal{W}}\widetilde{\mathcal{W}}^\top\|_F$. By Lemma 24 and $\widetilde{\mathcal{W}} = \widetilde{\boldsymbol{B}}_\perp\widetilde{\mathcal{E}}$, we have

$$\begin{aligned}\mathbb{E}\|\widetilde{\mathcal{W}}\widetilde{\mathcal{W}}^\top\|_F^2 &\leqslant \mathbb{E}\|\widetilde{\mathcal{E}}\widetilde{\mathcal{E}}^\top\|_F^2\\ &= \frac{1}{n^2}(p-d)H(p-d+H+1)\\ &\lesssim \frac{pH\,(p+H)}{n^2},\end{aligned} \tag{45}$$

and

$$\begin{aligned}\mathbb{E}\mathrm{Tr}\left(\widetilde{\mathcal{W}}^\top\widetilde{\mathcal{W}}\right) &\leqslant \frac{1}{n}(p-d)H\\ &\lesssim \frac{pH}{n}.\end{aligned}$$

Since on the event $\mathtt{E}$, $\|\widetilde{\mathcal{V}}^\top\widetilde{\mathcal{V}}\| < 2\kappa\lambda$,

$$\begin{aligned}1_{\mathtt{E}}\|\widetilde{\mathcal{V}}\widetilde{\mathcal{W}}^\top\|_F^2 &= 1_{\mathtt{E}}\mathrm{Tr}\left(\widetilde{\mathcal{V}}\widetilde{\mathcal{W}}^\top\widetilde{\mathcal{W}}\widetilde{\mathcal{V}}^\top\right)\\ &\leqslant 1_{\mathtt{E}}\|\widetilde{\mathcal{V}}^\top\widetilde{\mathcal{V}}\|\mathrm{Tr}\left(\widetilde{\mathcal{W}}^\top\widetilde{\mathcal{W}}\right)\\ &\leqslant 2\kappa\lambda\mathrm{Tr}\left(\widetilde{\mathcal{W}}^\top\widetilde{\mathcal{W}}\right),\end{aligned}$$

where the second inequality is due to the fact that $\mathrm{Tr}(\boldsymbol{A}\boldsymbol{L}) \leqslant \|\boldsymbol{A}\|\mathrm{Tr}(\boldsymbol{L})$ holds for any positive semi-definite matrices $\boldsymbol{L}$ and $\boldsymbol{A}$. It then follows that $\mathbb{E}1_{\mathtt{E}}\|\widetilde{\mathcal{V}}\widetilde{\mathcal{W}}^\top\|_F^2 \lesssim \frac{\lambda pH}{n}$. Combining this with Equation (45), we have

$$\begin{aligned}\mathbb{E}1_{\mathtt{E}}\|\widehat{\widetilde{\boldsymbol{\Lambda}}}_H - \widehat{\widetilde{\boldsymbol{\Lambda}}}_V\|_F^2 &\lesssim \frac{\lambda pH}{n} + \frac{pH\,(p+H)}{n^2}\\ &\lesssim \frac{\lambda pH}{n},\end{aligned} \tag{46}$$

because $p \vee H/\,(n\lambda)$ is small.

Note that $\widetilde{\mathcal{V}} = \widetilde{\boldsymbol{B}}\mathcal{Z}$ and $\mathcal{Z} = \widetilde{\boldsymbol{B}}^\top\widetilde{\mathcal{V}}$, we have $\mathrm{Tr}\left(\widetilde{\mathcal{V}}\widetilde{\mathcal{V}}^\top\right) = \mathrm{Tr}\left(\mathcal{Z}\mathcal{Z}^\top\right) \leqslant d\|\widetilde{\mathcal{V}}\widetilde{\mathcal{V}}^\top\|$ because $\widetilde{\boldsymbol{B}}^\top\widetilde{\boldsymbol{B}} = \mathbf{I}_d$. Therefore,

$$\begin{aligned}\|\widehat{\widetilde{\boldsymbol{\Lambda}}}_V\|_F^2 &= \mathrm{Tr}\left(\widetilde{\mathcal{V}}\widetilde{\mathcal{V}}^\top\widetilde{\mathcal{V}}\widetilde{\mathcal{V}}^\top\right)\\ &\leqslant \|\widetilde{\mathcal{V}}\widetilde{\mathcal{V}}^\top\|\mathrm{Tr}\left(\widetilde{\mathcal{V}}\widetilde{\mathcal{V}}^\top\right)\\ &\leqslant d\|\widetilde{\mathcal{V}}\widetilde{\mathcal{V}}^\top\|^2,\end{aligned}$$

which implies that $1_{\mathtt{E}}\|\widehat{\widetilde{\boldsymbol{\Lambda}}}_V\|_F^2 \leqslant 4\kappa^2\lambda^2 d$.

In view of inequalities (44),(46), and (43), we have

$$\mathbb{E}\left(1_{\mathtt{E}}\|(\widehat{\boldsymbol{\Sigma}}^{1/2}\widehat{\boldsymbol{B}})^{\otimes} - \widetilde{\boldsymbol{B}}^{\otimes}\|_F^2\right) \lesssim \frac{1}{\lambda^2}\mathbb{E}\left(1_{\mathtt{E}}\|\Delta\|_F^2\right)$$

$$
\begin{aligned}
&\lesssim \frac{1}{\lambda^2}\left(\frac{\lambda p H}{n} + \lambda^2 d\frac{p+\log(n\lambda)}{n}\right) \\
&\lesssim \frac{H[p+\log(n\lambda)]}{n\lambda}.
\end{aligned} \tag{47}
$$

Combining Equations (41), (47) and (42), one has

$$
\mathbb{E}\left(1_{\mathtt{E}}\|\widehat{\boldsymbol{B}}^{\otimes}-\boldsymbol{B}^{\otimes}\|_F^2\right) \lesssim \frac{H[p+\log(n\lambda)]}{n\lambda}. \tag{48}
$$

Step 3: Synthesis.

In addition, one has $\|\widehat{\boldsymbol{B}}^{\otimes}-\boldsymbol{B}^{\otimes}\|_F^2 \leqslant 2d$ and $\mathbb{P}(\mathtt{E}^c) \lesssim \frac{1}{n\lambda}$, one has $\mathbb{E}\left(1_{\mathtt{E}^c}\|\widehat{\boldsymbol{B}}^{\otimes}-\boldsymbol{B}^{\otimes}\|_F^2\right) \lesssim \frac{d}{n\lambda}$. We conclude that

$$
\mathbb{E}\|\widehat{\boldsymbol{B}}^{\otimes}-\boldsymbol{B}^{\otimes}\|_F^2 \lesssim \frac{H[p+\log(n\lambda)]}{n\lambda}.
$$

□

Proof of Lemma 8. Note that $\widehat{\widehat{\boldsymbol{\Sigma}}} := \boldsymbol{\Sigma}^{-1/2}\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1/2}$ is the empirical covariance matrix of $\widetilde{\boldsymbol{X}}_i$. By Lemma 27, with probability at least $1-2\exp(-nt^2/2)$, the eigenvalues of $\widehat{\widehat{\boldsymbol{\Sigma}}}$ lie between $(1-\sqrt{p/n}-t)^2$ and $(1+\sqrt{p/n}+t)^2$.

Fix $t \asymp \sqrt{2\log(n\lambda)/n}$ and assume the event happens.

If $\frac{p+\log(n\lambda)}{n}$ is sufficiently small (<1), the eigenvalues of $\widehat{\widehat{\boldsymbol{\Sigma}}}$ lie between $(1/2, 2)$. In this case, $\|\widehat{\boldsymbol{\Sigma}}^{-1}\| \leqslant \|\boldsymbol{\Sigma}^{1/2}\widehat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\Sigma}^{1/2}\|M < 2M$ because $\|\boldsymbol{\Sigma}^{-1}\| < M$ by assumption. We can also easily see that $\|\widehat{\mathbf{I}}\| \leqslant \|\widehat{\boldsymbol{\Sigma}}^{-1}\|^{1/2}\|\boldsymbol{\Sigma}\|^{1/2} < \sqrt{2}M$.

Furthermore, $\widehat{\mathbf{I}}-\mathbf{I}_p = \widehat{\boldsymbol{\Sigma}}^{-1/2}\left(\boldsymbol{\Sigma}^{1/2}-\widehat{\boldsymbol{\Sigma}}^{1/2}\right)$. By Schmitt [1992, Lemma 2.2] and the fact that $\boldsymbol{\Sigma} \gtrsim M^{-1}\mathbf{I}$ and $\widehat{\boldsymbol{\Sigma}} \gtrsim (2M)^{-1}\mathbf{I}$, we have

$$
\begin{aligned}
\|\boldsymbol{\Sigma}^{1/2}-\widehat{\boldsymbol{\Sigma}}^{1/2}\| &< 3\sqrt{M}\|\boldsymbol{\Sigma}-\widehat{\boldsymbol{\Sigma}}\| \\
&= 3\sqrt{M}\|\boldsymbol{\Sigma}^{1/2}(\mathbf{I}-\widehat{\widehat{\boldsymbol{\Sigma}}})\boldsymbol{\Sigma}^{1/2}\| \\
&\leqslant 3\sqrt{M}M\|\mathbf{I}-\widehat{\widehat{\boldsymbol{\Sigma}}}\|,
\end{aligned}
$$

where the last inequality is because $\|\boldsymbol{\Sigma}\| < M$ by assumption. Since $\|\mathbf{I}-\widehat{\widehat{\boldsymbol{\Sigma}}}\| = \max_{1\leqslant i\leqslant p}|1-\sigma_i(\widehat{\widehat{\boldsymbol{\Sigma}}})| \lesssim \sqrt{\frac{p+\log(n\lambda)}{n}}$, we have $\|\widehat{\mathbf{I}}-\mathbf{I}_p\| < 5M^2\|\mathbf{I}-\widehat{\widehat{\boldsymbol{\Sigma}}}\| \lesssim \sqrt{\frac{p+\log(n\lambda)}{n}}$. □

D.2 Proof of Theorem 8

Preliminaries

Let $\mathcal{L} = \{T \subset [p] : S \subset T, |T| \leqslant 2s\}$. For any $T \in \mathcal{L}$, we define some notations.

Let $\boldsymbol{J}_T$ be the matrix formed by the rows of $\mathbf{I}_p$ in T. Let $\boldsymbol{\Sigma}_{TT}$ be the sub-matrix of $\boldsymbol{\Sigma}$ with row indices and column indices both equal to T, i.e., $\boldsymbol{J}_T\boldsymbol{\Sigma}\boldsymbol{J}_T^{\top}$. Let $\boldsymbol{\Sigma}_{TT}^{1/2}$ be a square root of $\boldsymbol{\Sigma}_{TT}$. Note that it is different from the sub-matrix of a square root of $\boldsymbol{\Sigma}$. Let $\widetilde{\boldsymbol{B}}_{(T)} = \boldsymbol{\Sigma}_{TT}^{1/2}\boldsymbol{J}_T\boldsymbol{B}$. Then $\widetilde{\boldsymbol{B}}_{(T)}^{\top}\widetilde{\boldsymbol{B}}_{(T)} = \boldsymbol{B}^{\top}\boldsymbol{J}_T^{\top}\boldsymbol{\Sigma}_{TT}\boldsymbol{J}_T\boldsymbol{B} = \boldsymbol{B}^{\top}\boldsymbol{J}_T^{\top}\boldsymbol{J}_T\boldsymbol{\Sigma}\boldsymbol{J}_T^{\top}\boldsymbol{J}_T\boldsymbol{B} = \boldsymbol{B}^{\top}\boldsymbol{\Sigma}\boldsymbol{B} = \mathbf{I}_d$ because $S \subset T$. Let $\widetilde{\boldsymbol{B}}_{(T),\perp}$ is a $|T| \times (|T|-d)$ orthogonal matrix such that $\widetilde{\boldsymbol{B}}_{(T)}^{\top}\widetilde{\boldsymbol{B}}_{(T),\perp} = 0$.

For a pair of $(\boldsymbol{X}, Y)$ that is sampled from the distribution $\mathcal{M} \in \mathfrak{M}_s(p, d, \lambda)$, we introduce the following notations.

Let $\widetilde{\boldsymbol{X}}_{(T)} = \boldsymbol{\Sigma}_{TT}^{-1/2}\boldsymbol{J}_T\boldsymbol{X}$. Then $\widetilde{\boldsymbol{X}}_{(T)} \sim N(0, \mathbf{I}_{|T|})$ because $\boldsymbol{\Sigma}_{TT}^{-1/2}\boldsymbol{J}_T\boldsymbol{\Sigma}\boldsymbol{J}_T^{\top}\boldsymbol{\Sigma}_{TT}^{-1/2} = \mathbf{I}_{|T|}$.

Let $\boldsymbol{Z} = \boldsymbol{B}^\top \boldsymbol{X}$.Note that $\boldsymbol{Z} = \boldsymbol{B}^\top \boldsymbol{J}_T^\top \boldsymbol{J}_T \boldsymbol{X} = \widetilde{\boldsymbol{B}}_{(T)}^\top \widetilde{\boldsymbol{X}}_{(T)}$ because $S \subset T$. Let $\widetilde{\boldsymbol{E}}_{(T)} = \widetilde{\boldsymbol{B}}_{(T),\perp}^\top \widetilde{\boldsymbol{X}}_{(T)}$. Since $\widetilde{\boldsymbol{X}}_{(T)} \sim N(0, \mathbf{I}_{|T|})$, one has $\boldsymbol{Z} \sim N(0, \boldsymbol{I}_d)$ and $\widetilde{\boldsymbol{E}}_{(T)} \sim N(0, \boldsymbol{I}_{|T|-d})$. Furthermore, $\boldsymbol{Z} \perp\!\!\!\perp \widetilde{\boldsymbol{E}}_{(T)}$ and

$$\widetilde{\boldsymbol{X}}_{(T)} = \widetilde{\boldsymbol{B}}_{(T)} \widetilde{\boldsymbol{B}}_{(T)}^\top \widetilde{\boldsymbol{X}}_{(T)} + \widetilde{\boldsymbol{B}}_{(T),\perp} \widetilde{\boldsymbol{B}}_{(T),\perp}^\top \widetilde{\boldsymbol{X}}_{(T)} = \widetilde{\boldsymbol{B}}_{(T)} \boldsymbol{Z} + \widetilde{\boldsymbol{B}}_{(T),\perp} \widetilde{\boldsymbol{E}}_{(T)}.$$

Let $\widetilde{\boldsymbol{V}}_{(T)} = \widetilde{\boldsymbol{B}}_{(T)} \boldsymbol{Z}$ and $\widetilde{\boldsymbol{W}}_{(T)} = \widetilde{\boldsymbol{B}}_{(T),\perp} \widetilde{\boldsymbol{E}}_{(T)}$. Then $\widetilde{\boldsymbol{V}}_{(T)}^\top \widetilde{\boldsymbol{W}}_{(T)} = 0$. Let $\boldsymbol{\Lambda}_z = \operatorname{Cov}\left(\mathbb{E}[\boldsymbol{Z} \mid Y]\right)$ and let $\widetilde{\boldsymbol{\Lambda}}_{(T)} = \widetilde{\boldsymbol{B}}_{(T)} \boldsymbol{\Lambda}_z \widetilde{\boldsymbol{B}}_{(T)}^\top$. Then $\widetilde{\boldsymbol{\Lambda}}_{(T)} = \operatorname{Cov}\left(\mathbb{E}[\widetilde{\boldsymbol{X}}_{(T)} \mid Y]\right)$.

We next introduce the notation for the sliced samples. For example, we define $\overline{\widetilde{\boldsymbol{X}}}_{(T),h,\cdot}$ similarly to the definition of $\overline{\boldsymbol{X}}_{h,\cdot}$ in (3) and $\widetilde{\mathcal{X}}_{(T)} = \frac{1}{\sqrt{H}} \left[\overline{\widetilde{\boldsymbol{X}}}_{(T),1,\cdot}\ , \ \overline{\widetilde{\boldsymbol{X}}}_{(T),2,\cdot}, ..., \ \overline{\widetilde{\boldsymbol{X}}}_{(T),H,\cdot}\right]$.

Similarly, we define $\overline{\widetilde{\boldsymbol{V}}}_{(T),h,\cdot}$, $\overline{\widetilde{\boldsymbol{W}}}_{(T),h,\cdot}$, $\overline{\widetilde{\boldsymbol{E}}}_{(T),h,\cdot}$ and $\widetilde{\mathcal{V}}_{(T)}$, $\widetilde{\mathcal{W}}_{(T)}$, $\widetilde{\mathcal{E}}_{(T)}$. Then $\widetilde{\mathcal{V}}_{(T)} = \widetilde{\boldsymbol{B}}_{(T)} \mathcal{Z}$ and $\widetilde{\mathcal{W}}_{(T)} = \widetilde{\boldsymbol{B}}_{(T),\perp} \widetilde{\mathcal{E}}_{(T)}$. We see that $\widetilde{\mathcal{V}}_{(T)}^\top \widetilde{\mathcal{W}}_{(T)} = 0$. Since $\widetilde{\boldsymbol{E}}_{(T)} \sim N(0, \mathbf{I}_{|T|-d})$ and is independent of Y, we know that the entries $\widetilde{\mathcal{E}}_{(T),i,j}$ of $\widetilde{\mathcal{E}}_{(T)}$ are *i.i.d.* samples of $N(0, \frac{1}{n})$.

Define $\widehat{\widetilde{\boldsymbol{\Lambda}}}_{(T)} = \widetilde{\mathcal{X}}_{(T)} \widetilde{\mathcal{X}}_{(T)}^\top$, $\widehat{\widetilde{\boldsymbol{\Lambda}}}_{V,(T)} = \widetilde{\mathcal{V}}_{(T)} \widetilde{\mathcal{V}}_{(T)}^\top$ and $\widehat{\boldsymbol{\Lambda}}_z = \mathcal{Z}\mathcal{Z}^\top$. Recall the definition of the SIR estimate for $\operatorname{Cov}(\mathbb{E}(\boldsymbol{X} \mid Y))$: $\widehat{\boldsymbol{\Lambda}}_H = \mathcal{X}\mathcal{X}^\top$. Then $\boldsymbol{J}_T \widehat{\boldsymbol{\Lambda}}_H \boldsymbol{J}_T^\top = \boldsymbol{\Sigma}_{TT}^{1/2} \widehat{\widetilde{\boldsymbol{\Lambda}}}_{(T)} \boldsymbol{\Sigma}_{TT}^{1/2}$ and $\widehat{\widetilde{\boldsymbol{\Lambda}}}_{V,(T)} = \widetilde{\boldsymbol{B}}_{(T)} \widehat{\boldsymbol{\Lambda}}_z \widetilde{\boldsymbol{B}}_{(T)}^\top$.

We have the following decomposition

$$\widehat{\widetilde{\boldsymbol{\Lambda}}}_{(T)} = \widetilde{\mathcal{V}}_{(T)} \widetilde{\mathcal{V}}_{(T)}^\top + \widetilde{\mathcal{V}}_{(T)} \widetilde{\mathcal{W}}_{(T)}^\top + \widetilde{\mathcal{W}}_{(T)} \widetilde{\mathcal{V}}_{(T)}^\top + \widetilde{\mathcal{W}}_{(T)} \widetilde{\mathcal{W}}_{(T)}^\top \tag{49}$$

Since we have randomly divided the samples into two equal sets of samples, we have the corresponding statistics $\boldsymbol{\Lambda}_H^{(i)}, \widetilde{\boldsymbol{\Lambda}}_{(T)}^{(i)}, \widetilde{\boldsymbol{\Lambda}}_{V,(T)}^{(i)}, \widetilde{\mathcal{V}}_{(T)}^{(i)}, \widetilde{\mathcal{E}}_{(T)}^{(i)}, \widetilde{\mathcal{W}}_{(T)}^{(i)}$, and $\widehat{\boldsymbol{\Lambda}}_z^{(i)}$ for the ith set of samples ($i = 1, 2$) similar to the definition of $\widehat{\boldsymbol{\Lambda}}_H, \widehat{\widetilde{\boldsymbol{\Lambda}}}_{(T)}, \widehat{\widetilde{\boldsymbol{\Lambda}}}_{V,(T)}, \widetilde{\mathcal{V}}_{(T)}, \widetilde{\mathcal{E}}_{(T)}, \widetilde{\mathcal{W}}_{(T)}$, and $\widehat{\boldsymbol{\Lambda}}_z$ respectively.

Finally, we introduce an "oracle estimator", where the word *oracle* suggests this is an estimator only if we know S.

$$\begin{aligned} \widehat{\boldsymbol{B}}_O &:= \arg\max_{\boldsymbol{B}} \operatorname{Tr}(\boldsymbol{B}^\top \boldsymbol{\Lambda}_H^{(1)} \boldsymbol{B}) \\ \text{s.t.} &\quad \boldsymbol{B}^\top \widehat{\boldsymbol{\Sigma}}^{(1)} \boldsymbol{B} = \mathbf{I}_d, \quad \operatorname{supp}(\boldsymbol{B}) = S. \end{aligned} \tag{50}$$

Main part of the proof

For $i = 1, 2$, we define two events:

(i) $\widetilde{\mathtt{E}}_2^{(i)} := \left\{\max_{T \in \mathcal{L}} \left(\|\widetilde{\boldsymbol{\Lambda}}_{V,(T)}^{(i)} - \widetilde{\boldsymbol{\Lambda}}_{(T)}\|\right) \leqslant \frac{2\kappa\lambda}{3\nu}\right\}$,

(ii) $\widetilde{\mathtt{E}}_1^{(1)} = \left\{\ \max_{T \in \mathcal{L}} \left(\|\widetilde{\mathcal{W}}_{(T)}^{(1)} \widetilde{\mathcal{W}}_{(T)}^{(1),\top}\|\right) \leqslant 6 \frac{2s \vee H + s\log(ep/s) + \log(n\lambda)}{n}\ \right\}$.

Furthermore, define $\widetilde{\mathtt{E}} = \widetilde{\mathtt{E}}_1^{(1)} \cap \widetilde{\mathtt{E}}_2^{(1)} \cap \widetilde{\mathtt{E}}_2^{(2)}$.

We apply Lemma 27 to $\sqrt{n} \cdot \widetilde{\mathcal{E}}_{(T)}^{(i)}$ to conclude that

$$\mathbb{P}\left(\|\widetilde{\mathcal{W}}_{(T)}^{(i)} \widetilde{\mathcal{W}}_{(T)}^{(i),\top}\| > 6 \frac{\max(|T|, H) + t}{n}\right) \leqslant 2\exp\left(-t\right),$$

which implies that

$$\mathbb{P}\left(\exists T \in \mathcal{L}, \|\widetilde{\mathcal{W}}_{(T)}^{(i)} \widetilde{\mathcal{W}}_{(T)}^{(i),\top}\| > 6 \frac{\max(|T|, H) + t}{n}\right) \leqslant 2|\mathcal{L}|\exp\left(-t\right).$$

In view of Corollary 2, one has the following analogy.

Corollary 4. For $\nu \in (\kappa, 2\kappa]$, we can find constants C and $\widetilde{C}$, such that if

$$\kappa^2 H^2 \left(\log(nH) + \log\kappa + d\right) < Cn\lambda$$

and $\kappa\left(2s \vee H + s\log(ep/s) + \log(n\lambda)\right) < Cn\lambda$, then $\mathbb{P}\left(\widetilde{\mathtt{E}}^c\right) \leqslant \frac{\widetilde{C}}{n\lambda}$ and on the event $\widetilde{\mathtt{E}}$, the followings hold

a) $\frac{1}{3}\lambda \leqslant \lambda_d(\widehat{\boldsymbol{\Lambda}}_z^{(i)}) \leqslant \lambda_1(\widehat{\boldsymbol{\Lambda}}_z^{(i)}) \leqslant 2\kappa\lambda$.

b) $\|\widetilde{\boldsymbol{\Lambda}}^{(i)}_{(T)} - \widetilde{\boldsymbol{\Lambda}}^{(i)}_{V,(T)}\| \leqslant \lambda\sqrt{18\kappa\frac{2s\vee H + s\log(ep/s)+\log(n\lambda)}{n\lambda}} < \frac{1}{4}\lambda$, for any $T \in \mathcal{L}$.

c) $\lambda_{d+1}(\widetilde{\boldsymbol{\Lambda}}^{(i)}_{(T)}) < \frac{1}{4}\lambda$, for any $T \in \mathcal{L}$.

d) $\|\widetilde{\boldsymbol{\Lambda}}^{(i)}_{(T)}\| \leqslant 3\kappa\lambda$.

Proof of d):

Proof. Since $\widetilde{\boldsymbol{\Lambda}}^{(i)}_{V,(T)} = \widetilde{\boldsymbol{B}}^{(i)}_{(T)}\widehat{\boldsymbol{\Lambda}}_z^{(i)}\widetilde{\boldsymbol{B}}^{\top}_{(T)}$ and $\widetilde{\boldsymbol{B}}^{\top}_{(T)}\widetilde{\boldsymbol{B}}_{(T)} = \mathbf{I}_d$, one has:

$$\lambda_1(\widetilde{\boldsymbol{\Lambda}}^{(i)}_{V,(T)}) = \lambda_1(\widetilde{\boldsymbol{B}}_{(T)}\widehat{\boldsymbol{\Lambda}}_z^{(i)}\widetilde{\boldsymbol{B}}^{\top}_{(T)}) = \lambda_1(\widetilde{\boldsymbol{B}}^{\top}_{(T)}\widetilde{\boldsymbol{B}}_{(T)}\widehat{\boldsymbol{\Lambda}}_z^{(i)}) = \lambda_1(\widehat{\boldsymbol{\Lambda}}_z^{(i)}).$$

Combining a) and b) leads to that

$$\|\widetilde{\boldsymbol{\Lambda}}^{(i)}_{(T)}\| \leqslant \|\widetilde{\boldsymbol{\Lambda}}^{(i)}_{(T)} - \widetilde{\boldsymbol{\Lambda}}^{(i)}_{V,(T)}\| + \lambda_1(\widehat{\boldsymbol{\Lambda}}_z^{(i)}) \leqslant 3\kappa\lambda.$$

□

Let $\widehat{\boldsymbol{\Sigma}}_{TT} = \boldsymbol{J}_T\widehat{\boldsymbol{\Sigma}}^{(1)}\boldsymbol{J}_T^{\top}$, and $\widehat{\mathbf{I}}_T = \widehat{\boldsymbol{\Sigma}}_{TT}^{-1/2}\boldsymbol{\Sigma}_{TT}^{1/2}$.

Lemma 9. *There exist constants C and $\widetilde{C}$, such that if $s\log(ep/s) + \log(n\lambda) < Cn$, then it holds with probability at least $1 - \widetilde{C}/n\lambda$ that for all $T \in \mathcal{L}$, $\|\widehat{\mathbf{I}}_T\|^2 < 2M^2$, $\|\widehat{\boldsymbol{\Sigma}}_{TT}^{-1/2}\|^2 < 2M$ and $\|\widehat{\mathbf{I}}_T - \mathbf{I}_{|T|}\|^2 < \widetilde{C}\frac{s\log(ep/s)+\log(n\lambda)}{n}$.*

Proof. The proof follows the same argument in Lemma 8 by choosing $t \asymp \sqrt{2[s\log(ep/s) + \log(n\lambda)]/n}$ and is omitted. □

Let $\mathtt{E}$ be the intersection of $\tilde{\mathtt{E}}$ and the events in Lemma 9. On the event $\mathtt{E}$, the results stated in Corollary 4 and Lemma 9 uniformly hold for $T \in \mathcal{L}$, in particular for $\mathrm{supp}(\widehat{\boldsymbol{B}}) \cup S$.

In the following, we set T to be the random element $\mathrm{supp}(\widehat{\boldsymbol{B}}) \cup S$. Furthermore, we abbreviate $\widetilde{\boldsymbol{B}}_{(T)}$ by $\boldsymbol{F}$, i.e. $\boldsymbol{F} := \boldsymbol{\Sigma}_{TT}^{1/2}\boldsymbol{J}_T\boldsymbol{B}$. Similarly, we define $\widehat{\boldsymbol{F}}_O = \widehat{\boldsymbol{\Sigma}}_{TT}^{1/2}\boldsymbol{J}_T\widehat{\boldsymbol{B}}_O$ and $\widehat{\boldsymbol{F}} = \widehat{\boldsymbol{\Sigma}}_{TT}^{1/2}\boldsymbol{J}_T\widehat{\boldsymbol{B}}$. Then $\boldsymbol{F}$, $\widehat{\boldsymbol{F}}_O$ and $\widehat{\boldsymbol{F}}$ are all in $\mathbb{O}(|T|, d)$.

Let $\widehat{\boldsymbol{F}}_O^{\top}\boldsymbol{F} = \boldsymbol{U}_1\Delta\boldsymbol{U}_2^{\top}$ be the singular value decomposition of $\widehat{\boldsymbol{F}}_O^{\top}\boldsymbol{F}$ such that $\boldsymbol{U}_i \in \mathbb{O}(d,d)$ and Δ is a $d \times d$ diagonal matrix with non-negative entries. Let $\boldsymbol{M} := \boldsymbol{U}_2^{\top}\widehat{\boldsymbol{\Lambda}}_z^{(2)}\boldsymbol{U}_2$.

By the definition of $\widehat{\boldsymbol{B}}$ and $\boldsymbol{\Lambda}_H^{(2)}$ (the SIR estimator of $\boldsymbol{\Lambda}$ based on the second set of samples), one has $0 \geqslant \langle\boldsymbol{\Lambda}_H^{(2)}, \widehat{\boldsymbol{B}}_O\widehat{\boldsymbol{B}}_O^{\top} - \widehat{\boldsymbol{B}}\widehat{\boldsymbol{B}}^{\top}\rangle = \langle\boldsymbol{\Lambda}_H^{(2)}, \boldsymbol{J}_T^{\top}\boldsymbol{J}_T(\widehat{\boldsymbol{B}}_O\widehat{\boldsymbol{B}}_O^{\top} - \widehat{\boldsymbol{B}}\widehat{\boldsymbol{B}}^{\top})\boldsymbol{J}_T^{\top}\boldsymbol{J}_T\rangle = \langle\boldsymbol{J}_T\boldsymbol{\Lambda}_H^{(2)}\boldsymbol{J}_T^{\top}, \boldsymbol{J}_T(\widehat{\boldsymbol{B}}_O\widehat{\boldsymbol{B}}_O^{\top} - \widehat{\boldsymbol{B}}\widehat{\boldsymbol{B}}^{\top})\boldsymbol{J}_T^{\top}\rangle$ where the first equality comes from the fact that $\mathrm{supp}(\widehat{\boldsymbol{B}}) \cup \mathrm{supp}(\widehat{\boldsymbol{B}}_O) = T$.

Applying the Lemma 32 with the positive definite matrix $\boldsymbol{U}_1\boldsymbol{M}\boldsymbol{U}_1^{\top}$, one has

$$\begin{aligned}
\frac{\lambda_d(\widehat{\boldsymbol{\Lambda}}_z^{(2)})}{2}\|\widehat{\boldsymbol{F}}\widehat{\boldsymbol{F}}^{\top} - \widehat{\boldsymbol{F}}_O\widehat{\boldsymbol{F}}_O^{\top}\|_F^2 \leqslant & \langle\widehat{\boldsymbol{F}}_O\boldsymbol{U}_1\boldsymbol{M}\boldsymbol{U}_1^{\top}\widehat{\boldsymbol{F}}_O^{\top}, \widehat{\boldsymbol{F}}_O\widehat{\boldsymbol{F}}_O^{\top} - \widehat{\boldsymbol{F}}\widehat{\boldsymbol{F}}^{\top}\rangle \\
= & \langle\widehat{\boldsymbol{\Sigma}}_{TT}^{1/2}\widehat{\boldsymbol{F}}_O\boldsymbol{U}_1\boldsymbol{M}\boldsymbol{U}_1^{\top}\widehat{\boldsymbol{F}}_O^{\top}\widehat{\boldsymbol{\Sigma}}_{TT}^{1/2}, \boldsymbol{J}_T(\widehat{\boldsymbol{B}}_O\widehat{\boldsymbol{B}}_O^{\top} - \widehat{\boldsymbol{B}}\widehat{\boldsymbol{B}}^{\top})\boldsymbol{J}_T^{\top}\rangle \\
\leqslant & \langle\widehat{\boldsymbol{\Sigma}}_{TT}^{1/2}\widehat{\boldsymbol{F}}_O\boldsymbol{U}_1\boldsymbol{M}\boldsymbol{U}_1^{\top}\widehat{\boldsymbol{F}}_O^{\top}\widehat{\boldsymbol{\Sigma}}_{TT}^{1/2} - \boldsymbol{J}_T\boldsymbol{\Lambda}_H^{(2)}\boldsymbol{J}_T^{\top}, \boldsymbol{J}_T(\widehat{\boldsymbol{B}}_O\widehat{\boldsymbol{B}}_O^{\top} - \widehat{\boldsymbol{B}}\widehat{\boldsymbol{B}}^{\top})\boldsymbol{J}_T^{\top}\rangle
\end{aligned}$$

$$:= \mathrm{I} + \mathrm{II} \tag{51}$$

where

$$\begin{aligned}
\mathrm{I} =& \langle \widehat{\boldsymbol{\Sigma}}_{TT}^{1/2} \widehat{\boldsymbol{F}}_O \boldsymbol{U}_1 \boldsymbol{M} \boldsymbol{U}_1^\top \widehat{\boldsymbol{F}}_O^\top \widehat{\boldsymbol{\Sigma}}_{TT}^{1/2} - \widehat{\boldsymbol{\Sigma}}_{TT}^{1/2} \widetilde{\boldsymbol{\Lambda}}_{V,(T)}^{(2)} \widehat{\boldsymbol{\Sigma}}_{TT}^{1/2}, \boldsymbol{J}_T (\widehat{\boldsymbol{B}}_O \widehat{\boldsymbol{B}}_O^\top - \widehat{\boldsymbol{B}} \widehat{\boldsymbol{B}}^\top) \boldsymbol{J}_T^\top \rangle \\
\mathrm{II} =& \langle \widehat{\boldsymbol{\Sigma}}_{TT}^{1/2} \widetilde{\boldsymbol{\Lambda}}_{V,(T)}^{(2)} \widehat{\boldsymbol{\Sigma}}_{TT}^{1/2} - \boldsymbol{\Sigma}_{TT}^{1/2} \widetilde{\boldsymbol{\Lambda}}_{(T)}^{(2)} \boldsymbol{\Sigma}_{TT}^{1/2}, \boldsymbol{J}_T (\widehat{\boldsymbol{B}}_O \widehat{\boldsymbol{B}}_O^\top - \widehat{\boldsymbol{B}} \widehat{\boldsymbol{B}}^\top) \boldsymbol{J}_T^\top \rangle .
\end{aligned}$$

The last inequality holds because $\boldsymbol{J}_T \boldsymbol{\Lambda}_H^{(2)} \boldsymbol{J}_T^\top = \boldsymbol{\Sigma}_{TT}^{1/2} \widetilde{\boldsymbol{\Lambda}}_{(T)}^{(2)} \boldsymbol{\Sigma}_{TT}^{1/2}$.

For I:

We first rewrite I:

$$\begin{aligned}
\mathrm{I} =& \langle \widehat{\boldsymbol{\Sigma}}_{TT}^{1/2} \widehat{\boldsymbol{F}}_O \boldsymbol{U}_1 \boldsymbol{M} \boldsymbol{U}_1^\top \widehat{\boldsymbol{F}}_O^\top \widehat{\boldsymbol{\Sigma}}_{TT}^{1/2} - \widehat{\boldsymbol{\Sigma}}_{TT}^{1/2} \widetilde{\boldsymbol{\Lambda}}_{V,(T)}^{(2)} \widehat{\boldsymbol{\Sigma}}_{TT}^{1/2}, \boldsymbol{J}_T (\widehat{\boldsymbol{B}}_O \widehat{\boldsymbol{B}}_O^\top - \widehat{\boldsymbol{B}} \widehat{\boldsymbol{B}}^\top) \boldsymbol{J}_T^\top \rangle \\
=& \langle \widehat{\boldsymbol{F}}_O \boldsymbol{U}_1 \boldsymbol{M} \boldsymbol{U}_1^\top \widehat{\boldsymbol{F}}_O^\top - \widetilde{\boldsymbol{\Lambda}}_{V,(T)}^{(2)}, \widehat{\boldsymbol{F}}_O \widehat{\boldsymbol{F}}_O^\top - \widehat{\boldsymbol{F}} \widehat{\boldsymbol{F}}^\top \rangle .
\end{aligned}$$

Note that $\boldsymbol{M} = \boldsymbol{U}_2^\top \widehat{\boldsymbol{\Lambda}}_z^{(2)} \boldsymbol{U}_2$, so on the event $\mathtt{E}$, the eigenvalues of $\boldsymbol{M}$ are in $(\frac{1}{3}\lambda, 2\kappa\lambda)$. Following the same proof of (32), one has $1_{\mathtt{E}} |\mathrm{I}| \leqslant C\lambda \|\widehat{\boldsymbol{F}}_O \widehat{\boldsymbol{F}}_O^\top - \boldsymbol{F}\boldsymbol{F}^\top\|_F \|\widehat{\boldsymbol{F}}_O \widehat{\boldsymbol{F}}_O^\top - \widehat{\boldsymbol{F}} \widehat{\boldsymbol{F}}^\top\|_F$.

For II: Recall that $\widehat{\mathbf{I}}_T = \widehat{\boldsymbol{\Sigma}}_{TT}^{-1/2} \boldsymbol{\Sigma}_{TT}^{1/2}$.

$$\begin{aligned}
\mathrm{II} =& \langle \widehat{\boldsymbol{\Sigma}}_{TT}^{1/2} \widetilde{\boldsymbol{\Lambda}}_{V,(T)}^{(2)} \widehat{\boldsymbol{\Sigma}}_{TT}^{1/2} - \boldsymbol{\Sigma}_{TT}^{1/2} \widetilde{\boldsymbol{\Lambda}}_{(T)}^{(2)} \boldsymbol{\Sigma}_{TT}^{1/2}, \boldsymbol{J}_T (\widehat{\boldsymbol{B}}_O \widehat{\boldsymbol{B}}_O^\top - \widehat{\boldsymbol{B}} \widehat{\boldsymbol{B}}^\top) \boldsymbol{J}_T^\top \rangle \\
=& \langle \widetilde{\boldsymbol{\Lambda}}_{V,(T)}^{(2)} - \widehat{\boldsymbol{\Sigma}}_{TT}^{-1/2} \boldsymbol{\Sigma}_{TT}^{1/2} \widetilde{\boldsymbol{\Lambda}}_{(T)}^{(2)} \boldsymbol{\Sigma}_{TT}^{1/2} \widehat{\boldsymbol{\Sigma}}_{TT}^{-1/2}, \widehat{\boldsymbol{\Sigma}}_{TT}^{1/2} \boldsymbol{J}_T (\widehat{\boldsymbol{B}}_O \widehat{\boldsymbol{B}}_O^\top - \widehat{\boldsymbol{B}} \widehat{\boldsymbol{B}}^\top) \boldsymbol{J}_T^\top \widehat{\boldsymbol{\Sigma}}_{TT}^{1/2} \rangle \\
=& \langle \widetilde{\boldsymbol{\Lambda}}_{V,(T)}^{(2)} - \widehat{\mathbf{I}}_T \widetilde{\boldsymbol{\Lambda}}_{(T)}^{(2)} \widehat{\mathbf{I}}_T^\top, \widehat{\boldsymbol{F}}_O \widehat{\boldsymbol{F}}_O^\top - \widehat{\boldsymbol{F}} \widehat{\boldsymbol{F}}^\top \rangle .
\end{aligned}$$

We can bound the last expression using the next lemma, whose proof is deferred to the end of this section.

Lemma 10. *If $\lambda \leqslant \varpi_d \leqslant \frac{1}{d}$ and $n\lambda \leqslant e^s$, then on the event $\mathtt{E}$, we have:*

$$\langle \widetilde{\boldsymbol{\Lambda}}_{V,(T)}^{(2)} - \widehat{\mathbf{I}}_T \widetilde{\boldsymbol{\Lambda}}_{(T)}^{(2)} \widehat{\mathbf{I}}_T^\top, \widehat{\boldsymbol{F}}_O \widehat{\boldsymbol{F}}_O^\top - \widehat{\boldsymbol{F}} \widehat{\boldsymbol{F}}^\top \rangle \lesssim \lambda \epsilon_n \|\widehat{\boldsymbol{F}}_O \widehat{\boldsymbol{F}}_O^\top - \widehat{\boldsymbol{F}} \widehat{\boldsymbol{F}}^\top\|_F . \tag{52}$$

On the event $\mathtt{E}$, $\lambda_d(\widehat{\boldsymbol{\Lambda}}_z^{(2)}) \geqslant \lambda/3$. Equation (51) leads to

$$\|\widehat{\boldsymbol{F}} \widehat{\boldsymbol{F}}^\top - \widehat{\boldsymbol{F}}_O \widehat{\boldsymbol{F}}_O^\top\|_F^2 \lesssim \left(\|\widehat{\boldsymbol{F}}_O \widehat{\boldsymbol{F}}_O^\top - \boldsymbol{F}\boldsymbol{F}^\top\|_F + \sqrt{\epsilon_n^2} \right) \|\widehat{\boldsymbol{F}} \widehat{\boldsymbol{F}}^\top - \widehat{\boldsymbol{F}}_O \widehat{\boldsymbol{F}}_O^\top\|_F ,$$

which yields

$$\|\widehat{\boldsymbol{F}} \widehat{\boldsymbol{F}}^\top - \widehat{\boldsymbol{F}}_O \widehat{\boldsymbol{F}}_O^\top\|_F \lesssim \|\widehat{\boldsymbol{F}}_O \widehat{\boldsymbol{F}}_O^\top - \boldsymbol{F}\boldsymbol{F}^\top\|_F + \sqrt{\epsilon_n^2} . \tag{53}$$

By triangle inequality,

$$\|\widehat{\boldsymbol{F}} \widehat{\boldsymbol{F}}^\top - \boldsymbol{F}\boldsymbol{F}\|_F \leqslant \|\widehat{\boldsymbol{F}} \widehat{\boldsymbol{F}}^\top - \widehat{\boldsymbol{F}}_O \widehat{\boldsymbol{F}}_O^\top\|_F + \|\widehat{\boldsymbol{F}}_O \widehat{\boldsymbol{F}}_O^\top - \boldsymbol{F}\boldsymbol{F}^\top\|_F ,$$

and thus on the event $\mathtt{E}$,

$$\|\widehat{\boldsymbol{F}} \widehat{\boldsymbol{F}}^\top - \boldsymbol{F}\boldsymbol{F}\|_F^2 \lesssim \|\widehat{\boldsymbol{F}}_O \widehat{\boldsymbol{F}}_O^\top - \boldsymbol{F}\boldsymbol{F}^\top\|_F^2 + \epsilon_n^2$$

Following the proof of Equations (48) in Section D.1, there exists a set $\mathtt{E}_O$ such that $\mathbb{P}(\mathtt{E}_O^c) \lesssim \frac{1}{n\lambda}$ and

$$\mathbb{E}\left(1_{\mathtt{E}_O} \|\widehat{\boldsymbol{F}}_O \widehat{\boldsymbol{F}}_O^\top - \boldsymbol{F}\boldsymbol{F}^\top\|_F^2 \right) \lesssim \epsilon_n^2 , \tag{54}$$

$$\mathbb{E}\left(1_{\mathtt{E}_O} \|\widehat{\boldsymbol{B}}_O \widehat{\boldsymbol{B}}_O^\top - \boldsymbol{B}\boldsymbol{B}^\top\|_F^2 \right) \lesssim \epsilon_n^2 . \tag{55}$$

Therefore

$$\mathbb{E}\left(\ 1_{\mathtt{E}\cap\mathtt{E}_O}\|\widehat{\boldsymbol{F}}\widehat{\boldsymbol{F}}^\top - \boldsymbol{F}\boldsymbol{F}^\top\|_F^2\right) \lesssim \epsilon_n^2.$$

On the event $\mathtt{E}$, by Lemma 9 and $\|\widehat{\boldsymbol{\Sigma}}_{TT}^{-1/2}\|^2 < 2M$, one has

$$\begin{aligned}
&\|\widehat{\boldsymbol{B}}\widehat{\boldsymbol{B}}^\top - \widehat{\boldsymbol{B}}_O\widehat{\boldsymbol{B}}_O^\top\|_F \\
=&\|\boldsymbol{J}_T\left(\widehat{\boldsymbol{B}}\widehat{\boldsymbol{B}}^\top - \widehat{\boldsymbol{B}}_O\widehat{\boldsymbol{B}}_O^\top\right)\boldsymbol{J}_T^\top\|_F \\
\leqslant&\|\widehat{\boldsymbol{\Sigma}}_{TT}^{-1/2}\|^2\|\widehat{\boldsymbol{\Sigma}}_{TT}^{1/2}\boldsymbol{J}_T\left(\widehat{\boldsymbol{B}}\widehat{\boldsymbol{B}}^\top - \widehat{\boldsymbol{B}}_O\widehat{\boldsymbol{B}}_O^\top\right)\boldsymbol{J}_T^\top\widehat{\boldsymbol{\Sigma}}_{TT}^{1/2}\|_F \\
=&\|\widehat{\boldsymbol{\Sigma}}_{TT}^{-1/2}\|^2\|\widehat{\boldsymbol{F}}\widehat{\boldsymbol{F}}^\top - \widehat{\boldsymbol{F}}_O\widehat{\boldsymbol{F}}_O^\top\|_F \\
<&2M\left(\|\widehat{\boldsymbol{F}}_O\widehat{\boldsymbol{F}}_O^\top - \boldsymbol{F}\boldsymbol{F}^\top\|_F + \sqrt{\epsilon_n^2}\right),
\end{aligned} \tag{56}$$

where the first equation is due to the definition of T, the first inequality is due to Lemma 25 and the last inequality is because Equation (53).

By triangle inequality, one has $\|\widehat{\boldsymbol{B}}\widehat{\boldsymbol{B}}^\top - \boldsymbol{B}\boldsymbol{B}^\top\|_F \leqslant \|\widehat{\boldsymbol{B}}\widehat{\boldsymbol{B}}^\top - \widehat{\boldsymbol{B}}_O\widehat{\boldsymbol{B}}_O^\top\|_F + \|\widehat{\boldsymbol{B}}_O\widehat{\boldsymbol{B}}_O^\top - \boldsymbol{B}\boldsymbol{B}^\top\|_F$. This, together with Equations (54) to (56), yields

$$\mathbb{E}\left(\ 1_{\mathtt{E}\cap\mathtt{E}_O}\|\widehat{\boldsymbol{B}}\widehat{\boldsymbol{B}}^\top - \boldsymbol{B}\boldsymbol{B}^\top\|_F^2\right) \lesssim \epsilon_n^2.$$

Since $\mathbb{P}(\mathtt{E}^c \cup \mathtt{E}_O^c) \lesssim \frac{1}{n\lambda}$, one has

$$\mathbb{E}\|\widehat{\boldsymbol{B}}\widehat{\boldsymbol{B}}^\top - \boldsymbol{B}\boldsymbol{B}^\top\|_F^2 \lesssim \epsilon_n^2.$$

Proof of Lemma 10. First, we have

$$\begin{aligned}
\widetilde{\boldsymbol{\Lambda}}_{V,(T)}^{(2)} - \widehat{\mathbf{I}}_T\widetilde{\boldsymbol{\Lambda}}_{(T)}^{(2)}\widehat{\mathbf{I}}_T^\top =&\widetilde{\mathcal{V}}_{(T)}^{(2)}\widetilde{\mathcal{V}}_{(T)}^{(2),\top} - \widehat{\mathbf{I}}_T\widetilde{\mathcal{V}}_{(T)}^{(2)}\widetilde{\mathcal{V}}_{(T)}^{(2),\top}\widehat{\mathbf{I}}_T^\top - \\
&\left(\widehat{\mathbf{I}}_T\widetilde{\mathcal{V}}_{(T)}^{(2)}\widetilde{\mathcal{W}}_{(T)}^{(2),\top}\widehat{\mathbf{I}}_T^\top + \widehat{\mathbf{I}}_T\widetilde{\mathcal{W}}_{(T)}^{(2)}\widetilde{\mathcal{V}}_{(T)}^{(2),\top}\widehat{\mathbf{I}}_T^\top + \widehat{\mathbf{I}}_T\widetilde{\mathcal{W}}_{(T)}^{(2)}\widetilde{\mathcal{W}}_{(T)}^{(2),\top}\widehat{\mathbf{I}}_T^\top\right).
\end{aligned}$$

Following the same proof of Lemma 7 and notice that $1_{\mathtt{E}}\|\widehat{\mathbf{I}}_T\|^2 < 2M^2$, we have:

$$\langle\widehat{\mathbf{I}}_T\widetilde{\mathcal{V}}_{(T)}^{(2)}\widetilde{\mathcal{W}}_{(T)}^{(2),\top}\widehat{\mathbf{I}}_T^\top + \widehat{\mathbf{I}}_T\widetilde{\mathcal{W}}_{(T)}^{(2)}\widetilde{\mathcal{V}}_{(T)}^{(2),\top}\widehat{\mathbf{I}}_T^\top + \widehat{\mathbf{I}}_T\widetilde{\mathcal{W}}_{(T)}^{(2)}\widetilde{\mathcal{W}}_{(T)}^{(2),\top}\widehat{\mathbf{I}}_T^\top, \widehat{\boldsymbol{F}}_O\widehat{\boldsymbol{F}}_O^\top - \widehat{\boldsymbol{F}}\widehat{\boldsymbol{F}}^\top\rangle \lesssim \lambda\epsilon_n\|\widehat{\boldsymbol{F}}_O\widehat{\boldsymbol{F}}_O^\top - \widehat{\boldsymbol{F}}\widehat{\boldsymbol{F}}^\top\|_F.$$

Then we only need to show that

$$\langle\widetilde{\mathcal{V}}_{(T)}^{(2)}\widetilde{\mathcal{V}}_{(T)}^{(2),\top} - \widehat{\mathbf{I}}_T\widetilde{\mathcal{V}}_{(T)}^{(2)}\widetilde{\mathcal{V}}_{(T)}^{(2),\top}\widehat{\mathbf{I}}_T^\top, \widehat{\boldsymbol{F}}_O\widehat{\boldsymbol{F}}_O^\top - \widehat{\boldsymbol{F}}\widehat{\boldsymbol{F}}^\top\rangle \lesssim \lambda\epsilon_n\|\widehat{\boldsymbol{F}}_O\widehat{\boldsymbol{F}}_O^\top - \widehat{\boldsymbol{F}}\widehat{\boldsymbol{F}}^\top\|_F.$$

Using the inequality that $|\mathrm{Tr}(\boldsymbol{A})| \leqslant \sqrt{\mathrm{rank}(\boldsymbol{A})}\|\boldsymbol{A}\|_F$ and Lemma 25, we have

$$\begin{aligned}
&\langle\widetilde{\mathcal{V}}_{(T)}^{(2)}\widetilde{\mathcal{V}}_{(T)}^{(2),\top} - \widehat{\mathbf{I}}_T\widetilde{\mathcal{V}}_{(T)}^{(2)}\widetilde{\mathcal{V}}_{(T)}^{(2),\top}\widehat{\mathbf{I}}_T^\top, \widehat{\boldsymbol{F}}_O\widehat{\boldsymbol{F}}_O^\top - \widehat{\boldsymbol{F}}\widehat{\boldsymbol{F}}^\top\rangle^2 \\
\leqslant&d\|\left(\widehat{\boldsymbol{F}}_O\widehat{\boldsymbol{F}}_O^\top - \widehat{\boldsymbol{F}}\widehat{\boldsymbol{F}}^\top\right)\left(\widetilde{\mathcal{V}}_{(T)}^{(2)}\widetilde{\mathcal{V}}_{(T)}^{(2),\top} - \widehat{\mathbf{I}}_T\widetilde{\mathcal{V}}_{(T)}^{(2)}\widetilde{\mathcal{V}}_{(T)}^{(2),\top}\widehat{\mathbf{I}}_T^\top\right)\|_F^2 \\
\leqslant&d\|\widehat{\boldsymbol{F}}_O\widehat{\boldsymbol{F}}_O^\top - \widehat{\boldsymbol{F}}\widehat{\boldsymbol{F}}^\top\|_F^2\|\widetilde{\mathcal{V}}_{(T)}^{(2)}\widetilde{\mathcal{V}}_{(T)}^{(2),\top} - \widehat{\mathbf{I}}_T\widetilde{\mathcal{V}}_{(T)}^{(2)}\widetilde{\mathcal{V}}_{(T)}^{(2),\top}\widehat{\mathbf{I}}_T^\top\|^2 \\
\leqslant&d\|\widehat{\boldsymbol{F}}_O\widehat{\boldsymbol{F}}_O^\top - \widehat{\boldsymbol{F}}\widehat{\boldsymbol{F}}^\top\|_F^2(\|\widehat{\mathbf{I}}_T\| + 1)^2\|\widetilde{\mathcal{V}}_{(T)}^{(2)}\widetilde{\mathcal{V}}_{(T)}^{(2),\top}\|^2\|\widehat{\mathbf{I}}_T - \mathbf{I}_{|T|}\|^2,
\end{aligned} \tag{57}$$

where the last inequality is due to the following fact: for any two $m \times m$ matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, we can write $\boldsymbol{A} - \boldsymbol{B}\boldsymbol{A}\boldsymbol{B}^\top = (\boldsymbol{I} - \boldsymbol{B})\boldsymbol{A} + \boldsymbol{B}\boldsymbol{A}(\boldsymbol{I} - \boldsymbol{B})^\top$ and conclude the inequality that $\|\boldsymbol{A} - \boldsymbol{B}\boldsymbol{A}\boldsymbol{B}^\top\| \leqslant \|\boldsymbol{I} - \boldsymbol{B}\|\|\boldsymbol{A}\| + \|\boldsymbol{B}\|\|\boldsymbol{A}\|\|\boldsymbol{I} - \boldsymbol{B}\| = (\|\boldsymbol{B}\| + 1)\|\boldsymbol{A}\|\|\boldsymbol{I} - \boldsymbol{B}\|$.

By Corollary 4 and Lemma 9, we know that

$$\|\widehat{\mathbf{I}}_T\| \lesssim 1; \quad \|\widetilde{\mathcal{V}}^{(2)}_{(T)}\widetilde{\mathcal{V}}^{(2),\top}_{(T)}\| \lesssim \lambda; \quad \|\widehat{\mathbf{I}}_T - \mathbf{I}_{|T|}\|^2 \lesssim \frac{s\log(ep/s)+\log(n\lambda)}{n}.$$

Insert these equations into (57) and use the conditions that $\lambda \leqslant \frac{1}{d}$ and $n\lambda \leqslant e^s$, we have

$$\begin{aligned}
&\langle \widetilde{\mathcal{V}}^{(2)}_{(T)}\widetilde{\mathcal{V}}^{(2),\top}_{(T)} - \widehat{\mathbf{I}}_T\widetilde{\mathcal{V}}^{(2)}_{(T)}\widetilde{\mathcal{V}}^{(2),\top}_{(T)}\widehat{\mathbf{I}}_T^\top, \widehat{\boldsymbol{F}}_O\widehat{\boldsymbol{F}}_O^\top - \widehat{\boldsymbol{F}}\widehat{\boldsymbol{F}}^\top\rangle^2 \\
\lesssim & d\lambda^2\frac{s\log(ep/s)+\log(n\lambda)}{n}\|\widehat{\boldsymbol{F}}_O\widehat{\boldsymbol{F}}_O^\top - \widehat{\boldsymbol{F}}\widehat{\boldsymbol{F}}^\top\|_F^2 \\
\lesssim & \lambda^2\epsilon_n^2\|\widehat{\boldsymbol{F}}_O\widehat{\boldsymbol{F}}_O^\top - \widehat{\boldsymbol{F}}\widehat{\boldsymbol{F}}^\top\|_F^2.
\end{aligned}$$

□

E Proof of Theorem 4

We follow the the standard procedure of applying Fano's inequality to obtain the minimax lower bound. The following lemma is one version of the generalized Fano method.

Lemma 11 (Yu [1997])**.** *Let $N \geqslant 2$ be an integer and $\{\theta_1, \ldots, \theta_N\} \subset \Theta_0$ index a collection of probability measures $\mathbb{P}_{\theta_i}$ on a measurable space $(\mathcal{X}, \mathcal{A})$. Let ρ be a pseudometric on Θ_0 and suppose that for all $i \neq j$*

$$\rho(\theta_i, \theta_j) \geqslant \alpha_N, \quad \text{and} \quad KL(\mathbb{P}_{\theta_i}, \mathbb{P}_{\theta_j}) \leqslant \beta_N.$$

Then every $\mathcal{A}$-measurable estimator $\hat{\theta}$ satisfies

$$\max_i \mathbb{E}\rho(\hat{\theta}, \theta_i) \geqslant \frac{\alpha_N}{2}\left(1 - \frac{\beta_N + \log 2}{\log N}\right).$$

We first introduce the following packing set.

Lemma 12 (Packing Set)**.** *For any $\varepsilon \in (0, \sqrt{2(d \wedge (p-d))}]$ and any $\alpha \in (0,1)$, there exists a subset $\Theta \subset \mathbb{O}(p,d)$ such that*

$$|\Theta| \geqslant \left(\frac{c_0}{\alpha}\right)^{d(p-d)},$$

and for any $\boldsymbol{B}, \widetilde{\boldsymbol{B}} \in \Theta$,

$$\|\boldsymbol{B} - \widetilde{\boldsymbol{B}}\|_F \leqslant 2\varepsilon, \qquad \|\boldsymbol{B}\boldsymbol{B}^\top - \widetilde{\boldsymbol{B}}\widetilde{\boldsymbol{B}}^\top\|_F \geqslant \alpha\varepsilon,$$

where c_0 is an absolute constant.

Recall the Definition 4 that $\psi(z_1, \ldots, z_d)$ equals to the index i of the largest absolute values of the coordinates multiplied by $\text{sgn}(z_i)$ if $\|\boldsymbol{z}\|^2$ is less than m_d the median of χ^2_d the chi-squared distribution with d degrees of freedom and $A_i = \psi^{-1}(i)$ for all i in $\{\pm 1, \ldots, \pm d\}$, or more explicitly,

$$A_i \;=\; \{\boldsymbol{z} \in \mathbb{R}^d : \|\boldsymbol{z}\|^2 \leqslant m_d, \text{sgn}(z_{|i|}) = \text{sgn}(i), \text{ and } |z_{|i|}| > |z_j|, \forall j \neq i\}.$$

Essentially, the ball centered at the original with radius $\sqrt{m_d}$ in $\mathbb{R}^d$ is partitioned into $2d$ disjoint parts A_i's that have the same shape. For our later convenience, we define $A_0 = \{z \in \mathbb{R}^d : \|z\|^2 > m_d\}$ the complement of the ball. We have $\mathbb{P}(Z \in A_0) = \mathbb{P}(\psi(Z) = 0)$.

Define

$$\lambda_{0,d} := (2d)^{-1}\mathbb{E}\left(Z_1 \mid \boldsymbol{Z} \in A_1\right)^2, \text{ where } \boldsymbol{Z} \sim N(0, \boldsymbol{I}_d).$$

This number will be used in the following two propositions, which show that the joint distribution $(\boldsymbol{X}, Y)$ in (10) enjoys the desired properties (i) and (ii) stated in Section 4.2.1.

Proposition 2. *For any $\boldsymbol{B} \in \mathbb{O}(p,d)$ and $\mathbb{P}_{\boldsymbol{B}}$ constructed by Equation (10), Y can be represented as $f(\boldsymbol{B}^\top \boldsymbol{X}, \epsilon)$ for $\epsilon \sim N(0,1)$. Furthermore, $f(\boldsymbol{B}^\top \boldsymbol{X}, \epsilon)$ belongs to the class $\mathcal{F}_d(\lambda, \kappa, 8d)$ for any $\kappa \geqslant 1$ and $\lambda = \rho^2 \lambda_{0,d}$.*

Proposition 3. *Suppose $\boldsymbol{B}$ and $\widetilde{\boldsymbol{B}}$ are in $\mathbb{O}(p,d)$. Let $\mathbb{P}_{\boldsymbol{B}}$ and $\mathbb{P}_{\widetilde{\boldsymbol{B}}}$ be defined by Equation (10).*

1. *For any $\rho \in (0,1)$, it holds that*

$$KL(\mathbb{P}_{\boldsymbol{B}}, \mathbb{P}_{\widetilde{\boldsymbol{B}}}) \leqslant \frac{\rho^2}{2(1-\rho^2)} \|\boldsymbol{B} - \widetilde{\boldsymbol{B}}\|_F^2.$$

2. *There exist a universal constant C and a constant $\delta_d = \Theta(d^{-7.1/2})$ such that for any $\rho \in (0, \delta_d]$, it holds that*

$$KL(\mathbb{P}_{\boldsymbol{B}}, \mathbb{P}_{\widetilde{\boldsymbol{B}}}) \leqslant \frac{C\rho^2}{1-\rho^2} \lambda_{0,d} \|\boldsymbol{B} - \widetilde{\boldsymbol{B}}\|_F^2.$$

The proofs of Lemma 12, Propositions 2 and 3 will be given in the subsequent subsections. We can now prove Theorem 4.

Fix any $\alpha \in (0,1)$ (e.g. $\alpha = 1/2$) and take Θ to be the subset in Lemma 12. For each $\boldsymbol{B} \in \Theta$, define $\mathbb{P}_{\boldsymbol{B}}$ by Equation (10). Proposition 2 guarantees that $\mathbb{P}_{\boldsymbol{B}} \in \mathfrak{M}(p,d,\lambda)$. Denoted by $\varpi_d := \min(\delta_d^2, 1/2)\lambda_{0,d}$. Suppose $\lambda \leqslant \varpi_d$. Let $\rho = \sqrt{\lambda/\lambda_{0,d}}$. Then $1/(1-\rho^2) \leqslant 2$ and we can apply the second statement of Proposition 3 to bound the KL-divergence between each pairs of different populations $\mathbb{P}^n_{\boldsymbol{B}}$ and $\mathbb{P}^n_{\widetilde{\boldsymbol{B}}}$ for $\boldsymbol{B}, \widetilde{\boldsymbol{B}} \in \Theta$.

Let $\varepsilon^2 = c_1 \frac{d(p-d)}{Cn\lambda}$, where C is the constant in Proposition 3 and c_1 is a constant such that $c_1/(\log(c_0/\alpha) \leqslant 1/16$ for c_0 in Lemma 12. Then using Lemma 11, we have

$$\begin{aligned}
&\inf_{\widehat{\boldsymbol{B}}} \sup_{\boldsymbol{B} \in \Theta} \mathbb{E}\|\widehat{\boldsymbol{B}}\widehat{\boldsymbol{B}}^\top - \boldsymbol{B}\boldsymbol{B}^\top\|_F^2 && (58)\\
\geqslant & \min_{\boldsymbol{B}, \widetilde{\boldsymbol{B}} \in \Theta, \boldsymbol{B} \neq \widetilde{\boldsymbol{B}}} \|\boldsymbol{B}\boldsymbol{B}^\top - \widetilde{\boldsymbol{B}}\widetilde{\boldsymbol{B}}^\top\|_F^2 \left(1 - \frac{\max \mathrm{KL}(\mathbb{P}^n_{\boldsymbol{B}}, \mathbb{P}^n_{\widetilde{\boldsymbol{B}}}) + \log(2)}{\log(|\Theta|)}\right)\\
\geqslant & \alpha^2 \varepsilon^2 \left(1 - \frac{8Cn\rho^2\lambda_{0,d}\varepsilon^2 + \log 2}{\log(|\Theta|)}\right)\\
\geqslant & \alpha^2 c_1 \cdot \frac{d(p-d)}{n\lambda} \cdot \left(1 - \frac{8c_1 d(p-d)}{\log(|\Theta|)} - \frac{\log 2}{\log(|\Theta|)}\right).
\end{aligned}$$

Since $\log|\Theta| > d(p-d)\log(c_0/\alpha) \geqslant 16c_1 d(p-d)$, we have

$$\inf_{\widehat{\boldsymbol{B}}} \sup_{\boldsymbol{B} \in \Theta} \mathbb{E}_{\boldsymbol{B}}\|\widehat{\boldsymbol{B}}\widehat{\boldsymbol{B}}^\top - \boldsymbol{B}\boldsymbol{B}^\top\|_F^2 \gtrsim \frac{d(p-d)}{n\lambda}.$$

We complete the proof of Theorem 4.

Remark 3. If we apply the first statement of Proposition 3 to bound the KL-divergence between each pairs of different populations $\mathbb{P}^n_{\boldsymbol{B}}$ and $\mathbb{P}^n_{\widetilde{\boldsymbol{B}}}$, we do not need to require $\lambda \leqslant \varpi_d$, and following the same argument as above, we obtain the lower bound

$$\inf_{\widehat{\boldsymbol{B}}} \sup_{\boldsymbol{B} \in \Theta} \mathbb{E}_{\boldsymbol{B}}\|\widehat{\boldsymbol{B}}\widehat{\boldsymbol{B}}^\top - \boldsymbol{B}\boldsymbol{B}^\top\|_F^2 \gtrsim \frac{d(p-d)\lambda_{0,d}}{n\lambda}.$$

A similar yet more straightforward argument could result in a weaker lower bound at the rate of $\frac{p}{n\lambda}$, which does not impose the constraint $\lambda \leqslant \varpi_d$. Although this result does not exhibit the linear dependence of the minimax rate on the structural dimension d and is not as sharp, it still reflects the significant effect of the small gSNR on the estimation of the central space.

E.1 Proof of Lemma 12

We first state two lemmas from the literature.

Lemma 13 ([Cai et al., 2013, Lemma 1])**.** *For any $\varepsilon \in (0, \sqrt{2(d \wedge (p-d))}]$, any $\alpha \in (0,1)$ and any $\boldsymbol{A} \in \mathbb{O}(p,d)$, there exists a subset $\Theta \subset \mathbb{O}(p,d)$ such that*

$$|\Theta| \geqslant \left(\frac{c_0}{\alpha}\right)^{d(p-d)},$$

and for any $\boldsymbol{B}, \widetilde{\boldsymbol{B}} \in \Theta$,

$$\rho(\boldsymbol{A}\boldsymbol{A}^\top, \boldsymbol{B}\boldsymbol{B}^\top) \leqslant \varepsilon, \qquad \rho(\boldsymbol{B}\boldsymbol{B}^\top, \widetilde{\boldsymbol{B}}\widetilde{\boldsymbol{B}}^\top) \geqslant \alpha\varepsilon,$$

where c_0 is an absolute constant.

Lemma 14 ([Ma and Li, 2020, Lemma 6.5])**.** *For any matrices $\boldsymbol{A}_1, \boldsymbol{A}_2 \in \mathbb{O}(p,d)$, there exists some $\boldsymbol{Q} \in \mathbb{O}(d,d)$ such that*

$$\|\boldsymbol{A}_1 - \boldsymbol{A}_2\boldsymbol{Q}\|_F \leqslant \|\boldsymbol{A}_1\boldsymbol{A}_1^\top - \boldsymbol{A}_2\boldsymbol{A}_2^\top\|_F. \tag{59}$$

Proof of Lemma 12. Pick any $\boldsymbol{A} \in \mathbb{O}(p,d)$. Let Θ_0 be the subset in Lemma 13. For each $\boldsymbol{B}_0 \in \Theta_0$, we can find $\boldsymbol{Q}_{\boldsymbol{B}_0}$ such that $\|\boldsymbol{A} - \boldsymbol{B}_0\boldsymbol{Q}_{\boldsymbol{B}_0}\|_F \leqslant \epsilon$ due to Lemma 14. Define $\Theta = \{\boldsymbol{B}_0\boldsymbol{Q}_{\boldsymbol{B}_0} : \boldsymbol{B}_0 \in \Theta_0\}$. By the triangle inequality, it is easy to see that Θ satisfies the requirement of the lemma. Thus we complete the proof of Lemma 12. □

E.2 Proof of Proposition 2

To prove Proposition 2, we need the following lemma.

Lemma 15. *Suppose $(\boldsymbol{Z}, Y)$ is constructed as in Equation (10). Define $\boldsymbol{e}_0 = \boldsymbol{0}$ and for $i = 1, \ldots, d$, define $\boldsymbol{e}_{-i} = -\boldsymbol{e}_i$.*

1. *$P(W = i) = (4d)^{-1}$, for $i = \pm 1, \ldots, \pm d$ and $\mathbb{E}\left[\boldsymbol{e}_W\boldsymbol{e}_W^\top\right] = \frac{1}{2d}\boldsymbol{I}$.*
2. *If $y \in (i-\sigma, i+\sigma)$, then $\boldsymbol{l}(y) = \mathbb{E}[\boldsymbol{Z} \mid Y = y] = \mathbb{E}(Z_1|\psi(\boldsymbol{Z}) = 1)\boldsymbol{e}_i = \sqrt{2d\lambda_{0,d}}\boldsymbol{e}_i$ for $i = \pm 1, \ldots, \pm d$. If $y \in (-\sigma, \sigma)$, then $\boldsymbol{l}(y) = \boldsymbol{0}$.*
3. *$\mathrm{Cov}\left(\mathbb{E}[\boldsymbol{Z} \mid Y]\right) = \lambda_{0,d}\boldsymbol{I}_d$.*
4. *All eigenvalues of $\mathrm{Cov}(\mathbb{E}[\boldsymbol{Z} \mid Y])$ equal to $\lambda_{0,d}$.*
5. *$\boldsymbol{l}(y)$ satisfies the weak (K,τ)-sliced stable condition for any $K \geqslant 4d\max(1+2\gamma, \tau)$, where $\gamma \in (0,1)$ is defined as in Definition 2.*
6. *$\frac{1}{100d} \leqslant \lambda_{0,d} \leqslant \frac{4\log(2d)}{d}$ and $\frac{1}{8d\sqrt{2}} \leqslant \mathbb{E}\left(Z_1\ 1_{\boldsymbol{Z}\in A_1}\right) \leqslant \frac{\sqrt{2\log(2d)}}{2d}$. Furthermore, $\lambda_{0,d} \asymp \frac{\log(d)}{d}$ as $d \to \infty$.*

Proof of Lemma 15.

1: It is a direct corollary of the fact that $\boldsymbol{Z} \sim N(0, \mathbf{I}_d)$ and m_d is the median of $\|\boldsymbol{Z}\|^2$.

2: Since A_0 is the complement of a ball centered at $\mathbf{0}$, it is rationally invariant. By the symmetry of standard normal random vectors, $\mathbb{E}(\boldsymbol{Z} \mid \psi(\boldsymbol{Z}) = 0) \overset{a.s.}{=} \mathbb{E}(\boldsymbol{Z} \mid \boldsymbol{Z} \in A_0) = \mathbf{0}$.
Fix any $i = 1, \ldots, d$ and any $\Upsilon = \pm 1$. Under the condition that $y \in (\Upsilon i - \sigma, \Upsilon i + \sigma)$, one has

$$\begin{aligned}
&\mathbb{E}[\boldsymbol{Z}|Y = y] = \mathbb{E}[\boldsymbol{Z}|\psi(\boldsymbol{Z}) = \Upsilon i] = \mathbb{E}[\boldsymbol{Z}|\Upsilon Z_i = \max_{j\in[d]}|Z_j|],\\
&\mathbb{E}[Z_k|\Upsilon Z_i = \max_{j\in[d]}|Z_j|] = 0(\forall k \neq i),\\
&\mathbb{E}[Z_i|\Upsilon Z_i = \max_{j\in[d]}|Z_j|] = \Upsilon\mathbb{E}[Z_1 \mid Z_1 = \max_{j\in[d]}|Z_j|] = \Upsilon\mathbb{E}[Z_1 \mid \psi(\boldsymbol{Z}) = 1].
\end{aligned}$$

3 & 4: By the first and the second statements,

$$\begin{aligned}\mathrm{Cov}\left(\mathbb{E}[\boldsymbol{Z} \mid Y]\right) =&\mathbb{E}[\mathbb{E}[\boldsymbol{Z}|Y]\mathbb{E}[\boldsymbol{Z}^\top|Y]]\\ =&\frac{1}{4d}\sum_{i=-d}^{d}\mathbb{E}^2(Z_1|\psi(\boldsymbol{Z})=1)\boldsymbol{e}_i\boldsymbol{e}_i^\top\\ =&\lambda_{0,d}\boldsymbol{I}_d.\end{aligned}$$

5: In the following, fixed $\boldsymbol{\beta}\in\mathbb{R}^d$. Since $\ell(y)=\mathbf{0}$ for $y\in(-\sigma,\sigma)$, we can focus on the case where $|Y|>\sigma$. Fix any $\gamma\in(0,1)$ as defined in Definition 2.

Let $J_i=(i-\sigma,i+\sigma)$ for each $i=\pm1,\ldots,\pm d$. Suppose $\{\mathcal{S}_h=[a_{h-1},a_h):h=1,\ldots,H\}$ is a partition of $[-d-\sigma,d+\sigma]$ such that

$$\frac{1-\gamma}{H}\leqslant\mathbb{P}(Y\in\mathcal{S}_h)\leqslant\frac{1+\gamma}{H},\qquad\forall h=1,\ldots,H.$$

Since $(1+\gamma)/H\leqslant(1+\gamma)/K<(4d)^{-1}=P(Y\in J_i)$ for any $i=\pm1,\ldots,\pm d$, we conclude that $\mathcal{S}_h$ can overlap with at most two J_i's. If $\mathcal{S}_h$ is covered by some J_i, then $\mathrm{Cov}\left(\boldsymbol{\beta}^\top\boldsymbol{l}(Y)\middle|Y\in\mathcal{S}_h\right)=0$ because of the second statement. If $\mathcal{S}_h$ overlaps with J_i and J_k, then by the AM-GM inequality, it holds that

$$\begin{aligned}\mathrm{Cov}\left(\boldsymbol{\beta}^\top\boldsymbol{l}(Y)\middle|Y\in\mathcal{S}_h\right)&=\mathbb{P}(Y\in J_i\mid Y\in\mathcal{S}_h)\mathbb{P}(Y\in J_k\mid Y\in\mathcal{S}_h)\left(\boldsymbol{l}(i)^\top\boldsymbol{\beta}-\boldsymbol{l}(k)^\top\boldsymbol{\beta}\right)^2\\ &\leqslant2^{-1}\left(\left[\boldsymbol{l}(i)^\top\boldsymbol{\beta}\right]^2+\left[\boldsymbol{l}(k)^\top\boldsymbol{\beta}\right]^2\right).\end{aligned}$$

Summing over all h, one has

$$\begin{aligned}&\frac{1}{H}\sum_{h=1}^{H}\mathrm{Cov}\left(\boldsymbol{\beta}^\top\boldsymbol{l}(Y)\middle|Y\in\mathcal{S}_h\right)\\ \leqslant&\frac{1}{H}\sum_{h=1}^{H}\sum_{\substack{i\neq k,\\ i,k\in\{\pm1,\ldots,\pm d\}}}1_{\mathcal{S}_h\cap J_i\neq0}\ 1_{\mathcal{S}_h\cap J_k\neq0}2^{-1}\left(\left[\boldsymbol{l}(i)^\top\boldsymbol{\beta}\right]^2+\left[\boldsymbol{l}(k)^\top\boldsymbol{\beta}\right]^2\right)\\ \leqslant&\frac{1}{H}\sum_{i\in\{\pm1,\ldots,\pm d\}}\left[\boldsymbol{l}(i)^\top\boldsymbol{\beta}\right]^2\\ =&\frac{4d}{H}\mathrm{Cov}\left(\boldsymbol{\beta}^\top\boldsymbol{l}(Y)\right),\end{aligned}$$

where the last inequality is due to the fact that for each i, there are at most two values of h such that $\mathcal{S}_h$ is overlapped with J_i but not covered by J_i and thus $\left[\boldsymbol{l}(i)^\top\boldsymbol{\beta}\right]^2$ appears at most twice in the summation. Since $H/(4d)\geqslant K/(4d)\geqslant\tau$, we conclude that $\ell(y)$ is weak (K,τ)-sliced stable w.r.t. Y.

6: Suppose $\boldsymbol{Z}\sim N(0,\boldsymbol{I}_d)$. By symmetry,

$$\begin{aligned}\mathbb{E}\left(Z_1\ 1_{\boldsymbol{Z}\in A_1}\right)=&\mathbb{E}\left(Z_i\ 1_{\boldsymbol{Z}\in A_i}\right)\quad(\text{for any}\quad i\in[d])\\ =&(2d)^{-1}\sum_{i\in\{\pm1,\ldots,\pm d\}}\mathbb{E}\left(|Z_{|i|}|\ 1_{\boldsymbol{Z}\in A_i}\right)\\ =&(2d)^{-1}\sum_{i\in\{\pm1,\ldots,\pm d\}}\mathbb{E}\left(\max_{i=1,\ldots,d}|Z_i|\ 1_{\boldsymbol{Z}\in A_i}\right)\\ =&(2d)^{-1}\mathbb{E}\left(\max_{i=1,\ldots,d}|Z_i|\ 1_{\|\boldsymbol{Z}\|^2\leqslant m_d}\right).\end{aligned}$$

To obtain an upper bound, we note that

$$\begin{aligned}\mathbb{E}\left(\max_{i=1,\ldots,d}|Z_i|\ 1_{\|\boldsymbol{Z}\|^2\leqslant m_d}\right) \leqslant & \mathbb{E}\left(\max_{i=1,\ldots,d}|Z_i|\right)\\ \leqslant & \sqrt{2\log(2d)},\end{aligned}$$

where the last inequality is due to the maximal inequality of Gaussian r.v.s. (See Section 2.5 in Boucheron et al. [2013]).

To get a lower bound, we note that $\max_i |Z_i| \geqslant \sqrt{\frac{1}{d}\sum_i Z_i^2}$. Therefore

$$\begin{aligned}\mathbb{E}\left(\max_{i=1,\ldots,d}|Z_i|\ 1_{\|\boldsymbol{Z}\|^2\leqslant m_d}\right) &\geqslant d^{-1/2}\mathbb{E}\left(\|\boldsymbol{Z}\|\ 1_{\|\boldsymbol{Z}\|^2\leqslant m_d}\right)\\ &\geqslant d^{-1/2}\frac{1}{10}\mathbb{E}\|\boldsymbol{Z}\|\\ &\geqslant \frac{1}{10}\sqrt{\frac{d-\frac{1}{2}}{d}} \geqslant \frac{1}{10\sqrt{2}},\end{aligned}$$

where the second inequality is by part 2 of Lemma 19 and the third is due to a lower estimate used in the proof of that lemma. The two bounds together yield

$$\frac{1}{20d\sqrt{2}} \leqslant \mathbb{E}\left(Z_1\ 1_{\boldsymbol{Z}\in A_1}\right) \leqslant \frac{\sqrt{2\log(2d)}}{2d}.$$

Recall that $\lambda_{0,d} := (2d)^{-1}\mathbb{E}\left(Z_1 \mid \boldsymbol{Z}\in A_1\right)^2$. By part 1, $\mathbb{P}(\boldsymbol{Z}\in A_1) = (4d)^{-1}$. Therefore, $\frac{1}{100d} \leqslant \lambda_{0,d} \leqslant \frac{4\log(2d)}{d}$.

For d sufficiently large, we have $m_d > \sqrt{\log d}$. Following the same proof of Equation (3.14) in Ledoux and Talagrand [1991, Chapter 3.3], there is some positive constant c_0 such that

$$\mathbb{E}\left(\max_{i=1,\ldots,d}|Z_i|\ 1_{\|\boldsymbol{Z}\|^2\leqslant m_d}\right) \geqslant c_0\sqrt{\log d}.$$

Therefore, $\lambda_{0,d} \asymp \frac{\log d}{d}$.

□

Proof of Proposition 2. Note that any $k \geqslant 2$ independent uniform random variable sequence (r.v.s. for short) can be constructed from a single $U \sim \mathrm{Unif}(0,1)$ as follows. Represent U as $\sum_{j=1}^{\infty} k^{-j}a_j$ for $a_j \in \{0,1,\ldots,k-1\}$. Let $U^{(i)} = \sum_{j=1}^{\infty} k^{-j}a_{(j-1)k+i}$ for each $i = 1,\ldots,k$. Since a_j's are independent and identically distributed, one conclude that $U^{(i)}$'s are independent and identically distributed, each following $\mathrm{Unif}(0,1)$.

Let $U = \Phi(\epsilon)$, where Φ is the cumulative density function (C.D.F.) for the standard normal distribution. Since $\epsilon \sim N(0,1)$, one has $U \sim \mathrm{Unif}(0,1)$. Let $k = 1+d$. Using the above construction of $U^{(i)}$'s and let $\xi_j = \Phi^{-1}(U^{(j)})$ for $1 \leqslant j \leqslant d$ and $\eta = \sigma \cdot U^{(1+d)}$, we can represent Y as a function of $\boldsymbol{B}^\top\boldsymbol{X}$ and ϵ.

Since

$$\begin{pmatrix}\boldsymbol{X}\\ \boldsymbol{Z}\end{pmatrix} = \begin{pmatrix}\boldsymbol{I}_p & \boldsymbol{0}\\ \rho\boldsymbol{B}^\top & \sqrt{1-\rho^2}\boldsymbol{I}_d\end{pmatrix}\begin{pmatrix}\boldsymbol{X}\\ \xi\end{pmatrix},$$

the joint distribution of $(\boldsymbol{X},\boldsymbol{Z})$ is also normal. Thus by elementary results for normal distributions, one has

$$\boldsymbol{X}\mid\boldsymbol{Z} \sim N\left(\mathbb{E}[\boldsymbol{X}] + \mathrm{Cov}(\boldsymbol{X},\boldsymbol{Z})\mathrm{var}^{-1}(\boldsymbol{Z})(\boldsymbol{Z}-\mathbb{E}[\boldsymbol{Z}]), \mathrm{var}(\boldsymbol{X}) - \mathrm{Cov}(\boldsymbol{X},\boldsymbol{Z})\mathrm{var}^{-1}(\boldsymbol{Z})\mathrm{Cov}(\boldsymbol{Z},\boldsymbol{X})\right)$$

$$=N(\rho \boldsymbol{B}\boldsymbol{Z}, \boldsymbol{I}_p - \rho^2 \boldsymbol{B}\boldsymbol{B}^\top).$$

Hence

$$\mathbb{E}[\boldsymbol{X} \mid Y] = \mathbb{E}[\mathbb{E}[\boldsymbol{X} \mid \boldsymbol{Z}, \eta] \mid Y] = \mathbb{E}[\rho \boldsymbol{B}\boldsymbol{Z} \mid Y] = \rho \boldsymbol{B}\mathbb{E}[\boldsymbol{Z} \mid Y],$$

and

$$\mathrm{Cov}(\mathbb{E}[\boldsymbol{X} \mid Y]) = \rho^2 \boldsymbol{B}\mathrm{Cov}(\mathbb{E}[\boldsymbol{Z} \mid Y])\boldsymbol{B}^\top.$$

Lemma 15 shows that all eigenvalues of $\mathrm{Cov}(\mathbb{E}[\boldsymbol{Z} \mid Y])$ equal to $\lambda_{0,d}$.

Furthermore, $\boldsymbol{l}(y) = \mathbb{E}[\boldsymbol{Z} \mid Y = y]$ satisfies the weak (K, τ)-sliced stable condition for any $K \geqslant 4d \max(2, \tau)$. Therefore, $f(\boldsymbol{B}^\top \boldsymbol{X}, \epsilon)$ belongs to the class $\mathcal{F}_d(\lambda, \kappa, 8d)$ if we choose $\tau = 2$ and $\gamma = 1/2$. □

E.3 Proof of Proposition 3

The first statement is relatively simple to prove. By the construction in Equation (10), $\mathbb{P}_{\boldsymbol{B}}(Y \mid \boldsymbol{X}, \boldsymbol{Z}) = \mathbb{P}_{\widetilde{\boldsymbol{B}}}(Y \mid \boldsymbol{X}, \boldsymbol{Z})$, a.s. By basic properties of KL-divergence,

$$\begin{aligned}\mathrm{KL}(\mathbb{P}_{\boldsymbol{B}}, \mathbb{P}_{\widetilde{\boldsymbol{B}}}) \leqslant & \ \mathrm{KL}(\mathbb{P}_{\boldsymbol{B}}, \mathbb{P}_{\widetilde{\boldsymbol{B}}}) + \mathbb{E}_{\boldsymbol{X}, Y \sim \mathbb{P}_{\boldsymbol{B}}} \left(\mathrm{KL}(\mathbb{P}_{\boldsymbol{B}}(\boldsymbol{Z} \mid \boldsymbol{X}, Y), \mathbb{P}_{\widetilde{\boldsymbol{B}}}(\boldsymbol{Z} \mid \boldsymbol{X}, Y)\right) \\ = & \ \mathrm{KL}(\mathbb{P}_{\boldsymbol{B}}(\boldsymbol{X}, \boldsymbol{Z}, Y), \mathbb{P}_{\widetilde{\boldsymbol{B}}}(\boldsymbol{X}, \boldsymbol{Z}, Y)) \\ = & \ \mathrm{KL}(\mathbb{P}_{\boldsymbol{B}}(\boldsymbol{X}, \boldsymbol{Z}), \mathbb{P}_{\widetilde{\boldsymbol{B}}}(\boldsymbol{X}, \boldsymbol{Z})). \end{aligned} \tag{60}$$

Furthermore, let $\phi_p(\boldsymbol{x})$ be the density function for $N(0, \boldsymbol{I}_p)$. Then we have

$$\begin{aligned} & \mathrm{KL}(\mathbb{P}_{\boldsymbol{B}}(\boldsymbol{X}, W), \mathbb{P}_{\widetilde{\boldsymbol{B}}}(\boldsymbol{X}, W)) \\ = & \mathbb{E}_{\boldsymbol{B}}\left[\log\left(\frac{\phi_p(\boldsymbol{X})(1-\rho^2)^{-d/2}\phi_d(\boldsymbol{Z} - \rho \boldsymbol{B}^\top \boldsymbol{X})}{\phi_p(\boldsymbol{X})(1-\rho^2)^{-d/2}\phi_d(\boldsymbol{Z} - \rho \widetilde{\boldsymbol{B}}^\top \boldsymbol{X})}\right)\right] \\ = & \mathbb{E}_{\boldsymbol{B}}\left[\frac{1}{2(1-\rho^2)}\left(\|\boldsymbol{Z} - \rho \boldsymbol{B}^\top \boldsymbol{X}\|^2 - \|\boldsymbol{Z} - \rho \widetilde{\boldsymbol{B}}^\top \boldsymbol{X}\|^2\right)\right] \\ = & \frac{\rho^2}{2(1-\rho^2)}\mathbb{E}\left[\|(\boldsymbol{B} - \widetilde{\boldsymbol{B}})^\top \boldsymbol{X}\|^2\right] = \frac{\rho^2}{2(1-\rho^2)}\|\boldsymbol{B} - \widetilde{\boldsymbol{B}}\|_F^2. \end{aligned}$$

The rest of the proof is about the second statement. By the construction in Equation (10), $\mathbb{P}_{\boldsymbol{B}}(Y \mid \boldsymbol{X}, W) = \mathbb{P}_{\widetilde{\boldsymbol{B}}}(Y \mid \boldsymbol{X}, W)$, a.s. By basic properties of KL-divergence,

$$\begin{aligned}\mathrm{KL}(\mathbb{P}_{\boldsymbol{B}}, \mathbb{P}_{\widetilde{\boldsymbol{B}}}) \leqslant & \ \mathrm{KL}(\mathbb{P}_{\boldsymbol{B}}, \mathbb{P}_{\widetilde{\boldsymbol{B}}}) + \mathbb{E}_{\boldsymbol{X}, Y \sim \mathbb{P}_{\boldsymbol{B}}} \left(\mathrm{KL}(\mathbb{P}_{\boldsymbol{B}}(W \mid \boldsymbol{X}, Y), \mathbb{P}_{\widetilde{\boldsymbol{B}}}(W \mid \boldsymbol{X}, Y)\right) \\ = & \ \mathrm{KL}(\mathbb{P}_{\boldsymbol{B}}(\boldsymbol{X}, W, Y), \mathbb{P}_{\widetilde{\boldsymbol{B}}}(\boldsymbol{X}, W, Y)) \\ = & \ \mathrm{KL}(\mathbb{P}_{\boldsymbol{B}}(\boldsymbol{X}, W), \mathbb{P}_{\widetilde{\boldsymbol{B}}}(\boldsymbol{X}, W)). \end{aligned} \tag{61}$$

Furthermore, let $\phi_p(\boldsymbol{x})$ be the density function for $N(0, \boldsymbol{I}_p)$. Then we have

$$\begin{aligned} & \mathrm{KL}(\mathbb{P}_{\boldsymbol{B}}(\boldsymbol{X}, W), \mathbb{P}_{\widetilde{\boldsymbol{B}}}(\boldsymbol{X}, W)) \\ = & \mathbb{E}_{\boldsymbol{B}}\left[\log\left(\frac{\phi_p(\boldsymbol{X})\mathbb{P}_{\boldsymbol{B}}(\boldsymbol{Z} \in A_W \mid \boldsymbol{X})}{\phi_p(\boldsymbol{X})\mathbb{P}_{\widetilde{\boldsymbol{B}}}(\boldsymbol{Z} \in A_W \mid \boldsymbol{X})}\right)\right] \\ = & \mathbb{E}_{\boldsymbol{B}}\left[\log\left(\frac{\mathbb{P}_{\boldsymbol{B}}(\boldsymbol{Z} \in A_W \mid \boldsymbol{X})}{\mathbb{P}_{\widetilde{\boldsymbol{B}}}(\boldsymbol{Z} \in A_W \mid \boldsymbol{X})}\right)\right]. \end{aligned}$$

Since we need to analyze the probability $\mathbb{P}_{\widetilde{\boldsymbol{B}}}(\boldsymbol{Z} \in A_W \mid \boldsymbol{X})$, it is convenient to express it as functions of $(\widetilde{\boldsymbol{B}} - \boldsymbol{B})^\top \boldsymbol{X}$. In the following, $\boldsymbol{\xi}$ is a generic random vector that is independent with everything else and follows $N(0, \boldsymbol{I}_d)$. Let $\Delta \boldsymbol{B} = \widetilde{\boldsymbol{B}} - \boldsymbol{B}$. For any fixed $\boldsymbol{\mu} \in \mathbb{R}^d$, $w \in \{-d, \dots, d\}$, define $g_w^{\boldsymbol{\mu}}(\boldsymbol{t}) = \mathbb{P}(\rho \boldsymbol{\mu} + \rho \boldsymbol{t} + \sqrt{1-\rho^2}\boldsymbol{\xi} \in$

A_w) for $\boldsymbol{t} \in \mathbb{R}^d$. Now, we have $\mathbb{P}_{\boldsymbol{B}}(\boldsymbol{Z} \in A_W \mid \boldsymbol{X}) = g_W^{\boldsymbol{B}^\top \boldsymbol{X}}(\boldsymbol{0})$ and $\mathbb{P}_{\widetilde{\boldsymbol{B}}}(\boldsymbol{Z} \in A_W \mid \boldsymbol{X}) = g_W^{\boldsymbol{B}^\top \boldsymbol{X}}(\Delta \boldsymbol{B}^\top \boldsymbol{X})$. Furthermore,

$$\mathrm{KL}(\mathbb{P}_{\boldsymbol{B}}(\boldsymbol{X}, W), \mathbb{P}_{\widetilde{\boldsymbol{B}}}(\boldsymbol{X}, W)) = -\mathbb{E}_{\boldsymbol{B}}\left[\log g_W^{\boldsymbol{B}^\top \boldsymbol{X}}(\Delta \boldsymbol{B}^\top \boldsymbol{X}) - \log g_W^{\boldsymbol{B}^\top \boldsymbol{X}}(\boldsymbol{0})\right]. \tag{62}$$

It is pedagogical to provide an overview of our argument in obtaining an upper bound of the KL-divergence. We will apply a second-order Taylor expansion to $\log g_W^{\boldsymbol{B}^\top \boldsymbol{X}}(\boldsymbol{t})$ around $\boldsymbol{0}$. Since the first-order derivative has a zero expectation, we only need a careful examination of the second derivative. At the end, we can show that the KL-divergence is close enough to $\frac{\rho^2}{1-\rho^2}\mathrm{Tr}\left(\Delta \boldsymbol{B}^\top \Delta \boldsymbol{B} \mathbb{E}_{\boldsymbol{B}}\left[\mathbb{E}\left(\boldsymbol{\xi} \mid \psi(\boldsymbol{\xi}) = W\right)^{\otimes}\right]\right)$ when ρ is sufficiently small. By our construction in Equation (10),

$$\mathbb{E}_{\boldsymbol{B}}\left[\mathbb{E}\left(\boldsymbol{\xi} \mid \psi(\boldsymbol{\xi}) = W\right)^{\otimes}\right] = \lambda_{0,d} \boldsymbol{I}_d,$$

which implies that the KL-divergence is closed to $\frac{\rho^2 \lambda_{0,d}}{1-\rho^2}\mathrm{Tr}\left(\Delta \boldsymbol{B}^\top \Delta \boldsymbol{B}\right)$.

I. Taylor expansion. By Taylor expansion with an integral remainder, i.e., $f(\boldsymbol{t}) = f(\boldsymbol{0}) + \nabla f(\boldsymbol{0})\boldsymbol{t} + \int_0^1 \boldsymbol{t}^\top \nabla^2 f(s\boldsymbol{t})\boldsymbol{t}(1-s)\,\mathrm{d}\,s$, one has

$$\begin{aligned}
&\log g_w^{\boldsymbol{B}^\top \boldsymbol{x}}(\Delta \boldsymbol{B}^\top \boldsymbol{x}) - \log g_w^{\boldsymbol{B}^\top \boldsymbol{x}}(\boldsymbol{0}) \\
=&\langle \Delta \boldsymbol{B}^\top \boldsymbol{x}, \nabla \log g_w^{\boldsymbol{B}^\top \boldsymbol{x}}(\boldsymbol{0})\rangle \ldots \\
&+ \int_0^1 \boldsymbol{x}^\top \Delta \boldsymbol{B}\left(\nabla^2 \log g_w^{\boldsymbol{B}^\top \boldsymbol{x}}(\alpha \Delta \boldsymbol{B}^\top \boldsymbol{x})\right) \Delta \boldsymbol{B}^\top \boldsymbol{x}(1-\alpha)\,\mathrm{d}\,\alpha.
\end{aligned} \tag{63}$$

Lemma 16. *The derivative of* $\log g_w^{\boldsymbol{\mu}}(\boldsymbol{t})$ *is*

$$\nabla \log g_w^{\boldsymbol{\mu}}(\boldsymbol{t}) = \sqrt{\frac{\rho^2}{1-\rho^2}}\mathbb{E}\left(\boldsymbol{\xi} \mid \rho\boldsymbol{\mu} + \rho\boldsymbol{t} + \sqrt{1-\rho^2}\boldsymbol{\xi} \in A_w\right), \tag{64}$$

and the second order derivative is

$$\begin{aligned}
\nabla^2 \log g_w^{\boldsymbol{\mu}}(\boldsymbol{t}) = \frac{\rho^2}{1-\rho^2}\Big\{&-\boldsymbol{I}_d + \mathbb{E}\left(\boldsymbol{\xi}^{\otimes} \mid \rho\boldsymbol{\mu} + \rho\boldsymbol{t} + \sqrt{1-\rho^2}\boldsymbol{\xi} \in A_w\right) \\
&-\mathbb{E}\left(\boldsymbol{\xi} \mid \rho\boldsymbol{\mu} + \rho\boldsymbol{t} + \sqrt{1-\rho^2}\boldsymbol{\xi} \in A_w\right)^{\otimes}\Big\}.
\end{aligned} \tag{65}$$

By Equation (64),

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{B}}\langle \Delta \boldsymbol{B}^\top \boldsymbol{X}, \nabla \log g_W^{\boldsymbol{B}^\top \boldsymbol{X}}(\boldsymbol{0})\rangle \\
=&\sqrt{\frac{\rho^2}{1-\rho^2}}\mathbb{E}_{\boldsymbol{B}}\left\langle \Delta \boldsymbol{B}^\top \boldsymbol{X}, \mathbb{E}\left(\boldsymbol{\xi} \mid \rho\boldsymbol{B}^\top \boldsymbol{X} + \sqrt{1-\rho^2}\boldsymbol{\xi} \in A_W, \boldsymbol{X}, W\right)\right\rangle.
\end{aligned}$$

We split the expectation into parts given by $\{W = w\}$ and use properties of conditional expectation to obtain

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{B}}\left\langle \Delta \boldsymbol{B}^\top \boldsymbol{X}, \mathbb{E}\left(\boldsymbol{\xi} \mid \rho\boldsymbol{B}^\top \boldsymbol{X} + \sqrt{1-\rho^2}\boldsymbol{\xi} \in A_W, \boldsymbol{X}, W\right)\right\rangle \\
=&\sum_{w=-d}^{d}\mathbb{E}_{\boldsymbol{B}}\left(1_{W=w}\left\langle \Delta \boldsymbol{B}^\top \boldsymbol{X}, \mathbb{E}\left(\boldsymbol{\xi} \mid \rho\boldsymbol{B}^\top \boldsymbol{X} + \sqrt{1-\rho^2}\boldsymbol{\xi} \in A_w, \boldsymbol{X}\right)\right\rangle\right)
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{w=-d}^{d} \mathbb{E}_{\boldsymbol{B}} \left(\mathbb{P}\left(W = w \mid \boldsymbol{X}\right) \left\langle \Delta \boldsymbol{B}^\top \boldsymbol{X}, \frac{\mathbb{E}\left(\boldsymbol{\xi}\ 1_{\rho \boldsymbol{B}^\top \boldsymbol{X} + \sqrt{1-\rho^2}\boldsymbol{\xi} \in A_w} \mid \boldsymbol{X}\right)}{\mathbb{P}\left(\rho \boldsymbol{B}^\top \boldsymbol{X} + \sqrt{1-\rho^2}\boldsymbol{\xi} \in A_w \mid \boldsymbol{X}\right)} \right\rangle \right) \\
&= \sum_{w=-d}^{d} \mathbb{E}_{\boldsymbol{B}} \left(\left\langle \Delta \boldsymbol{B}^\top \boldsymbol{X}, \mathbb{E}\left(\boldsymbol{\xi}\ 1_{\rho \boldsymbol{B}^\top \boldsymbol{X} + \sqrt{1-\rho^2}\boldsymbol{\xi} \in A_w} \mid \boldsymbol{X}\right) \right\rangle \right) \\
&= \sum_{w=-d}^{d} \mathbb{E}_{\boldsymbol{B}} \left(\left\langle \Delta \boldsymbol{B}^\top \boldsymbol{X}, \boldsymbol{\xi}\ 1_{\rho \boldsymbol{B}^\top \boldsymbol{X} + \sqrt{1-\rho^2}\boldsymbol{\xi} \in A_w} \right\rangle \right) \\
&= \mathbb{E}_{\boldsymbol{B}} \left(\langle \Delta \boldsymbol{B}^\top \boldsymbol{X}, \boldsymbol{\xi} \rangle \right) = 0,
\end{aligned}
$$

where the second equation is due to fact that the conditional distribution of $Z \mid \boldsymbol{X} \overset{d}{=} \rho \boldsymbol{B}^\top \boldsymbol{X} + \sqrt{1-\rho^2}\boldsymbol{\xi}$ and the last equation is because ξ is independent with $\boldsymbol{X}$. We thus showed that

$$
\begin{aligned}
&\mathbb{E}_{\boldsymbol{B}} \langle \Delta \boldsymbol{B}^\top \boldsymbol{X}, \nabla \log g_W^{\boldsymbol{B}^\top \boldsymbol{X}}(\boldsymbol{0}) \rangle \\
&= \sqrt{\frac{\rho^2}{1-\rho^2}} \mathbb{E}_{\boldsymbol{B}} \left(\langle \Delta \boldsymbol{B}^\top \boldsymbol{X}, \boldsymbol{\xi} \rangle \right) = 0.
\end{aligned}
$$

Therefore, it suffices to focus on the second order term in Equation (63).

II. Analysis of the second order term. In the following, we fix any $\alpha \in (0,1)$.

As a shorthand, we write $J(w,\rho) := (1-\rho^2)^{-1/2}(A_w - \rho \boldsymbol{B}^\top \boldsymbol{X} - \rho\alpha\Delta \boldsymbol{B}^\top \boldsymbol{X})$, which is a random set that depends on $\boldsymbol{X}$ with parameters w and ρ.

By Equation (65), one has

$$
\begin{aligned}
&- \mathbb{E}_{\boldsymbol{B}} \left[\boldsymbol{X}^\top \Delta \boldsymbol{B}\ \nabla^2 \log g_W^{\boldsymbol{B}^\top \boldsymbol{X}} (\alpha \Delta \boldsymbol{B}^\top \boldsymbol{X})\ \Delta \boldsymbol{B}^\top \boldsymbol{X} \right] \\
&= \frac{\rho^2}{1-\rho^2} \operatorname{Tr}\left(\mathbb{E}_{\boldsymbol{B}} \left\{ \Delta \boldsymbol{B}^\top \boldsymbol{X}\boldsymbol{X}^\top \Delta \boldsymbol{B} \left[\boldsymbol{I}_d - \mathbb{E}\left(\boldsymbol{\xi}^{\otimes} \mid \boldsymbol{\xi} \in J(W,\rho), \boldsymbol{X}, W\right) \right] \right\} \right) \\
&\quad + \frac{\rho^2}{1-\rho^2} \operatorname{Tr}\left(\mathbb{E}_{\boldsymbol{B}} \left\{ \Delta \boldsymbol{B}^\top \boldsymbol{X}\boldsymbol{X}^\top \Delta \boldsymbol{B} \left[\mathbb{E}\left(\boldsymbol{\xi} \mid \boldsymbol{\xi} \in J(W,\rho), \boldsymbol{X}, W\right)^{\otimes} \right] \right\} \right).
\end{aligned}
\tag{66}
$$

The rest of the proof is dedicated to bounding the two terms in (66) by dropping the factor $\frac{\rho^2}{1-\rho^2}$.

Some intuitions.

Before we move on, it is worth checking the limits of these two terms as $\rho \to 0$. In this case, $J(W,\rho) \to A_W$.

(a). The inner conditional expectation in the first term

$$
\mathbb{E}\left(\boldsymbol{\xi}^{\otimes} \mid \boldsymbol{\xi} \in J(W,\rho), \boldsymbol{X}, W\right) \to \mathbb{E}\left(\boldsymbol{\xi}^{\otimes} \mid \boldsymbol{\xi} \in A_W\right) = \frac{\mathbb{E}\left(\boldsymbol{\xi}^{\otimes}\ 1_{\boldsymbol{\xi} \in A_W}\right)}{\mathbb{P}\left(\boldsymbol{\xi} \in A_W\right)}.
$$

Furthermore, when $\rho = 0$, W and $\boldsymbol{X}$ becomes independent, and the distribution of W is the same as $\psi(\xi)$. Therefore, the expectation of the last equation w.r.t. W equals to $\sum_w \mathbb{E}\left(\boldsymbol{\xi}^{\otimes}\ 1_{\boldsymbol{\xi} \in A_w}\right) = \mathbb{E}\left(\boldsymbol{\xi}^{\otimes}\right) = \boldsymbol{I}_d$, from which we conclude that the first term converges to 0 as $\rho \to 0$.

(b). Similarly, the inner conditional expectation in the second term

$$
\mathbb{E}\left(\boldsymbol{\xi} \mid \boldsymbol{\xi} \in J(W,\rho), \boldsymbol{X}, W\right) \to \mathbb{E}\left(\boldsymbol{\xi} \mid \boldsymbol{\xi} \in A_W\right) = \sqrt{2d\lambda_{0,d}}\, \boldsymbol{e}_W,
$$

where $\{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_d\}$ is the standard basis of $\mathbb{R}^d$, $\boldsymbol{e}_0 = \boldsymbol{0}$, and $\boldsymbol{e}_{-i} = -\boldsymbol{e}_i$ for $i = 1, \ldots, d$. Therefore,

$$
\mathbb{E}\left[\mathbb{E}\left(\boldsymbol{\xi} \mid \boldsymbol{\xi} \in J(W,\rho), \boldsymbol{X}, W\right)^{\otimes} \right] \to \lambda_{0,d} \boldsymbol{I}_d.
$$

Thus, the second term converges to $\frac{\rho^2 \lambda_{0,d}}{1-\rho^2} \operatorname{Tr}\left(\mathbb{E}_{\boldsymbol{B}} \left\{ \Delta \boldsymbol{B}^\top \boldsymbol{X}\boldsymbol{X}^\top \Delta \boldsymbol{B} \right\} \right) = \frac{\rho^2 \lambda_{0,d}}{1-\rho^2} \|\Delta \boldsymbol{B}\|_F^2$.

These intuitions can be justified rigorously by using the continuous dependence of the probability measure on ρ. We state the result in the next two lemmas, whose proofs are deferred.

Lemma 17. *Let $\epsilon = 100\lambda_{0,d}$. There exists a constant $\delta_d^{(1)}$ such that for any $\rho \in (0, \delta_d^{(1)})$ and any $\alpha \in (0,1)$, any $\boldsymbol{B}, \widetilde{\boldsymbol{B}} \in \mathbb{O}(p,d)$,*

$$Tr\left(\mathbb{E}_{\boldsymbol{B}}\left\{\Delta\boldsymbol{B}^\top \boldsymbol{X}\boldsymbol{X}^\top \Delta\boldsymbol{B}\left[\boldsymbol{I}_d - \mathbb{E}\left(\boldsymbol{\xi}^{\otimes} \mid \boldsymbol{\xi} \in J(W,\rho), \boldsymbol{X}, W\right)\right]\right\}\right) \leqslant 2\epsilon\|\Delta\boldsymbol{B}\|_F^2.$$

Furthermore, $\delta_d^{(1)}$ can be taken as $c'd^{-5/2}$ where c' is the constant in Lemma 20.

Lemma 18. *Let $\epsilon = 100\lambda_{0,d}$. There exist a constant $\delta_d^{(2)}$ and a universal constant C such that for any $\rho \in (0, \delta_d^{(2)})$ and any $\alpha \in (0,1)$, any $\boldsymbol{B}, \widetilde{\boldsymbol{B}} \in \mathbb{O}(p,d)$,*

$$Tr\left(\mathbb{E}_{\boldsymbol{B}}\left\{\Delta\boldsymbol{B}^\top \boldsymbol{X}\boldsymbol{X}^\top \Delta\boldsymbol{B}\left[\mathbb{E}\left(\boldsymbol{\xi} \mid \boldsymbol{\xi} \in J(W,\rho), \boldsymbol{X}, W\right)^{\otimes}\right]\right\}\right) < \left[\lambda_{0,d} + C\epsilon\right]\|\Delta\boldsymbol{B}\|_F^2.$$

Furthermore, $\delta_d^{(2)}$ can be taken as $c'd^{-7/2-\varsigma}$ for some universal constant c' and any positive number ς.

We apply Lemmas 17 and 18, and let $\delta_d = \min\left(\delta_d^{(1)}, \delta_d^{(2)}\right)$. In view of Equation (66), we conclude that for any $\rho \in (0, \delta_d)$, any $\alpha \in (0,1)$, it holds that

$$-\mathbb{E}_{\boldsymbol{B}}\left[\boldsymbol{X}^\top \Delta\boldsymbol{B}\ \nabla^2 \log g_W^{\boldsymbol{B}^\top \boldsymbol{X}}(\alpha\Delta\boldsymbol{B}^\top \boldsymbol{X})\ \Delta\boldsymbol{B}^\top \boldsymbol{X}\right] \leqslant C\rho^2\lambda_{0,d}\|\Delta\boldsymbol{B}\|_F^2.$$

Combining Equations (61), (62), and (63), we conclude that

$$\mathrm{KL}(\mathbb{P}_{\boldsymbol{B}}, \mathbb{P}_{\widetilde{\boldsymbol{B}}}) \leqslant C\rho^2\lambda_{0,d}\|\Delta\boldsymbol{B}\|_F^2.$$

Therefore, we complete the proof of the proposition.

E.4 Lemmas for uniform controls on d-dimensional Gaussian measures

This section collects some results about d-dimensional Gaussian random vectors. These results will be used in proving Lemmas 17 and 18.

Lemma 19 (Tail probability for Chi-square)**.** *Suppose X is a chi-squared random variable with d degrees of freedom (χ_d^2).*

1. *For any constant $k > 0$, there exists a constant $C_k = O(k)$, such that $\mathbb{P}\left(X \geqslant C_k d\right) \leqslant d^{-k}$ and $\mathbb{E}\left(X^a\ 1_{X > C_k d}\right) \leqslant Cd^{-k}$ for $a = 1, 2$, where C is a universal constant.*
2. *Let m_d be the median of χ_d^2. It holds that*

$$\mathbb{E}\left(\sqrt{X}\ 1_{X \leqslant m_d}\right) \geqslant \frac{1}{10}\mathbb{E}\left(\sqrt{X}\right).$$

3. *There exists a universal constant $\pi_0 > 0$, such that $\mathbb{P}(X \geqslant \frac{m_d}{1-\rho^2}) \geqslant \pi_0$ whenever $\rho^2 < \left(3de^{1/3}\right)^{-1}$.*

Proof of Lemma 19.

Statement 1.

If $d = 1$, it is trivial. Suppose $d \geqslant 2$. By Equation (4.3) in Laurent and Massart [2000],

$$\mathbb{P}(X - d \geqslant 2\sqrt{dx} + 2x) \leqslant \exp(-x). \tag{67}$$

Choosing $x = k\log(d)$, we have $\mathbb{P}(X \geqslant d + 2\sqrt{kd\log d} + 2k\log(d)) \leqslant d^{-k}$.

Let $C_0 \geqslant k(\inf_{d\geqslant 2} \frac{d}{\log d})^{-1}$. Then $2\sqrt{kd\log d} + 2k\log(d) \leqslant 2(\sqrt{C_0} + C_0)d$. For any $C \geqslant 1 + 2(\sqrt{C_0} + C_0) = O(k)$, it holds that $\mathbb{P}(X \geqslant Cd) \leqslant d^{-k}$.

The inequality (67) also implies that for any $t > d$, we have

$$\mathbb{P}(X > t) \leqslant \exp\left(-(t-d)^2/(4t)\right).$$

If $t > 2d$, then $t - d > t/2$ and $(t-d)^2/(4t) \geqslant t/16$. By Fubini's theorem, for any $r > 2d$,

$$\begin{aligned}
\mathbb{E}\left(X^a\ 1_{X>r}\right) &= r^a\mathbb{P}(X > r) + \int_r^\infty at^{a-1}\mathbb{P}(X > t)\,\mathrm{d}\,t \\
&= r^a\mathbb{P}(X > r) + \int_r^\infty at^{a-1}e^{-t/16}\,\mathrm{d}\,t \\
&\leqslant e^{-r/16}\left(r^a + 16 + \ 1_{a=2}\left(240 + 32r\right)\right) \\
&\leqslant C_1 r^2 e^{-r/16}.
\end{aligned}$$

For $r = 2d + 16k\log(d) = d\cdot O(k)$, $C_1 r^2 e^{-r/16} \leqslant \frac{1}{d^k}C_1(2d + 16k\log(d))^2 e^{-d/8} \leqslant C_2\frac{1}{d^k}$ because $de^{-d/16}$ is bounded.

Statement 2.

Since m_d is the median of χ_d^2, $\mathbb{P}(X \leqslant m_d) = 1/2$.

Note that $\mathbb{E}\left(\sqrt{X}\ 1_{X\leqslant m_d}\right) = \mathbb{E}\left(\sqrt{X}\right) - \mathbb{E}\left(\sqrt{X}\ 1_{X>m_d}\right)$ and by the Cauchy–Schwarz inequality, $\mathbb{E}\left(\sqrt{X}\ 1_{X>m_d}\right) \leqslant \sqrt{\mathbb{E}X\mathbb{P}\left(X > m_d\right)}$. Since m_d is the median and $\mathbb{E}X = d$, we have

$$\frac{\mathbb{E}\left(\sqrt{X}\ 1_{X\leqslant m_d}\right)}{\mathbb{E}\left(\sqrt{X}\right)} \geqslant 1 - \frac{\sqrt{d/2}}{\mathbb{E}\left(\sqrt{X}\right)}.$$

For $d = 1$, $\mathbb{E} = \sqrt{\frac{2}{\pi}}$ and thus the RHS is no less than $1 - \sqrt{\pi}/2 > 0.1$. A direct calculation yields

$$\begin{aligned}
\mathbb{E}\left(\sqrt{X}\right) &= \int_0^\infty \frac{x^{(d+1)/2-1}e^{-x/2}}{2^{d/2}\Gamma(d/2)}dx \\
&= 2^{1/2}\frac{\Gamma((1+d)/2)}{\Gamma(d/2)} \\
&\geqslant \sqrt{2\left(\frac{d}{2} - \frac{1}{4}\right)},
\end{aligned}$$

where in the last inequality we have applied a bound on the ratio of gamma functions that $\sqrt{x - \frac{1}{4}} < \frac{\Gamma(x+1/2)}{\Gamma(x)}$ proved by Watson [1959]. Since $d/(d-1/2)$ is decreasing in d and $d \geqslant 2$, we conclude that $\frac{\mathbb{E}(\sqrt{X}\ 1_{X\leqslant m_d})}{\mathbb{E}(\sqrt{X})} \geqslant 1/10$.

Statement 3.

By Corollary 3 in Zhang and Zhou [2020], there exist uniform constants $C, c > 0$ such that

$$\mathbb{P}(X - d \geqslant x) \geqslant c\exp\left(-Cx \wedge \frac{x^2}{k}\right), \quad \forall x > 0.$$

By the continuity of measure, we have $\mathbb{P}(X \geqslant d) \geqslant c$. It remains to check that $\frac{m_d}{1-\rho^2} < d$ for ρ^2 small.

Berg and Pedersen [2006] have proved that $m_d \leqslant de^{-1/(3d)}$. Note that

$$\frac{e^{-1/(3d)}}{1-\rho^2} \leqslant 1 \Leftrightarrow e^{-1/(3d)} \leqslant 1-\rho^2 \Leftrightarrow \rho^2 \leqslant 1 - e^{-1/(3d)}.$$

By simple calculus, we have an elementary inequality that $1 - e^{-x} > xe^{-1/3}$ for any $x \in (0, 1/3]$. Therefore, $1 - e^{-1/(3d)} > e^{-1/3}/(3d)$, which is greater than ρ^2. In other words, we have $\frac{m_d}{1-\rho^2} < d$.

□

Lemma 20. *Fix any $w \in \{-d, \dots, 0, 1, \dots, d\}$ and any $\boldsymbol{\mu} \in \mathbb{R}^d$. Denote by $\mathcal{N}_1$ the distribution $N(\rho\boldsymbol{\mu}; (1-\rho^2)\boldsymbol{I}_d)$ and by $\mathcal{N}_0$ the distribution $N(\boldsymbol{0}; \boldsymbol{I}_d)$. Denote by $\boldsymbol{Z}$ the element in the sample space.*

1. *For $\rho^2 < 1/2$, it holds that $|\mathbb{P}_{\mathcal{N}_1}(\boldsymbol{Z} \in A_w) - \mathbb{P}_{\mathcal{N}_0}(\boldsymbol{Z} \in A_w)| \leqslant \rho\|\boldsymbol{\mu}\|$.*
2. *Let $c_0 = \log(2)/4$. For any $k > 0$, there is a positive constant c_k such that for any $\rho < c_k d^{-k-3/2}$, it holds that if $\rho\|\boldsymbol{\mu}\|^2 < c_0$ then $|\mathbb{E}_{\mathcal{N}_1}(\boldsymbol{Z}\ 1_{\boldsymbol{Z}\in A_w}) - \mathbb{E}_{\mathcal{N}_0}(\boldsymbol{Z}\ 1_{\boldsymbol{Z}\in A_w})| \leqslant C\left(d^{-k} + (\rho^{1/3}\|\boldsymbol{\mu}\|)^2 d^{-k/3}\right) + \rho\|\boldsymbol{\mu}\|$.*
3. *In particular, if $\|\boldsymbol{\mu}\| \leqslant C'd^{1/2}$, then there exists a constant $c' > 0$ depending only on k, so that*
   - *(a) for any $\rho \leqslant c'd^{-k-1/2}$, it holds that $|\mathbb{P}_{\mathcal{N}_1}(\boldsymbol{Z} \in A_w) - \mathbb{P}_{\mathcal{N}_0}(\boldsymbol{Z} \in A_w)| \leqslant 8^{-1}d^{-k}$;*
   - *(b) for any $\rho \leqslant c'd^{-k-3/2}$, it holds that $\|\mathbb{E}_{\mathcal{N}_1}(\boldsymbol{Z}\ 1_{\boldsymbol{Z}\in A_w}) - \mathbb{E}_{\mathcal{N}_0}(\boldsymbol{Z}\ 1_{\boldsymbol{Z}\in A_w})\| \leqslant O(d^{-k})$.*

Proof of Lemma 20.

Statement 1.

Using the formula for the KL-divergence between Gaussian measures (see for example, Equation A.23 in Williams and Rasmussen [2006]), we have

$$\begin{aligned}\mathrm{KL}\left(\mathcal{N}_1 || \mathcal{N}_0\right) =& \frac{1}{2}\log\left|\Sigma_0\Sigma_1^{-1}\right| + \\ & \frac{1}{2}\operatorname{tr}\Sigma_0^{-1}\left((\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^\top + \Sigma_1 - \Sigma_0\right) \\ =& \frac{1}{2}\left(-d\log(1-\rho^2) + \|\rho\boldsymbol{\mu}\|^2 + d(1-\rho^2) - d\right) \\ \leqslant& \frac{\rho^2}{2}\|\boldsymbol{\mu}\|^2,\end{aligned}$$

where the last inequality is because $x + \log(1-x) \geqslant 0$ for any $x \in (0, 1/2)$.

By the definition of total variation, we have

$$\begin{aligned}&\left|\mathbb{P}(\rho\boldsymbol{\mu} + \sqrt{1-\rho^2}\boldsymbol{\xi} \in A_w) - \mathbb{P}(\boldsymbol{\xi} \in A_w)\right| \\ \leqslant& TV(\mathcal{N}_1, \mathcal{N}_0) \leqslant \sqrt{\frac{1}{2}\mathrm{KL}\left(\mathcal{N}_1 || \mathcal{N}_0\right)} \leqslant \frac{\rho\|\boldsymbol{\mu}\|}{2},\end{aligned}$$

where the second inequality is Pinsker's inequality (see for example Lemma 2.5 in Tsybakov [2008]).

Statement 2.

By Lemma 19, one can choose any $r \geqslant \sqrt{C_k d}$ such that

$$\mathbb{E}_{\mathcal{N}_0}(\|\boldsymbol{Z}\|\ 1_{\|\boldsymbol{Z}\|>r}) \leqslant \sqrt{\mathbb{P}(\|\boldsymbol{Z}\| > r)\mathbb{E}_{\mathcal{N}_0}(\|\boldsymbol{Z}\|^2\ 1_{\|\boldsymbol{Z}\|>r})} \leqslant Cd^{-k}. \tag{68}$$

Similarly, we can choose any $r \geqslant c_0 + \sqrt{2C_k d}$ such that

$$\mathbb{E}_{\mathcal{N}_1}(\|\boldsymbol{Z}\|\ 1_{\|\boldsymbol{Z}\|>r}) \tag{69}$$

$$
\begin{aligned}
&\leqslant \mathbb{E}_{\mathcal{N}_0}\left(\|\sqrt{1-\rho^2}\boldsymbol{Z}+\rho\boldsymbol{\mu}\| \ 1_{\|\sqrt{1-\rho^2}\boldsymbol{Z}+\rho\boldsymbol{\mu}\|>r}\right)\\
&\leqslant \rho\|\boldsymbol{\mu}\| + \sqrt{1-\rho^2}\mathbb{E}_{\mathcal{N}_0}\left(\|\boldsymbol{Z}\| \ 1_{\|\boldsymbol{Z}\|>(r-\rho\|\boldsymbol{\mu}\|)}/\sqrt{1-\rho^2}\right)\\
&\leqslant \rho\|\boldsymbol{\mu}\| + \mathbb{E}_{\mathcal{N}_0}\left(\|\boldsymbol{Z}\| \ 1_{\|\boldsymbol{Z}\|>\sqrt{C_k d}}\right)\\
&\leqslant \rho\|\boldsymbol{\mu}\| + Cd^{-k}
\end{aligned}
$$

Below, we fix $r = c_0 + \sqrt{2C_k d}$.

Denote by $\phi_1(\boldsymbol{z})$ and $\phi_0(\boldsymbol{z})$ the density functions of $\mathcal{N}_1$ and $\mathcal{N}_0$, respectively. Note that

$$
\frac{\phi_1(\boldsymbol{z})}{\phi_0(\boldsymbol{z})} = \exp\left(\frac{-1}{2(1-\rho^2)}\left(\rho\|\boldsymbol{z}\|^2 - 2\rho\boldsymbol{\mu}^\top\boldsymbol{z} + \rho^2\|\boldsymbol{\mu}\|^2\right)\right)
$$

By calculus, we have an elementary inequality that $t \leqslant e^t - 1 \leqslant 2t$ for any $|t| < \log 2$. Note that $\left|\rho\|\boldsymbol{z}\|^2 - 2\rho\boldsymbol{\mu}^\top\boldsymbol{z} + \rho^2\|\boldsymbol{\mu}\|^2\right| \leqslant 2\rho\|\boldsymbol{z}\|^2 + (\rho+\rho^2)\|\boldsymbol{\mu}\|^2 \leqslant 2\rho(\|\boldsymbol{z}\|^2 + \|\boldsymbol{\mu}\|^2)$.

Recall that $c_0 = (\log 2)/4$. If $\rho\|\boldsymbol{\mu}\|^2 < c_0$ and $\rho r^2 < c_0$, then

$$
\begin{aligned}
&\left\|\mathbb{E}_{\mathcal{N}_1}(\boldsymbol{Z} \ 1_{\boldsymbol{Z}\in A_w, \|\boldsymbol{Z}\|\leqslant r}) - \mathbb{E}_{\mathcal{N}_0}(\boldsymbol{Z} \ 1_{\boldsymbol{Z}\in A_w, \|\boldsymbol{Z}\|\leqslant r})\right\| \qquad (70)\\
\leqslant &\int \boldsymbol{z} \ 1_{\boldsymbol{z}\in A_w, \|\boldsymbol{Z}\|\leqslant r}\phi_0(\boldsymbol{z})\left|\frac{\phi_1(\boldsymbol{z})}{\phi_0(\boldsymbol{z})} - 1\right| \mathrm{d}\,\boldsymbol{z}\\
\leqslant &\int \boldsymbol{z} \ 1_{\boldsymbol{z}\in A_w, \|\boldsymbol{Z}\|\leqslant r}\phi_0(\boldsymbol{z})2\left(2\rho\|\boldsymbol{z}\|^2 + (\rho+\rho^2)\|\boldsymbol{\mu}\|^2\right) \mathrm{d}\,\boldsymbol{z}\\
\leqslant &2\rho r(r^2 + \|\boldsymbol{\mu}\|^2)\int \phi_0(\boldsymbol{z})\,\mathrm{d}\,\boldsymbol{z}.
\end{aligned}
$$

Combining Equations (68), (69), and (70), we have if $\rho r^2 < c_0$, then

$$
\|\mathbb{E}_{\mathcal{N}_1}(\boldsymbol{Z} \ 1_{\boldsymbol{Z}\in A_w}) - \mathbb{E}_{\mathcal{N}_0}(\boldsymbol{Z} \ 1_{\boldsymbol{Z}\in A_w})\| \leqslant \rho\|\boldsymbol{\mu}\| + 2Cd^{-k} + 2\rho r(r^2 + \|\boldsymbol{\mu}\|^2).
$$

Recall that $r = c_0 + \sqrt{2C_k d}$ and $r^2 = O(C_k d)$, one can choose c_k small enough to guarantee $c_k d^{-k-3/2} r^2 < c_0$ and $c_k d^{-k-3/2} r^3 < Cd^{-k}$, from which the desired inequality follows.

Statement 3 is a direct corollary of Statement 2. □

Lemma 21. *Suppose $\boldsymbol{\xi} \sim N(0, \boldsymbol{I}_d)$ and $\rho^2 < (3d)^{-1}$. For any $\boldsymbol{t} \in \mathbb{R}^d$ and any $w \in \{\pm 1, \ldots, \pm d\}$, there exists a universal constant C it holds that*

$$
\|\mathbb{E}\left(\boldsymbol{\xi} \ \mid \sqrt{1-\rho^2}\boldsymbol{\xi} + \boldsymbol{t} \in A_w\right)\| \leqslant C(\sqrt{d} + \|\boldsymbol{t}\|).
$$

Proof. For any $w = \pm 1, \ldots, \pm d$, by definition of A_w, $\sqrt{1-\rho^2}\boldsymbol{\xi} + \boldsymbol{t} \in A_w$ implies that $\|\sqrt{1-\rho^2}\boldsymbol{\xi} + \boldsymbol{t}\| \leqslant \sqrt{m_d} < \sqrt{d}$. Thus $\|\mathbb{E}\left(\boldsymbol{\xi} \ \mid \sqrt{1-\rho^2}\boldsymbol{\xi} + \boldsymbol{t} \in A_w\right)\| \leqslant (\sqrt{d} + \|\boldsymbol{t}\|)/\sqrt{1-\rho^2}$.

For $w = 0$, by the famous Anderson inequality [Anderson, 1955], we have

$$
\mathbb{P}(\|\sqrt{1-\rho^2}\boldsymbol{\xi} + \boldsymbol{t}\| \leqslant \sqrt{m_d}) \leqslant \mathbb{P}(\|\sqrt{1-\rho^2}\boldsymbol{\xi}\| \leqslant \sqrt{m_d}) \leqslant 1 - \pi_0 < 1,
$$

where the second inequality is due to Lemma 19 and $\rho^2 < (3d)^{-1}$. Therefore, $\mathbb{P}(\sqrt{1-\rho^2}\boldsymbol{\xi} + \boldsymbol{t} \in A_0) \geqslant \pi_0$. Note that $\|\mathbb{E}\left(\boldsymbol{\xi} \ 1_{\sqrt{1-\rho^2}\boldsymbol{\xi}+\boldsymbol{t}\in A_0}\right)\| \leqslant \sqrt{\mathbb{E}\|\boldsymbol{\xi}\|^2} = \sqrt{d}$, we conclude that

$$
\|\mathbb{E}\left(\boldsymbol{\xi} \ \mid \sqrt{1-\rho^2}\boldsymbol{\xi} + \boldsymbol{t} \in A_0\right)\| \leqslant \sqrt{d}/\pi_0.
$$

□

E.5 Proofs of Lemmas in Section E.3

We first prove Lemma 17 and Lemma 18. The proof of Lemma 16 is straightforward and can be found at the end of this section.

Following the intuitions in Section E.3, we will make use of the continuous dependence of the probability measure on ρ. Such continuous dependence is uniform if the set $J(W,\rho)$ is bounded. Therefore, we consider dividing the sample space into two parts: one where $J(W,\rho)$ is bounded, and the other where the expectations are negligible.

Let $\boldsymbol{\Pi}$ be the projection matrix onto a $(2d)$-dimensional subspace of $\mathbb{R}^p$ formed by the columns of $\boldsymbol{B}$ and $\widetilde{\boldsymbol{B}}$; if the rank of $[\boldsymbol{B},\widetilde{\boldsymbol{B}}]$ is smaller than $2d$, we can always add some extra columns provided that $2d<p$. Then $\boldsymbol{B}^\top\boldsymbol{X}=\boldsymbol{B}^\top\boldsymbol{\Pi}\boldsymbol{X}$, $\widetilde{\boldsymbol{B}}^\top\boldsymbol{X}=\widetilde{\boldsymbol{B}}^\top\boldsymbol{\Pi}\boldsymbol{X}$, and $\Delta\boldsymbol{B}^\top\boldsymbol{X}=\Delta\boldsymbol{B}^\top\boldsymbol{\Pi}\boldsymbol{X}$. We also have $\|\boldsymbol{B}^\top\boldsymbol{X}+\alpha\Delta\boldsymbol{B}^\top\boldsymbol{X}\|\leqslant\|\boldsymbol{\Pi}\boldsymbol{X}\|$.

In the proofs, we will fix some positive real number $R=O(d)$ and define two events $E_1:=\{\|\boldsymbol{\Pi}\boldsymbol{X}\|\leqslant R\}$ and $E_2:=\{\|\boldsymbol{\Pi}\boldsymbol{X}\|>R\}$. We will also make use of the following facts

1. By Lemma 19, for any $k>0$, we can choose $R=\sqrt{2C_kd}$ so that $\mathbb{P}(E_2)\leqslant(2d)^{-k}$.
2. By Lemma 15, $\epsilon=100\lambda_{0,d}\geqslant d^{-1}$.
3. On the event E_1,
$$\|\boldsymbol{B}^\top\boldsymbol{X}+\rho\alpha(\widetilde{\boldsymbol{B}}-\boldsymbol{B})^\top\boldsymbol{X}\|\leqslant R. \tag{71}$$
4. For any symmetric matrix $\boldsymbol{A}$ of dimension d, $\|\boldsymbol{A}\|\boldsymbol{I}_d-\boldsymbol{A}$ is positive definite. For any two positive definite matrices $\boldsymbol{B}$ and $\boldsymbol{C}$, $\operatorname{Tr}(\boldsymbol{B}\boldsymbol{C})\geqslant0$. Therefore, $\operatorname{Tr}(\boldsymbol{A}\boldsymbol{B})\leqslant\|\boldsymbol{A}\|\operatorname{Tr}(\boldsymbol{B})$.

Proof of Lemma 17. We write

$$\mathbb{E}_{\boldsymbol{B}}\left\{\Delta\boldsymbol{B}^\top\boldsymbol{X}\boldsymbol{X}^\top\Delta\boldsymbol{B}\boldsymbol{I}_d\right\}=\mathbb{E}_{\boldsymbol{B}}\left\{\,1_{E_1}\Delta\boldsymbol{B}^\top\boldsymbol{X}\boldsymbol{X}^\top\Delta\boldsymbol{B}\right\}+\mathbb{E}_{\boldsymbol{B}}\left\{\,1_{E_2}\Delta\boldsymbol{B}^\top\boldsymbol{X}\boldsymbol{X}^\top\Delta\boldsymbol{B}\right\}$$

and

$$\begin{aligned}&-\operatorname{Tr}\left(\mathbb{E}_{\boldsymbol{B}}\left\{\Delta\boldsymbol{B}^\top\boldsymbol{X}\boldsymbol{X}^\top\Delta\boldsymbol{B}\mathbb{E}\left(\boldsymbol{\xi}^{\otimes}\mid\boldsymbol{\xi}\in J(W,\rho),\boldsymbol{X},W\right)\right\}\right)\\ \leqslant&-\operatorname{Tr}\left(\mathbb{E}_{\boldsymbol{B}}\left\{\,1_{E_1}\Delta\boldsymbol{B}^\top\boldsymbol{X}\boldsymbol{X}^\top\Delta\boldsymbol{B}\mathbb{E}\left(\boldsymbol{\xi}^{\otimes}\mid\boldsymbol{\xi}\in J(W,\rho),\boldsymbol{X},W\right)\right\}\right).\end{aligned}$$

We first control the expectation on E_2. Since $\Delta\boldsymbol{B}^\top\boldsymbol{X}=\Delta\boldsymbol{B}^\top\Pi\boldsymbol{X}$, we have

$$\begin{aligned}\operatorname{Tr}\left[\Delta\boldsymbol{B}^\top\boldsymbol{X}\boldsymbol{X}^\top\Delta\boldsymbol{B}\right]&=\operatorname{Tr}\left[(\boldsymbol{\Pi}\boldsymbol{X})^{\otimes}(\Delta\boldsymbol{B})^{\otimes}\right]\\&\leqslant\|(\boldsymbol{\Pi}\boldsymbol{X})^{\otimes}\|\operatorname{Tr}\left[(\Delta\boldsymbol{B})^{\otimes}\right].\end{aligned}\tag{72}$$

Thus

$$\operatorname{Tr}\left[\mathbb{E}\left(\,1_{E_2}\Delta\boldsymbol{B}^\top\boldsymbol{X}\boldsymbol{X}^\top\Delta\boldsymbol{B}\right)\right]\leqslant\|\Delta B\|_F^2\mathbb{E}\left(\|\boldsymbol{\Pi}\boldsymbol{X}\|^2\,1_{E_2}\right).$$

We fix $R=2C_4d$ and obtain $\mathbb{P}(E_2)<(2d)^{-4}$. Note that the second moment of χ^2_{2d} is $4d(d+1)$. By Cauchy–Schwarz inequality, $\mathbb{E}\left(\|\boldsymbol{\Pi}\boldsymbol{X}\|^2\,1_{E_2}\right)\leqslant\mathbb{E}\left(\|\boldsymbol{\Pi}\boldsymbol{X}\|^4\right)^{1/2}\mathbb{P}(E_2)^{1/2}=2\sqrt{d(d+1)}\mathbb{P}(E_2)^{1/2}<d^{-1}$. By Lemma 15, $\epsilon=100\lambda_{0,d}\geqslant d^{-1}$, and thus

$$\operatorname{Tr}\left[\mathbb{E}\left(\,1_{E_2}\Delta\boldsymbol{B}^\top\boldsymbol{X}\boldsymbol{X}^\top\Delta\boldsymbol{B}\right)\right]\leqslant\epsilon\|\Delta B\|_F^2.\tag{73}$$

We next control the two expectations on E_1.

We can split the expectation into parts given by $\{W=w\}$ and use properties of conditional expectation to obtain

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{B}}\left\{\,1_{E_1}\Delta\boldsymbol{B}^\top\boldsymbol{X}\boldsymbol{X}^\top\Delta\boldsymbol{B}\left[-\mathbb{E}\left(\boldsymbol{\xi}^{\otimes}\mid\boldsymbol{\xi}\in J(W,\rho),\boldsymbol{X},W\right)\right]\right\}\\=&\sum_{w=-d}^{d}\mathbb{E}_{\boldsymbol{B}}\left\{\,1_{E_1}\Delta\boldsymbol{B}^\top\boldsymbol{X}\boldsymbol{X}^\top\Delta\boldsymbol{B}\,1_{W=w}\left[-\mathbb{E}\left(\boldsymbol{\xi}^{\otimes}\mid\boldsymbol{\xi}\in J(w,\rho),\boldsymbol{X}\right)\right]\right\}\end{aligned}$$

$$= - \sum_{w=-d}^{d} \mathbb{E}\left\{ 1_{E_1} \Delta \boldsymbol{B}^\top \boldsymbol{X}\boldsymbol{X}^\top \Delta \boldsymbol{B} \mathbb{P}_{\boldsymbol{B}}(W = w \mid \boldsymbol{X}) \left[\mathbb{E}\left(\boldsymbol{\xi}^{\otimes} \mid \boldsymbol{\xi} \in J(w,\rho), \boldsymbol{X}\right)\right]\right\}.$$

In addition, we have

$$\begin{aligned}\boldsymbol{I}_d =& \mathbb{E}\left(\boldsymbol{\xi}^{\otimes} \sum_{w=-d}^{d} 1_{\boldsymbol{\xi} \in J(w,\rho)} \mid \boldsymbol{X}\right)\\ =& \mathbb{E}\left(\boldsymbol{\xi}^{\otimes} \, 1_{\boldsymbol{\xi} \in J(w,\rho)} \mid \boldsymbol{X}\right)\\ =& \sum_{w=-d}^{d} \mathbb{P}(\boldsymbol{\xi} \in J(w,\rho) \mid \boldsymbol{X}) \mathbb{E}\left(\boldsymbol{\xi}^{\otimes} \mid \boldsymbol{\xi} \in J(w,\rho), \boldsymbol{X}\right).\end{aligned}$$

From the last two expressions, we can write

$$\begin{aligned}&\operatorname{Tr}\left(\mathbb{E}_{\boldsymbol{B}}\left\{ 1_{E_1} \Delta \boldsymbol{B}^\top \boldsymbol{X}\boldsymbol{X}^\top \Delta \boldsymbol{B}\left[\boldsymbol{I}_d - \mathbb{E}\left(\boldsymbol{\xi}^{\otimes} \mid \boldsymbol{\xi} \in J(W,\rho), \boldsymbol{X}, W\right)\right]\right\}\right)\\ =& \sum_{w=-d}^{d} \operatorname{Tr}\left(\mathbb{E}\left\{ 1_{E_1} \Delta \boldsymbol{B}^\top \boldsymbol{X}\boldsymbol{X}^\top \Delta \boldsymbol{B} \mathbb{E}\left(\boldsymbol{\xi}^{\otimes} \mid \boldsymbol{\xi} \in J(w,\rho), \boldsymbol{X}\right)\right.\right.\\ &\left.\left.\left[\mathbb{P}(\boldsymbol{\xi} \in J(w,\rho) \mid \boldsymbol{X}) - \mathbb{P}_{\boldsymbol{B}}(W = w \mid \boldsymbol{X})\right]\right\}\right).\end{aligned}$$

On the event E_1, we apply Lemma 20(3) together with the inequality in (71), and conclude that for any $\rho \leqslant c' d^{-5/2}$ and for any $w \in \{-d, \dots, d\}$,

$$|\mathbb{P}(\boldsymbol{\xi} \in J(w,\rho) \mid \boldsymbol{X}) - \mathbb{P}_{\boldsymbol{B}}(W = w \mid \boldsymbol{X})| < \frac{1}{8d^2}.$$

For any $w \neq 0$, since $P_{\boldsymbol{B}}(W = w \mid \boldsymbol{X}) = (4d)^{-1}$, it then follows that $P(\boldsymbol{\xi} \in J(w,\rho) \mid \boldsymbol{X}) \geqslant (8d)^{-1}$ and

$$\begin{aligned}|\mathbb{P}(\boldsymbol{\xi} \in J(w,\rho) \mid \boldsymbol{X}) - \mathbb{P}_{\boldsymbol{B}}(W = w \mid \boldsymbol{X})| &< d^{-1} P(\boldsymbol{\xi} \in J(w,\rho) \mid \boldsymbol{X})\\ &\leqslant \epsilon \mathbb{P}(\boldsymbol{\xi} \in J(w,\rho) \mid \boldsymbol{X}).\end{aligned}$$

The same holds for $w = 0$.

Using the last inequality, we have

$$\begin{aligned}&\operatorname{Tr}\left(\mathbb{E}_{\boldsymbol{B}}\left\{ 1_{E_1} \Delta \boldsymbol{B}^\top \boldsymbol{X}\boldsymbol{X}^\top \Delta \boldsymbol{B}\left[\boldsymbol{I}_d - \mathbb{E}\left(\boldsymbol{\xi}^{\otimes} \mid \boldsymbol{\xi} \in J(W,\rho), \boldsymbol{X}, W\right)\right]\right\}\right)\\ \leqslant& \epsilon \sum_{w=-d}^{d} \operatorname{Tr}\left(\mathbb{E}\left\{ 1_{E_1} \Delta \boldsymbol{B}^\top \boldsymbol{X}\boldsymbol{X}^\top \Delta \boldsymbol{B} \mathbb{P}(\boldsymbol{\xi} \in J(W,\rho) \mid \boldsymbol{X}) \mathbb{E}\left(\boldsymbol{\xi}^{\otimes} \mid \boldsymbol{\xi} \in J(W,\rho), \boldsymbol{X}\right)\right\}\right)\\ =& \epsilon \operatorname{Tr}\left\{\mathbb{E}\left[1_{E_1} \Delta \boldsymbol{B}^\top \boldsymbol{X}\boldsymbol{X}^\top \Delta \boldsymbol{B} \mathbb{E}\left(\boldsymbol{\xi}^{\otimes} \sum_{w=-d}^{d} 1_{\boldsymbol{\xi} \in J(W,\rho)} \mid \boldsymbol{X}\right)\right]\right\}\\ =& \epsilon \operatorname{Tr}\left\{\mathbb{E}\left[1_{E_1} \Delta \boldsymbol{B}^\top \boldsymbol{X}\boldsymbol{X}^\top \Delta \boldsymbol{B} \boldsymbol{I}_d\right]\right\}\\ \leqslant& \epsilon \operatorname{Tr}\left\{\mathbb{E}\left[\Delta \boldsymbol{B}^\top \boldsymbol{X}\boldsymbol{X}^\top \Delta \boldsymbol{B}\right]\right\}\\ =& \epsilon \|\Delta \boldsymbol{B}\|_F^2.\end{aligned}$$

Combining the last inequality with (73), we complete the proof. □

Proof of Lemma 18. In the following, we use C to denote any universal constant, whose value may change from line to line.

We first control the expectation on E_2. Without loss of generality, we can assume $\rho^2 < (3d)^{-1}$. By an elementary trace inequality that $\operatorname{Tr}\left(\boldsymbol{u}\boldsymbol{u}^\top \boldsymbol{v}\boldsymbol{v}^\top\right) \leqslant \|\boldsymbol{v}\|^2 \|\boldsymbol{u}\|^2$, one has

$$\operatorname{Tr}\left(\Delta \boldsymbol{B}^\top \boldsymbol{X}\boldsymbol{X}^\top \Delta \boldsymbol{B}\left[\mathbb{E}\left(\boldsymbol{\xi} \mid \boldsymbol{\xi} \in J(W,\rho), \boldsymbol{X}, W\right)^{\otimes}\right]\right) \tag{74}$$

$$
\begin{aligned}
&\leqslant \|\Delta \boldsymbol{B}^\top \boldsymbol{X}\|^2 \left\|\mathbb{E}\left(\boldsymbol{\xi} \mid \boldsymbol{\xi} \in J(W,\rho), \boldsymbol{X}, W\right)^{\otimes}\right\|^2 \\
&\leqslant \| \ \Delta \boldsymbol{B}^\top \boldsymbol{X}\|^2 C(d+\rho^2\|\boldsymbol{\Pi}\boldsymbol{X}\|^2) \\
&\leqslant \|\Delta \boldsymbol{B}\|_F^2 \|\boldsymbol{\Pi}\boldsymbol{X}\|^2 C(d+\rho^2\|\boldsymbol{\Pi}\boldsymbol{X}\|^2),
\end{aligned}
$$

where the second inequality is due to Lemma 21 and the inequality in (71), and the third is due to the inequality in (72).

Denote by $U := \|\boldsymbol{\Pi}\boldsymbol{X}\|$. Then $U^2 \sim \chi^2_{2d}$. We can pick $R = \sqrt{2C_1 d}$ so that $\mathbb{E}\left(\ 1_{U>R} U^2(d+U^2)\right) \leqslant C d^{-1}$. Then (74) implies that

$$
\begin{aligned}
&\operatorname{Tr}\left(\mathbb{E}_{\boldsymbol{B}}\left\{ \ 1_{E_2} \Delta \boldsymbol{B}^\top \boldsymbol{X}\boldsymbol{X}^\top \Delta \boldsymbol{B}\left[\mathbb{E}\left(\boldsymbol{\xi} \mid \boldsymbol{\xi} \in J(W,\rho), \boldsymbol{X}, W\right)^{\otimes}\right]\right\}\right) \\
\leqslant & C\|\Delta \boldsymbol{B}\|_F^2 \mathbb{E}\left\{ \ 1_{U>R}\left(U^2(d+\rho^2 U^2\right)\right\} \\
\leqslant & C d^{-1}\|\Delta \boldsymbol{B}\|_F^2.
\end{aligned}
$$

We next control the expectation on E_1.

$$
\begin{aligned}
&\mathbb{E}_{\boldsymbol{B}}\left[\mathbb{E}\left(\boldsymbol{\xi} \mid \boldsymbol{\xi} \in J(W,\rho), \boldsymbol{X}, W\right)^{\otimes} \mid \boldsymbol{X}\right] \\
=&\sum_{w=-d}^{d} \mathbb{E}_{\boldsymbol{B}}\left[\ 1_{W=w} \ \mathbb{E}\left(\boldsymbol{\xi} \mid \boldsymbol{\xi} \in J(w,\rho), \boldsymbol{X}\right)^{\otimes} \mid \boldsymbol{X}\right] \\
=&\sum_{w=-d}^{d} \mathbb{E}_{\boldsymbol{B}}\left[\ \mathbb{P}_{\boldsymbol{B}}(W=w \mid \boldsymbol{X}) \mathbb{E}\left(\boldsymbol{\xi} \mid \boldsymbol{\xi} \in J(w,\rho), \boldsymbol{X}\right)^{\otimes} \mid \boldsymbol{X}\right] \\
=&\sum_{w=-d}^{d} \mathbb{E}_{\boldsymbol{B}}\left[\ \mathbb{P}_{\boldsymbol{B}}(W=w \mid \boldsymbol{X}) \mathbb{P}\left(\boldsymbol{\xi} \in J(W,\rho) \mid \boldsymbol{X}\right)^{-2} \mathbb{E}\left(\boldsymbol{\xi} \ 1_{\boldsymbol{\xi} \in J(w,\rho)} \mid \boldsymbol{X}\right)^{\otimes} \mid \boldsymbol{X}\right].
\end{aligned}
$$

Note that as $\rho \to 0$, the limit of $\mathbb{P}\left(\boldsymbol{\xi} \in J(W,\rho) \mid \boldsymbol{X}\right)$ is $\mathbb{P}(W=w \mid \boldsymbol{X})$ and the limit of $\mathbb{E}\left(\boldsymbol{\xi} \ 1_{\boldsymbol{\xi} \in J(w,\rho)} \mid \boldsymbol{X}\right)$ is $\mathbb{E}\left(\boldsymbol{\xi} \ 1_{\boldsymbol{\xi} \in A_w}\right)$. Below we shall bound the deviation of these terms from their respective limits.

We write $\boldsymbol{\mu} = \boldsymbol{B}^\top \boldsymbol{X} + \rho\alpha(\widetilde{\boldsymbol{B}} - \boldsymbol{B})^\top \boldsymbol{X}$. On the event E_1, the inequality in (71) shows that $\|\boldsymbol{\mu}\| \leqslant R$. Denote by $\mathcal{N}_1$ the distribution $N(\rho\boldsymbol{\mu}; (1-\rho^2)\boldsymbol{I}_d)$ and by $\mathcal{N}_0$ the distribution $N(\boldsymbol{0}; \boldsymbol{I}_d)$. We have

$$
\begin{aligned}
\mathbb{E}\left(\boldsymbol{\xi} \ 1_{\boldsymbol{\xi} \in J(w,\rho)} \mid \boldsymbol{X}\right) &= \left[\mathbb{E}\left(\rho\boldsymbol{\mu} + \sqrt{1-\rho^2}\boldsymbol{\xi} \ 1_{\boldsymbol{\xi} \in J(w,\rho)} \mid \boldsymbol{X}\right) - \rho\boldsymbol{\mu}\right](1-\rho^2)^{-1/2} \\
&= \left[\mathbb{E}_{\mathcal{N}_1}(\boldsymbol{Z} \ 1_{\boldsymbol{Z} \in A_w}) - \rho\boldsymbol{\mu}\right](1-\rho^2)^{-1/2}.
\end{aligned}
$$

We apply Lemma 20 and conclude that for any $\rho \leqslant c' d^{-4}$ and for any $w \in \{\pm 1, \ldots, \pm d\}$,

$$
|\mathbb{P}(\boldsymbol{\xi} \in J(w,\rho) \mid \boldsymbol{X}) - \mathbb{P}_{\boldsymbol{B}}(W=w \mid \boldsymbol{X})| < \frac{1}{8d^{7/2}}
$$

and

$$
\begin{aligned}
&\|\mathbb{E}\left(\boldsymbol{\xi} \ 1_{\boldsymbol{\xi} \in J(w,\rho)} \mid \boldsymbol{X}\right) - \mathbb{E}(\boldsymbol{\xi} \ 1_{\boldsymbol{\xi} \in A_w})\| \\
\leqslant &(1-\rho^2)^{-1/2}\left|\mathbb{E}_{\mathcal{N}_1}(\boldsymbol{Z} \ 1_{\boldsymbol{Z} \in A_w}) - \mathbb{E}_{\mathcal{N}_0}(\boldsymbol{Z} \ 1_{\boldsymbol{Z} \in A_w})\right| \quad \ldots \\
&\qquad + \rho(1-\rho^2)^{-1/2}\|\boldsymbol{\mu}\| + (1-(1-\rho^2)^{-1/2})\|\mathbb{E}(\boldsymbol{\xi} \ 1_{\boldsymbol{\xi} \in A_w})\| \\
\leqslant & O(d^{-5/2}) + 2\rho\|\boldsymbol{\mu}\| + \rho^2\sqrt{\log(2d)}/d = O(d^{-5/2}),
\end{aligned}
$$

where we have used $R = O(\sqrt{d})$ and $\|\mathbb{E}(\boldsymbol{\xi} \ 1_{\boldsymbol{\xi} \in A_w})\| = \mathbb{E}(\boldsymbol{\xi}_1 \ 1_{\boldsymbol{\xi} \in A_1}) \leqslant \frac{\sqrt{2\log(2d)}}{2d}$ given in Lemma 15. It also follows that

$$
\|\mathbb{E}\left(\boldsymbol{\xi} \ 1_{\boldsymbol{\xi} \in J(W,\rho)} \mid \boldsymbol{X}\right)\| \leqslant C\sqrt{\log(2d)}/d \text{ and } (8d)^{-1} \leqslant \mathbb{P}(\boldsymbol{\xi} \in J(w,\rho) \mid \boldsymbol{X}) \leqslant (2d)^{-1}.
$$

To utilize the bounds that have been established, we present the following elementary facts. For any positive numbers b, B and any vectors $\boldsymbol{u}$, $\boldsymbol{U}$, we have

1. $b^{-2}\boldsymbol{u}^{\otimes} - B^{-2}\boldsymbol{U}^{\otimes} = b^{-2}\left(\boldsymbol{u}^{\otimes} - \boldsymbol{U}^{\otimes}\right) + (b^{-2} - B^{-2})\boldsymbol{U}^{\otimes}$
2. $\|\boldsymbol{u}^{\otimes} - \boldsymbol{U}^{\otimes}\|_F \leqslant \|\boldsymbol{u}(\boldsymbol{u}-\boldsymbol{U})^\top\|_F + \|(\boldsymbol{u}-\boldsymbol{U})\boldsymbol{U}^\top\|_F \leqslant (\|\boldsymbol{u}\| + \|\boldsymbol{U}\|)\|\boldsymbol{u}-\boldsymbol{U}\|$
3. $|b^{-2} - B^{-2}| \leqslant b^{-2}B^{-2}(b+B)|b-B|$

Now let $B = P_{\boldsymbol{B}}(W = w \mid \boldsymbol{X})$, $b = \mathbb{P}(\boldsymbol{\xi} \in J(w,\rho) \mid \boldsymbol{X})$, $\boldsymbol{U} = \mathbb{E}(\boldsymbol{\xi}\ 1_{\boldsymbol{\xi}\in A_w})$, and $\boldsymbol{u} = \mathbb{E}\left(\boldsymbol{\xi}\ 1_{\boldsymbol{\xi}\in J(W,\rho)} \mid \boldsymbol{X}\right)$. We conclude that on E_1,

$$\begin{aligned}&\left\|\mathbb{P}\left(\boldsymbol{\xi} \in J(w,\rho) \mid \boldsymbol{X}\right)^{-2}\left[\mathbb{E}\left(\boldsymbol{\xi}\ 1_{\boldsymbol{\xi}\in J(w,\rho)} \mid \boldsymbol{X}\right)^{\otimes}\right] - \mathbb{P}_{\boldsymbol{B}}(W = w \mid \boldsymbol{X})^{-2}\left[\mathbb{E}(\boldsymbol{\xi}\ \ 1_{\boldsymbol{\xi}\in A_w})^{\otimes}\right]\right\|_F\\ \leqslant& O(d\sqrt{\log(2d)}d^{-5/2} + d\log(2d)d^{-7/2}) = O(d^{-1}).\end{aligned}$$

A similar result can be obtained for $w = 0$.

Since $\epsilon \geqslant d^{-1}$, there is some universal constant C such that

$$\begin{aligned}&\operatorname{Tr}\left(\mathbb{E}_{\boldsymbol{B}}\left\{\ 1_{E_1}\Delta\boldsymbol{B}^\top\boldsymbol{X}\boldsymbol{X}^\top\Delta\boldsymbol{B}\left[\mathbb{E}\left(\boldsymbol{\xi} \mid \boldsymbol{\xi} \in J(W,\rho), \boldsymbol{X}, W\right)^{\otimes}\right]\right\}\right)\\ \leqslant& \sum_{w=-d}^{d}\operatorname{Tr}\left(\mathbb{E}_{\boldsymbol{B}}\left\{\ 1_{E_1}\Delta\boldsymbol{B}^\top\boldsymbol{X}\boldsymbol{X}^\top\Delta\boldsymbol{B}\mathbb{P}_{\boldsymbol{B}}(W = w \mid \boldsymbol{X})\left[\mathbb{E}(\boldsymbol{\xi}\ \ 1_{\boldsymbol{\xi}\in A_w})^{\otimes}\right]\right\}\right) \quad \dots\\ &\qquad + C\epsilon\mathbb{E}_{\boldsymbol{B}}\left\{\ 1_{E_1}\|\Delta\boldsymbol{B}^\top\boldsymbol{X}\|^2\right\}\\ \leqslant& C\epsilon\|\Delta\boldsymbol{B}\|_F^2 + \operatorname{Tr}\left(\mathbb{E}_{\boldsymbol{B}}\left\{\ 1_{E_1}\Delta\boldsymbol{B}^\top\boldsymbol{X}\boldsymbol{X}^\top\Delta\boldsymbol{B}\Sigma_{0,d}\right\}\right)\\ \leqslant& (C\epsilon + \|\Sigma_{0,d}\|)\|\Delta\boldsymbol{B}\|_F^2,\end{aligned}$$

where $\Sigma_{0,d} = \sum_{w=-d}^{d}\mathbb{P}_{\boldsymbol{B}}(W = w \mid \boldsymbol{X})\left[\mathbb{E}(\boldsymbol{\xi}\ \ 1_{\boldsymbol{\xi}\in A_w})\right]^{\otimes} = \operatorname{Cov}\left[\mathbb{E}\left(\boldsymbol{\xi} \mid \psi(\boldsymbol{\xi})\right)\right]$. By Lemma 15, $\|\Sigma_{0,d}\| = \lambda_{0,d}$.

Remark 4. In the proof, we choose $\rho \leqslant c'd^{-4}$ to simplify the result. In fact, it is sufficient to choose $\rho \leqslant c'd^{-7/2-\xi'}$ for any small $\xi' > 0$.

□

Proof of Lemma 16. It is straightforward to compute the derivative and Hessian of $\log g_w^{\boldsymbol{\mu}}(\boldsymbol{t})$ as follows:

- Let $\phi_d(\boldsymbol{z};\boldsymbol{a};\boldsymbol{M})$ be the p.d.f. of $N(\boldsymbol{a},\boldsymbol{M})$ for $\boldsymbol{a} \in \mathbb{R}^d$ and $\boldsymbol{M}$ be a $d \times d$ positive definite matrix. Then $\frac{\mathrm{d}}{\mathrm{d}\,\boldsymbol{a}}\phi_d(\boldsymbol{z};\boldsymbol{a};\boldsymbol{M}) = \boldsymbol{M}^{-1}(\boldsymbol{z}-\boldsymbol{a})\phi_d(\boldsymbol{z};\boldsymbol{a};\boldsymbol{M})$.
- $g_w^{\boldsymbol{\mu}}(\boldsymbol{t}) = \int\ 1_{\boldsymbol{z}\in A_w}\phi_d\left(\boldsymbol{z};\rho\boldsymbol{\mu}+\rho\boldsymbol{t};(1-\rho^2)\boldsymbol{I}_d\right)\mathrm{d}\,\boldsymbol{z}$. This is obtained by transforming $\boldsymbol{Z} = \rho\boldsymbol{\mu} + \rho\boldsymbol{t} + \sqrt{1-\rho^2}\boldsymbol{\xi}$.
- Then $\nabla g_w^{\boldsymbol{\mu}}(\boldsymbol{t}) = \int\ 1_{\boldsymbol{z}\in A_w}\frac{\rho}{1-\rho^2}(\boldsymbol{z}-\rho\boldsymbol{\mu}-\rho\boldsymbol{t})\phi_d\left(\boldsymbol{z};\rho\boldsymbol{\mu}+\rho\boldsymbol{t};(1-\rho^2)\boldsymbol{I}_d\right)\mathrm{d}\,\boldsymbol{z}$. This can also be written as $\nabla g_w^{\boldsymbol{\mu}}(\boldsymbol{t}) = \sqrt{\frac{\rho^2}{1-\rho^2}}\mathbb{E}\left(\boldsymbol{\xi}\ 1_{\rho\boldsymbol{\mu}+\rho\boldsymbol{t}+\sqrt{1-\rho^2}\boldsymbol{\xi}\in A_w}\right)$.
- $$\begin{aligned}\nabla^2 g_w^{\boldsymbol{\mu}}(\boldsymbol{t}) =& \int\ 1_{\boldsymbol{z}\in A_w}\left(-\frac{\rho^2}{1-\rho^2}\boldsymbol{I}_d + \frac{\rho^2}{(1-\rho^2)^2}(\boldsymbol{z}-\rho\boldsymbol{\mu}-\rho\boldsymbol{t})^{\otimes}\right)\dots\\ &\times\phi_d\left(\boldsymbol{z};\rho\boldsymbol{\mu}+\rho\boldsymbol{t};(1-\rho^2)\boldsymbol{I}_d\right)\mathrm{d}\,\boldsymbol{z}\\ =& -\frac{\rho^2}{1-\rho^2}\boldsymbol{I}_d g_w^{\boldsymbol{\mu}}(\boldsymbol{t})\dots\\ &+\frac{\rho^2}{(1-\rho^2)^2}\int_{A_w}(\boldsymbol{z}-\rho\boldsymbol{\mu}-\rho\boldsymbol{t})^{\otimes}\phi_d\left(\boldsymbol{z};\rho\boldsymbol{\mu}+\rho\boldsymbol{t};(1-\rho^2)\boldsymbol{I}_d\right)\mathrm{d}\,\boldsymbol{z}.\end{aligned}$$

- $\nabla \log g_w^{\boldsymbol{\mu}}(\boldsymbol{t}) = \frac{1}{g_w^{\boldsymbol{\mu}}(\boldsymbol{t})} \nabla g_w^{\boldsymbol{\mu}}(\boldsymbol{t})$.
- $\nabla^2 \log g_w^{\boldsymbol{\mu}}(\boldsymbol{t}) = \frac{1}{g_w^{\boldsymbol{\mu}}(\boldsymbol{t})} \nabla^2 g_w^{\boldsymbol{\mu}}(\boldsymbol{t}) - \left(\frac{1}{g_w^{\boldsymbol{\mu}}(\boldsymbol{t})} \nabla g_w^{\boldsymbol{\mu}}(\boldsymbol{t}) \right)^{\otimes}$.
- By transforming $\boldsymbol{\xi} = (\boldsymbol{Z} - \rho\boldsymbol{\mu} - \rho\boldsymbol{t})/\sqrt{1-\rho^2}$, one has

$$\nabla \log g_w^{\boldsymbol{\mu}}(\boldsymbol{t}) = \sqrt{\frac{\rho^2}{1-\rho^2}} \mathbb{E}\left(\boldsymbol{\xi} \mid \rho\boldsymbol{\mu} + \rho\boldsymbol{t} + \sqrt{1-\rho^2}\boldsymbol{\xi} \in A_w \right)$$

and

$$\frac{1}{g_w^{\boldsymbol{\mu}}(\boldsymbol{t})} \nabla^2 g_w^{\boldsymbol{\mu}}(\boldsymbol{t}) = -\frac{\rho^2}{1-\rho^2} \boldsymbol{I}_d + \frac{\rho^2}{1-\rho^2} \mathbb{E}\left(\boldsymbol{\xi}^{\otimes} \mid \rho\boldsymbol{\mu} + \rho\boldsymbol{t} + \sqrt{1-\rho^2}\boldsymbol{\xi} \in A_w \right).$$

Therefore

$$\begin{aligned} \nabla^2 \log g_w^{\boldsymbol{\mu}}(\boldsymbol{t}) = \frac{\rho^2}{1-\rho^2} \Big\{ & -\boldsymbol{I}_d + \mathbb{E}\left(\boldsymbol{\xi}^{\otimes} \mid \rho\boldsymbol{\mu} + \rho\boldsymbol{t} + \sqrt{1-\rho^2}\boldsymbol{\xi} \in A_w \right) \\ & - \mathbb{E}\left(\boldsymbol{\xi} \mid \rho\boldsymbol{\mu} + \rho\boldsymbol{t} + \sqrt{1-\rho^2}\boldsymbol{\xi} \in A_w \right)^{\otimes} \Big\}. \end{aligned} \tag{75}$$

□

F Proof of Theorem 7

Proof. We only need to prove the lower bounds

$$\inf_{\widehat{\boldsymbol{B}}} \sup_{\mathcal{M} \in \mathfrak{M}_s(p,d,\lambda)} \mathbb{E}_{\mathcal{M}} \left\| \widehat{\boldsymbol{B}} \widehat{\boldsymbol{B}}^{\top} - \boldsymbol{B}\boldsymbol{B}^{\top} \right\|_{\mathrm{F}}^2 \gtrsim \frac{d(s-d)}{n\lambda}. \tag{76}$$

and

$$\inf_{\widehat{\boldsymbol{B}}} \sup_{\mathcal{M} \in \mathfrak{M}_s(p,d,\lambda)} \mathbb{E}_{\mathcal{M}} \left\| \widehat{\boldsymbol{B}} \widehat{\boldsymbol{B}}^{\top} - \boldsymbol{B}\boldsymbol{B}^{\top} \right\|_{\mathrm{F}}^2 \gtrsim \frac{(s-d)\log \frac{e(p-d)}{s-d}}{n\lambda}. \tag{77}$$

To prove the inequality (76), consider the following sub-model which assumes the support of the indices matrix is $\{1, 2, \ldots, s\}$.

$$\widetilde{\mathfrak{M}}_s(p,d,\lambda) := \left\{ \begin{array}{c|l} \text{distribution of} & \boldsymbol{X} \sim N(0, \boldsymbol{I}_p), \epsilon \sim N(0,1) \text{ is independent of } \boldsymbol{X}, \\ \left(\boldsymbol{X}, Y = f(\boldsymbol{B}^{\top}\boldsymbol{X}, \epsilon)\right) & \boldsymbol{B} \text{ is a } p \times d \text{ matrix}, \boldsymbol{B}^{\top}\boldsymbol{B} = \mathbf{I}_d, \quad \operatorname{supp}(\boldsymbol{B}) = [s], \\ & f \in \mathcal{F}_d(\lambda, \kappa, K), K = C_0 d. \end{array} \right\}, \tag{78}$$

A sufficient statistic for estimating $\boldsymbol{B}$ in this submodel is the data of Y and $\boldsymbol{X}_{1:s}$. Because $\widetilde{\mathfrak{M}}_s(p,d,\lambda) \subset \mathfrak{M}_s(p,d,\lambda)$ and $\widetilde{\mathfrak{M}}_s(p,d,\lambda)$ is essentially the same as a submodel of $\mathfrak{M}(s,d,\lambda)$ for $(\boldsymbol{X}_{1:s}, Y)$ with $\operatorname{Cov}(X_{1:s}) = \boldsymbol{I}_s$ assumed known. Following the exact same proof of Theorem 6, we obtain the inequality (76).

To prove the inequality (77), we apply the Fano method (Lemma 11) and the construction of $\mathbb{P}_{\boldsymbol{B}}$ in Equation (10). The main challenge is to construct a rich packing set $\Theta \subset \mathbb{O}_s(p,d)$. The rest of the proof is nearly identical to the proof of Theorem 4 in Lin et al. [2021], which in turn follows from the argument in [Vu and Lei, 2012, Theorem 2.1] and [Cai et al., 2013, Theorem 3]. We present it here for the sake of completeness. For any $\varepsilon \in (0,1]$, [Vu and Lei, 2012, Lemma 3.1.2] have constructed a set $\Theta_0 \subset \mathbb{O}_{s-d+1}(p-d+1, 1)$, such that

1. $\varepsilon/\sqrt{2} < \|\theta_1 - \theta_2\| \leqslant \sqrt{2}\varepsilon$ for all distinct pairs $\theta_1, \theta_2 \in \Theta_0$,
2. $\log|\Theta_0| \geqslant c(s-d)[\log(p-d) - \log(s-d)]$, where c is a positive constant.

For each $\theta \in \Theta_0$, define

$$\boldsymbol{B}_\theta = \begin{pmatrix} \theta & \mathbf{0}_{(p-d+1)\times(d-1)} \\ \mathbf{0}_{(d-1)\times 1} & \mathbf{I}_{d-1} \end{pmatrix}.$$

Then $\boldsymbol{B}_\theta \in \mathbb{O}_s(p,d)$. Let $\Theta = \{\boldsymbol{B}_\theta : \theta \in \Theta_0\}$. We then use the construction in Equation (10) to obtain a family of distributions with the indices matrix $\boldsymbol{B}_\theta$ for each $\theta \in \Theta_0$. Applying the argument in Equation (58) with $\varepsilon^2 = c_1' \frac{(s-d)[\log(p-d)-\log(s-d)]}{n\lambda}$, we obtain Equation (77).

Since $2d < s$, the right hand sides on the inequalities (76) and (77) can be reduced to $\frac{ds}{n\lambda}$ and $\frac{s\log(ep/s)}{n\lambda}$, respectively. □

G Proofs of results in Section 3.1

The gSNR of the distribution in (7) equals to $d^{-1}\left(\mathbb{E}[\max|Z_i|]\right)^2$, which is at the same order as $\frac{\log(d)}{d}$. A proof of this statement is similar to the proof of Proposition 2 in Appendix E.2 and is omitted here.

To prove Theorem 2, we need the following proposition in addition to Theorem 3.

Proposition 4. *Suppose $\boldsymbol{Z}$ is a d-random vector, Y is a random element. Suppose $\{S_h : h = 1, \ldots, H\}$ is a partition of the range of Y such that for some $\gamma > 0$ and $\tau > 1+\gamma$, $\mathbb{P}(Y \in S_h) \leqslant (1+\gamma)H^{-1}$, and for any $\boldsymbol{\beta} \in \mathbb{S}^{d-1}$,*

$$\frac{1}{H}\sum_{h=1}^{H} \operatorname{var}\left(\boldsymbol{\beta}^\top \mathbb{E}\left(\boldsymbol{Z} \mid Y\right) \middle| Y \in S_h\right) \leqslant \frac{1}{\tau}\operatorname{var}\left(\boldsymbol{\beta}^\top \mathbb{E}\left(\boldsymbol{Z} \mid Y\right)\right). \tag{79}$$

Let $W = \sum_{h=1}^{H} h\, 1_{Y \in S_h}$. It holds that for any $\boldsymbol{\beta} \in \mathbb{S}^{d-1}$,

$$\left(1 - \frac{1+\gamma}{\tau}\right)\operatorname{var}\left(\boldsymbol{\beta}^\top \mathbb{E}\left(\boldsymbol{Z} \mid Y\right)\right) \leqslant \operatorname{var}\left(\boldsymbol{\beta}^\top \mathbb{E}\left(\boldsymbol{Z} \mid W\right)\right). \tag{80}$$

Proof of Theorem 2. Since $\boldsymbol{m}(y)$ is weak (K,τ)-sliced stable, Proposition 4 indicates that the smallest eigenvalue of $\operatorname{Cov}\left[\mathbb{E}\left(\boldsymbol{Z} \mid Y\right)\right]$ can be bounded by that of $\operatorname{Cov}\left[\mathbb{E}\left(\boldsymbol{Z} \mid W\right)\right]$ for a discrete random variable W with K outcomes. Then by Theorem 2, the smallest eigenvalue of $\operatorname{Cov}\left[\mathbb{E}\left(\boldsymbol{Z} \mid W\right)\right]$ is bounded by $O(d^{-1}\log K)$. Since $K = O(d)$, this bound becomes $O(\frac{\log(d)}{d})$. □

Proof of Proposition 4. Fix $\boldsymbol{\beta} \in \mathbb{R}^d$ and let $U = \boldsymbol{\beta}^\top \boldsymbol{Z}$. By the law of total variance, we have

$$\begin{aligned}
\operatorname{var}\left(\mathbb{E}\left(U \mid Y\right)\right) &= \mathbb{E}\left\{\operatorname{var}\left[\mathbb{E}\left(U \mid Y\right) \mid W\right]\right\} + \operatorname{var}\left\{\mathbb{E}\left[\mathbb{E}\left(U \mid Y\right) \mid W\right]\right\} \\
&= \sum_{h=1}^{H} \mathbb{P}(W = h)\operatorname{var}\left[\mathbb{E}\left(U \mid Y\right) \mid W = h\right] + \operatorname{var}\left[\mathbb{E}\left(U \mid W\right)\right] \\
&\leqslant \frac{1+\gamma}{H}\sum_{h=1}^{H} \operatorname{var}\left[\mathbb{E}\left(U \mid Y\right) \mid Y \in S_h\right] + \operatorname{var}\left[\mathbb{E}\left(U \mid W\right)\right] \\
&\leqslant \frac{1+\gamma}{\tau}\operatorname{var}\left(\mathbb{E}\left(U \mid Y\right)\right) + \operatorname{var}\left[\mathbb{E}\left(U \mid W\right)\right],
\end{aligned}$$

where the two inequalities are due to the premises. It then follows that

$$\left(1 - \frac{1+\gamma}{\tau}\right)\operatorname{var}\left(\mathbb{E}\left(U \mid Y\right)\right) \leqslant \operatorname{var}\left(\mathbb{E}\left(U \mid W\right)\right)$$

and the proposition is proved. □

Proof of Theorem 3. Suppose $\boldsymbol{\beta} \sim N(0, \boldsymbol{I}_d)$ that is independent with $(\boldsymbol{Z}, W)$. For any symmetric matrix $\boldsymbol{M}$, by the definition of the smallest eigenvalue, it holds that

$$\lambda_{\min}(\boldsymbol{M}) \leqslant \mathbb{E}\left(\frac{\boldsymbol{\beta}^\top}{\|\boldsymbol{\beta}\|}\boldsymbol{M}\frac{\boldsymbol{\beta}}{\|\boldsymbol{\beta}\|}\right).$$

By the normality of $\boldsymbol{\beta}$, we have $1 = d^{-1}\mathbb{E}\|\boldsymbol{\beta}\|^2$ and $\|\boldsymbol{\beta}\|$ is independent of $\boldsymbol{\beta}/\|\boldsymbol{\beta}\|$, and thus

$$\lambda_{\min}(\boldsymbol{M}) \leqslant d^{-1}\mathbb{E}\left(\boldsymbol{\beta}^\top \boldsymbol{M}\boldsymbol{\beta}\right).$$

Therefore, it is sufficient to show

$$\mathbb{E}\left(\boldsymbol{\beta}^\top \mathrm{Cov}\left[\mathbb{E}\left(\boldsymbol{Z} \mid W\right)\right]\boldsymbol{\beta}\right) \leqslant 37 \ \mathrm{Ent}(W).$$

For any $\boldsymbol{u} \in \mathbb{R}^d$, we have $\mathbb{E}\left[\mathbb{E}\left(\boldsymbol{u}^\top \boldsymbol{Z} \mid W\right)\right] = 0$ and

$$\begin{aligned}
&\boldsymbol{u}^\top \mathrm{Cov}\left[\mathbb{E}\left(\boldsymbol{Z} \mid W\right)\right]\boldsymbol{u} \\
=&\mathrm{Cov}\left[\mathbb{E}\left(\boldsymbol{u}^\top \boldsymbol{Z} \mid W\right)\right] \\
=&\mathbb{E}\left(\left[\mathbb{E}\left(\boldsymbol{u}^\top \boldsymbol{Z} \mid W\right)\right]^2\right) \\
=&\mathbb{E}\left(\sum_{i=1}^d\sum_{j=1}^d \mathbb{E}\left(u_iZ_i \mid W\right)\mathbb{E}\left(u_jZ_j \mid W\right)\right) \\
=&\sum_{i=1}^d\sum_{j=1}^d u_iu_j\mathbb{E}\left[\mathbb{E}\left(Z_i \mid W\right)\mathbb{E}\left(Z_j \mid W\right)\right].
\end{aligned}$$

Therefore, using the equation that $\mathbb{E}(\boldsymbol{\beta}_i\boldsymbol{\beta}_j) = 1_{i=j}$ and the independence between $\boldsymbol{\beta}$ and $(\boldsymbol{Z}, W)$, we have

$$\mathbb{E}\left(\boldsymbol{\beta}^\top \mathrm{Cov}\left[\mathbb{E}\left(\boldsymbol{Z} \mid W\right)\right]\boldsymbol{\beta}\right) = \sum_{i=1}^d \mathbb{E}\left(\mathbb{E}\left(Z_i \mid W\right)^2\right).$$

Fixed any w in the support of W. Lemma 22 shows that

$$\sum_{i=1}^d \mathbb{E}\left(Z_i \mid W = w\right)^2 \leqslant \mathbb{P}(W = w)^{-2}\min_\theta\left[\theta\mathbb{P}(W = w) + \mathbb{E}(Z_1 - \theta)_+\right]^2.$$

We now choose θ such that $\mathbb{P}(W = w) = \mathbb{P}(Z_1 > \theta)$. Here $\theta > 0$ because $\mathbb{P}(W = w) < 1/2$. Since $\mathbb{E}(Z_1 - \theta)_+ = \mathbb{E}(Z_1 - \theta)1_{Z_1>\theta} = \mathbb{E}\left(Z_1 1_{Z_1>\theta}\right) - \theta\mathbb{P}(Z_1 > \theta)$, we have

$$\sum_{i=1}^d \mathbb{E}\left(Z_i \mid W = w\right)^2 \leqslant \left[\mathbb{E}\left(Z_1 \mid Z_1 > \theta\right)\right]^2.$$

By Lemma 23, the right hand side of the last inequality is bounded by $37 \ \log\frac{1}{\mathbb{P}(W=w)}$. Hence, we have

$$\mathbb{E}\left(\boldsymbol{\beta}^\top \mathrm{Cov}\left[\mathbb{E}\left(\boldsymbol{Z} \mid W\right)\right]\boldsymbol{\beta}\right) \leqslant \sum_w \mathbb{P}(W = w)\cdot 37 \ \log\frac{1}{\mathbb{P}(W = w)} = 37 \ \mathrm{Ent}(W),$$

and thus complete the proof of Equation G.

In particular, when the support of W has K elements, the maximum entropy of W is maximized by a uniform distribution on these K elements and $\max_W\left\{\mathrm{Ent}(W)\right\} = \log K$. □

Lemma 22. *Suppose the distribution of $\boldsymbol{Z}$ is invariant to orthogonal transformations and W is a discrete random variable. For any w in the support of W and any number θ, it holds that*

$$\|\mathbb{E}(\boldsymbol{Z}1_{W=w})\|^2 \leqslant [\theta\mathbb{P}(W=w) + \mathbb{E}(Z_1-\theta)_+]^2 .$$

Proof of Lemma 22. Let $f(\boldsymbol{z}) = \mathbb{P}(W=w \mid \boldsymbol{Z}=\boldsymbol{z})$ be the conditional probability of $W=w$ given $\boldsymbol{Z}=\boldsymbol{z}$ and $\mathbb{P}(W=w) \in (0,1)$. Then $f(\boldsymbol{z}) \in [0,1]$ and $\mathbb{E}f(\boldsymbol{Z}) = \mathbb{P}(W=w)$.

Let $\boldsymbol{\alpha} = \mathbb{E}\left(\boldsymbol{Z}1_{W=w}\right)$, which equals to $\mathbb{E}\left(\boldsymbol{Z}\mathbb{E}[1_{W=w} \mid \boldsymbol{Z}]\right) = \mathbb{E}\left(\boldsymbol{Z}f(\boldsymbol{Z})\right)$ by the law of total expectation. If $\boldsymbol{\alpha} = \boldsymbol{0}$, the lemma holds trivially.

Assume $\boldsymbol{\alpha} \neq \boldsymbol{0}$. Let $\boldsymbol{V} = [\boldsymbol{V}_1, \ldots, \boldsymbol{V}_d]$ be a d-dimensional orthogonal matrix such that $\boldsymbol{V}_1 = \boldsymbol{\alpha}/\|\boldsymbol{\alpha}\|$. Let $\boldsymbol{U} = \boldsymbol{V}^\top \boldsymbol{Z}$. $\boldsymbol{U}$ has the same distribution as $\boldsymbol{Z}$ because its distribution is invariant to orthogonal transformations. So $\mathbb{E}f(\boldsymbol{U}) = \mathbb{E}f(\boldsymbol{Z}) = \mathbb{P}(W=w)$.

Note that $\boldsymbol{\alpha}^\top \boldsymbol{Z} = \|\boldsymbol{\alpha}\| U_1$. We have $\|\boldsymbol{\alpha}\|^2 = \mathbb{E}\left[\boldsymbol{\alpha}^\top \boldsymbol{Z} f(\boldsymbol{Z})\right] = \|\boldsymbol{\alpha}\|\mathbb{E}\left[U_1 f(\boldsymbol{V}\boldsymbol{U})\right]$.

For any number θ, we have

$$\begin{aligned}
&\|\boldsymbol{\alpha}\| - \theta\mathbb{P}(W=w) \\
=&\mathbb{E}\left[U_1 f(\boldsymbol{V}\boldsymbol{U})\right] - \theta\mathbb{E}f(\boldsymbol{U}) \\
=&\mathbb{E}\left[Z_1 f(\boldsymbol{V}\boldsymbol{Z})\right] - \mathbb{E}\left[\theta f(\boldsymbol{V}\boldsymbol{Z})\right] \\
=&\mathbb{E}\left[(Z_1-\theta) f(\boldsymbol{V}\boldsymbol{Z})\right] \\
\leqslant&\mathbb{E}\left[(Z_1-\theta)_+\right],
\end{aligned}$$

where the last inequality is because $f(\boldsymbol{V}\boldsymbol{Z}) \in [0,1]$ and $ab \leqslant \max(0,a)$ for any $b \in [0,1]$. Therefore, $\|\boldsymbol{\alpha}\| \leqslant \theta\mathbb{P}(W=w) + \mathbb{E}\left[(Z_1-\theta)_+\right]$. □

Lemma 23. *If $Z \sim N(0,1)$ and $\theta > 0$, then*

$$\mathbb{E}(Z \mid Z > \theta)^2 \leqslant 37 \ \log \frac{1}{\mathbb{P}(Z > \theta)}.$$

Proof of Lemma 23. It is well known that for any $t > 0$,

$$(2\pi)^{-1/2}\frac{t}{t^2+1}e^{-t^2/2} \leqslant \mathbb{P}(Z > t) \leqslant e^{-t^2/2}. \tag{81}$$

By direct calculation,

$$\mathbb{E}\left(Z1_{Z>\theta}\right) = (2\pi)^{-1/2}e^{-\theta^2/2}. \tag{82}$$

We consider the value of θ separately in two cases:

1. $\theta \geqslant 1$: Using the first inequality in Equation (81) and Equation (82), we have $\mathbb{E}(Z \mid Z > \theta)^2 \leqslant (\theta + 1/\theta)^2 \leqslant 4\theta^2$. Using the second inequality in Equation (81), we have $\log\frac{1}{\mathbb{P}(Z>\theta)} \geqslant \theta^2/2$.

2. $\theta \in (0,1)$: Then $\mathbb{E}\left(Z1_{Z>\theta}\right) \leqslant \mathbb{E}\left(Z1_{Z>0}\right)$ and $\mathbb{P}(Z > \theta) > \mathbb{P}(Z > 1)$. We have $\mathbb{E}(Z \mid Z > \theta)^2 \leqslant 4e$. Also note that $\mathbb{P}(Z > \theta) < \mathbb{P}(Z > 0) = 1/2$, we have $\log\frac{1}{\mathbb{P}(Z>\theta)} \geqslant \log 2$.

Since $4 < 37/2$ and $4e < 37 \ \log 2$, we conclude the desired inequality for both cases. □

H Assisting Lemmas

Lemma 24. *Let $\boldsymbol{K}$ be an $a \times b$ matrix with each entry being i.i.d. standard normal random variables. Then* $\mathbb{E}[\|\boldsymbol{K}\boldsymbol{K}^\top\|_F^2] = ab(a+b+1)$, $\mathbb{E}[\|\boldsymbol{K}\|_F^2] = ab$ *and* $\mathbb{E}[\|\boldsymbol{K}\|_F^4] = a^2b^2 + 2ab$.

Lemma 25. *Let $\boldsymbol{A}$, $\boldsymbol{B}$ be $l \times m$ and $m \times n$ matrices, respectively. Then one has $\|\boldsymbol{A}\boldsymbol{B}\|_F \leqslant \|\boldsymbol{A}\|\|\boldsymbol{B}\|_F$, where $\|\boldsymbol{A}\|$ denotes the largest singular value of $\boldsymbol{A}$.*

Lemma 26 (Weyl's Inequality)**.** *Let $\boldsymbol{A}$, $\boldsymbol{B}$ be $m \times n$ matrices for some $1 \leqslant m \leqslant n$. Then for all $1 \leqslant i \leqslant m$,*

$$|\sigma_i(\boldsymbol{A}+\boldsymbol{B}) - \sigma_i(\boldsymbol{A})| \leqslant \|\boldsymbol{B}\|_{op}.$$

If $\boldsymbol{A}$, $\boldsymbol{B}$ are $m \times m$ symmetric matrices, then for all $1 \leqslant i \leqslant m$,

$$|\lambda_i(\boldsymbol{A}+\boldsymbol{B}) - \lambda_i(\boldsymbol{A})| \leqslant \|\boldsymbol{B}\|_{op}.$$

Proof. See, e.g., [Tao, 2012, Chapter 1.3]. □

Lemma 27 (Vershynin [2010])**.** *Let $\boldsymbol{A}$ be a $p \times H$ matrix ($p \geqslant H$), whose entries are independent standard normal random variables. Then for every $t \geqslant 0$, with probability at least $1 - 2\exp(-t^2/2)$, one has that*

$$\sqrt{p} - \sqrt{H} - t \leqslant \sigma_H(\boldsymbol{A}) \leqslant \sigma_1(\boldsymbol{A}) \leqslant \sqrt{p} + \sqrt{H} + t.$$

Lemma 28 (Sin-Theta Theorem, Cai et al. [2013])**.** *Let $\boldsymbol{A}$ and $\boldsymbol{A}+\boldsymbol{E}$ be symmetric matrices satisfying*

$$\boldsymbol{A} = [\boldsymbol{F}_0, \boldsymbol{F}_1] \begin{bmatrix} \boldsymbol{A}_0 & 0 \\ 0 & \boldsymbol{A}_1 \end{bmatrix} \begin{bmatrix} \boldsymbol{F}_0^\top \\ \boldsymbol{F}_1^\top \end{bmatrix} \qquad \boldsymbol{A}+\boldsymbol{E} = [\boldsymbol{G}_0, \boldsymbol{G}_1] \begin{bmatrix} \boldsymbol{\Lambda}_0 & 0 \\ 0 & \boldsymbol{\Lambda}_1 \end{bmatrix} \begin{bmatrix} \boldsymbol{G}_0^\top \\ \boldsymbol{G}_1^\top \end{bmatrix}$$

where $[\boldsymbol{F}_0, \boldsymbol{F}_1]$ and $[\boldsymbol{G}_0, \boldsymbol{G}_1]$ are orthogonal matrices. If the eigenvalues of $\boldsymbol{A}_0$ lie in an interval (a, b) and the eigenvalues of $\boldsymbol{\Lambda}_1$ are excluded from the interval $(a - \delta, b + \delta)$ for some $\delta > 0$, then

$$\|\boldsymbol{F}_0\boldsymbol{F}_0^\top - \boldsymbol{G}_0\boldsymbol{G}_0^\top\| \leqslant \frac{\min(\|\boldsymbol{F}_1^\top \boldsymbol{E}\boldsymbol{G}_0\|, \|\boldsymbol{F}_0^\top \boldsymbol{E}\boldsymbol{G}_1\|)}{\delta},$$

and

$$\frac{1}{\sqrt{2}}\|\boldsymbol{F}_0\boldsymbol{F}_0^\top - \boldsymbol{G}_0\boldsymbol{G}_0^\top\|_F \leqslant \frac{\min(\|\boldsymbol{F}_1^\top \boldsymbol{E}\boldsymbol{G}_0\|_F, \|\boldsymbol{F}_0^\top \boldsymbol{E}\boldsymbol{G}_1\|_F)}{\delta}.$$

Lemma 29 (Cai et al. [2013])**.** *Let $\boldsymbol{K} \in \mathbb{R}^{p\times p}$ be symmetric such that $Tr(\boldsymbol{K}) = 0$ and $\|\boldsymbol{K}\|_F \leqslant 1$. Let $\boldsymbol{Z}$ be an $H \times p$ matrix consisting of independent standard normal entries. Then for any $t > 0$, one has*

$$\mathbb{P}\left(\left|\left\langle \boldsymbol{Z}^\top \boldsymbol{Z}, \boldsymbol{K}\right\rangle\right| \geqslant 2\sqrt{H}t + 2t^2\right) \leqslant 2\exp\left(-t^2\right). \tag{83}$$

We remind that this lemma is a trivial modification of Lemma 4 in Cai et al. [2013], where they assumed $\|\boldsymbol{K}\|_F = 1$.

Lemma 30 (Cai et al. [2013])**.** *Let $X_1, ..., X_N$ be random variables such that each satisfies*

$$\mathbb{P}(|X_i| \geqslant at + bt^2) \leqslant c\exp\left(-t^2\right) \tag{84}$$

where $a, b, c > 0$. Then

$$\mathbb{E}\max|X_i|^2 \leqslant (2a^2 + 8b^2)\log(ecN) + 2b^2\log^2(cN). \tag{85}$$

Lemma 31. *Let $\boldsymbol{A}$, $\boldsymbol{B}$ be $m \times l$ orthogonal matrices, i.e., $\boldsymbol{A}^\top\boldsymbol{A} = \mathbf{I}_l = \boldsymbol{B}^\top\boldsymbol{B}$ and $\boldsymbol{M}$ be an $l \times l$ positive definite matrix with eigenvalues d_j such as $0 < \lambda \leqslant d_l \leqslant d_{l-1} \leqslant ... \leqslant d_1 \leqslant \kappa\lambda$. If $\boldsymbol{A}^\top\boldsymbol{B}$ is a diagonal matrix with non-negative entries, then there exists a constant C which only depends on κ such that $\|\boldsymbol{A}\boldsymbol{M}\boldsymbol{A}^\top - \boldsymbol{B}\boldsymbol{M}\boldsymbol{B}^\top\|_F \leqslant C\lambda\|\boldsymbol{A}\boldsymbol{A}^\top - \boldsymbol{B}\boldsymbol{B}^\top\|_F$.*

Proof. We first show that for any symmetric matrix $\boldsymbol{C}$ and positive semi-definite matrix $\boldsymbol{D}$, one has

$$\lambda_{\min}(\boldsymbol{C})\lambda_i(\boldsymbol{D}) \leqslant \lambda_i(\boldsymbol{C}\boldsymbol{D}) \leqslant \lambda_{\max}(\boldsymbol{C})\lambda_i(\boldsymbol{D}).$$

This is because the Courant–Fischer min-max theorem [Tao, 2012, Theorem 1.3.2]:

$$\lambda_i(\boldsymbol{C}\boldsymbol{D}) = \lambda_i(\sqrt{\boldsymbol{D}}\boldsymbol{C}\sqrt{\boldsymbol{D}}) = \inf_{\substack{F\subset\mathbb{R}^n,\\ \dim(F)=n-i+1}} \sup_{x\in F\setminus\{0\}} \frac{(\boldsymbol{C}\sqrt{\boldsymbol{D}}x, \sqrt{\boldsymbol{D}}x)}{(\sqrt{\boldsymbol{D}}x, \sqrt{\boldsymbol{D}}x)} \frac{(\boldsymbol{D}x, x)}{(x, x)}. \tag{86}$$

Let $\Delta = \mathbf{I}_l - \boldsymbol{B}^\top \boldsymbol{A}$, then $0 \leqslant \Delta_{ii} \leqslant 1$ for $1 \leqslant i \leqslant l$, so $\mathrm{Tr}(\Delta^2) \leqslant \mathrm{Tr}(\Delta)$. $4\mathrm{Tr}(\Delta) - 2\mathrm{Tr}(\Delta^2) = \|\boldsymbol{A}\boldsymbol{A}^\top - \boldsymbol{B}\boldsymbol{B}^\top\|_F^2$.

If $C^2 > 2\kappa^2 - 1$, then $(C^2-1)\mathrm{Tr}(\Delta^2) \leqslant (C^2-1)\mathrm{Tr}(\Delta) \leqslant 2(C^2-\kappa^2)\mathrm{Tr}(\Delta)$, that is,

$$2\kappa^2\mathrm{Tr}(\Delta) - \mathrm{Tr}(\Delta^2) \leqslant 2C^2\mathrm{Tr}(\Delta) - 2C^2\mathrm{Tr}(\Delta^2). \tag{87}$$

We have

$$\begin{aligned} \|\boldsymbol{A}\boldsymbol{M}\boldsymbol{A}^\top - \boldsymbol{B}\boldsymbol{M}\boldsymbol{B}^\top\|_F^2 = 4\mathrm{Tr}(\boldsymbol{M}^2\Delta) - 2\mathrm{Tr}(\boldsymbol{M}\Delta\boldsymbol{M}\Delta) &\overset{(a)}{\leqslant} 4\kappa^2\lambda^2\mathrm{Tr}(\Delta) - 2\lambda^2\mathrm{Tr}(\Delta^2) \\ &\overset{(b)}{\leqslant} 2C^2\lambda^2(2\mathrm{Tr}(\Delta) - \mathrm{Tr}(\Delta^2)) = C^2\lambda^2\|\boldsymbol{A}\boldsymbol{A}^\top - \boldsymbol{B}\boldsymbol{B}^\top\|_F^2, \end{aligned}$$

where (a) is obtained by applying (86) for three times with $(\boldsymbol{C}, \boldsymbol{D}) = (\boldsymbol{M}^2, \Delta)$, $(\boldsymbol{C}, \boldsymbol{D}) = (\boldsymbol{M}, \Delta\boldsymbol{M}\Delta)$, and $(\boldsymbol{C}, \boldsymbol{D}) = (\boldsymbol{M}, \Delta^2)$, respectively, while (b) comes from (87). □

Lemma 32. *For a positive definite matrix $\boldsymbol{M}$ with eigenvalue $\lambda_1 \geqslant \ldots \geqslant \lambda_d > 0$ and orthogonal matrices $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{E}, \boldsymbol{F}$, i.e., $\boldsymbol{A}^\top\boldsymbol{A} = \boldsymbol{B}^\top\boldsymbol{B} = \boldsymbol{E}^\top\boldsymbol{E} = \boldsymbol{F}^\top\boldsymbol{F} = \mathbf{I}_d$, one has*

$$\frac{\lambda_d}{2}\|\boldsymbol{A}\boldsymbol{B}^\top - \boldsymbol{E}\boldsymbol{F}^\top\|_F^2 \leqslant \langle \boldsymbol{A}\boldsymbol{M}\boldsymbol{B}^\top, \boldsymbol{A}\boldsymbol{B}^\top - \boldsymbol{E}\boldsymbol{F}^\top \rangle \leqslant \frac{\lambda_1}{2}\|\boldsymbol{A}\boldsymbol{B}^\top - \boldsymbol{E}\boldsymbol{F}^\top\|_F^2.$$

Proof. It is a direct corollary of the Lemma 8 in Gao et al. [2015]. □

I Additional simulation results

The section contains (i) the detailed procedures of sampling from a GP and the additional simulation results in the 'Gaussian process' part of Section 3.2; (ii) the additional simulation results in the 'dependence of error w.r.t. d and λ' part of Section 4.2.2.

I.1 Detailed procedures of sampling from a GP

(i) generate $\boldsymbol{X}_i \overset{iid}{\sim} N(0, I_p)$ and take $\boldsymbol{x}_i = \boldsymbol{B}^\top\boldsymbol{X}_i, i \in [n]$;

(ii) generate $(f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_n))$ from the n-dimensional normal distribution

$$N\left(\{\mu(\boldsymbol{x}_i)\}_{i\in[n]}, \{\Sigma(\boldsymbol{x}_i, \boldsymbol{x}_j)\}_{i,j\in[n]}\right);$$

(iii) generate $\epsilon_i \overset{iid}{\sim} N(0,1), i \in [n]$ and take $Y_i = f(\boldsymbol{x}_i) + \epsilon_i$.

I.2 Average logarithm of gSNR for $H = 10, 20, 30, 50$

This subsection contains average logarithm of estimated gSNR as a function of n for various values of d and as a function of d for various values of n. The H is chosen to be $10, 20, 30, 50$ respectively.

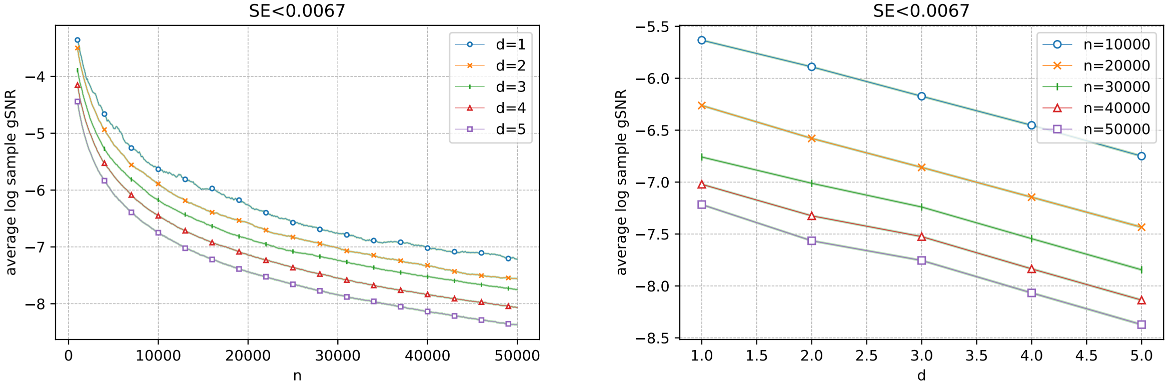


Figure 4: Average logarithm of gSNR with increasing n (left) and increasing d (right) for $H = 10$.

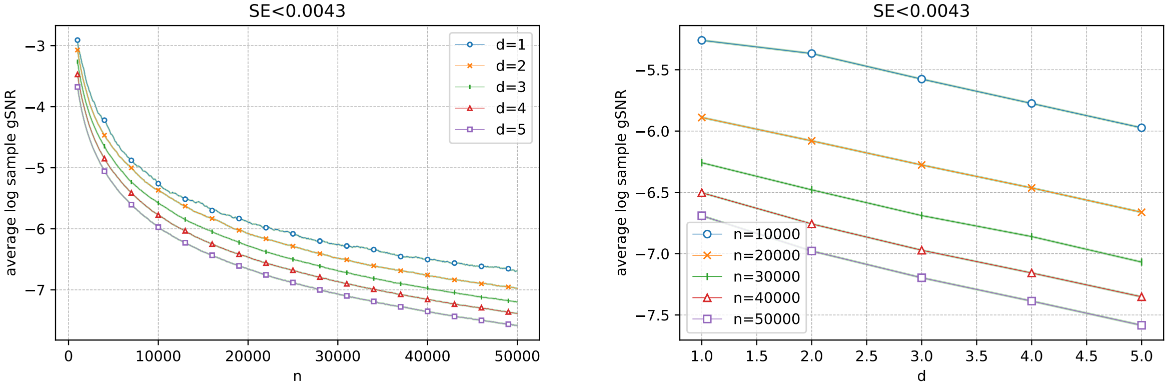


Figure 5: Average logarithm of gSNR with increasing n (left) and increasing d (right) for $H = 20$.

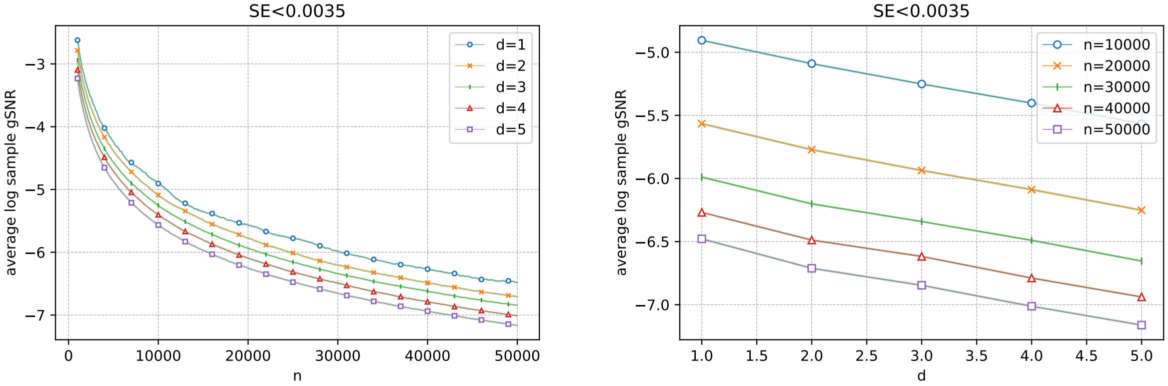


Figure 6: Average logarithm of gSNR with increasing n (left) and increasing d (right) for $H = 30$.

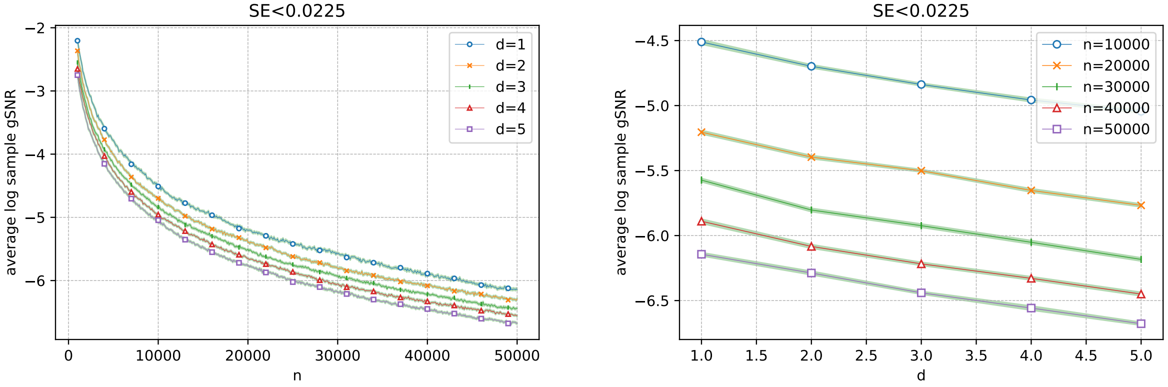
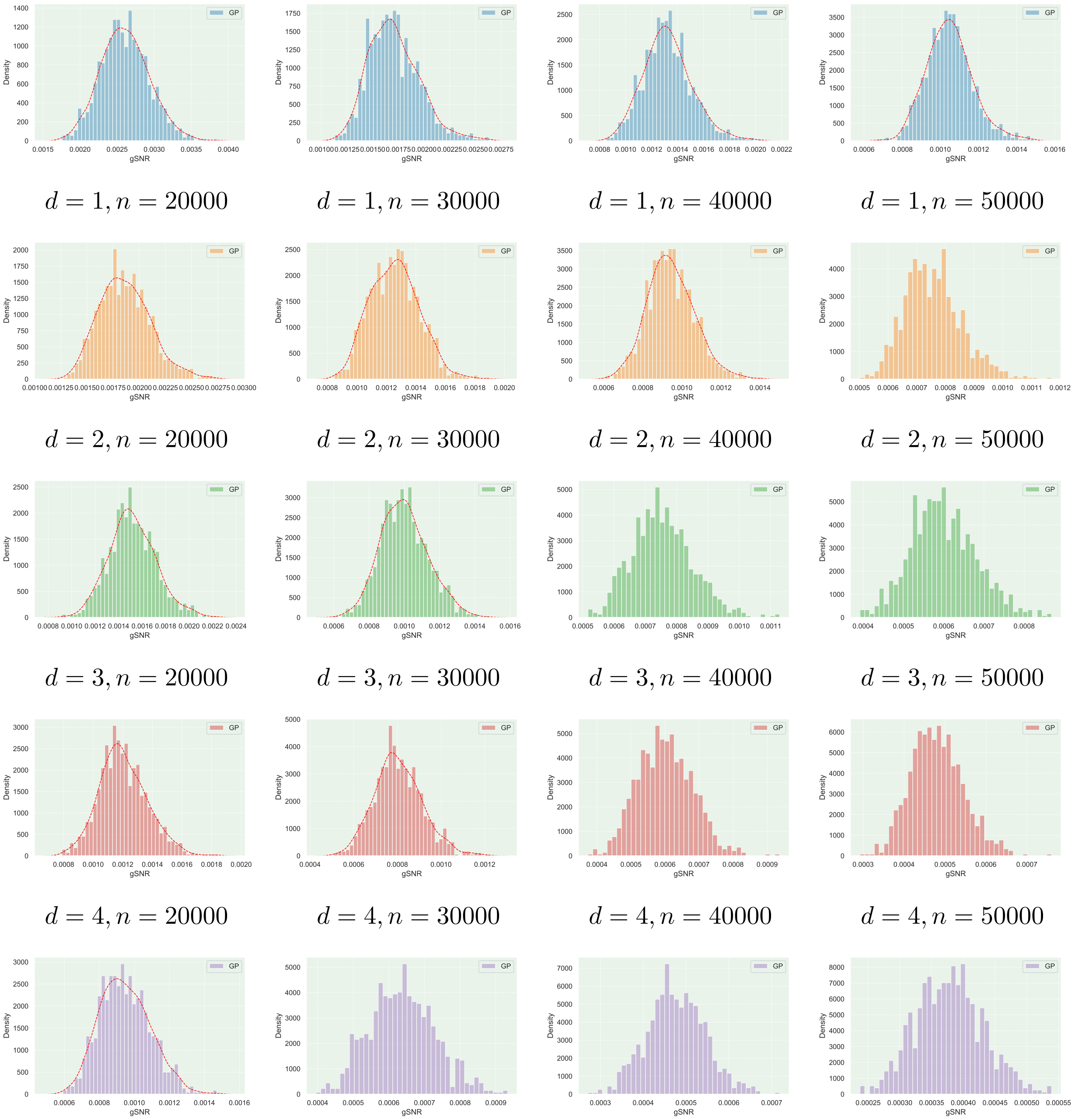


Figure 7: Average logarithm of gSNR with increasing n (left) and increasing d (right) for $H = 50$.

I.3 Histogram of gSNR of GP and pure noise

This subsection contains the histogram of gSNR over $1,000$ random functions with different d and n in Figure 8.



$d = 1, n = 20000$ $d = 1, n = 30000$ $d = 1, n = 40000$ $d = 1, n = 50000$

$d = 2, n = 20000$ $d = 2, n = 30000$ $d = 2, n = 40000$ $d = 2, n = 50000$

$d = 3, n = 20000$ $d = 3, n = 30000$ $d = 3, n = 40000$ $d = 3, n = 50000$

$d = 4, n = 20000$ $d = 4, n = 30000$ $d = 4, n = 40000$ $d = 4, n = 50000$

$d = 5, n = 20000$ $d = 5, n = 30000$ $d = 5, n = 40000$ $d = 5, n = 50000$

Figure 8: Histogram of gSNR of GP

I.4 Additional simulation results in Section 4.2.2

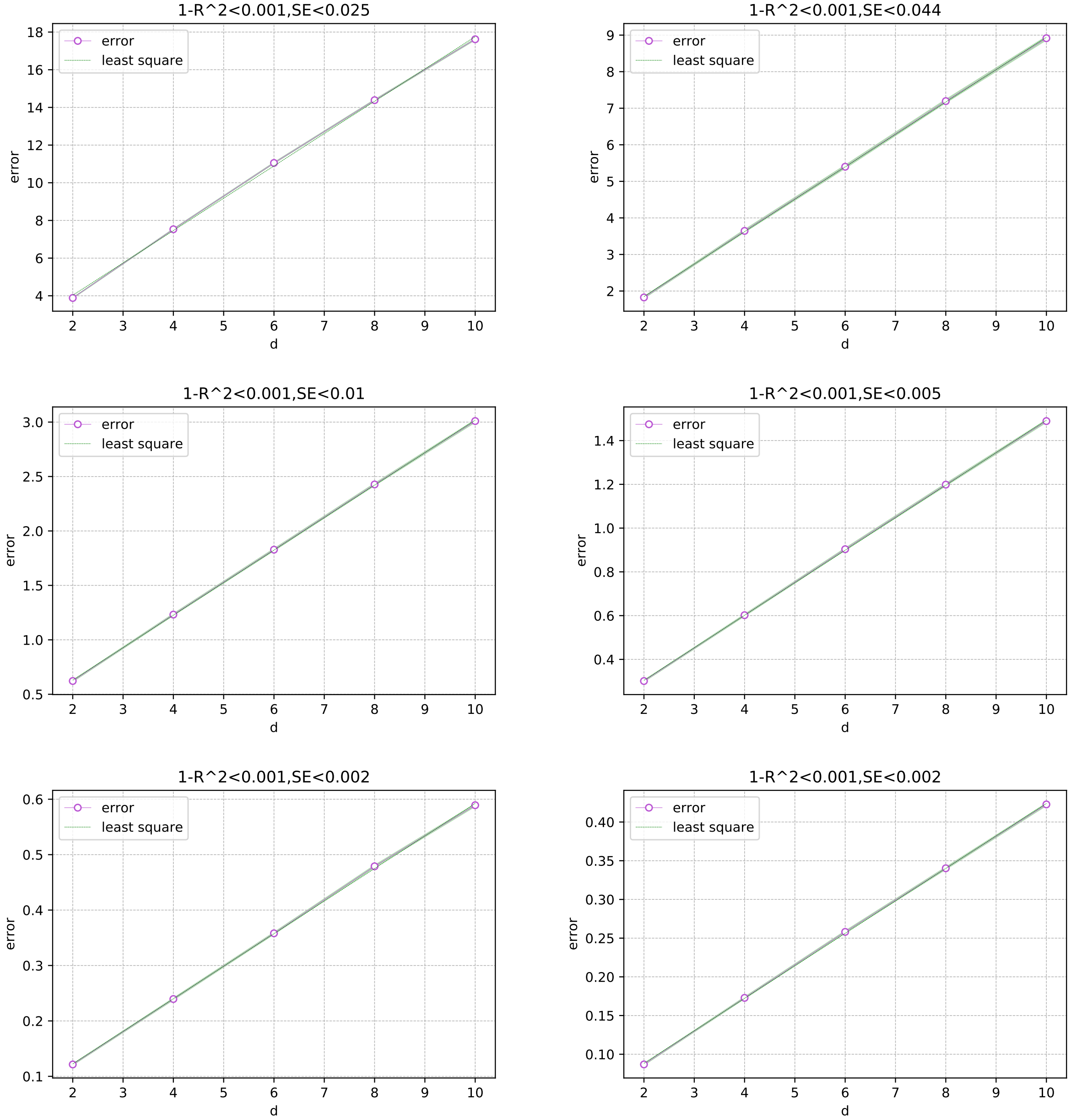


Figure 9: Error with increasing d for $\delta \in \{0.01, 0.02, 0.03, 0.04, 0.06, 0.07\}$

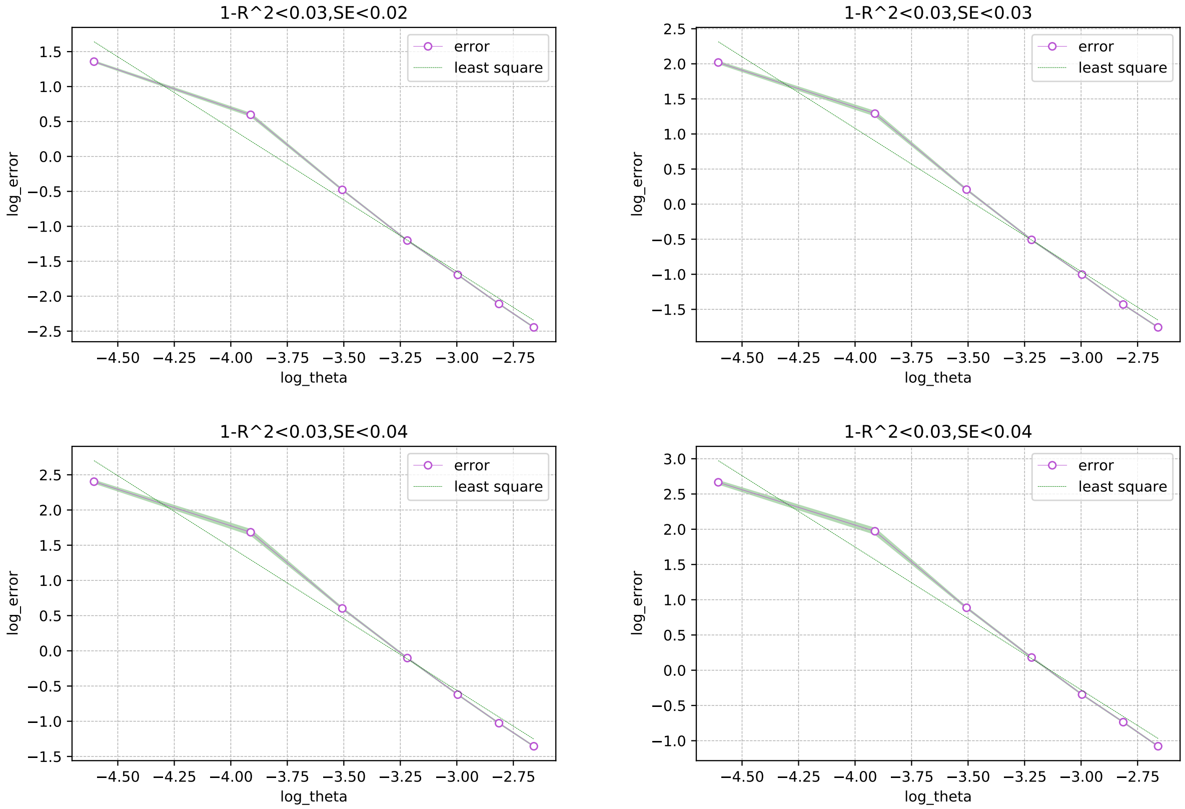


Figure 10: Error with increasing δ for $d \in \{2, 4, 6, 8\}$. The slopes are $-2.049, -2.04, -2.032, -2.026$ respectively.