Journal of Machine Learning Research 22 (2021) 1-45 Submitted 6/20; Revised 12/20; Published 1/21

Optimal Structured Principal Subspace Estimation: Metric Entropy and Minimax Rates

Tony Cai TCAI@WHARTON.UPENN.EDU
Department of Statistics
University of Pennsylvania
Philadelphia, PA 19104, USA

Hongzhe Li HONGZHE@PENNMEDICINE.UPENN.EDU
Department of Biostatistics, Epidemiology and Informatics
University of Pennsylvania
Philadelphia, PA 19104, USA

Rong Ma RONGM@UPENN.EDU
Department of Biostatistics, Epidemiology and Informatics
University of Pennsylvania
Philadelphia, PA 19104, USA

Editor: Ambuj Tewari

Abstract

Driven by a wide range of applications, several principal subspace estimation problems have been studied individually under different structural constraints. This paper presents a unified framework for the statistical analysis of a general structured principal subspace estimation problem which includes as special cases sparse PCA/SVD, non-negative PCA/SVD, subspace constrained PCA/SVD, and spectral clustering. General minimax lower and upper bounds are established to characterize the interplay between the information-geometric complexity of the constraint set for the principal subspaces, the signal-to-noise ratio (SNR), and the dimensionality. The results yield interesting phase transition phenomena concerning the rates of convergence as a function of the SNRs and the fundamental limit for consistent estimation. Applying the general results to the specific settings yields the minimax rates of convergence for those problems, including the previous unknown optimal rates for sparse SVD, non-negative PCA/SVD and subspace constrained PCA/SVD.

Keywords: Low-rank matrix; Metric entropy; Minimax risk; Principal component analysis; Singular value decomposition

1. Introduction

Spectral methods such as the principal component analysis (PCA) and the singular value decomposition (SVD) are a ubiquitous technique in modern data analysis with a wide range of applications in many fields including statistics, machine learning, applied mathematics, and engineering. As a fundamental tool for dimension reduction, the spectral methods aim to extract the low-dimensional structures embedded in the high-dimensional data. In many of these modern applications, the complexity of the data sets and the need of incorporating the existing knowledge from subject areas require the data analysts to take into account the prior structural information on the statistical objects of interest in their analysis. In

particular, many interesting problems in high-dimensional data analysis can be formulated as a structured principal subspace estimation problem where one has the prior knowledge that the underlying principal subspace satisfies certain structural conditions (see Section 1.2 for a list of related problems).

The present paper aims to provide a unified treatment of the structured principal subspace estimation problems that have attracted much recent interest in both theory and practice.

1.1 Problem Setup

To fix ideas, we consider two generic models that have been extensively studied in the literature, namely, the *matrix denoising model* and the *spiked Wishart model*; see, for example, Johnstone (2001); Baik and Silverstein (2006); Paul (2007); Bai and Yao (2008); Cai et al. (2013); Donoho and Gavish (2014); Wang and Fan (2017); Choi et al. (2017); Donoho et al. (2018); Perry et al. (2018); Bao et al. (2018), among many others.

Definition 1 (Matrix Denoising Model) *Let $\mathbf{Y} \in \mathbb{R}^{p_1 \times p_2}$ be the observed data matrix generated from the model $\mathbf{Y} = \mathbf{U}\mathbf{\Gamma}\mathbf{V}^\top + \mathbf{Z}$ where $\mathbf{Z} \in \mathbb{R}^{p_1 \times p_2}$ has i.i.d. entries from $N(0, \sigma^2)$, $\mathbf{\Gamma} \in \mathbb{R}^{r \times r}$ is a diagonal matrix with ordered diagonal entries $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_r > 0$ for $1 \leq r \leq \min\{p_1, p_2\}$, $\mathbf{U} \in O(p_1, r)$, and $\mathbf{V} \in O(p_2, r)$ with $O(p, r) = \{\mathbf{W} \in \mathbb{R}^{p \times r} : \mathbf{W}^\top \mathbf{W} = \mathbf{I}_r\}$ being the set of all $p \times r$ orthonormal matrices.*

Definition 2 (Spiked Wishart Model) *Let $\mathbf{Y} \in \mathbb{R}^{n \times p}$ be the observed data matrix whose rows $Y_i \in \mathbb{R}^p$, $i = 1, \ldots, n$, are independently generated from $N(\boldsymbol{\mu}, \mathbf{U}\mathbf{\Gamma}\mathbf{U}^\top + \sigma^2 \mathbf{I}_p)$ where $\mathbf{U} \in O(p, r)$ with $1 \leq r \leq p$, and $\mathbf{\Gamma} \in \mathbb{R}^{r \times r}$ is diagonal with ordered diagonal entries $\lambda_1 \geq ... \geq \lambda_r > 0$. Equivalently, Y_i can be viewed as $Y_i = X_i + \epsilon_i$ where $X_i \sim N(\boldsymbol{\mu}, \mathbf{U}\mathbf{\Gamma}\mathbf{U}^\top)$, $\epsilon_i \sim N(0, \sigma^2 \mathbf{I}_p)$, and $X_1, \ldots, X_n$ and $\epsilon_1, \ldots, \epsilon_n$ are independent.*

These two models have attracted substantial practical and theoretical interest, and have been studied in different contexts in statistics, probability, and machine learning. The present paper addresses the problem of optimal estimation of the principal (eigen/singular) subspaces spanned by the orthonormal columns of $\mathbf{U}$ (denoted as $\mathrm{span}(\mathbf{U})$), based on the data matrix $\mathbf{Y}$ and the prior structural knowledge on $\mathbf{U}$. Specifically, we aim to uncover the deep connections between the statistical limit of the estimation problem as measured by the minimax risk and the geometric complexity of the parameter spaces as characterized by functions of certain entropy measures.

Since the principal subspaces can be uniquely identified with their associated projection matrices, estimating $\mathrm{span}(\mathbf{U})$ is equivalent to estimating $\mathbf{U}\mathbf{U}^\top$. A commonly used metric for gauging the distance between two linear subspaces $\mathrm{span}(\mathbf{U}_1)$ and $\mathrm{span}(\mathbf{U}_2)$ is

$$d(\mathbf{U}_1, \mathbf{U}_2) = \|\mathbf{U}_1\mathbf{U}_1^\top - \mathbf{U}_2\mathbf{U}_2^\top\|_F.$$

In this paper, we use $d(\cdot, \cdot)$ as the loss function and measure the performance of an estimator $\widehat{\mathbf{U}}$ of $\mathbf{U}$ by the risk

$$\mathcal{R}(\widehat{\mathbf{U}}, \mathbf{U}) = \mathbb{E}[d(\widehat{\mathbf{U}}, \mathbf{U})].$$

1.2 Related Works

The problem considered in this paper can be viewed as a generalization and unification of many interesting problems in high-dimensional statistics and machine learning. We first present a few examples to demonstrate the richness of the structured principal subspace estimation problem and its connections to the existing literature.

1. *Sparse PCA/SVD.* The goal of sparse PCA/SVD is to recover span($\mathbf{U}$) under the assumption that columns of $\mathbf{U}$ are sparse. Sparse PCA has been extensively studied in the past two decades under the spiked Wishart model (see, for example, d'Aspremont et al. (2005); Zou et al. (2006); Shen and Huang (2008); Witten et al. (2009); Yang et al. (2011); Vu and Lei (2012); Cai et al. (2013); Ma (2013); Birnbaum et al. (2013); Cai et al. (2015), among many others). In particular, the exact minimax rates of convergence under the loss $d(\cdot, \cdot)$ was established by Cai et al. (2013) in the general rank-r setting. In contrast, theoretical analysis for the sparse SVD is relatively scarce, and the minimax rate of convergence remains unknown.

2. *Non-negative PCA/SVD.* Non-negative PCA/SVD aims to estimate span($\mathbf{U}$) under the assumption that entries of $\mathbf{U}$ are non-negative. This problem has been studied by Deshpande et al. (2014) and Montanari and Richard (2015) under the rank-one matrix denoising model (r=1), where the statistical limit and certain sharp asymptotics were carefully established. However, it is still unclear what are the minimax rates of convergence for estimating span($\mathbf{U}$) under either rank-one or general rank-r settings under either the spiked Wishart model or matrix denoising model.

3. *Subspace Constrained PCA/SVD.* The subspace constrained PCA/SVD assumes the columns of $\mathbf{U}$ are in some low-dimensional linear subspaces of $\mathbb{R}^p$. In other words, $\mathbf{U} \in \mathcal{C}_A(p,k) = \{\mathbf{U} \in O(p,r) : A\mathbf{U}_{\cdot j} = 0 \text{ for all } 1 \le j \le r\}$ for some rank $(p-k)$ matrix $A \in \mathbb{R}^{p\times(p-k)}$ where $r < k < p$. Estimating the principal subspaces under various linear subspace constraints has been considered in many applications such as network clustering (Wang and Davidson, 2010; Kawale and Boley, 2013; Kleindessner et al., 2019). However, the minimax rates of convergence for subspace constrained PCA/SVD remain unknown.

4. *Spectral Clustering.* Suppose we observe $Y_i \sim N(\boldsymbol{\theta}_i, \sigma^2\mathbf{I}_p)$ independently, where $\boldsymbol{\theta}_i \in \{\boldsymbol{\theta}, -\boldsymbol{\theta}\} \subset \mathbb{R}^p$ for $i = 1, ..., n$. Let $\mathbf{Y} \in \mathbb{R}^{n\times p}$ such that Y_i is the i-th row of $\mathbf{Y}$. We have $\mathbf{Y} = \mathbf{h}\boldsymbol{\theta}^\top + \mathbf{Z}$ where $\mathbf{h} \in \{\pm 1\}^n$ and $\mathbf{Z}$ has i.i.d. entries from $N(0, \sigma^2)$. The spectral clustering of $\{Y_i\}_{1\le i\le n}$ aims to recover the class labels in $\mathbf{h}$. Equivalently, the spectral clustering can be treated as estimating the leading left singular vector $\mathbf{u} = \mathbf{h}/\|\mathbf{h}\|_2$ in the matrix denoising model with $\mathbf{u} \in \mathcal{C}_{\pm}^n = \{\mathbf{u} \in \mathbb{R}^n : \|\mathbf{u}\|_2 = 1, u_i \in \{\pm n^{-1/2}\}\}$. See Azizyan et al. (2013); Jin and Wang (2016); Lu and Zhou (2016); Jin et al. (2017); Cai and Zhang (2018); Giraud and Verzelen (2018); Ndaoud (2018); Löffler et al. (2019) and references therein for recent theoretical results.

In addition to the aforementioned problems, there are many other interesting problems that share the same generic form as the structured principal subspace estimation problem. For example, motivated by applications in the statistical analysis of metagenomics data,

Ma et al. (2019, 2020) considered an approximately rank-one matrix denoising model where the leading singular vector satisfies the monotonicity constraint. In a special case of matrix denoising model, namely, the Gaussian Wigner model $\mathbf{Y} = \lambda \mathbf{u}\mathbf{u}^\top + \mathbf{Z} \in \mathbb{R}^{n\times n}$, where $\mathbf{Z}$ has i.i.d. entries (up to symmetry) drawn from a Gaussian distribution, the Gaussian $\mathbb{Z}/2$ synchronization problem (Javanmard et al., 2016; Perry et al., 2018) aims to recover the leading singular vector $\mathbf{u}$ where $\mathbf{u} \in \{\mathbf{u} \in \mathbb{R}^n : \|\mathbf{u}\|_2 = 1, u_i \in \{\pm n^{-1/2}\}\}$. These important applications provide motivations for a unified framework to study the fundamental difficulty and optimality of these estimation problems.

On the other hand, investigations of metric entropy as a measure of statistical complexity has been one of the central topics in theoretical statistics, ranging from nonparametric function estimation (Yatracos, 1988; Haussler and Opper, 1997b; Yang and Barron, 1999; Yang, 1999; Wu and Yang, 2016), high-dimensional statistical inference (Raskutti et al., 2011; Verzelen, 2012; Vu and Lei, 2012; Cai et al., 2013; Ma, 2013) to statistical learning theory (Haussler and Opper, 1997a; Lugosi and Nobel, 1999; Bousquet et al., 2002; Bartlett and Mendelson, 2002; Koltchinskii, 2006; Lecué and Mendelson, 2009; Cai et al., 2016; Rakhlin et al., 2017). Among them, interesting connections between the complexity of the parameter space and the fundamental difficulty of the statistical problem as quantified by certain minimax risk have been carefully established. In this sense, the current work stands as a step along this direction in the context of principal subspace estimation under some general random matrix models.

1.3 Main Contribution

The main contribution of this paper is three-fold. Firstly, a unified framework is introduced for the study of structured principal subspace estimation problems under both the matrix denoising model and the spiked Wishart model. Novel generic minimax lower bounds and risk upper bounds are established to characterize explicitly the interplay between the information-geometric complexity of the structural set for the principal subspaces, the signal-to-noise ratio (SNR), and the dimensionality of the parameter spaces. The results yield interesting phase transition phenomena concerning the rates of convergence as functions of the SNRs and the fundamental limit for consistent estimation. The general lower and upper bounds reduce determination of the minimax optimal rates for many interesting problems to mere calculations of certain information-geometric quantities. Secondly, to obtain the general risk upper bounds, new technical tools are developed for the analysis of the proposed estimators in their general forms. In addition, the minimax lower bounds rely on careful constructions of multiple composite hypotheses about the structured parameter spaces, and non-trivial calculations of the Kullback-Leibler (KL) divergence between certain mixture probability measures, which can be of independent interest. Thirdly, by directly applying our general results to the specific problems discussed in Section 1.2, we establish the minimax optimal rates for those problems. Among them, the minimax rates for sparse SVD, non-negative PCA/SVD and subspace constrained PCA/SVD, are to our knowledge previously unknown.

1.4 Organization and Notation

The rest of the paper is organized as follows. After introducing the notation at the end of this section, we characterize in Section 2 a minimax lower bound under the matrix denoising model using local metric entropy measures. A general estimator is introduced in Section 3 and its risk upper bound is obtained via certain global metric entropy measures. In Section 4, the spiked Wishart model is discussed in detail and generic risk lower and upper bounds are obtained. The general results are applied in Section 5 to specific settings and minimax optimal rates are established by explicitly calculating the local and global metric-entropic quantities. In Section 6, we address the computational issues of the proposed estimators and discuss some extensions and make connections to some other interesting problems.

For a vector $\mathbf{a} = (a_1, ..., a_n)^\top \in \mathbb{R}^n$, we denote $\mathrm{diag}(a_1, ..., a_n) \in \mathbb{R}^{n\times n}$ as the diagonal matrix whose i-th diagonal entry is a_i, and define the ℓ_p norm $\|\mathbf{a}\|_p = \big(\sum_{i=1}^n a_i^p\big)^{1/p}$. We write $a \wedge b = \min\{a, b\}$ and $a \vee b = \max\{a, b\}$. For a matrix $\mathbf{A} = (a_{ij}) \in \mathbb{R}^{p_1\times p_2}$, we define its Frobenius norm as $\|\mathbf{A}\|_F = \sqrt{\sum_{i=1}^{p_1}\sum_{j=1}^{p_2} a_{ij}^2}$ and its spectral norm as $\|\mathbf{A}\| = \sup_{\|\mathbf{x}\|_2\le 1}\|\mathbf{A}\mathbf{x}\|_2$; we also denote $\mathbf{A}_{\cdot i} \in \mathbb{R}^{p_1}$ as its i-th column and $\mathbf{A}_{i\cdot} \in \mathbb{R}^{p_2}$ as its i-th row. Let $O(p, k) = \{\mathbf{V} \in \mathbb{R}^{p\times k} : \mathbf{V}^\top\mathbf{V} = \mathbf{I}_k\}$ be the set of all $p \times k$ orthonormal matrices and $O_p = O(p, p)$, the set of p-dimensional orthonormal matrices. For a rank r matrix $\mathbf{A} \in \mathbb{R}^{p_1\times p_2}$ with $1 \le r \le p_1 \wedge p_2$, its SVD is denoted as $\mathbf{A} = \mathbf{U}\mathbf{\Gamma}\mathbf{V}^\top$ where $\mathbf{U} \in O(p_1, r)$, $\mathbf{V} \in O(p_2, r)$, and $\mathbf{\Gamma} = \mathrm{diag}(\lambda_1(\mathbf{A}), \lambda_2(\mathbf{A}), ..., \lambda_r(\mathbf{A}))$ with $\lambda_{\max}(\mathbf{A}) = \lambda_1(\mathbf{A}) \ge \lambda_2(\mathbf{A}) \ge ... \ge \lambda_{p_1\wedge p_2}(\mathbf{A}) = \lambda_{\min}(\mathbf{A}) \ge 0$ being the ordered singular values of $\mathbf{A}$. The columns of $\mathbf{U}$ and the columns of $\mathbf{V}$ are the left singular vectors and right singular vectors associated to the non-zero singular values of $\mathbf{A}$, respectively. For a given set S, we denote its cardinality as $|S|$. For sequences $\{a_n\}$ and $\{b_n\}$, we write $a_n = o(b_n)$ or $a_n \ll b_n$ if $\lim_n a_n/b_n = 0$, and write $a_n = O(b_n)$, $a_n \lesssim b_n$ or $b_n \gtrsim a_n$ if there exists a constant C such that $a_n \le Cb_n$ for all n. We write $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $a_n \gtrsim b_n$. Lastly, $c, C, C_0, C_1, ...$ are constants that may vary from place to place.

2. Minimax Lower Bounds via Local Packing

We start with the matrix denoising model. Without loss of generality, we focus on estimating the structured left singular subspace $\mathrm{span}(\mathbf{U})$. Specifically, for a given subset $\mathcal{C} \subset O(p_1, r)$, we consider the parameter space

$$\mathcal{Y}(\mathcal{C}, t, p_1, p_2, r) = \left\{(\mathbf{\Gamma}, \mathbf{U}, \mathbf{V}) : \begin{array}{l}\mathbf{\Gamma} = \mathrm{diag}(\lambda_1, ..., \lambda_r), \mathbf{U} \in \mathcal{C}, \mathbf{V} \in O(p_2, r)\\ Lt \ge \lambda_1 \ge ... \ge \lambda_r \ge t/L > 0\end{array}\right\}, \tag{1}$$

for some fixed constant $L > 1$. For any $\mathbf{U} \in O(p_1, r)$ and $\epsilon \in (0, 1)$, the ϵ-ball centered at $\mathbf{U}$ is defined as

$$\mathbb{B}(\mathbf{U}, \epsilon) = \{\mathbf{U}' \in O(p_1, r) : d(\mathbf{U}', \mathbf{U}) \le \epsilon\},$$

and for any given subset $\mathcal{C} \subset O(p_1, r)$, we define

$$\mathrm{diam}(\mathcal{C}) = \sup_{\mathbf{U}_1,\mathbf{U}_2\in\mathcal{C}} d(\mathbf{U}_1, \mathbf{U}_2).$$

We introduce the concepts of packing and covering of a given set before stating a general minimax lower bound.

Definition 3 (ϵ-packing and ϵ-covering) *Let (V, d) be a metric space and $M \subset V$. We say that $G(M, d, \epsilon) \subset M$ is an* ϵ-packing *of M if for any $m_i, m_j \in G(M, d, \epsilon)$ with $m_i \neq m_j$, it holds that $d(m_i, m_j) > \epsilon$. We say that $H(M, d, \epsilon) \subset M$ is an* ϵ-covering *of M if for any $m \in M$, there exists an $m' \in H(M, d, \epsilon)$ such that $d(m, m') < \epsilon$. We denote $\mathcal{M}(M, d, \epsilon) = \max\{|G(M, d, \epsilon)|\}$ and $\mathcal{N}(M, d, \epsilon) = \min\{|H(M, d, \epsilon)|\}$ as the* ϵ-packing number *and the* ϵ-covering number *of M, respectively.*

Following Yang and Barron (1999), we also define the metric entropy of a given set.

Definition 4 (packing and covering ϵ-entropy) *Let $\mathcal{M}(M, d, \epsilon)$ and $\mathcal{N}(M, d, \epsilon)$ be the ϵ-packing and ϵ-covering number of M, respectively. We call $\log \mathcal{M}(M, d, \epsilon)$ the packing ϵ-entropy and $\log \mathcal{N}(M, d, \epsilon)$ the covering ϵ-entropy of M.*

The following theorem gives a minimax lower bound for estimating $\mathrm{span}(\mathbf{U})$ over $\mathcal{Y}(\mathcal{C}, t, p_1, p_2, r)$, as a function of the cardinality of a local packing set of $\mathcal{C}$, the magnitude of the leading singular values (t), the noise level (σ^2), the rank (r), and the dimension (p_2) of the right singular vectors in $\mathbf{V}$.

Theorem 5 *Under the matrix denoising model $\mathbf{Y} = \mathbf{U}\mathbf{\Gamma}\mathbf{V}^\top + \mathbf{Z}$ where $(\mathbf{\Gamma}, \mathbf{U}, \mathbf{V}) \in \mathcal{Y}(\mathcal{C}, t, p_1, p_2, r)$, suppose there exist some $\mathbf{U}_0 \in \mathcal{C}$, $\epsilon_0 > 0$ and $\alpha \in (0, 1)$ such that a local packing set $G = G(\mathbb{B}(\mathbf{U}_0, \epsilon_0) \cap \mathcal{C}, d, \alpha\epsilon_0)$ satisfies*

$$\epsilon_0 = \left[\left(\frac{\sqrt{c\sigma^2(t^2 + \sigma^2 p_2)}}{t^2} \sqrt{\log |G|} \right) \wedge \mathrm{diam}(\mathcal{C}) \right], \tag{2}$$

for some $c \in (0, 1/640]$. Then, as long as $|G| \geq 2$, it holds that, for $\theta = (\mathbf{\Gamma}, \mathbf{U}, \mathbf{V})$,

$$\inf_{\widehat{\mathbf{U}}} \sup_{\theta \in \mathcal{Y}(\mathcal{C}, t, p_1, p_2, r)} \mathcal{R}(\widehat{\mathbf{U}}, \mathbf{U}) \gtrsim \left[\left(\frac{\sigma\sqrt{t^2 + \sigma^2 p_2}}{t^2} \sqrt{\log |G|} \right) \wedge \mathrm{diam}(\mathcal{C}) \right], \tag{3}$$

where the infimum is over all the estimators based on the observation $\mathbf{Y}$.

The above theorem, to the best of our knowledge, is the first minimax lower bound result for the matrix denoising model under the general parameter space (1). Its proof is separated into two parts. In the strong signal regime ($t^2 \gtrsim \sigma^2 p_2$), the minimax lower bound can be obtained by generalizing the ideas in Vu and Lei (2012, 2013) and Cai et al. (2013), where a general lower bound for testing multiple hypotheses (Lemma 30) is applied to obtain (3). In contrast, the analysis is much more complicated in the weak signal regime ($t^2 \lesssim \sigma^2 p_2$) due to the asymmetry between $\mathbf{U}$ and $\mathbf{V}$: the dependence on p_2 needs to be captured by extra efforts in the lower bound construction (Cai and Zhang, 2018), which is different from the aforementioned works on sparse PCA. To achieve this, our analysis relies on a generalized Fano's method for testing multiple composite hypotheses (Lemma 31) and a nontrivial calculation of the pairwise KL divergence between certain mixture probability measures (Lemma 32).

For general $\mathcal{C}$, the existence of ϵ_0 satisfying (2) is not guaranteed by itself – it has to be determined case by case. In addition, for different constructions of the local packing set $G(\mathbb{B}(\mathbf{U}_0, \epsilon_0) \cap \mathcal{C}, d, \alpha\epsilon_0)$ satisfying (2), the magnitude of ϵ_0 could also vary. Therefore, with

our general results, obtaining a sharp minimax lower bound for a specific problem reduces to identifying a local packing set with largest possible ϵ_0 such that equation (2) holds. See Section 5 for specific examples of calculating such local packing entropy.

A key observation from the above theorem is the role of the local packing set $G(\mathbb{B}(\mathbf{U}_0, \epsilon_0) \cap \mathcal{C}, d, \alpha\epsilon_0)$ and its entropy measure $\log |G(\mathbb{B}(\mathbf{U}_0, \epsilon_0) \cap \mathcal{C}, d, \alpha\epsilon_0)|$ in characterizing the fundamental difficulty of the estimation problem. Similar phenomena connecting the local packing numbers to the minimax lower bounds has been observed in, for example, nonparametric function estimation (Yang and Barron, 1999), high-dimensional linear regression (Raskutti et al., 2011; Verzelen, 2012), and sparse principal component analysis (Vu and Lei, 2012; Cai et al., 2013). Compared with global packing, local packing reflects the fact that the fundamental difficulty of the estimation problem as quantified by the minimax risk is usually determined by the *local* geometry of the parameter space around certain worst-case scenario. In particular, most of the current techniques for deriving the minimax lower bound rely on constructing hypotheses about the parameter of interest that are as separated as possible, while their corresponding probability measures are asymptotically indistinguishable (i.e., having bounded KL divergence or χ^2 divergence). In this case, after careful constructions, the parameters corresponding to the local packing set oftentimes meet this simultaneous requirement of separateness and closeness, whereas it is difficult to achieve based on the global packing set.

In Cai and Zhang (2018), the minimax rate for estimating $\mathrm{span}(\mathbf{U})$ under the unstructured matrix denoising models (i.e., $\mathcal{C} = O(p_1, r)$) was shown to be

$$\inf_{\widehat{\mathbf{U}}} \sup_{(\mathbf{\Gamma},\mathbf{U},\mathbf{V})\in\mathcal{Y}(O(p_1,r),t,p_1,p_2,r)} \mathcal{R}(\widehat{\mathbf{U}}, \mathbf{U}) \asymp \left(\frac{\sigma\sqrt{(t^2+\sigma^2 p_2)rp_1}}{t^2} \wedge \sqrt{r}\right). \tag{4}$$

In light of the packing number estimates for the orthogonal group (Lemma 1 of Cai et al. (2013)), one can show that, for $\mathcal{C} = O(p_1, r)$, there exists a local packing set $G(\mathbb{B}(\mathbf{U}_0, \epsilon_0) \cap \mathcal{C}, d, \alpha\epsilon_0)$ satisfying (2) such that $\log |G(\mathbb{B}(\mathbf{U}_0, \epsilon_0)\cap\mathcal{C}, d, \alpha\epsilon_0)| \asymp p_1 r$. Hence, the lower bound in (4) is a direct consequence of (3). In addition, comparing the lower bound (3) to (4), we observe that the information-geometric quantity $\log |G(\mathbb{B}(\mathbf{U}_0, \epsilon_0) \cap \mathcal{C}, d, \alpha\epsilon_0)|$ essentially quantifies the *intrinsic statistical dimension* of the set $\mathcal{C}$.

3. Risk Upper Bound using Dudley's Entropy Integral

In this section, we consider a general singular subspace estimator and study its theoretical properties. Specifically, we obtain its risk upper bound which, analogous to the minimax lower bound, can be expressed as a function of certain entropic measures related to the structural constraint $\mathcal{C}$.

Under the matrix denoising model, with the parameters $(\mathbf{\Gamma}, \mathbf{U}, \mathbf{V}) \in \mathcal{Y}(\mathcal{C}, t, p_1, p_2, r)$ for some given set $\mathcal{C} \subset O(p_1, r)$, we consider the structured singular subspace estimator

$$\widehat{\mathbf{U}} = \arg\max_{\mathbf{U}\in\mathcal{C}} \mathrm{tr}(\mathbf{U}^\top \mathbf{Y}\mathbf{Y}^\top \mathbf{U}), \tag{5}$$

which is also the constrained maximum likelihood estimator. Before stating our main theorem, we need to make more definitions about quantities that play important roles in our subsequent discussions.

Definition 6 *For given $\mathcal{C} \subset O(p_1, r)$ and any $\mathbf{U} \in \mathcal{C}$, we define the set*

$$\mathcal{T}(\mathcal{C}, \mathbf{U}) = \left\{ \frac{\mathbf{W}\mathbf{W}^\top - \mathbf{U}\mathbf{U}^\top}{\|\mathbf{W}\mathbf{W}^\top - \mathbf{U}\mathbf{U}^\top\|_F} \in \mathbb{R}^{p_1 \times p_1} : \mathbf{W} \in \mathcal{C} \setminus \{\mathbf{U}\} \right\},$$

equipped with the Frobenius distance d_2, where for any $\mathbf{D}_1, \mathbf{D}_2 \in \mathcal{T}(\mathcal{C}, \mathbf{U})$, we define $d_2(\mathbf{D}_1, \mathbf{D}_2) = \|\mathbf{D}_1 - \mathbf{D}_2\|_F$.

Definition 7 (Dudley's entropy integral) *For a metric space (T, d) and a subset $A \subset T$, Dudley's entropy integral of A is defined as $D(A, d) = \int_0^\infty \sqrt{\log \mathcal{N}(A, d, \epsilon)} d\epsilon$. Moreover, we define $D'(A, d) = \int_0^\infty \log \mathcal{N}(A, d, \epsilon) d\epsilon$.*

The Dudley's entropy integral measures the geometric complexity of a given set. Its geometric properties as well as its relationships with other geometric complexity measures such as the Gaussian width and the Sudakov minoration estimate (Cai et al., 2016) have been carefully studied in literature. See, for example, the well-celebrated Sudakov minoration theorem (Ledoux and Talagrand, 2013) and the Dudley's theorem (Dudley, 2010).

Theorem 8 *Under the matrix denoising model, for any given subset $\mathcal{C} \subset O(p_1, r)$ and the parameter space $\mathcal{Y}(\mathcal{C}, t, p_1, p_2, r)$, if $t/\sigma \gtrsim \sup_{\mathbf{U} \in \mathcal{C}} [D'(\mathcal{T}(\mathcal{C}, \mathbf{U}), d_2)/D(\mathcal{T}(\mathcal{C}, \mathbf{U}), d_2)]$, it holds that*

$$\sup_{(\mathbf{\Gamma}, \mathbf{U}, \mathbf{V}) \in \mathcal{Y}(\mathcal{C}, t, p_1, p_2, r)} \mathcal{R}(\widehat{\mathbf{U}}, \mathbf{U}) \lesssim \left(\frac{\sigma \Delta(\mathcal{C}) \sqrt{t^2 + \sigma^2 p_2}}{t^2} \wedge \mathrm{diam}(\mathcal{C}) \right), \tag{6}$$

where $\Delta(\mathcal{C}) = \sup_{\mathbf{U} \in \mathcal{C}} D(\mathcal{T}(\mathcal{C}, \mathbf{U}), d_2)$.

The proof of the above theorem, as it concerns the generic estimator (5) under some arbitrary structural set $\mathcal{C}$, is involved and very different from the existing works such as Cai et al. (2013), Deshpande et al. (2014), Cai and Zhang (2018) and Zhang et al. (2018), where specific examples of $\mathcal{C}$ are considered. The argument relies on careful analyses of the supremum of a Gaussian chaos of order two, and the supremum of a Gaussian process. In the latter case, we applied Dudley's integral inequality (Theorem 23) and the invariance property of the covering numbers with respect to Lipschitz maps (Lemma 24), whereas in the former case, the Arcones-Giné decoupling inequality (Theorem 25) as well as a generic chaining argument (Theorem 28) were used to obtain the desired upper bounds. Many technical tools concatenated for the proof of this theorem can be of independent interest. See more details in Section A.1.

Interestingly, both the risk upper bound (6) and the minimax lower bound (3) indicate two phase transitions when treated as a function of the SNR t/σ, with the first critical point

$$\frac{t}{\sigma} \asymp \sqrt{p_2}, \tag{7}$$

and the second critical point

$$\frac{t}{\sigma} \asymp \left[\frac{\zeta}{\mathrm{diam}^2(\mathcal{C})} + \sqrt{\frac{\zeta p_2}{\mathrm{diam}^2(\mathcal{C})}} \right]^{1/2}, \tag{8}$$

where in the upper bound $\zeta = \Delta^2(\mathcal{C})$ and in the lower bound $\zeta = \log|G(\mathbb{B}(\mathbf{U}_0, \epsilon_0) \cap \mathcal{C}, d, \alpha\epsilon_0)|$. Specifically, the phase transition at the first critical point highlights the role of the dimensionality of the right singular vectors ($\mathbf{V}$) and the change of the rates of convergence from an inverse quadratic function ($\sigma^2\sqrt{p_2\zeta}/t^2$) to an inverse linear function ($\sigma\sqrt{\zeta}/t$) of t/σ. The message from the second phase transition concerns the statistical limit of the estimation problem: consistent estimation is possible only when the SNR exceeds the critical point (8) asymptotically. See Figure 1 for a graphical illustration. As for the implications of the condition

$$t/\sigma \gtrsim \sup_{\mathbf{U}\in\mathcal{C}}[D'(\mathcal{T}(\mathcal{C}, \mathbf{U}), d_2)/D(\mathcal{T}(\mathcal{C}, \mathbf{U}), d_2)] \tag{9}$$

required by Theorem 8, it can be seen in Section 5 that, for many specific problems, a sufficient condition for (9) is that t/σ is above the second critical point (8), which is mild and natural since the latter condition characterizes the region where $\widehat{\mathbf{U}}$ is consistent and more generally where consistent estimation is possible.

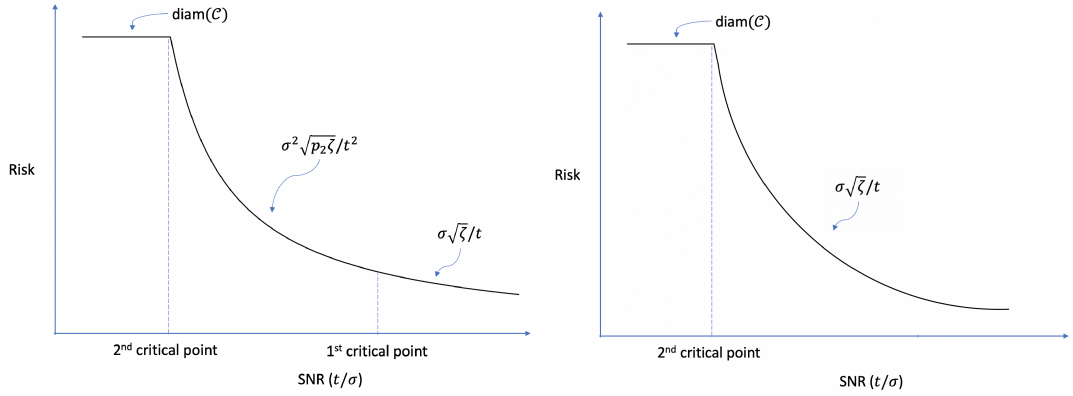

Figure 1: Graphical illustrations of the phase transitions of the risk as a function of the SNRs under the matrix denoising model. Left: $\zeta/\mathrm{diam}^2(\mathcal{C}) \ll p_2$; Right: $\zeta/\mathrm{diam}^2(\mathcal{C}) \gg p_2$

Another interesting phenomena demonstrated by our analysis concerns the relationship between the two critical points. Specifically, when $\zeta/\mathrm{diam}^2(\mathcal{C}) \ll p_2$, the second critical point becomes $\big(\zeta p_2/\mathrm{diam}^2(\mathcal{C})\big)^{1/4}$, which is much smaller than the first critical point $\sqrt{p_2}$ (Figure 1, left); when $\zeta/\mathrm{diam}^2(\mathcal{C}) \gg p_2$, the second critical point becomes $\big(\zeta/\mathrm{diam}^2(\mathcal{C})\big)^{1/2}$, which is much larger than $\sqrt{p_2}$ so that in this case the first critical point disappears (Figure 1, right). In general, the above discrepancy has deep implications as to the fundamental difficulty of the estimation problem. For example, in the unstructured case, it can be shown that $\zeta/\mathrm{diam}^2(\mathcal{C}) \asymp p_1$, so that the behavior of minimax rates for estimating $\mathrm{span}(\mathbf{U})$ relies heavily on the relative magnitude between p_1 and p_2 (Cai and Zhang, 2018).

Comparing our risk upper bound (6) to the minimax lower bound (3), we can observe the similar role played by the information-geometric quantities that characterize the intrinsic statistical dimension of the sets $\mathcal{C}$ or $\mathcal{T}(\mathcal{C}, \mathbf{U})$. Specifically, in (6), the quantity $\Delta(\mathcal{C})$ is related to the global covering entropy, whereas in (3), the quantity $\sqrt{\log|G(\mathbb{B}(\mathbf{U}_0, \epsilon_0) \cap \mathcal{C}, d, \alpha\epsilon_0)|}$

is associated to the local packing entropy. To obtain the minimax optimal rate of convergence, we need to compare the above two quantities and show

$$\Delta^2(\mathcal{C}) \asymp \log |G(\mathbb{B}(\mathbf{U}_0, \epsilon_0) \cap \mathcal{C}, d, \alpha\epsilon_0)|. \tag{10}$$

For example, under the unconstrained setting, direct calculation yields $\Delta^2(\mathcal{C}) \asymp p_1 r$, which coincides with the local packing entropy and thus leads to the sharp rate in (4). However, proving the above equation in its general form is difficult. Alternatively, we briefly discuss the affinity between these two geometric quantities yielded by information theory and leave more detailed discussions in the context of some specific examples in Section 5.

By definition of the packing numbers, we have the relationship

$$\log |G(\mathbb{B}(\mathbf{U}_0, \epsilon_0) \cap \mathcal{C}, d, \alpha\epsilon_0)| \leq \log \mathcal{M}(\mathbb{B}(\mathbf{U}_0, \epsilon_0) \cap \mathcal{C}, d, \alpha\epsilon_0), \tag{11}$$

that links $\log |G(\mathbb{B}(\mathbf{U}_0, \epsilon_0) \cap \mathcal{C}, d, \alpha\epsilon_0)|$ to the local packing entropy. A well known fact about the equivalence between the packing and the covering number of a set M is that

$$\mathcal{M}(M, d, 2\epsilon) \leq \mathcal{N}(M, d, \epsilon) \leq \mathcal{M}(M, d, \epsilon). \tag{12}$$

Moreover, Yang and Barron (1999) obtained a very interesting result connecting the local and the global (covering) metric entropies. Specifically, let $\mathbf{U}$ be any element from M, then

$$\log \mathcal{M}(M, d, \epsilon/2) - \log \mathcal{M}(M, d, \epsilon) \leq \log \mathcal{M}(\mathbb{B}(\mathbf{U}, \epsilon) \cap M, d, \epsilon/2) \leq \log \mathcal{M}(M, d, \epsilon). \tag{13}$$

In Section 5, by focusing on some specific examples of $\mathcal{C}$ that are widely considered in practice, we show that equation (10) holds, which along with our generic lower and upper bounds recovers some existing minimax rates, and more importantly, helps to establish some previously unknown rates.

4. Structured Eigen Subspace Estimation in the Spiked Wishart Model

We turn the focus in this section to the spiked Wishart model where one has $i.i.d.$ observations $Y_i \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma} = \mathbf{U}\boldsymbol{\Gamma}\mathbf{V}^\top + \sigma^2\mathbf{I}$, which is usually referred as the spiked covariance. Similar to the matrix denoising model, a minimax lower bound based on some local packing set and a risk upper bound based on the Dudley's entropy integral can be obtained.

4.1 Minimax Lower Bound

For any given subset $\mathcal{C} \subset O(p, r)$, we consider the parameter space

$$\mathcal{Z}(\mathcal{C}, t, p, r) = \{(\boldsymbol{\Gamma}, \mathbf{U}) : \boldsymbol{\Gamma} = \mathrm{diag}(\lambda_1, ..., \lambda_r), Lt \geq \lambda_1 \geq ... \geq \lambda_r \geq t/L > 0, \mathbf{U} \in \mathcal{C}\},$$

where $L > 1$ is some fixed constant. The following theorem provides the minimax lower bound for estimating $\mathrm{span}(\mathbf{U})$ over $\mathcal{Z}(\mathcal{C}, t, p, r)$ under the spiked Wishart model.

Theorem 9 *Under the spiked Wishart model where* $(\mathbf{\Gamma}, \mathbf{U}) \in \mathcal{Z}(\mathcal{C}, t, p, r)$*, suppose there exist some* $\mathbf{U}_0 \in \mathcal{C}$*,* $\epsilon_0 > 0$ *and* $\alpha \in (0,1)$ *such that a local packing set* $G = G(\mathbb{B}(\mathbf{U}_0, \epsilon_0) \cap \mathcal{C}, d, \alpha\epsilon_0)$ *satisfies*

$$\epsilon_0 = \left[\left(\frac{\sigma\sqrt{c(\sigma^2+t)}}{t\sqrt{n}}\sqrt{\log|G|}\right) \wedge \operatorname{diam}(\mathcal{C})\right], \tag{14}$$

for some $c \in (0, 1/32]$*. Then, as long as* $|G(\mathbb{B}(\mathbf{U}_0, \epsilon_0) \cap \mathcal{C}, d, \alpha\epsilon_0)| \geq 2$*, it holds that*

$$\inf_{\widehat{\mathbf{U}}} \sup_{(\mathbf{\Gamma},\mathbf{U})\in\mathcal{Z}(\mathcal{C},t,p,r)} \mathcal{R}(\widehat{\mathbf{U}}, \mathbf{U}) \gtrsim \left[\left(\frac{\sigma\sqrt{\sigma^2+t}}{t\sqrt{n}}\sqrt{\log|G|}\right) \wedge \operatorname{diam}(\mathcal{C})\right], \tag{15}$$

where the infimum is over all the estimators based on the observation $\mathbf{Y}$*.*

In Zhang et al. (2018), the sharp minimax rate for estimating $\operatorname{span}(\mathbf{U})$ under the unstructured spiked Wishart model was obtained as

$$\inf_{\widehat{\mathbf{U}}} \sup_{(\mathbf{\Gamma},\mathbf{U})\in\mathcal{Z}(O(p,r),t,p,r)} \mathcal{R}(\widehat{\mathbf{U}}, \mathbf{U}) \asymp \left(\frac{\sigma\sqrt{(\sigma^2+t)rp}}{t\sqrt{n}} \wedge \sqrt{r}\right), \tag{16}$$

whose lower bound immediately follows from (15). Comparing the general lower bound (15) to (16), we observe again that the local entropic quantity $\log|G(\mathbb{B}(\mathbf{U}_0, \epsilon_0) \cap \mathcal{C}, d, \alpha\epsilon_0)|$ characterizes the intrinsic statistical dimension (which is rp in the case of $\mathcal{C} = O(p,r)$) of the set $\mathcal{C}$. See Section 5 for more examples.

4.2 Risk Upper Bound

Under the spiked Wishart model, to estimate the eigen subspace $\operatorname{span}(\mathbf{U})$ under the structural constraint $\mathbf{U} \in \mathcal{C}$, we start with the sample covariance matrix

$$\hat{\mathbf{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})(Y_i - \bar{Y})^\top,$$

where $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ and Y_i is the i-th row of the observed data matrix $\mathbf{Y} \in \mathbb{R}^{n\times p}$. Since $\hat{\mathbf{\Sigma}}$ is invariant to any translation on $\mathbf{Y}$, we assume $\boldsymbol{\mu} = 0$ without loss of generality.

Similar to the matrix denoising model, for the spiked Wishart model, with a slight abuse of notation, we define the eigen subspace estimator as

$$\widehat{\mathbf{U}} = \arg\max_{\mathbf{U}\in\mathcal{C}} \operatorname{tr}(\mathbf{U}^\top \hat{\mathbf{\Sigma}}\mathbf{U}). \tag{17}$$

The following theorem provides the risk upper bound of $\widehat{\mathbf{U}}$.

Theorem 10 *Under the spiked Wishart model, for any given* $\mathcal{C} \subset O(p,r)$ *and the parameter space* $\mathcal{Z}(\mathcal{C}, t, p, r)$*, suppose* $n \gtrsim \max\{\log\frac{t}{\sigma^2}, r\}$ *and* $\sqrt{t}/\sigma \gtrsim \sup_{\mathbf{U}\in\mathcal{C}}[D'(\mathcal{T}(\mathcal{C},\mathbf{U}), d_2)/D(\mathcal{T}(\mathcal{C},\mathbf{U}), d_2)]$*, then*

$$\sup_{(\mathbf{\Gamma},\mathbf{U})\in\mathcal{Z}(\mathcal{C},t,p,r)} \mathcal{R}(\widehat{\mathbf{U}}, \mathbf{U}) \lesssim \left(\frac{\sigma\Delta(\mathcal{C})\sqrt{t+\sigma^2}}{t\sqrt{n}} \wedge \operatorname{diam}(\mathcal{C})\right),$$

where $\Delta(\mathcal{C})$ *is defined in Theorem 8.*

Similar to the matrix denoising model, the above risk upper bound has a great affinity to the minimax lower bound (15), up to a difference in the information-geometric (metric-entropic) measure of $\mathcal{C}$; the sharpness of our results relies on the relative magnitude between the pair of quantities $\Delta^2(\mathcal{C})$ and $\log|G(\mathbb{B}(\mathbf{U}_0,\epsilon_0)\cap\mathcal{C},d,\alpha\epsilon_0)|$. In addition, the phase transitions in the rates of the lower and upper bounds as functions of the SNR t/σ^2 can be observed with the first critical point at

$$\frac{t}{\sigma^2}\asymp 1, \tag{18}$$

and the second critical point at

$$\frac{t}{\sigma^2}\asymp \frac{\zeta}{n\cdot\mathrm{diam}^2(\mathcal{C})}+\sqrt{\frac{\zeta}{n\cdot\mathrm{diam}^2(\mathcal{C})}}, \tag{19}$$

where in the lower bound $\zeta=\log|G(\mathbb{B}(\mathbf{U}_0,\epsilon_0)\cap\mathcal{C},d,\alpha\epsilon_0)|$ and in the upper bound $\zeta=\Delta^2(\mathcal{C})$. Again, the phase transition at the first critical point reflects the change of the speed of the rates of convergence, whereas the phase transition at the second critical point characterizes the statistical limit of the estimation problem. See Figure 2 for a graphical illustration. Finally, it will be seen in Section 5 that for many specific problems, the condition $t/\sigma^2\gtrsim \sup_{\mathbf{U}\in\mathcal{C}}[D'^2(\mathcal{T}(\mathcal{C},\mathbf{U}),d_2)/D^2(\mathcal{T}(\mathcal{C},\mathbf{U}),d_2)]$ required by Theorem 10 is mild and in fact necessary for consistent estimation.

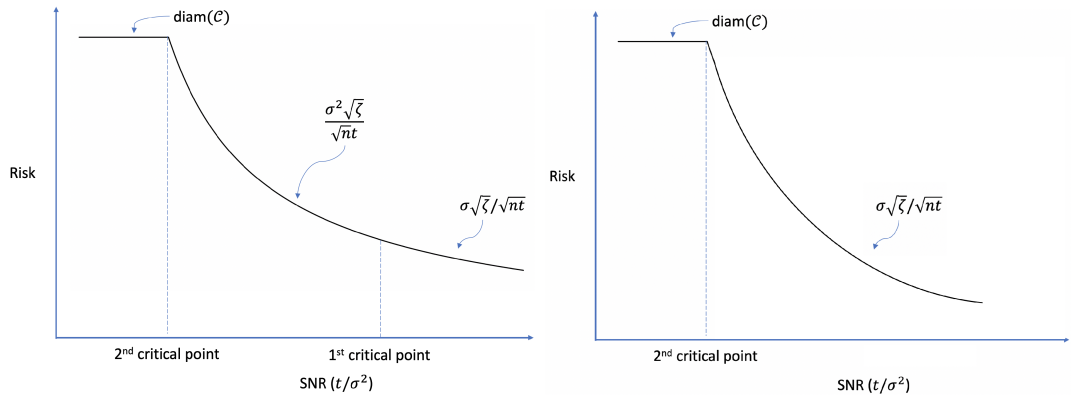

Figure 2: Graphical illustrations of the phase transitions in risks as a function of the SNRs under the spiked Wishart model. Left: $\zeta/\mathrm{diam}^2(\mathcal{C})\ll n$; Right: $\zeta/\mathrm{diam}^2(\mathcal{C})\gg n$.

5. Applications

In the following, building upon the minimax lower bounds and the risk upper bounds established in the previous sections, we obtain minimax rates and fundamental limits for various structural principal subspace estimation problems of broad interest. Specifically, in light of our generic results, we show the asymptotic equivalence of various local and global entropic measures associated to some specific examples of $\mathcal{C}$. Previous discussions under the general settings such as the phase transition phenomena also apply to each of the examples.

5.1 Sparse PCA/SVD

We start with the sparse PCA/SVD where the columns of $\mathbf{U}$ are sparse vectors. Suppose $\mathcal{C}_S(p,r,k)$ is the k-sparse subset of $O(p,r)$ for some $k \le p$, i.e., $\mathcal{C}_S(p,r,k) = \{\mathbf{U} \in O(p,r) : \max_{1\le i\le r} \|\mathbf{U}_{\cdot i}\|_0 \le k\}$. The following proposition concerns some estimates about the local and global entropic quantities associated with the set $\mathcal{C}_S(p,r,k)$. For simplicity, we denote $\mathcal{C}_S(k) = \mathcal{C}_S(p,r,k)$ when there is no confusion.

Proposition 11 *Under the matrix denoising model where $(\mathbf{\Gamma}, \mathbf{U}, \mathbf{V}) \in \mathcal{Y}(\mathcal{C}_S(k), t, p_1, p_2, r)$ with $k = o(p_1)$ and $r = o(k)$, there exist some $(\mathbf{U}_0, \epsilon_0, \alpha)$ and a local packing set $G(\mathbb{B}(\mathbf{U}_0, \epsilon_0) \cap \mathcal{C}_S(p_1, r, k), d, \alpha\epsilon_0)$ satisfying (2) such that*

$$\Delta^2(\mathcal{C}_S(p_1,k,r)) \lesssim k\log(ep_1/k) + kr \lesssim \log|G(\mathbb{B}(\mathbf{U}_0,\epsilon_0)\cap\mathcal{C}_S(p_1,r,k),d,\alpha\epsilon_0)|.$$

Similarly, under the spiked Wishart model where $(\mathbf{\Gamma}, \mathbf{V}) \in \mathcal{Z}(\mathcal{C}_S(k), t, p, r)$ with $k = o(p)$ and $r = o(k)$, there exist some $(\mathbf{U}_0, \epsilon_0, \alpha)$ and a local packing set $G(\mathbb{B}(\mathbf{U}_0, \epsilon_0) \cap \mathcal{C}_S(p, r, k), d, \alpha\epsilon_0)$ satisfying (14) such that

$$\Delta^2(\mathcal{C}_S(p,k,r)) \lesssim k\log(ep/k) + rk \lesssim \log|G(\mathbb{B}(\mathbf{U}_0,\epsilon_0)\cap\mathcal{C}_S(p,k,r),d,\alpha\epsilon_0)|.$$

In light of our lower and upper bounds under both the matrix denoising model (Theorem 5 and 8) and the spiked Wishart model (Theorem 9 and 10), with Proposition 11, we are able to establish sharp minimax rates of convergence for sparse PCA/SVD.

Theorem 12 *Under the matrix denoising model with $\mathbf{U} \in \mathcal{C}_S(p_1, r, k)$ where $k = o(p_1)$ and $r = o(k)$, for $t/\sigma \gtrsim \sqrt{k\log(ep_1/k)} + \sqrt{rk}$, it holds that*

$$\inf_{\widehat{\mathbf{U}}} \sup_{\mathcal{Y}(\mathcal{C}_S(k),t,p_1,p_2,r)} \mathcal{R}(\widehat{\mathbf{U}}, \mathbf{U}) \asymp \left[\frac{\sigma\sqrt{t^2+\sigma^2 p_2}}{t^2}\left(\sqrt{k\log\frac{ep_1}{k}} + \sqrt{rk}\right)\right] \wedge \sqrt{r}, \tag{20}$$

where the minimax rate is achieved by (5). Similarly, under the spiked Wishart model with $\mathbf{U} \in \mathcal{C}_S(p,r,k)$ where $k = o(p)$ and $r = o(k)$, if $n \gtrsim \max\{\log\frac{t}{\sigma^2}, r\}$ and $\sqrt{t}/\sigma \gtrsim \sqrt{k\log(ep/k)} + \sqrt{rk}$, then

$$\inf_{\widehat{\mathbf{U}}} \sup_{\mathcal{Z}(\mathcal{C}_S(k),t,p,r)} \mathcal{R}(\widehat{\mathbf{U}}, \mathbf{U}) \asymp \left[\frac{\sigma\sqrt{t+\sigma^2}}{t\sqrt{n}}\left(\sqrt{k\log\frac{ep}{k}} + \sqrt{rk}\right)\right] \wedge \sqrt{r}, \tag{21}$$

where the minimax rate is achieved by (17). In particular, for $r = O(1)$, both estimators (5) and (17) are rate-optimal whenever consistent estimation is possible.

The minimax rate (21) for the spiked Wishart model (sparse PCA) recovers the ones obtained by Vu and Lei (2012) and Cai et al. (2013) under either finite rank or $r = o(k)$ settings. In contrast, the result (20) for the matrix denoising model (sparse SVD), to the best of our knowledge, has not been established.

5.2 Non-Negative PCA/SVD

We now turn to the non-negative PCA/SVD under either the matrix denoising model (SVD) or the spiked Wishart model (PCA) where $\mathbf{U} \in \mathcal{C}_N(p,r) = \{\mathbf{U} = (u_{ij}) \in O(p,r) : u_{ij} \geq 0 \text{ for all } i,j\}$. The following proposition provides estimates about the local and global entropic quantities related to the set $\mathcal{C}_N(p,r)$.

Proposition 13 *Under the matrix denoising model where $(\mathbf{\Gamma}, \mathbf{U}, \mathbf{V}) \in \mathcal{Y}(\mathcal{C}_N(p_1,r), t, p_1, p_2, r)$, there exist some $(\mathbf{U}_0, \epsilon_0, \alpha)$ and a local packing set $G(\mathbb{B}(\mathbf{U}_0,\epsilon_0) \cap \mathcal{C}_N(p_1,r), d, \alpha\epsilon_0)$ satisfying (2) such that*

$$\log |G(\mathbb{B}(\mathbf{U}_0,\epsilon_0) \cap \mathcal{C}_N(p_1,r), d, \alpha\epsilon_0)| \gtrsim p_1, \quad \text{and} \quad \Delta^2(\mathcal{C}_N(p_1,r)) \lesssim p_1 r.$$

Similarly, under the spiked Wishart model where $(\mathbf{\Gamma}, \mathbf{U}) \in \mathcal{Z}(\mathcal{C}_N(p,r), t, p, r)$, there exist some $(\mathbf{U}_0, \epsilon_0, \alpha)$ and a local packing set $G(\mathbb{B}(\mathbf{U}_0,\epsilon_0) \cap \mathcal{C}_N(p,r), d, \alpha\epsilon_0)$ satisfying (14) such that

$$\log |G(\mathbb{B}(\mathbf{U}_0,\epsilon_0) \cap \mathcal{C}_N(p,r), d, \alpha\epsilon_0)| \gtrsim p, \quad \text{and} \quad \Delta^2(\mathcal{C}_N(p,r)) \lesssim pr.$$

Remark 14 *Unfortunately, unlike the previous example, the estimates of the local and global entropic quantities provided by the above proposition are not as precise. Between the estimates of two geometric quantities there is a gap of factor r, which implies matching minimax rates only when $r = O(1)$ (see Theorem 15 below). In particular, it is unclear whether such discrepancy is intrinsic to the corresponding geometric quantities or due to the limitations of our technical tools for deriving sharp estimates of them.*

Proposition 13 enables us to establish the following minimax rates using the general lower and upper bounds from the previous sections.

Theorem 15 *Under the matrix denoising model with $\mathbf{U} \in \mathcal{C}_N(p_1,r)$ where $r = O(1)$, it holds that*

$$\inf_{\widehat{\mathbf{U}}} \sup_{\mathcal{Y}(\mathcal{C}_N(p_1,r),t,p_1,p_2,r)} \mathcal{R}(\widehat{\mathbf{U}}, \mathbf{U}) \asymp \left(\frac{\sigma\sqrt{(t^2+\sigma^2 p_2)p_1}}{t^2} \wedge 1 \right), \tag{22}$$

and the estimator (5) is rate-optimal whenever consistent estimation is possible. Similarly, for the spiked Wishart model with $\mathbf{U} \in \mathcal{C}_N(p,r)$ where $r = O(1)$, if $n \gtrsim \max\{\log \frac{t}{\sigma^2}, r\}$, then

$$\inf_{\widehat{\mathbf{U}}} \sup_{\mathcal{Z}(\mathcal{C}_N(p,r),t,p,r)} \mathcal{R}(\widehat{\mathbf{U}}, \mathbf{U}) \asymp \left(\frac{\sigma\sqrt{(t+\sigma^2)p}}{t\sqrt{n}} \wedge 1 \right), \tag{23}$$

where the estimator (17) is rate-optimal whenever consistent estimation is possible.

The minimax rates for non-negative PCA/SVD, which were previously unknown, turn out to be the same as the rates for the ordinary unstructured SVD (Cai and Zhang, 2018) and PCA (Zhang et al., 2018). This is due to the fact claimed in Proposition 13 that, under the finite rank scenarios, as a much smaller subset of $O(p,r)$, $\mathcal{C}_N(p,r)$ has asymptotically the same geometric complexity as $O(p,r)$.

Remark 16 *Deshpande et al. (2014) considered the rank-one Gaussian Wigner model* $\mathbf{Y} = \lambda \mathbf{u}\mathbf{u}^\top + \mathbf{Z} \in \mathbb{R}^{p_1 \times p_1}$*, which can be treated as a special case of the matrix denoising model. Specifically, it was shown that, for* $\widehat{\mathbf{u}} = \arg\max_{\mathbf{u} \in \mathcal{C}_N(p,1)} \mathbf{u}^\top \mathbf{Y}\mathbf{u}$*, it holds that*

$$\sup_{(\lambda,\mathbf{u}) \in \mathcal{Z}(\mathcal{C}_N(p,1),t,p,1)} \mathbb{E}[1 - |\widehat{\mathbf{u}}^\top \mathbf{u}|] \lesssim \left(\frac{\sigma\sqrt{p}}{t} \wedge 1 \right),$$

which, by the fact that $1 - |\widehat{\mathbf{u}}^\top \mathbf{u}| \le d(\widehat{\mathbf{u}}, \mathbf{u})$*, can be implied by our result (see also Section 6.2). Similar problems were studied in Montanari and Richard (2015) under the setting where* $p_1/p_2 \to \alpha \in (0, \infty)$*. However, their focus is to unveil the asymptotic behavior of* $\widehat{\mathbf{u}}^\top \mathbf{u}$ *as well as the analysis of an approximate message passing algorithm, which is different from ours.*

5.3 Subspace Constrained PCA/SVD

In some applications such as network clustering (Wang and Davidson, 2010; Kawale and Boley, 2013; Kleindessner et al., 2019), it is of interest to estimate principal subspaces with certain linear subspace constraints. For example, under the matrix denoising model, for some fixed $A \in \mathbb{R}^{p_1 \times (p_1 - k)}$ of rank $(p_1 - k)$ where $r < k < p_1$, a k-dimensional subspace constraint on the singular subspace span($\mathbf{U}$) could be $\mathbf{U} \in \mathcal{C}_A(p_1, r, k) = \{\mathbf{U} \in O(p_1, r) : A\mathbf{U}_{\cdot i} = 0, \forall 1 \le i \le r\}$. Again, subspace constrained PCA/SVD can be solved based on the general results obtained in the previous sections.

Proposition 17 *For given* $A \in \mathbb{R}^{p_1 \times (p_1 - k)}$ *of rank* $(p_1 - k)$*, under the matrix denoising model where* $(\mathbf{\Gamma}, \mathbf{U}, \mathbf{V}) \in \mathcal{Y}(\mathcal{C}_A(p_1, r, k), t, p_1, p_2, r)$*, there exist some* $(\mathbf{U}_0, \epsilon_0, \alpha)$ *and a local packing set* $G(\mathbb{B}(\mathbf{U}_0, \epsilon_0) \cap \mathcal{C}_A(p_1, r, k), d, \alpha\epsilon_0)$ *satisfying (2) such that*

$$\Delta^2(\mathcal{C}_A(p_1, r, k)) \lesssim rk \lesssim \log |G(\mathbb{B}(\mathbf{u}_0, \epsilon_0) \cap \mathcal{C}_A(p_1, r, k), d, \alpha\epsilon_0)|.$$

Similarly, for given $B \in \mathbb{R}^{p \times (p-k)}$ *of rank* $(p - k)$*, under the spiked Wishart model with* $(\mathbf{\Gamma}, \mathbf{U}) \in \mathcal{Z}(\mathcal{C}_B(p, r, k), t, p, r)$*, there exist some* $(\mathbf{U}_0, \epsilon_0, \alpha)$ *and a local packing set* $G(\mathbb{B}(\mathbf{U}_0, \epsilon_0) \cap \mathcal{C}_B(p, r, k), d, \alpha\epsilon_0)$ *satisfying (14) such that*

$$\Delta^2(\mathcal{C}_B(p, r, k)) \lesssim rk \lesssim \log |G(\mathbb{B}(\mathbf{U}_0, \epsilon_0) \cap \mathcal{C}_B(p, r, k), d, \alpha\epsilon_0)|.$$

Theorem 18 *Under the matrix denoising model with* $\mathbf{U} \in \mathcal{C}_A(p_1, r, k)$ *where* $r < k < p_1$ *and* $A \in \mathbb{R}^{p_1 \times (p_1 - k)}$ *is of rank* $(p_1 - k)$*, for* $t/\sigma \gtrsim \sqrt{rk}$*, it holds that*

$$\inf_{\widehat{\mathbf{U}}} \sup_{\mathcal{Y}(\mathcal{C}_A(p_1,r,k),t,p_1,p_2,r)} \mathcal{R}(\widehat{\mathbf{U}}, \mathbf{U}) \asymp \left(\frac{\sigma\sqrt{(t^2 + \sigma^2 p_2)rk}}{t^2} \wedge \sqrt{r} \right), \tag{24}$$

where the minimax rate is achieved by (5). Similarly, under the spiked Wishart model with $\mathbf{U} \in \mathcal{C}_B(p, r, k)$*, where* $r < k < p$ *and* $B \in \mathbb{R}^{p \times (p-k)}$ *is of rank* $(p - k)$*, if* $n \gtrsim \max\{\log \frac{t}{\sigma^2}, r\}$ *and* $\sqrt{t}/\sigma \gtrsim \sqrt{rk}$*, then*

$$\inf_{\widehat{\mathbf{U}}} \sup_{\mathcal{Z}(\mathcal{C}_B(p,r,k),t,p,r)} \mathcal{R}(\widehat{\mathbf{U}}, \mathbf{U}) \asymp \left(\frac{\sigma\sqrt{(t + \sigma^2)rk}}{t\sqrt{n}} \wedge \sqrt{r} \right), \tag{25}$$

where the minimax rate is achieved by (17). In particular, if $r = O(1)$*, then both estimators (5) and (17) are rate-optimal whenever consistent estimation is possible.*

5.4 Spectral Clustering

As discussed in Section 1.2, spectral clustering can be treated as estimation of the structural eigenvector under the rank-one matrix denoising model $\mathbf{Y} = \lambda \mathbf{u}\mathbf{v}^\top + \mathbf{Z} \in \mathbb{R}^{n\times p}$ where $\lambda = \|\mathbf{h}\|_2\|\boldsymbol{\theta}\|_2 = \sqrt{n}\|\boldsymbol{\theta}\|_2$ is the global signal strength, $\mathbf{u} = \mathbf{h}/\|\mathbf{h}\|_2 \in \mathcal{C}_\pm^n = \{\mathbf{u} \in \mathbb{R}^n : \|\mathbf{u}\|_2 = 1, u_i \in \{\pm n^{-1/2}\}\}$ indicates the group labels, and $\mathbf{Z}$ has i.i.d. entries from $N(0, \sigma^2)$. As a result, important insights about the clustering problem can be obtained by calculating the entropic quantities related to $\mathcal{C}_\pm^n$ and applying the general results from the previous sections.

Proposition 19 *Under the matrix denoising model where $(\lambda, \mathbf{u}, \mathbf{v}) \in \mathcal{Y}(\mathcal{C}_\pm^n, t, n, p, 1)$, it holds that $\Delta^2(\mathcal{C}_\pm^n) \lesssim n$. In addition, if $t^2 = C\sigma^2(\sqrt{pn} + n)$ for some constant $C > 0$, then there exist some $(\mathbf{u}_0, \epsilon_0, \alpha)$ and a local packing set $G(\mathbb{B}(\mathbf{u}_0, \epsilon_0) \cap \mathcal{C}_\pm^n, d, \alpha\epsilon_0)$ satisfying (2) such that $\log|G(\mathbb{B}(\mathbf{u}_0, \epsilon_0) \cap \mathcal{C}_\pm^n, d, \alpha\epsilon_0)| \asymp n$.*

Theorem 20 *Under the spectral clustering model defined in Section 1.2, or equivalently, the matrix denoising model $\mathbf{Y} = \lambda\mathbf{u}\mathbf{v}^\top + \mathbf{Z} \in \mathbb{R}^{n\times p}$ where $\mathbf{u} \in \mathcal{C}_\pm^n$, the estimator $\widehat{\mathbf{u}} = \arg\max_{\mathbf{u}\in\mathcal{C}_\pm^n} \mathbf{u}^\top\mathbf{Y}\mathbf{Y}^\top\mathbf{u}$ satisfies*

$$\sup_{(\lambda,\mathbf{u},\mathbf{v})\in\mathcal{Y}(\mathcal{C}_\pm^n,t,n,p,1)} \mathcal{R}(\widehat{\mathbf{u}}, \mathbf{u}) \lesssim \left(\frac{\sigma\sqrt{(t^2 + \sigma^2 p)n}}{t^2} \wedge 1\right). \tag{26}$$

In addition, if $t^2 \lesssim \sigma^2(n + \sqrt{np})$, then

$$\inf_{\widehat{\mathbf{u}}} \sup_{(\lambda,\mathbf{u},\mathbf{v})\in\mathcal{Y}(\mathcal{C}_\pm^n,t,n,p,1)} \mathcal{R}(\widehat{\mathbf{u}}, \mathbf{u}) \gtrsim C \tag{27}$$

for some absolute constant $C > 0$.

Intuitively, the fundamental difficulty for clustering relies on the interplay between the global signal strength t, which reflects both the sample size (n) and the distance between the two clusters ($\|\boldsymbol{\theta}\|_2$), the noise level ($\sigma^2$), and the dimensionality (p). In particular, the lower bound from the above theorem shows that one needs $\|\boldsymbol{\theta}\|_2^2 \gtrsim \sigma^2(\sqrt{p/n} + 1)$ in order to have consistent clustering. Moreover, the risk upper bound implies that, whenever $\|\boldsymbol{\theta}\|_2^2 \gtrsim \sigma^2(\sqrt{p/n} + 1)$, the estimator $\widehat{\mathbf{u}}$ is consistent. Theorem 20 thus establishes the fundamental statistical limit for the minimal global signal strength for consistent clustering. Similar phenomena have been observed by Azizyan et al. (2013) and Cai and Zhang (2018).

Nevertheless, it should be noted that, despite the fundamental limits for consistent recovery yielded by Theorem 20, the upper bound (26) is sub-optimal and can be further improved through a variant of Lloyd's iterations (Lu and Zhou, 2016; Ndaoud, 2018), or a hollowing method (Abbe et al., 2020).

6. Discussions

In this paper, we studied a collection of structural principal subspace estimation problems in a unified framework by exploring the deep connections between the difficulty of statistical estimation and the geometric complexity of the parameter spaces. Minimax optimal rates

of convergence for a collection of structured PCA/SVD problems are established. In this section, we discuss the computational issues of the proposed estimators as well as the extensions and connections to other problems.

6.1 Computationally Efficient Algorithms and the Iterative Projection Method

In general, the constrained optimization problems that define the estimators in (5) and (17) are computationally intractable. However, in practice, many iterative algorithms have been developed to approximate such estimators.

For example, under the matrix denoising model, given the data matrix $\mathbf{Y}$, the set $\mathcal{C}$, and an initial estimator $\mathbf{U}_0 \in O(p_1, r)$, an iterative algorithm for the constrained optimization problem $\arg\max_{\mathbf{U}\in\mathcal{C}} \operatorname{tr}(\mathbf{U}^\top \mathbf{Y}\mathbf{Y}^\top \mathbf{U})$ can be realized through iterations over the following updates for $t \geq 1$:

1. Multiplication: $\mathbf{G}_t = \mathbf{Y}\mathbf{Y}^\top \mathbf{U}_t$;

2. QR factorization: $\mathbf{U}'_{t+1}\mathbf{W}_{t+1} = \mathbf{G}_t$ where $\mathbf{U}'_{t+1}$ is $p_1 \times r$ orthonormal and $\mathbf{W}_{t+1}$ is $r \times r$ upper triangular;

3. Projection: $\mathbf{U}_{t+1} = \mathbb{P}_{\mathcal{C}}(\mathbf{U}'_{t+1})$.

Here the projection operator $\mathbb{P}_{\mathcal{C}}(\cdot)$ is defined as $\mathbb{P}_{\mathcal{C}}(\mathbf{U}) = \arg\min_{\mathbf{G}\in\mathcal{C}} d(\mathbf{U}, \mathbf{G})$. The above algorithm generalizes the ideas of the projected power method (see, for example, Boumal (2016); Chen and Candès (2018); Onaran and Villar (2017)) and the orthogonal iteration method (Golub and Van Loan, 2012; Ma, 2013).

The computational efficiency of this iterative algorithm relies on the complexity of the projection operator $\mathbb{P}_{\mathcal{C}}$ for a given $\mathcal{C}$. In the rank-one case (r=1), Ferreira et al. (2013) pointed out that, whenever the set $\mathcal{C}$ is an intersection of a convex cone and the unit sphere, the projection operator $\mathbb{P}_{\mathcal{C}}(\cdot)$ admits an explicit formula and can be computed efficiently. This class of spherical convex sets includes many of the above examples such as non-negative PCA/SVD and subspace constrained PCA/SVD. The case of spectral clustering, under the rank-one setting, is also straightforward as the projection has a simple expression $\mathbb{P}_{\mathcal{C}^n_\pm}(\mathbf{u}) = \operatorname{sgn}(\mathbf{u})/\sqrt{n}$ (see Ndaoud (2018) and Löffler et al. (2019) for more in depth discussions). As for sparse PCA/SVD, the computational side of the problem is much more complicated and has been extensively studied in literature (Shen and Huang, 2008; d'Aspremont et al., 2008; Witten et al., 2009; Journée et al., 2010; Ma, 2013; Vu et al., 2013; Yuan and Zhang, 2013; Deshpande and Montanari, 2014).

In addition to the iterative projection method discussed above, there are several other computationally efficient algorithms such as convex (semidefinite in particular) relaxations (Singer, 2011; Deshpande et al., 2014; Bandeira et al., 2017) and the approximate message passing algorithms (Deshpande and Montanari, 2014; Deshpande et al., 2014; Montanari and Richard, 2015; Rangan and Fletcher, 2012), that have been considered to solve the structured eigenvector problems. However, the focuses of these algorithms are still rank-one matrices, and it remains to be understood how well these algorithms generalize to the general rank-r cases. We leave further investigations along these directions to future work.

6.2 Extensions and Future Work

As mentioned in Section 1.2, an important special case of matrix denoising model is the Gaussian Wigner model (Deshpande et al., 2014; Montanari and Richard, 2015; Perry et al., 2018), where the data matrix $\mathbf{Y} = \mathbf{U}\mathbf{\Gamma}\mathbf{U}^\top + \mathbf{Z} \in \mathbb{R}^{p\times p}$ is symmetric, and the noise matrix $\mathbf{Z}$ has i.i.d. entries (up to symmetry) drawn from $N(0,\sigma^2)$. Consider the parameter space $\mathcal{Z}(\mathcal{C},t,p,r)$ defined in Section 4.1. It can be shown that, under similar conditions to those of Theorem 5,

$$\inf_{\widehat{\mathbf{U}}} \sup_{(\mathbf{\Gamma},\mathbf{U})\in\mathcal{Z}(\mathcal{C},t,p,r)} \mathcal{R}(\widehat{\mathbf{U}},\mathbf{U}) \gtrsim \left[\left(\frac{\sigma}{t}\sqrt{\log|G(\mathbb{B}(\mathbf{U}_0,\epsilon_0)\cap\mathcal{C},d,\alpha\epsilon_0)|}\right)\wedge \mathrm{diam}(\mathcal{C})\right]. \tag{28}$$

Moreover, if we define $\widehat{\mathbf{U}} = \arg\max_{\mathbf{U}\in\mathcal{C}} \mathrm{tr}(\mathbf{U}^\top\mathbf{Y}\mathbf{U})$, then its risk upper bound can be obtained as

$$\sup_{(\mathbf{\Gamma},\mathbf{U})\in\mathcal{Z}(\mathcal{C},t,p,r)} \mathcal{R}(\widehat{\mathbf{U}},\mathbf{U}) \lesssim \left(\frac{\sigma\Delta(\mathcal{C})}{t}\wedge \mathrm{diam}(\mathcal{C})\right). \tag{29}$$

These general bounds combined with the entropic quantities calculated in Section 5 would yield many other interesting optimality results. For instance, recall that the Gaussian $\mathbb{Z}/2$ synchronization problem can be treated as a rank-one Gaussian Wigner model $\mathbf{Y} = \lambda\mathbf{u}\mathbf{u}^\top + \mathbf{Z}$ where $\mathbf{u}\in\mathcal{C}_\pm^n$. In this case, we have, for $t \lesssim \sigma\sqrt{n}$

$$\inf_{\widehat{\mathbf{u}}} \sup_{(\lambda,\mathbf{u})\in\mathcal{Z}(\mathcal{C}_\pm^n,t,n,1)} \mathcal{R}(\widehat{\mathbf{u}},\mathbf{u}) \gtrsim C.$$

and, for $\widehat{\mathbf{u}} = \arg\max_{\mathbf{u}\in\mathcal{C}_\pm^n} \mathbf{u}^\top\mathbf{Y}\mathbf{u}$,

$$\sup_{(\lambda,\mathbf{u})\in\mathcal{Z}(\mathcal{C}_\pm^n,t,n,1)} \mathcal{R}(\widehat{\mathbf{u}},\mathbf{u}) \lesssim \left(\frac{\sigma\sqrt{n}}{t}\wedge 1\right).$$

This implies that, about Gaussian $\mathbb{Z}/2$ synchronization, to have consistent estimation/recovery, one needs $\lambda \gtrsim \sigma\sqrt{n}$, and the estimator $\widehat{\mathbf{u}}$ is consistent whenever $\lambda \gtrsim \sigma\sqrt{n}$. These results make interesting connections to the existing works (Javanmard et al., 2016; Perry et al., 2018) concerning the so-called critical threshold or fundamental limit for the SNRs in $\mathbb{Z}/2$ synchronization problems.

In the present paper, under the matrix denoising model, we only focused on the cases where the prior structural knowledge on the targeted singular subspace span($\mathbf{U}$) is available. However, in some applications, structural knowledge on the other singular subspace span($\mathbf{V}$) can also be available. An interesting question is whether and how much the prior knowledge on span($\mathbf{V}$) will help in the estimation of span($\mathbf{U}$). Some preliminary thinking suggests that novel phenomena might exist in such settings. For example, in an extreme case, if $\mathbf{V}$ is completely known a priori, then after a simple transform $\mathbf{Y}\mathbf{V} = \mathbf{U}\mathbf{\Gamma} + \mathbf{Z}\mathbf{V}$, estimation of span($\mathbf{U}$) can be reduced to a Gaussian mean estimation problem, whose minimax rate is clearly independent of the dimension of the columns in $\mathbf{V}$ and therefore quite different from the rates obtained in this paper. The problem again bears important concrete examples in statistics and machine learning. The present work provides a theoretical foundation for studying these problems.

Acknowledgement

The authors are grateful to the editors and four anonymous referees for their comments that improved the presentation of the paper. The research of Tony Cai was supported in part by NSF grants DMS-1712735 and DMS-2015259 and NIH grants R01-GM129781 and R01-GM123056. The research of Hongzhe Li was supported in part by NSF grant DMS-1712735 and NIH grants R01-GM129781 and R01-GM123056. R.M. would like to thank Shulei Wang, Sai Li, Rui Duan and Jiasheng Shi for helpful discussions.

Appendix A. Proof of the Main Theorems

In this section, we prove Theorems 5, 8, 9 and 10.

A.1 Risk Upper Bounds

This section proves Theorems 8 and 10. Throughout, for any $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{p_1 \times p_2}$, we denote $\langle \mathbf{X}, \mathbf{Y} \rangle = \mathrm{tr}(\mathbf{X}^\top \mathbf{Y})$. We recall Lemma 1 in Cai and Zhang (2018), which concerns the relationships between different distance measures.

Lemma 21 *For $\mathbf{H}_1, \mathbf{H}_2 \in O(p,r)$, $\|\mathbf{H}_1\mathbf{H}_1^\top - \mathbf{H}_2\mathbf{H}_2^\top\|_F = \sqrt{2(r - \|\mathbf{H}_1^\top \mathbf{H}_2\|_F^2)}$, and $\frac{1}{\sqrt{2}}\|\mathbf{H}_1\mathbf{H}_1^\top - \mathbf{H}_2\mathbf{H}_2^\top\|_F \le \inf_{\mathbf{O}\in O(r)} \|\mathbf{H}_1 - \mathbf{H}_2\mathbf{O}\|_F \le \|\mathbf{H}_1\mathbf{H}_1^\top - \mathbf{H}_2\mathbf{H}_2^\top\|_F$.*

Proof of Theorem 8. We begin by stating a useful lemma, whose proof is delayed to Section C.

Lemma 22 *Let $\mathbf{U} \in O(p,r)$, and $\mathbf{\Gamma} = \mathrm{diag}(\lambda_1, ..., \lambda_r)$. Then for any $\mathbf{W} \in O(p,r)$, we have $\frac{\lambda_r^2}{2}\|\mathbf{U}\mathbf{U}^\top - \mathbf{W}\mathbf{W}^\top\|_F^2 \le \langle \mathbf{U}\mathbf{\Gamma}^2\mathbf{U}^\top, \mathbf{U}\mathbf{U}^\top - \mathbf{W}\mathbf{W}^\top\rangle \le \frac{\lambda_1^2}{2}\|\mathbf{U}\mathbf{U}^\top - \mathbf{W}\mathbf{W}^\top\|_F^2$. If in addition we define $\mathbf{\Sigma} = \sigma^2\mathbf{I}_p + \mathbf{U}\mathbf{\Gamma}\mathbf{U}^\top$. Then for any $\mathbf{W} \in O(p,r)$, we have $\frac{\lambda_r}{2}\|\mathbf{U}\mathbf{U}^\top - \mathbf{W}\mathbf{W}^\top\|_F^2 \le \langle \mathbf{\Sigma}, \mathbf{U}\mathbf{U}^\top - \mathbf{W}\mathbf{W}^\top\rangle \le \frac{\lambda_1}{2}\|\mathbf{U}\mathbf{U}^\top - \mathbf{W}\mathbf{W}^\top\|_F^2$.*

By Lemma 22 and the fact that $\mathrm{tr}(\widehat{\mathbf{U}}^\top \mathbf{Y}\mathbf{Y}^\top \widehat{\mathbf{U}}) \ge \mathrm{tr}(\mathbf{U}^\top \mathbf{Y}\mathbf{Y}^\top \mathbf{U})$, or equivalently $\langle \mathbf{Y}\mathbf{Y}^\top, \mathbf{U}\mathbf{U}^\top - \widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top\rangle \le 0$, we have

$$\|\widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top - \mathbf{U}\mathbf{U}^\top\|_F^2 \le \frac{2}{\lambda_r^2}\langle \mathbf{U}\mathbf{\Gamma}^2\mathbf{U}^\top - \mathbf{Y}\mathbf{Y}^\top, \mathbf{U}\mathbf{U}^\top - \widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top\rangle.$$

Since $\mathbf{Y} = \mathbf{U}\mathbf{\Gamma}\mathbf{V}^\top + \mathbf{Z}$, we have $\mathbf{Y}\mathbf{Y}^\top = \mathbf{U}\mathbf{\Gamma}^2\mathbf{U}^\top + \mathbf{Z}\mathbf{V}\mathbf{\Gamma}\mathbf{U}^\top + \mathbf{U}\mathbf{\Gamma}\mathbf{V}^\top\mathbf{Z}^\top + \mathbf{Z}\mathbf{Z}^\top$. Thus

$$\begin{aligned}\|\widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top - \mathbf{U}\mathbf{U}^\top\|_F^2 &\le \frac{2}{\lambda_r^2}\big[\langle \mathbf{U}\mathbf{\Gamma}\mathbf{V}^\top\mathbf{Z}^\top, \widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top - \mathbf{U}\mathbf{U}^\top\rangle + \langle \mathbf{Z}\mathbf{V}\mathbf{\Gamma}\mathbf{U}^\top, \widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top - \mathbf{U}\mathbf{U}^\top\rangle \\ &\quad + \langle \mathbf{Z}\mathbf{Z}^\top, \widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top - \mathbf{U}\mathbf{U}^\top\rangle\big] \\ &\equiv \frac{2}{\lambda_r^2}(H_1 + H_2 + H_3).\end{aligned}$$

For H_1, if we set

$$\mathbf{G}_{\mathbf{W}} = \frac{\mathbf{W}\mathbf{W}^\top - \mathbf{U}\mathbf{U}^\top}{\|\mathbf{W}\mathbf{W}^\top - \mathbf{U}\mathbf{U}^\top\|_F}, \qquad \mathbf{W} \in O(p_1, r) \setminus \{\mathbf{U}\}, \tag{30}$$

we can write

$$H_1 = \langle \mathbf{U}\mathbf{\Gamma}\mathbf{V}^\top\mathbf{Z}^\top, \widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top - \mathbf{U}\mathbf{U}^\top\rangle = \|\widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top - \mathbf{U}\mathbf{U}^\top\|_F \cdot \langle \mathbf{U}\mathbf{\Gamma}\mathbf{V}^\top\mathbf{Z}^\top, \mathbf{G}_{\widehat{\mathbf{U}}}\rangle$$
$$\leq \|\widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top - \mathbf{U}\mathbf{U}^\top\|_F \cdot \sup_{\mathbf{W}\in\mathcal{C}} \operatorname{tr}(\mathbf{Z}\mathbf{V}\mathbf{\Gamma}\mathbf{U}^\top\mathbf{G}_{\mathbf{W}}).$$

Similarly, we have $H_2 \leq \|\widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top - \mathbf{U}\mathbf{U}^\top\|_F \cdot \sup_{\mathbf{W}\in\mathcal{C}} \operatorname{tr}(\mathbf{U}\mathbf{\Gamma}\mathbf{V}^\top\mathbf{Z}^\top\mathbf{G}_{\mathbf{W}})$, and $H_3 \leq \|\widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top - \mathbf{U}\mathbf{U}^\top\|_F \cdot \sup_{\mathbf{W}\in\mathcal{C}} \operatorname{tr}(\mathbf{Z}^\top\mathbf{G}_{\mathbf{W}}\mathbf{Z})$. It then follows that

$$\|\widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top - \mathbf{U}\mathbf{U}^\top\|_F \leq \frac{2}{\lambda_r^2}\bigg(\sup_{\mathbf{W}\in\mathcal{C}} \operatorname{tr}(\mathbf{Z}\mathbf{V}\mathbf{\Gamma}\mathbf{U}^\top\mathbf{G}_{\mathbf{W}}) + \sup_{\mathbf{W}\in\mathcal{C}} \operatorname{tr}(\mathbf{U}\mathbf{\Gamma}\mathbf{V}^\top\mathbf{Z}^\top\mathbf{G}_{\mathbf{W}}) + \sup_{\mathbf{W}\in\mathcal{C}} \operatorname{tr}(\mathbf{Z}^\top\mathbf{G}_{\mathbf{W}}\mathbf{Z})\bigg). \tag{31}$$

The above inequality decomposes the error $\mathbb{E}\|\widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top - \mathbf{U}\mathbf{U}^\top\|_F$ into three terms

$$\mathbb{E}\frac{2}{\lambda_r^2}\sup_{\mathbf{W}\in\mathcal{C}} \operatorname{tr}(\mathbf{Z}\mathbf{V}\mathbf{\Gamma}\mathbf{U}^\top\mathbf{G}_{\mathbf{W}}), \qquad \mathbb{E}\frac{2}{\lambda_r^2}\sup_{\mathbf{W}\in\mathcal{C}} \operatorname{tr}(\mathbf{U}\mathbf{\Gamma}\mathbf{V}^\top\mathbf{Z}^\top\mathbf{G}_{\mathbf{W}}), \qquad \mathbb{E}\frac{2}{\lambda_r^2}\sup_{\mathbf{W}\in\mathcal{C}} \operatorname{tr}(\mathbf{Z}^\top\mathbf{G}_{\mathbf{W}}\mathbf{Z}).$$

In fact, the errors due to the first two terms correspond to the risk in the strong SNR regime, and the last term contributes to the risk in the weak SNR regime. The rest of the proof is separated into three parts. In Part I, we show that the first two terms are bounded by $\sigma\lambda_1 D(\mathcal{T}(\mathcal{C}, \mathbf{U}), d_2)/\lambda_r^2$, whereas in Part II we show that the third term can be bounded by $\sigma^2\sqrt{p_2}D(\mathcal{T}(\mathcal{C}, \mathbf{U}), d_2)/\lambda_r^2 + \sigma D'(\mathcal{T}(\mathcal{C}, \mathbf{U}), d_2)/\lambda_r^2$. In Part III, we obtain the desired risk upper bound.

Part I. For the term $\sup_{\mathbf{W}\in\mathcal{C}} \operatorname{tr}(\mathbf{Z}\mathbf{V}\mathbf{\Gamma}\mathbf{U}^\top\mathbf{G}_{\mathbf{W}})$, we have

$$\sup_{\mathbf{W}\in\mathcal{C}} \operatorname{tr}(\mathbf{Z}\mathbf{V}\mathbf{\Gamma}\mathbf{U}^\top\mathbf{G}_{\mathbf{W}}) = \sup_{\mathbf{W}\in\mathcal{C}} \operatorname{tr}(\mathbf{U}^\top\mathbf{G}_{\mathbf{W}}\mathbf{Z}\mathbf{V}\mathbf{\Gamma}) = \sup_{\mathbf{W}\in\mathcal{C}} \sum_{i=1}^{r} \lambda_i \cdot (\mathbf{U}^\top\mathbf{G}_{\mathbf{W}}\mathbf{Z}\mathbf{V})_{ii}$$
$$\leq \lambda_1 \sup_{\mathbf{W}\in\mathcal{C}} \operatorname{tr}(\mathbf{V}\mathbf{U}^\top\mathbf{G}_{\mathbf{W}}\mathbf{Z}) \leq \lambda_1 \sup_{\mathbf{G}\in\mathcal{T}'(\mathcal{C},\mathbf{U},\mathbf{V})} \langle \mathbf{G}, \mathbf{Z}\rangle,$$

where we defined $\mathcal{T}'(\mathcal{C}, \mathbf{U}, \mathbf{V}) = \{\mathbf{G}_{\mathbf{W}}\mathbf{U}\mathbf{V}^\top \in \mathbb{R}^{p_1\times p_2} : \mathbf{W} \in \mathcal{C} \setminus \{\mathbf{U}\}\}$. To control the expected suprema of the Gaussian process $\sup_{\mathbf{G}\in\mathcal{T}'(\mathcal{C},\mathbf{U},\mathbf{V})}\langle \mathbf{G}, \mathbf{Z}\rangle$, we use the following Dudley's integral inequality (see, for example, Vershynin 2018, pp. 188).

Theorem 23 (Dudley's Integral Inequality) *Let $\{X_t\}_{t\in T}$ be a Gaussian process, that is, a jointly Gaussian family of centered random variables indexed by T, where T is equipped with the canonical distance $d(s,t) = \sqrt{\mathbb{E}(X_s - X_t)^2}$. For some universal constant L, we have $\mathbb{E}\sup_{t\in T} X_t \leq L\int_0^\infty \sqrt{\log \mathcal{N}(T, d, \epsilon)}d\epsilon$.*

For the Gaussian process $\sup_{\mathbf{G}\in\mathcal{T}'(\mathcal{C},\mathbf{U},\mathbf{V})}\langle \mathbf{G}, \mathbf{Z}\rangle$, the canonical distance defined over the set $\mathcal{T}'(\mathcal{C}, \mathbf{U}, \mathbf{V})$ can be obtained as follows. For any $\mathbf{G}_1, \mathbf{G}_2 \in \mathcal{T}(\mathcal{C}, \mathbf{U}, \mathbf{V})$, the canonical distance between $\mathbf{G}_1$ and $\mathbf{G}_2$, by definition, is $\sqrt{\mathbb{E}\langle \mathbf{G}_1 - \mathbf{G}_2, \mathbf{Z}\rangle^2} = \|\mathbf{G}_1 - \mathbf{G}_2\|_F \equiv d_2(\mathbf{G}_1, \mathbf{G}_2)$. Theorem 23 yields

$$\mathbb{E}\sup_{\mathbf{G}\in\mathcal{T}'(\mathcal{C},\mathbf{U},\mathbf{V})} \langle \mathbf{G}, \mathbf{Z}\rangle \leq C\sigma\int_0^\infty \sqrt{\log \mathcal{N}(\mathcal{T}'(\mathcal{C}, \mathbf{U}, \mathbf{V}), d_2, \epsilon)}d\epsilon, \tag{32}$$

for some universal constant $C > 0$. Next, for any $\mathbf{G}_1, \mathbf{G}_2 \in \mathcal{T}'(\mathcal{C}, \mathbf{U}, \mathbf{V})$, without loss of generality, if we assume $\mathbf{G}_1 = \mathbf{G}_{\mathbf{W}_1}\mathbf{U}\mathbf{V}^\top$ and $\mathbf{G}_2 = \mathbf{G}_{\mathbf{W}_2}\mathbf{U}\mathbf{V}^\top$, where $\mathbf{W}_1, \mathbf{W}_2 \in \mathcal{C} \setminus \{\mathbf{U}\}$, then it holds that

$$\begin{aligned} d_2(\mathbf{G}_1, \mathbf{G}_2) &= \|\mathbf{G}_1 - \mathbf{G}_2\|_F \leq \|\mathbf{G}_{\mathbf{W}_1} - \mathbf{G}_{\mathbf{W}_2}\|_F \|\mathbf{U}\| \|\mathbf{V}\| \\ &\leq \|\mathbf{G}_{\mathbf{W}_1} - \mathbf{G}_{\mathbf{W}_2}\|_F = d_2(\mathbf{G}_{\mathbf{W}_1}, \mathbf{G}_{\mathbf{W}_2}), \end{aligned} \tag{33}$$

where we used the fact that $\|\mathbf{H}\mathbf{G}\|_F \leq \|\mathbf{H}\|_F \|\mathbf{G}\|$. The next lemma, obtained by Szarek (1998), concerns the invariance property of the covering numbers with respect to Lipschitz maps.

Lemma 24 (Szarek (1998)) *Let (M, d) and (M_1, d_1) be metric spaces, $K \subset M$, $\Phi : M \to M_1$, and let $L > 0$. If Φ satisfies $d_1(\Phi(x), \Phi(y)) \leq L d(x, y)$ for $x, y, \in M$, then, for every $\epsilon > 0$, we have $\mathcal{N}(\Phi(K), d_1, L\epsilon) \leq \mathcal{N}(K, d, \epsilon)$.*

Define the set $\mathcal{T}(\mathcal{C}, \mathbf{U}) = \{\mathbf{G}_{\mathbf{W}} : \mathbf{W} \in \mathcal{C} \setminus \{\mathbf{U}\}\}$. Equation (33) and Lemma 24 imply

$$\log \mathcal{N}(\mathcal{T}'(\mathcal{C}, \mathbf{U}, \mathbf{V}), d_2, \epsilon) \leq \log \mathcal{N}(\mathcal{T}(\mathcal{C}, \mathbf{U}), d_2, \epsilon), \tag{34}$$

which means

$$\sup_{\mathbf{W} \in \mathcal{C}} \operatorname{tr}(\mathbf{Z}\mathbf{V}\mathbf{\Gamma}\mathbf{U}^\top \mathbf{G}_{\mathbf{W}}) \leq C\lambda_1 \sigma \int_0^\infty \sqrt{\log \mathcal{N}(\mathcal{T}(\mathcal{C}, \mathbf{U}), d_2, \epsilon)} d\epsilon. \tag{35}$$

Applying the same argument to $\sup_{\mathbf{W} \in \mathcal{C}} \operatorname{tr}(\mathbf{U}\mathbf{\Gamma}\mathbf{V}^\top \mathbf{Z}^\top \mathbf{G}_{\mathbf{W}})$ leads to

$$\sup_{\mathbf{W} \in \mathcal{C}} \operatorname{tr}(\mathbf{U}\mathbf{\Gamma}\mathbf{V}^\top \mathbf{Z}^\top \mathbf{G}_{\mathbf{W}}) \leq C\lambda_1 \sigma \int_0^\infty \sqrt{\log \mathcal{N}(\mathcal{T}(\mathcal{C}, \mathbf{U}), d_2, \epsilon)} d\epsilon. \tag{36}$$

Part II. To bound $\sup_{\mathbf{W} \in \mathcal{C}} \operatorname{tr}(\mathbf{Z}^\top \mathbf{G}_{\mathbf{W}} \mathbf{Z})$, note that $\operatorname{tr}(\mathbf{Z}^\top \mathbf{G}_{\mathbf{W}} \mathbf{Z}) = \operatorname{vec}(\mathbf{Z})^\top \mathbf{D}_{\mathbf{W}} \operatorname{vec}(\mathbf{Z})$, where $\operatorname{vec}(\mathbf{Z}) = (Z_{11}, ..., Z_{p_1 1}, Z_{12}, ..., Z_{p_1 2}, ..., Z_{1p_2}, ..., Z_{p_1 p_2})^\top$, and

$$\mathbf{D}_{\mathbf{W}} = \begin{bmatrix} \mathbf{G}_{\mathbf{W}} & & \\ & \ddots & \\ & & \mathbf{G}_{\mathbf{W}} \end{bmatrix} \in \mathbb{R}^{p_1 p_2 \times p_1 p_2}, \tag{37}$$

It suffices to control the expected supremum of the following Gaussian chaos of order 2,

$$\sup_{\mathbf{D} \in \mathcal{P}(\mathcal{C}, \mathbf{U})} \operatorname{vec}(\mathbf{Z})^\top \mathbf{D} \operatorname{vec}(\mathbf{Z}), \tag{38}$$

where $\mathcal{P}(\mathcal{C}, \mathbf{U}) = \{\mathbf{D}_{\mathbf{W}} \in \mathbb{R}^{p_1 p_2 \times p_1 p_2} : \mathbf{W} \in \mathcal{C} \setminus \{\mathbf{U}\}\}$. To analyze the above Gaussian chaos, a powerful tool from empirical process theory is the decoupling technique. In particular, we apply the following decoupling inequality obtained by Arcones and Giné (1993) (see also Theorem 2.5 of Krahmer et al. (2014)).

Theorem 25 (Arcones-Gené Decoupling Inequality) *Let $\{g_i\}_{1 \leq i \leq n}$ be a sequence of independent standard Gaussian variables and let $\{g_i'\}_{1 \leq i \leq n}$ be an independent copy of $\{g_i\}_{1 \leq i \leq n}$. Let $\mathcal{B}$ be a collection of $n \times n$ symmetric matrices. Then for all $p \geq 1$, there exists an absolute constant C such that*

$$\mathbb{E} \sup_{B \in \mathcal{B}} \left| \sum_{1 \leq j \neq k \leq n} B_{jk} g_j g_k + \sum_{j=1}^n B_{jj}(g_j^2 - 1) \right|^p \leq C^p \mathbb{E} \sup_{B \in \mathcal{B}} \left| \sum_{1 \leq j,k \leq n} B_{jk} g_j g_k' \right|^p.$$

From Theorem 25 and the fact that for given $\mathbf{W} \in \mathcal{C}\backslash\{\mathbf{U}\}$ we have $\mathbb{E}\mathrm{vec}(\mathbf{Z})^\top \mathbf{D}_{\mathbf{W}}\mathrm{vec}(\mathbf{Z}) = 0$, then

$$\mathbb{E}\sup_{\mathbf{D}\in\mathcal{P}(\mathcal{C},\mathbf{U})}[\mathrm{vec}(\mathbf{Z})^\top\mathbf{D}\mathrm{vec}(\mathbf{Z})] \le C\mathbb{E}\sup_{\mathbf{D}\in\mathcal{P}(\mathcal{C},\mathbf{U})}[\mathrm{vec}(\mathbf{Z})^\top\mathbf{D}\mathrm{vec}(\mathbf{Z}')] \tag{39}$$

where $\mathbf{Z}'$ is an independent copy of $\mathbf{Z}$. The upper bound of the right hand side of (39) can be obtained by using a generic chaining argument developed by Talagrand (2014). To state the result, we make the following definitions that characterize the complexity of a set in a metric space.

Definition 26 (admissible sequence) *Given a set T in the metric space (S, d), an admissible sequence is an increasing sequence $\{\mathcal{A}_n\}$ of partitions of T such that $|\mathcal{A}_0| = 1$ and $|\mathcal{A}_n| \le 2^{2^n}$ for $n \ge 1$.*

Definition 27 ($\gamma_\alpha(T, d)$) *Given $\alpha > 0$ and a set T in the metric space (S, d), we define $\gamma_\alpha(T, d) = \inf \sup_{t\in T}\sum_{n\ge 0} 2^{n/\alpha}\mathrm{diam}(A_n(t))$, where $A_n(t)$ is the unique element of $\mathcal{A}_n$ which contains t and the infimum is taken over all admissible sequences.*

The following theorem from (Talagrand, 2014, pp. 246) provides an important upper bound of the general decoupled Gaussian chaos of order 2.

Theorem 28 (Talagrand (2014)) *Let $\mathbf{g}, \mathbf{g}' \in \mathbb{R}^n$ be independent standard Gaussian vectors, and $\mathbf{Q} = \{q_{ij}\}_{1\le i,j\le n} \in \mathbb{R}^{n\times n}$. Given a set $T \subset \mathbb{R}^{n\times n}$ equipped with two distances $d_\infty(\mathbf{Q}_1, \mathbf{Q}_2) = \|\mathbf{Q}_1 - \mathbf{Q}_2\|$ and $d_2(\mathbf{Q}_1, \mathbf{Q}_2) = \|\mathbf{Q}_1 - \mathbf{Q}_2\|_F$,*

$$\mathbb{E}\sup_{\mathbf{Q}\in T}\mathbf{g}^\top\mathbf{Q}\mathbf{g}' \le L(\gamma_1(T, d_\infty) + \gamma_2(T, d_2)),$$

for some absolute constant $L \ge 0$.

A direct consequence of Theorem 28 is

$$\mathbb{E}\sup_{\mathbf{D}\in\mathcal{P}(\mathcal{C},\mathbf{U})}[\mathrm{vec}(\mathbf{Z})^\top\mathbf{D}\mathrm{vec}(\mathbf{Z}')] \le C\sigma^2(\gamma_1(\mathcal{P}(\mathcal{C},\mathbf{U}), d_\infty) + \gamma_2(\mathcal{P}(\mathcal{C},\mathbf{U}), d_2)). \tag{40}$$

Our next lemma obtains estimates of the functionals $\gamma_1(\mathcal{P}(\mathcal{C},\mathbf{U}), d_\infty)$ and $\gamma_2(\mathcal{P}(\mathcal{C},\mathbf{U}), d_2)$.

Lemma 29 *Let $\mathcal{T}(\mathcal{C},\mathbf{U}) = \{\mathbf{G}_{\mathbf{W}} \in \mathbb{R}^{p_1\times p_1} : \mathbf{W} \in \mathcal{C}\setminus\{\mathbf{U}\}\}$ be equipped with distances d_∞ and d_2 defined in Theorem 28. It holds that*

$$\gamma_1(\mathcal{P}(\mathcal{C},\mathbf{U}), d_\infty) \le C_1\int_0^\infty \log\mathcal{N}(\mathcal{T}(\mathcal{C},\mathbf{U}), d_2, \epsilon)d\epsilon, \tag{41}$$

$$\gamma_2(\mathcal{P}(\mathcal{C},\mathbf{U}), d_2) \le C_2\sqrt{p_2}\int_0^\infty \sqrt{\log\mathcal{N}(\mathcal{T}(\mathcal{C},\mathbf{U}), d_2, \epsilon)}d\epsilon. \tag{42}$$

Combining the above results, we have

$$\mathbb{E}\sup_{\mathbf{W}\in\mathcal{C}}\mathrm{tr}(\mathbf{Z}^\top\mathbf{G}_{\mathbf{W}}\mathbf{Z}) \lesssim \sigma^2\sqrt{p_2}\int_0^\infty \sqrt{\log\mathcal{N}(\mathcal{T}(\mathcal{C},\mathbf{U}), d_2, \epsilon)}d\epsilon + \sigma^2\int_0^\infty \log\mathcal{N}(\mathcal{T}(\mathcal{C},\mathbf{U}), d_2, \epsilon)d\epsilon. \tag{43}$$

Part III. By (31) (35) (36) and (43), we have, for any $(\mathbf{\Gamma}, \mathbf{U}, \mathbf{V}) \in \mathcal{Y}(\mathcal{C}, t, p_1, p_2, r)$, whenever $t \gtrsim \sigma D'(\mathcal{T}(\mathcal{C}, \mathbf{U}), d_2)/D(\mathcal{T}(\mathcal{C}, \mathbf{U}), d_2)$,

$$\begin{aligned}\mathbb{E}\|\widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top - \mathbf{U}\mathbf{U}^\top\|_F &\lesssim \frac{\sigma\lambda_1 D(\mathcal{T}(\mathcal{C}, \mathbf{U}), d_2)}{\lambda_r^2} + \frac{\sigma^2\sqrt{p_2}D(\mathcal{T}(\mathcal{C}, \mathbf{U}), d_2) + \sigma^2 D'(\mathcal{T}(\mathcal{C}, \mathbf{U}), d_2)}{\lambda_r^2} \\ &\lesssim \frac{\sigma\Delta(\mathcal{C})\sqrt{t^2 + \sigma^2 p_2}}{t^2}.\end{aligned}$$

The final result then follows by noticing the trivial upper bound of $\mathrm{diam}(\mathcal{C})$.

Proof of Theorem 10. Note that $\mathbf{Y} = \mathbf{X}\mathbf{\Gamma}^{1/2}\mathbf{U}^\top + \mathbf{Z} \in \mathbb{R}^{n\times p}$ where $\mathbf{\Gamma}^{1/2} = \mathrm{diag}(\lambda_1^{1/2}, ..., \lambda_r^{1/2})$, $\mathbf{X} \in \mathbb{R}^{n\times r}$ has i.i.d. entries from $\sim N(0, 1)$, and $\mathbf{Z}$ has i.i.d. entries from $N(0, \sigma^2)$. We can write

$$\begin{aligned}\hat{\mathbf{\Sigma}} = \frac{1}{n}\mathbf{Y}^\top\mathbf{Y} - \bar{Y}\bar{Y}^\top = \frac{1}{n}(&\mathbf{U}\mathbf{\Gamma}^{1/2}\mathbf{X}^\top\mathbf{X}\mathbf{\Gamma}^{1/2}\mathbf{U}^\top + \mathbf{Z}^\top\mathbf{X}\mathbf{\Gamma}^{1/2}\mathbf{U}^\top + \mathbf{U}\mathbf{\Gamma}^{1/2}\mathbf{X}^\top\mathbf{Z} + \mathbf{Z}^\top\mathbf{Z}) \\ &- (\mathbf{U}\mathbf{\Gamma}^{1/2}\bar{X}\bar{X}^\top\mathbf{\Gamma}^{1/2}\mathbf{U}^\top + \mathbf{U}\mathbf{\Gamma}^{1/2}\bar{X}\bar{Z}^\top + \bar{Z}\bar{X}^\top\mathbf{\Gamma}^{1/2}\mathbf{U}^\top + \bar{Z}\bar{Z}^\top),\end{aligned}$$

where $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i \in \mathbb{R}^r$ and $\bar{Z} = \frac{1}{n}\sum_{i=1}^n Z_i \in \mathbb{R}^p$. Now since $\mathrm{tr}(\widehat{\mathbf{U}}^\top\hat{\mathbf{\Sigma}}\widehat{\mathbf{U}}) \geq \mathrm{tr}(\mathbf{U}^\top\hat{\mathbf{\Sigma}}\mathbf{U})$, or equivalently $\langle\hat{\mathbf{\Sigma}}, \mathbf{U}\mathbf{U}^\top - \widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top\rangle \leq 0$, by Lemma 22, we have

$$\|\widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top - \mathbf{U}\mathbf{U}^\top\|_F^2 \leq \frac{2}{\lambda_r}\langle\mathbf{\Sigma} - \hat{\mathbf{\Sigma}}, \mathbf{U}\mathbf{U}^\top - \widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top\rangle.$$

Hence,

$$\begin{aligned}&\|\widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top - \mathbf{U}\mathbf{U}^\top\|_F^2 \\ &\leq \frac{2}{\lambda_r}\big[\langle n^{-1}\mathbf{Z}^\top\mathbf{X}\mathbf{\Gamma}^{1/2}\mathbf{U}^\top, \widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top - \mathbf{U}\mathbf{U}^\top\rangle + \langle n^{-1}\mathbf{U}\mathbf{\Gamma}^{1/2}\mathbf{X}^\top\mathbf{Z}, \widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top - \mathbf{U}\mathbf{U}^\top\rangle \\ &\quad + \langle n^{-1}\mathbf{U}\mathbf{\Gamma}^{1/2}\mathbf{X}^\top\mathbf{X}\mathbf{\Gamma}^{1/2}\mathbf{U}^\top - \mathbf{U}\mathbf{\Gamma}\mathbf{U}^\top, \widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top - \mathbf{U}\mathbf{U}^\top\rangle + \langle n^{-1}\mathbf{Z}^\top\mathbf{Z} - \mathbf{I}_p, \widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top - \mathbf{U}\mathbf{U}^\top\rangle \\ &\quad - \langle\mathbf{U}\mathbf{\Gamma}^{1/2}\bar{X}\bar{X}^\top\mathbf{\Gamma}^{1/2}\mathbf{U}^\top, \widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top - \mathbf{U}\mathbf{U}^\top\rangle - \langle\mathbf{U}\mathbf{\Gamma}^{1/2}\bar{X}\bar{Z}^\top, \widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top - \mathbf{U}\mathbf{U}^\top\rangle \\ &\quad - \langle\bar{Z}\bar{X}^\top\mathbf{\Gamma}^{1/2}\mathbf{U}^\top, \widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top - \mathbf{U}\mathbf{U}^\top\rangle - \langle\bar{Z}\bar{Z}^\top, \widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top - \mathbf{U}\mathbf{U}^\top\rangle\big] \\ &\equiv \frac{2}{\lambda_r}(H_1 + H_2 + H_3 + H_4 - H_5 - H_6 - H_7 - H_8).\end{aligned}$$

To control H_1, using the same notations in (30), we have

$$H_1 \leq \frac{1}{n}\|\widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top - \mathbf{U}\mathbf{U}^\top\|_F \cdot \sup_{\mathbf{W}\in\mathcal{C}}\mathrm{tr}(\mathbf{U}\mathbf{\Gamma}^{1/2}\mathbf{X}^\top\mathbf{Z}\mathbf{G}_\mathbf{W}).$$

Similarly, it holds that

$$
\begin{aligned}
H_2 &\le \frac{1}{n}\|\widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top - \mathbf{U}\mathbf{U}^\top\|_F \cdot \sup_{\mathbf{W}\in\mathcal{C}} \mathrm{tr}(\mathbf{Z}^\top\mathbf{X}\mathbf{\Gamma}^{1/2}\mathbf{U}^\top\mathbf{G}_{\mathbf{W}}),\\
H_3 &\le \langle \mathbf{\Gamma}^{1/2}(n^{-1}\mathbf{X}^\top\mathbf{X} - \mathbf{I}_r)\mathbf{\Gamma}^{1/2}, \mathbf{U}^\top\widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top\mathbf{U} - \mathbf{I}_r\rangle\\
&\le \|\mathbf{\Gamma}^{1/2}(n^{-1}\mathbf{X}^\top\mathbf{X} - \mathbf{I}_r)\mathbf{\Gamma}^{1/2}\| \cdot |\mathrm{tr}(\mathbf{U}^\top\widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top\mathbf{U} - \mathbf{I}_r)|\\
&\le \frac{\lambda_1}{2}\|n^{-1}\mathbf{X}^\top\mathbf{X} - \mathbf{I}_r\|\|\mathbf{U}\mathbf{U}^\top - \widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top\|_F^2,\\
H_4 &\le \|\widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top - \mathbf{U}\mathbf{U}^\top\|_F \cdot \sup_{\mathbf{W}\in\mathcal{C}} \mathrm{tr}((n^{-1}\mathbf{Z}^\top\mathbf{Z} - \mathbf{I}_p)\mathbf{G}_{\mathbf{W}}),\\
H_5 &\le \|\mathbf{\Gamma}^{1/2}\bar{X}\bar{X}^\top\mathbf{\Gamma}^{1/2}\| \cdot |\mathrm{tr}(\mathbf{U}^\top\widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top\mathbf{U} - \mathbf{I}_r)| \le \frac{\lambda_1}{2}\|\bar{X}\bar{X}^\top\|\|\mathbf{U}\mathbf{U}^\top - \widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top\|_F^2,\\
H_6 &\le \|\widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top - \mathbf{U}\mathbf{U}^\top\|_F \cdot \sup_{\mathbf{W}\in\mathcal{C}} \mathrm{tr}(\mathbf{U}\mathbf{\Gamma}^{1/2}\bar{X}\bar{Z}^\top\mathbf{G}_{\mathbf{W}}),\\
H_7 &\le \|\widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top - \mathbf{U}\mathbf{U}^\top\|_F \cdot \sup_{\mathbf{W}\in\mathcal{C}} \mathrm{tr}(\bar{Z}^\top\bar{X}\mathbf{\Gamma}^{1/2}\mathbf{U}^\top\mathbf{G}_{\mathbf{W}}),\\
H_8 &\le \|\widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top - \mathbf{U}\mathbf{U}^\top\|_F \cdot \sup_{\mathbf{W}\in\mathcal{C}} \mathrm{tr}(\bar{Z}\bar{Z}^\top\mathbf{G}_{\mathbf{W}}).
\end{aligned}
$$

Combining the above inequalities, we have

$$
\begin{aligned}
&\|\widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top - \mathbf{U}\mathbf{U}^\top\|_F\\
&\le \frac{2}{\lambda_r(1 - \frac{\lambda_1}{\lambda_r}\|n^{-1}\mathbf{X}^\top\mathbf{X} - \mathbf{I}_r\| - \frac{\lambda_1}{\lambda_r}\|\bar{X}\bar{X}^\top\|)}\Big(n^{-1}\sup_{\mathbf{W}\in\mathcal{C}} \mathrm{tr}(\mathbf{U}\mathbf{\Gamma}^{1/2}\mathbf{X}^\top\mathbf{Z}\mathbf{G}_{\mathbf{W}})\\
&\quad + n^{-1}\sup_{\mathbf{W}\in\mathcal{C}} \mathrm{tr}(\mathbf{Z}^\top\mathbf{X}\mathbf{\Gamma}^{1/2}\mathbf{U}^\top\mathbf{G}_{\mathbf{W}}) + \sup_{\mathbf{W}\in\mathcal{C}} \mathrm{tr}((n^{-1}\mathbf{Z}^\top\mathbf{Z} - \mathbf{I}_p)\mathbf{G}_{\mathbf{W}}) + \sup_{\mathbf{W}\in\mathcal{C}} \mathrm{tr}(\mathbf{U}\mathbf{\Gamma}^{1/2}\bar{X}\bar{Z}^\top\mathbf{G}_{\mathbf{W}})\\
&\quad + \sup_{\mathbf{W}\in\mathcal{C}} \mathrm{tr}(\bar{Z}^\top\bar{X}\mathbf{\Gamma}^{1/2}\mathbf{U}^\top\mathbf{G}_{\mathbf{W}}) + \sup_{\mathbf{W}\in\mathcal{C}} \mathrm{tr}(\bar{Z}\bar{Z}^\top\mathbf{G}_{\mathbf{W}})\Big)
\end{aligned}
\tag{44}
$$

The rest of the proof is separated into four parts, with the first three parts controlling the right-hand side of the inequality (44), and the last part deriving the final risk upper bound.

Part I. Note that

$$
\begin{aligned}
&\sup_{\mathbf{W}\in\mathcal{C}} \mathrm{tr}(\mathbf{U}\mathbf{\Gamma}^{1/2}\mathbf{X}^\top\mathbf{Z}\mathbf{G}_{\mathbf{W}}) = \sup_{\mathbf{W}\in\mathcal{C}} \mathrm{tr}(\mathbf{X}^\top\mathbf{Z}\mathbf{G}_{\mathbf{W}}\mathbf{U}\mathbf{\Gamma}^{1/2}) \le \lambda_1^{1/2}\sup_{\mathbf{W}\in\mathcal{C}} \mathrm{tr}(\mathbf{Z}\mathbf{G}_{\mathbf{W}}\mathbf{U}\mathbf{X}^\top/\|\mathbf{X}\|)\|\mathbf{X}\|\\
&\le \lambda_1^{1/2}\sup_{\mathbf{G}\in\mathcal{T}_0(\mathcal{C},\mathbf{U},\mathbf{X})} \langle \mathbf{Z}^\top, \mathbf{G}\rangle\|\mathbf{X}\|,
\end{aligned}
$$

where $\mathcal{T}_0(\mathcal{C},\mathbf{U},\mathbf{X}) = \big\{\frac{\mathbf{G}_{\mathbf{W}}\mathbf{U}\mathbf{X}^\top}{\|\mathbf{X}\|} : \mathbf{W}\in\mathcal{C}\setminus\{\mathbf{U}\}\big\}$. By Theorem 23, we have

$$
\mathbb{E}\bigg[\sup_{\mathbf{G}\in\mathcal{T}_0(\mathcal{C},\mathbf{U},\mathbf{X})} \langle \mathbf{Z}^\top, \mathbf{G}\rangle\bigg|\mathbf{X}\bigg] \le C\sigma\int_0^\infty \sqrt{\log\mathcal{N}(\mathcal{T}_0(\mathcal{C},\mathbf{U},\mathbf{X}), d_2, \epsilon)}d\epsilon.
$$

For any $\mathbf{G}_1, \mathbf{G}_2 \in \mathcal{T}_0(\mathcal{C},\mathbf{U},\mathbf{X})$, without loss of generality, if we assume $\mathbf{G}_1 = \|\mathbf{X}\|^{-1}\mathbf{G}_{\mathbf{W}_1}\mathbf{U}\mathbf{X}^\top$ and $\mathbf{G}_2 = \|\mathbf{X}\|^{-1}\mathbf{G}_{\mathbf{W}_2}\mathbf{U}\mathbf{X}^\top$ where $\mathbf{W}_1, \mathbf{W}_2 \in \mathcal{C}\setminus\{\mathbf{U}\}$, then

$$
d_2(\mathbf{G}_1,\mathbf{G}_2) \le \|\mathbf{G}_{\mathbf{W}_1} - \mathbf{G}_{\mathbf{W}_2}\|_F\|\mathbf{U}\| \le \|\mathbf{G}_{\mathbf{W}_1} - \mathbf{G}_{\mathbf{W}_2}\|_F = d_2(\mathbf{G}_{\mathbf{W}_1}, \mathbf{G}_{\mathbf{W}_2}). \tag{45}
$$

Again, recall the set $\mathcal{T}(\mathcal{C}, \mathbf{U})$ defined in the proof of Theorem 8, by Lemma 24, we have $\log \mathcal{N}(\mathcal{T}_0(\mathcal{C}, \mathbf{U}, \mathbf{X}), d_2, \epsilon) \leq \log \mathcal{N}(\mathcal{T}(\mathcal{C}, \mathbf{U}), d_2, \epsilon)$, which implies

$$\mathbb{E} \sup_{\mathbf{W} \in \mathcal{C}} \mathrm{tr}(\mathbf{U}\mathbf{\Gamma}^{1/2}\mathbf{X}^\top \mathbf{Z}\mathbf{G}_{\mathbf{W}}) \leq C\lambda_1^{1/2} \mathbb{E}\|\mathbf{X}\|\sigma \int_0^\infty \sqrt{\log \mathcal{N}(\mathcal{T}(\mathcal{C}, \mathbf{U}), d_2, \epsilon)} d\epsilon. \tag{46}$$

Now by Theorem 5.32 of Vershynin (2010), we have $\mathbb{E}\|\mathbf{X}\| \leq \sqrt{n} + \sqrt{r}$, then

$$\mathbb{E} n^{-1} \sup_{\mathbf{W} \in \mathcal{C}} \mathrm{tr}(\mathbf{U}\mathbf{\Gamma}^{1/2}\mathbf{X}^\top \mathbf{Z}\mathbf{G}_{\mathbf{W}}) \leq C\lambda_1^{1/2}\sigma(1/\sqrt{n} + \sqrt{r}/n) \int_0^\infty \sqrt{\log \mathcal{N}(\mathcal{T}(\mathcal{C}, \mathbf{U}), d_2, \epsilon)} d\epsilon. \tag{47}$$

Similarly, we can derive

$$\mathbb{E} n^{-1} \sup_{\mathbf{W} \in \mathcal{C}} \mathrm{tr}(\mathbf{Z}^\top \mathbf{X}\mathbf{\Gamma}^{1/2}\mathbf{U}^\top \mathbf{G}_{\mathbf{W}}) \leq C\lambda_1^{1/2}\sigma(1/\sqrt{n} + \sqrt{r}/n) \int_0^\infty \sqrt{\log \mathcal{N}(\mathcal{T}(\mathcal{C}, \mathbf{U}), d_2, \epsilon)} d\epsilon. \tag{48}$$

One the other hand, since $\sup_{\mathbf{W} \in \mathcal{C}} \mathrm{tr}(\mathbf{U}\mathbf{\Gamma}^{1/2}\bar{X}\bar{Z}^\top \mathbf{G}_{\mathbf{W}}) = \sup_{\mathbf{W} \in \mathcal{C}} \mathrm{tr}(\bar{X}\bar{Z}^\top \mathbf{G}_{\mathbf{W}}\mathbf{U}\mathbf{\Gamma}^{1/2}) \leq \lambda_1^{1/2} \sup_{\mathbf{W} \in \mathcal{C}} \mathrm{tr}(\bar{Z}^\top \mathbf{G}_{\mathbf{W}}\mathbf{U}\bar{X}/\|\bar{X}\|_2)\|\bar{X}\|_2 \leq \lambda_1^{1/2}\|\bar{X}\|_2 \sup_{\mathbf{g} \in \mathcal{T}_1} \langle \bar{Z}, \mathbf{g} \rangle$, where $\mathcal{T}_1(\mathcal{C}, \mathbf{U}, \mathbf{X}) = \{\frac{\mathbf{G}_{\mathbf{W}}\mathbf{U}\bar{X}}{\|\bar{X}\|_2} : \mathbf{W} \in \mathcal{C} \setminus \{\mathbf{U}\}\}$ is equipped with the Euclidean ℓ_2 distance. By Theorem 23, we have

$$\mathbb{E}\bigg[\sup_{\mathbf{g} \in \mathcal{T}_1(\mathcal{C}, \mathbf{U}, \mathbf{X})} \langle \bar{Z}, \mathbf{g} \rangle \bigg| \mathbf{X}\bigg] \leq \frac{C\sigma}{\sqrt{n}} \int_0^\infty \sqrt{\log \mathcal{N}(\mathcal{T}_1(\mathcal{C}, \mathbf{U}, \mathbf{X}), d_2, \epsilon)} d\epsilon.$$

Now for any $\mathbf{g}_1, \mathbf{g}_2 \in \mathcal{T}_1(\mathcal{C}, \mathbf{U}, \mathbf{X})$, without loss of generality, if we assume $\mathbf{g}_1 = \|\bar{X}\|_2^{-1}\mathbf{G}_{\mathbf{W}_1}\mathbf{U}\bar{X}$ and $\mathbf{g}_2 = \|\bar{X}\|_2^{-1}\mathbf{G}_{\mathbf{W}_2}\mathbf{U}\bar{X}$, where $\mathbf{W}_1, \mathbf{W}_2 \in \mathcal{C} \setminus \{\mathbf{U}\}$, then $\|\mathbf{g}_1 - \mathbf{g}_2\|_2 \leq \|\bar{X}\|_2^{-1}\|\mathbf{G}_{\mathbf{W}_1}\mathbf{U}\bar{X} - \mathbf{G}_{\mathbf{W}_2}\mathbf{U}\bar{X}\|_2 \leq d_\infty(\mathbf{G}_{\mathbf{W}_1}, \mathbf{G}_{\mathbf{W}_2}) \leq d_2(\mathbf{G}_{\mathbf{W}_1}, \mathbf{G}_{\mathbf{W}_2})$. Lemma 24 implies $\log \mathcal{N}(\mathcal{T}_1(\mathcal{C}, \mathbf{U}, \mathbf{X}), d_2, \epsilon) \leq \log \mathcal{N}(\mathcal{T}(\mathcal{C}, \mathbf{U}), d_2, \epsilon)$, which along with the fact that $\mathbb{E}\|\bar{X}\|_2 \lesssim \sqrt{r/n}$ implies

$$\mathbb{E} \sup_{\mathbf{W} \in \mathcal{C}} \mathrm{tr}(\mathbf{U}\mathbf{\Gamma}^{1/2}\bar{X}\bar{Z}^\top \mathbf{G}_{\mathbf{W}}) \leq \frac{C\sigma\sqrt{r}\lambda_1^{1/2}}{n} \int_0^\infty \sqrt{\log \mathcal{N}(\mathcal{T}(\mathcal{C}, \mathbf{U}), d_2, \epsilon)} d\epsilon. \tag{49}$$

Similarly, we have

$$\sup_{\mathbf{W} \in \mathcal{C}} \mathrm{tr}(\bar{Z}^\top \bar{X}\mathbf{\Gamma}^{1/2}\mathbf{U}^\top \mathbf{G}_{\mathbf{W}}) \leq \frac{C\sigma\sqrt{r}\lambda_1^{1/2}}{n} \int_0^\infty \sqrt{\log \mathcal{N}(\mathcal{T}(\mathcal{C}, \mathbf{U}), d_2, \epsilon)} d\epsilon. \tag{50}$$

Part II. Note that $\mathrm{tr}((n^{-1}\mathbf{Z}^\top \mathbf{Z} - \sigma^2 \mathbf{I}_p)\mathbf{G}_{\mathbf{W}}) = \mathrm{tr}(n^{-1}\mathbf{Z}^\top \mathbf{Z}\mathbf{G}_{\mathbf{W}}) - \sigma^2 \mathrm{tr}(\mathbf{G}_{\mathbf{W}}) = n^{-1}\mathrm{vec}(\mathbf{Z})^\top \mathbf{D}_{\mathbf{W}}\mathrm{vec}(\mathbf{Z})$, where $\mathbf{D}_{\mathbf{W}}$ is defined in (37). By the similar chaining argument in Part II of the proof of Theorem 8, we have

$$\mathbb{E} \sup_{\mathbf{W} \in \mathcal{C}} \mathrm{tr}((n^{-1}\mathbf{Z}^\top \mathbf{Z} - \mathbf{I}_p)\mathbf{G}_{\mathbf{W}}) \lesssim \frac{\sigma^2}{\sqrt{n}} \int_0^\infty \sqrt{\log \mathcal{N}(\mathcal{T}(\mathcal{C}, \mathbf{U}), d_2, \epsilon)} d\epsilon + \frac{\sigma^2}{n} \int_0^\infty \log \mathcal{N}(\mathcal{T}(\mathcal{C}, \mathbf{U}), d_2, \epsilon) d\epsilon \tag{51}$$

Similarly, since $\sup_{\mathbf{W} \in \mathcal{C}} \mathrm{tr}(\bar{Z}\bar{Z}^\top \mathbf{G}_{\mathbf{W}}) = \sup_{\mathbf{W} \in \mathcal{C}} \bar{Z}^\top \mathbf{G}_{\mathbf{W}}\bar{Z}$, we also have

$$\mathbb{E} \sup_{\mathbf{W} \in \mathcal{C}} \mathrm{tr}(\bar{Z}\bar{Z}^\top \mathbf{G}_{\mathbf{W}}) \lesssim \frac{\sigma^2}{n} \int_0^\infty \sqrt{\log \mathcal{N}(\mathcal{T}(\mathcal{C}, \mathbf{U}), d_2, \epsilon)} d\epsilon + \frac{\sigma^2}{n} \int_0^\infty \log \mathcal{N}(\mathcal{T}(\mathcal{C}, \mathbf{U}), d_2, \epsilon) d\epsilon. \tag{52}$$

Part III. Define the event $E = \{\|n^{-1}\mathbf{X}^\top\mathbf{X} - \mathbf{I}_r\| \le 1/(4L^2), \|\bar{X}\bar{X}^\top\| \le 1/(4L^2)\}$, where L is the constant in $\mathcal{Z}(\mathcal{C}, t, p, r)$. By Proposition D.1 in the Supplementary Material of Ma (2013),

$$P(\|n^{-1}\mathbf{X}^\top\mathbf{X} - \mathbf{I}_r\| \le 2(\sqrt{r/n} + t) + (\sqrt{r/n} + t)^2) \ge 1 - 2e^{-nt^2/2},$$

which implies $P(\|n^{-1}\mathbf{X}^\top\mathbf{X} - \mathbf{I}_r\| \le 1/(4L^2)) \ge 1 - 2e^{-cn}$. In addition, since $\|\bar{X}\bar{X}^\top\| \le \|\bar{X}\|_2^2 = \frac{1}{n}\sum_{i=1}^r g_i^2$, where $g_i \sim_{i.i.d.} N(0,1)$, it follows from the concentration inequality for independent exponential variables (Vershynin, 2010, Proposition 5.16) that $P(\|\bar{X}\bar{X}^\top\| \le 1/(4L^2)) \ge 1 - 2e^{-cn}$. Thus, it follows that

$$P(E^c) \le P(\|n^{-1}\mathbf{X}^\top\mathbf{X} - \mathbf{I}_r\| \ge 1/(4L^2)) + P(\|\bar{X}\bar{X}^\top\| \ge 1/(4L^2)) \le 4e^{-cn}.$$

Part IV. Note that $\mathbb{E}d(\mathbf{U}, \widehat{\mathbf{U}}) = \mathbb{E}[d(\mathbf{U}, \widehat{\mathbf{U}})|E] + \mathbb{E}[d(\mathbf{U}, \widehat{\mathbf{U}})|E^c]$. It follows from (44) and the inequalities (47)-(52) from Parts I and II that

$$\begin{aligned}
&\sup_{(\mathbf{\Gamma},\mathbf{U})\in\mathcal{Z}(\mathcal{C},t,p,t)} \mathbb{E}[d(\mathbf{U}, \widehat{\mathbf{U}})|E] \\
&\le \frac{C}{t}\left[\sqrt{t}\sigma\left(\frac{1}{\sqrt{n}} + \frac{\sqrt{r}}{n}\right)D(\mathcal{T}(\mathcal{C}, \mathbf{U}), d_2) + \frac{\sigma^2 D(\mathcal{T}(\mathcal{C}, \mathbf{U}), d_2)}{\sqrt{n}} + \frac{\sigma^2 D'(\mathcal{T}(\mathcal{C}, \mathbf{U}), d_2)}{n}\right] \\
&\le \frac{C\sigma\Delta(\mathcal{C})\sqrt{t(1 + r/n) + \sigma^2}}{\sqrt{n}t},
\end{aligned}$$

where the last inequality holds whenever $\sqrt{t}/\sigma \gtrsim \sup_{\mathbf{U}\in\mathcal{C}}[D'(\mathcal{T}(\mathcal{C}, \mathbf{U}), d_2)/D(\mathcal{T}(\mathcal{C}, \mathbf{U}), d_2)]$. On the other hand, by Part III, $\mathbb{E}[d(\mathbf{U}, \widehat{\mathbf{U}})|E^c] \le \mathrm{diam}(\mathcal{C}) \cdot P(E^c) \le C\sqrt{r}e^{-cn}$. Consequently as long as $n \gtrsim \max\{\log\frac{t}{\sigma^2}, r\}$ and $\sqrt{t}/\sigma \gtrsim \sup_{\mathbf{U}\in\mathcal{C}}[D'(\mathcal{T}(\mathcal{C}, \mathbf{U}), d_2)/D(\mathcal{T}(\mathcal{C}, \mathbf{U}), d_2)]$, we have

$$\sup_{(\mathbf{\Gamma},\mathbf{U})\in\mathcal{Z}(\mathcal{C},t,p,t)} \mathbb{E}d(\mathbf{U}, \widehat{\mathbf{U}}) \le \frac{C\sigma\Delta(\mathcal{C})\sqrt{t + \sigma^2}}{\sqrt{n}t}.$$

The final result then follows by noticing the trivial upper bound of $\mathrm{diam}(\mathcal{C})$.

A.2 Minimax Lower Bounds

Proof of Theorem 5. The proof is divided into two parts, the strong signal regime ($t^2 \ge \sigma^2 p_2/4$) and the weak signal regime ($t^2 < \sigma^2 p_2/4$). In the strong signal regime, the minimax lower bound does not depend on the dimensionality p_2 of $\mathbf{V}$, so it suffices to fix $\mathbf{V} = \mathbf{V}_0$ and perturb $\mathbf{U}$ around some $\mathbf{U}_0 \in \mathcal{C}$ to construct a parameter set whose corresponding probability measures are sufficiently close to each other. The minimax lower bound can be obtained by applying the general lower bound for testing *multiple simple hypotheses* (Lemma 30 below). However, in the weak signal regime, the minimax lower bound does depend on p_2. In this case, the construction of the probability measures should reflect the fundamental difficulty of estimating $\mathbf{U}$ due to the coupling effect of the unknown $\mathbf{V}$. In order to do so, we not only perturbed $\mathbf{U}$ around some $\mathbf{U}_0 \in \mathcal{C}$, but also considered an approximately uniformly distributed prior of $\mathbf{V}$ over $O(p_2, r)$ subject to certain spectral constraint. The minimax lower bound is then obtained by using a generalized Fano's method (Lemma 31) for testing *multiple composite hypotheses*. This mixing technique helps to obtain a sharper minimax lower bound (see also Cai and Zhang (2018)).

Part I. Strong Signal Regime. The following general lower bound for testing multiple hypotheses Tsybakov (2009) are needed.

Lemma 30 (Tsybakov (2009)) *Assume that $M \geq 2$ and suppose that (Θ, d) contains elements $\theta_0, \theta_1, ..., \theta_M$ such that: (i) $d(\theta_j, \theta_k) \geq 2s > 0$ for any $0 \leq j < k \leq M$; (ii) it holds that $\frac{1}{M}\sum_{j=1}^{M} D(P_j, P_0) \leq \alpha \log M$ with $0 < \alpha < 1/8$ and $P_j = P_{\theta_j}$ for $j = 0, 1, ..., M$, where $D(P_j, P_0) = \int \log \frac{dP_j}{dP_0} dP_j$ is the KL divergence between P_j and P_0. Then*

$$\inf_{\hat{\theta}} \sup_{\theta \in \Theta} P_\theta(d(\hat{\theta}, \theta) \geq s) \geq \frac{\sqrt{M}}{1+\sqrt{M}} \left(1 - 2\alpha - \sqrt{\frac{2\alpha}{\log M}}\right).$$

Let $\mathbf{V}_0 \in O(p_2, r)$ be fixed and $\mathbf{U}_0 \in \mathcal{C}$. Denote the ϵ-ball $B(\mathbf{U}_0, \epsilon) = \{\mathbf{U} \in O(p_1, r) : d(\mathbf{U}, \mathbf{U}_0) \leq \epsilon\}$. For some $\delta < \epsilon$, we consider the local δ-packing set $G_\delta = G(B(\mathbf{U}_0, \epsilon) \cap \mathcal{C}, d, \delta)$ such that for any pair $\mathbf{U}, \mathbf{U}' \in B(\mathbf{U}_0, \epsilon) \cap \mathcal{C}$, it holds that $d(\mathbf{U}, \mathbf{U}') = \|\mathbf{U}\mathbf{U}^\top - \mathbf{U}'\mathbf{U}'^\top\|_F \geq \delta$. We denote the elements of G_δ as $\mathbf{U}_i$ for $1 \leq i \leq |G_\delta|$. Lemma 21 shows that, for any i, we can find $\mathbf{O}_i \in O_r$ such that $\|\mathbf{U}_0 - \mathbf{U}_i\mathbf{O}_i\|_F \leq d(\mathbf{U}_0, \mathbf{U}_i) \leq \epsilon$. Set $\mathbf{U}_i' = \mathbf{U}_i\mathbf{O}_i$ and denote $G_\delta' = \{\mathbf{U}_i'\}$. For given $t > 0$, we consider the subset

$$\mathcal{X}(t, \epsilon, \delta, \mathbf{U}_0, \mathbf{V}_0) = \{(\mathbf{\Gamma}, \mathbf{U}, \mathbf{V}) : \mathbf{U} \in G_\delta', \mathbf{V} = \mathbf{V}_0, \mathbf{\Gamma} = t\mathbf{I}_r\} \subset \mathcal{Y}(\mathcal{C}, t, p_1, p_2, r).$$

In particular, the above construction admits $|\mathcal{X}(t, \epsilon, \delta, \mathbf{U}_0, \mathbf{V}_0)| = |G_\delta|$.

Moreover, for any $(\mathbf{\Gamma}, \mathbf{U}_i, \mathbf{V}_0) \in \mathcal{X}(t, \epsilon, \delta, \mathbf{U}_0, \mathbf{V}_0)$, let P_i be the probability measure of $\mathbf{Y} = \mathbf{U}_i\mathbf{\Gamma}\mathbf{V}_0^\top + \mathbf{Z}$ where $\mathbf{Z}$ has i.i.d. entries from $N(0, \sigma^2)$. We have, for $1 \leq i \neq j \leq |G_\delta|$,

$$D(P_i, P_j) = \frac{\|(\mathbf{U}_i' - \mathbf{U}_j')\mathbf{\Gamma}\mathbf{V}_0^\top\|_F^2}{2\sigma^2} \leq \frac{t^2\|\mathbf{U}_i' - \mathbf{U}_j'\|_F^2}{2\sigma^2} \leq \frac{2t^2\epsilon^2}{\sigma^2}.$$

Now set $\epsilon = \epsilon_0$ and $\delta = \alpha\epsilon$ for some $\alpha \in (0, 1)$. By assumption,

$$\left(\frac{c\sigma^2}{t^2} \log |G_{\alpha\epsilon_0}| \wedge \mathrm{diam}^2(\mathcal{C})\right) \leq \epsilon_0^2 \leq \left(\frac{\sigma^2}{32t^2} \log |G_{\alpha\epsilon_0}| \wedge \mathrm{diam}^2(\mathcal{C})\right) \tag{53}$$

for some $c \in (0, 1/32)$, it holds that $D(P_i, P_j) \leq \frac{1}{16} \log |G_{\alpha\epsilon_0}|$. Now by Lemma 30, it holds that, for $\theta = (\mathbf{\Gamma}, \mathbf{U}, \mathbf{V})$,

$$\inf_{\widehat{\mathbf{U}}} \sup_{\theta \in \mathcal{X}(t, \epsilon, \delta, \mathbf{U}_0, \mathbf{V}_0)} P_\theta(d(\widehat{\mathbf{U}}, \mathbf{U}) \geq \alpha\epsilon_0/2) \geq \frac{\sqrt{|G_{\alpha\epsilon_0}|}}{1 + \sqrt{|G_{\alpha\epsilon_0}|}} \left(\frac{7}{8} - \frac{1}{\sqrt{8 \log |G_{\alpha\epsilon_0}|}}\right).$$

By Markov's inequality, we have

$$\inf_{\widehat{\mathbf{U}}} \sup_{\theta \in \mathcal{X}(t, \epsilon, \delta, \mathbf{U}_0, \mathbf{V}_0)} \mathbb{E}_\theta d(\widehat{\mathbf{U}}, \mathbf{U}) \geq \frac{\alpha\epsilon_0\sqrt{|G_{\alpha\epsilon_0}|}}{2(1 + \sqrt{|G_{\alpha\epsilon_0}|})} \left(\frac{7}{8} - \frac{1}{\sqrt{8 \log |G_{\alpha\epsilon_0}|}}\right) \geq C\alpha\epsilon_0,$$

for some $C > 0$ as long as $|G_{\alpha\epsilon_0}| \geq 2$. Therefore, it holds that

$$\begin{aligned}
&\inf_{\widehat{\mathbf{U}}} \sup_{\theta \in \mathcal{Y}(\mathcal{C}, t, p_1, p_2, r)} \mathbb{E}_\theta d(\widehat{\mathbf{U}}, \mathbf{U}) \geq \inf_{\widehat{\mathbf{U}}} \sup_{\theta \in \mathcal{X}(t, \epsilon, \delta, \mathbf{U}_0, \mathbf{V}_0)} \mathbb{E}_\theta d(\widehat{\mathbf{U}}, \mathbf{U}) \\
&\gtrsim (\sigma t^{-1}\sqrt{\log |G_{\alpha\epsilon_0}|} \wedge \mathrm{diam}(\mathcal{C})) \gtrsim \left(\frac{\sigma\sqrt{t^2 + \sigma^2 p_2}}{t^2} \sqrt{\log |G_{\alpha\epsilon_0}|} \wedge \mathrm{diam}(\mathcal{C})\right).
\end{aligned}$$

Part II. Weak Signal Regime. The proof relies on the following generalized Fano's method, obtained by Ma et al. (2019), about testing multiple composite hypotheses.

Lemma 31 (Generalized Fano's Method) *Let $\mu_0, \mu_1, ..., \mu_M$ be $M+1$ priors on the parameter spaces Θ of the family $\{P_\theta\}$, and let P_j be the posterior probability measures on $(\mathcal{X}, \mathcal{A})$ such that*

$$P_j(S) = \int P_\theta(S)\mu_j(d\theta), \quad \forall S \in \mathcal{A}, \quad j = 0, 1, ..., M.$$

Let $F : \Theta \to (\mathbb{R}^d, d)$. If (i) there exist some sets $B_0, B_1, ..., B_M \subset \mathbb{R}^d$ such that $d(B_i, B_j) \geq 2s$ for some $s > 0$ for all $0 \leq i \neq j \leq M$ and $\mu_j(\theta \in \Theta : F(\theta) \in B_j) = 1$; and (ii) it holds that $\frac{1}{M}\sum_{j=1}^M D(P_j, P_0) \leq \alpha \log M$ with $0 < \alpha < 1/8$. Then

$$\inf_{\hat{F}} \sup_{\theta \in \Theta} P_\theta(d(\hat{F}, F(\theta)) \geq s) \geq \frac{\sqrt{M}}{1+\sqrt{M}}\left(1 - 2\alpha - \sqrt{\frac{2\alpha}{\log M}}\right).$$

To use the above lemma, we need to construct a collection of priors over the set $\mathcal{Y}(\mathcal{C}, t, p_1, p_2, r)$. Specifically, recall the previously constructed δ-packing set $G_\delta = \{\mathbf{U}_i : 1 \leq i \leq |G_\delta|\}$. Inspired by Cai and Zhang (2018), we consider the prior probability measure μ_i over $\mathcal{Y}(\mathcal{C}, t, p_1, p_2, r)$, whose definition is given as follows. Let $\mathbf{W}$ be a random matrix on $\mathbb{R}^{p_2 \times r}$, whose probability density is given by

$$p(\mathbf{W}) = C\left(\frac{p_2}{2\pi}\right)^{rp_2/2} \exp(-p_2\|\mathbf{W}\|_F^2/2) \cdot 1\{1/2 \leq \lambda_{\min}(\mathbf{W}) \leq \lambda_{\max}(\mathbf{W}) \leq 2\},$$

where C is a normalizing constant; then, if we denote $\tilde{\mathbf{U}}_i\tilde{\mathbf{\Gamma}}_i\tilde{\mathbf{V}}_i^\top$ as the SVD of $t\mathbf{U}_i\mathbf{W}^\top \in \mathbb{R}^{p_1 \times p_2}$ where $\mathbf{U}_i \in G_\delta$ and $\mathbf{W} \sim p(\mathbf{W})$, then μ_i is defined as the joint distribution of $(\tilde{\mathbf{\Gamma}}_i, \tilde{\mathbf{U}}_i, \tilde{\mathbf{V}}_i)$. By definition of $\mathbf{U}_i$, one can easily verify that μ_i is a well-defined probability measure on $\mathcal{Y}(\mathcal{C}, t, p_1, p_2, r)$. Note that, for any $\theta_i = (\tilde{\mathbf{\Gamma}}_i, \tilde{\mathbf{U}}_i, \tilde{\mathbf{V}}_i) \in \mathrm{supp}(\mu_i)$ and $\theta_j = (\tilde{\mathbf{\Gamma}}_j, \tilde{\mathbf{U}}_j, \tilde{\mathbf{V}}_j) \in \mathrm{supp}(\mu_j)$ with $1 \leq i \neq j \leq |G_\delta|$, it holds that $d(\tilde{\mathbf{U}}_i, \tilde{\mathbf{U}}_j) = d(\mathbf{U}_i, \mathbf{U}_j) \geq \delta$.

Consequently, the joint distribution of $\mathbf{Y} = \mathbf{U}\mathbf{\Gamma}\mathbf{V}^\top + \mathbf{Z}$ with $(\mathbf{\Gamma}, \mathbf{U}, \mathbf{V}) \sim \mu_i$ and $Z_{ij} \sim N(0, \sigma^2)$ can be expressed as

$$P_i(\mathbf{Y}) = C\int_{1/2 \leq \lambda_{\min}(\mathbf{W}) \leq \lambda_{\max}(\mathbf{W}) \leq 2} \frac{\sigma^{-p_1p_2}}{(2\pi)^{p_1p_2/2}} \exp(-\|\mathbf{Y} - t\mathbf{U}_i\mathbf{W}^\top\|_F^2/(2\sigma^2)) \times \left(\frac{p_2}{2\pi}\right)^{rp_2/2} \exp(-p_2\|\mathbf{W}\|_F^2/2)d\mathbf{W},$$

and it remains to control the pairwise KL divergence $D(P_i, P_j)$ for any $1 \leq i \neq j \leq |G_\delta|$. This is done by the next lemma, whose proof, which is involved, is delayed to Section C.

Lemma 32 *Under the assumption of the theorem, for any $1 \leq i \neq j \leq |G_\delta|$, we have $D(P_i, P_j) \leq \frac{C_1 t^4 d^2(\mathbf{U}_i, \mathbf{U}_j)}{\sigma^2(4t^2 + \sigma^2 p_2)} + C_2$ where $C_1, C_2 > 0$ are some uniform constant and $\{\mathbf{U}_i\}$ are elements of G_δ.*

Again, set $\epsilon = \epsilon_0$ and $\delta = \alpha\epsilon$ for some $\alpha \in (0, 1)$. By assumption,

$$\left(\frac{c\sigma^2(t^2 + \sigma^2 p_2)}{t^4} \log|G_\delta| \wedge \mathrm{diam}(\mathcal{C})\right) \leq \epsilon_0^2 \leq \left(\frac{\sigma^2(t^2 + \sigma^2 p_2)}{640t^4} \log|G_\delta| \wedge \mathrm{diam}(\mathcal{C})\right),$$

for some $c \in (0, 1/640]$. It then follows that $D(P_i, P_j) \le C \log |G_{\alpha\epsilon_0}| + C_2$. Now let $\mathcal{X}'(t, \epsilon, \delta, \mathbf{U}_0) = \bigcup_{1 \le i \le |G_{\alpha\epsilon_0}|} \mathrm{supp}(\mu_i)$. By Lemma 31 and Markov's inequality, we have, for $\theta = (\mathbf{\Gamma}, \mathbf{U}, \mathbf{V})$,

$$\inf_{\widehat{\mathbf{U}}} \sup_{\theta \in \mathcal{X}'(t, \epsilon_0, \alpha\epsilon_0, \mathbf{U}_0)} \mathbb{E}_\theta d(\widehat{\mathbf{U}}, \mathbf{U}) \ge \frac{\alpha\epsilon_0 \sqrt{|G_{\alpha\epsilon_0}|}}{2(1 + \sqrt{|G_{\alpha\epsilon_0}|})} \left(\frac{7}{8} - \frac{1}{\sqrt{8 \log |G_{\alpha\epsilon_0}|}} \right) \ge C\alpha\epsilon_0,$$

for some $C > 0$ as long as $|G_{\alpha\epsilon_0}| \ge 2$. Hence,

$$\begin{aligned} \inf_{\widehat{\mathbf{U}}} \sup_{\theta \in \mathcal{Y}(\mathcal{C}, t, p_1, p_2, r)} \mathbb{E}_\theta d(\widehat{\mathbf{U}}, \mathbf{U}) &\gtrsim \inf_{\widehat{\mathbf{U}}} \sup_{\theta \in \mathcal{X}'(t, \epsilon_0, \alpha\epsilon_0, \mathbf{U}_0)} \mathbb{E}_\theta d(\widehat{\mathbf{U}}, \mathbf{U}) \\ &\gtrsim \left(\frac{\sigma \sqrt{4t^2 + \sigma^2 p_2}}{t^2} \sqrt{\log |G_{\alpha\epsilon_0}|} \wedge \mathrm{diam}(\mathcal{C}) \right). \end{aligned}$$

Proof of Theorem 9. For some $\mathbf{U}_0 \in \mathcal{C}$, similar to the proof of Theorem 5, we consider the δ-packing set $G_\delta = G(B(\mathbf{U}_0, \epsilon) \cap \mathcal{C}, d, \delta)$, where for any $\mathbf{U}_i, \mathbf{U}_j \in G_\delta$, $d(\mathbf{U}_i, \mathbf{U}_j) = \|\mathbf{U}_i \mathbf{U}_i^\top - \mathbf{U}_j \mathbf{U}_j^\top\|_F \ge \delta$. Then, for given $t > 0$, we consider the subset $\mathcal{Z}'(t, \epsilon, \delta, \mathbf{U}_0) = \{(\mathbf{\Gamma}, \mathbf{U}) \in \mathcal{Z}(\mathcal{C}, t, p, r) : \mathbf{U} \in G_\delta, \mathbf{\Gamma} = t\mathbf{I}_r\}$, so that $|\mathcal{Z}'(t, \epsilon, \delta, \mathbf{U}_0)| = |G_\delta|$. Let P_i be the joint probability measure of $Y_k \sim_{i.i.d.} N(0, \mathbf{\Sigma}_i)$ with $k = 1, ..., n$ and $\mathbf{\Sigma}_i = t\mathbf{U}_i \mathbf{U}_i^\top + \sigma^2 \mathbf{I}_p$. We have, for any $1 \le i \ne j \le |G_\delta|$,

$$\begin{aligned} D(P_i, P_j) &= \frac{n}{2} \left(\mathrm{tr}(\mathbf{\Sigma}_j^{-1} \mathbf{\Sigma}_i) - p + \log \left(\frac{\det \mathbf{\Sigma}_i}{\det \mathbf{\Sigma}_j} \right) \right) \\ &= \frac{n}{2} \mathrm{tr} \bigg(- \frac{t}{t + \sigma^2} \mathbf{U}_i \mathbf{U}_i^\top + \frac{t}{\sigma^2} \mathbf{U}_j \mathbf{U}_j^\top - \frac{t^2}{\sigma^2(t + \sigma^2)} \mathbf{U}_i \mathbf{U}_i^\top \mathbf{U}_j \mathbf{U}_j^\top \bigg) \\ &= \frac{nt^2}{2\sigma^2(\sigma^2 + t)} (r - \|\mathbf{U}_i^\top \mathbf{U}_j\|_F^2) \le \frac{nt^2 d^2(\mathbf{U}_i, \mathbf{U}_j)}{\sigma^2(\sigma^2 + t)} \le \frac{nt^2 \epsilon^2}{\sigma^2(\sigma^2 + t)}, \end{aligned}$$

where the second equation follows from the Woodbury matrix identity and the second last inequality follows from Lemma 21. Now let $\epsilon = \epsilon_0$ and $\delta = \alpha\epsilon$ for some $\alpha \in (0, 1)$. By assumption,

$$\left(\frac{c\sigma^2(\sigma^2 + t)}{nt^2} \log |G_{\alpha\epsilon_0}| \wedge \mathrm{diam}^2(\mathcal{C}) \right) \le \epsilon_0^2 \le \left(\frac{\sigma^2(\sigma^2 + t)}{32nt^2} \log |G_{\alpha\epsilon_0}| \wedge \mathrm{diam}^2(\mathcal{C}) \right),$$

for some $c \in (0, 1/32)$. It holds that $D(P_i, P_j) \le \frac{1}{16} \log |G_{\alpha\epsilon_0}|$. Now by Lemma 30, it holds that, for $\theta = (\mathbf{\Gamma}, \mathbf{U})$,

$$\inf_{\widehat{\mathbf{U}}} \sup_{\theta \in \mathcal{Z}'(t, \epsilon_0, \alpha\epsilon_0, \mathbf{U}_0)} P_\theta(d(\widehat{\mathbf{U}}, \mathbf{U}) \ge \alpha\epsilon_0/2) \ge \frac{\sqrt{|G_{\alpha\epsilon_0}|}}{1 + \sqrt{|G_{\alpha\epsilon_0}|}} \left(\frac{7}{8} - \frac{1}{\sqrt{8 \log |G_{\alpha\epsilon_0}|}} \right).$$

By Markov's inequality, as long as $|G_{\alpha\epsilon_0}| \ge 2$, we have

$$\inf_{\widehat{\mathbf{U}}} \sup_{\theta \in \mathcal{Z}'(t, \epsilon_0, \alpha\epsilon_0, \mathbf{U})} \mathbb{E}_\theta d(\widehat{\mathbf{U}}, \mathbf{U}) \ge C\alpha\epsilon_0,$$

for some $C > 0$. Therefore, since $\mathcal{Z}'(t, \epsilon_0, \alpha\epsilon_0, \mathbf{U}) \subset \mathcal{Z}(\mathcal{C}, t, p, r)$,

$$\begin{aligned} \inf_{\widehat{\mathbf{U}}} \sup_{\theta \in \mathcal{Z}(\mathcal{C}, t, p, r)} \mathcal{R}(\widehat{\mathbf{U}}, \mathbf{U}) &\ge \inf_{\widehat{\mathbf{U}}} \sup_{\theta \in \mathcal{Z}'(t, \epsilon_0, \alpha\epsilon_0, \mathbf{U}_0)} \mathbb{E}_\theta d(\widehat{\mathbf{U}}, \mathbf{U}) \\ &\gtrsim \left(\frac{\sigma \sqrt{\sigma^2 + t}}{t\sqrt{n}} \sqrt{\log |G_{\alpha\epsilon_0}|} \wedge \mathrm{diam}(\mathcal{C}) \right). \end{aligned}$$

Appendix B. Calculation of Metric Entropies

In this section, we prove the results in Section 5 by calculating metric entropies of some specific sets. The calculation relies on the following useful lemmas.

Lemma 33 (Varshamov-Gilbert Bound) *Let $\Omega = \{0,1\}^n$ and $1 \le d \le n/4$. Then there exists a subset $\{\omega^{(1)}, ..., \omega^{(M)}\}$ of Ω such that $\|\omega^{(j)}\|_0 = d$ for all $1 \le j \le M$ and $\|\omega^{(j)} - \omega^{(k)}\|_0 \ge \frac{d}{2}$ for $0 \le j < k \le M$, and $\log M \ge cd \log \frac{n}{d}$ where $c \ge 0.233$.*

The proof of the above version of Varshamov-Gilbert bound can be found, for example, in Lemma 4.10 in Massart (2007)). The next two lemmas concern estimates of the covering/packing numbers of the orthogonal group.

Lemma 34 (Candes and Plan 2011) *Define $\mathcal{P}_0 = \{\bar{\mathbf{U}}\bar{\mathbf{\Gamma}}\bar{\mathbf{V}}^\top : \bar{\mathbf{U}}, \bar{\mathbf{V}} \in O(p, 2r), \|(\bar{\mathbf{\Gamma}}_{ii})_{1\le i\le 2r}\|_2 = 1\}$. Then for any $\epsilon \in (0, \sqrt{2})$, there exists an ϵ-covering set $H(\mathcal{P}_0, d_2, \epsilon)$ such that $|H(\mathcal{P}_0, d_2, \epsilon)| \le (c/\epsilon)^{2(2p+1)r}$ for some constant $c > 0$.*

Lemma 35 *For any $V \in O(k, r)$, identifying the subspace span(V) with its projection matrix $VV^\top$, define the metric on the Grassmannian manifold $G(k, r)$ by $\rho(VV^\top, UU^\top) = \|VV^\top - UU^\top\|_F$. Then for any $\epsilon \in (0, \sqrt{2(r \wedge (k-r))})$,*

$$\left(\frac{c_0}{\epsilon}\right)^{r(k-r)} \le \mathcal{N}(G(k,r), \rho, \epsilon) \le \left(\frac{c_1}{\epsilon}\right)^{r(k-r)},$$

where $\mathcal{N}(E, \epsilon)$ is the ϵ-covering number of E and c_0, c_1 are absolute constants. Moreover, for any $V \in O(k, r)$ and any $\alpha \in (0, 1)$, it holds that

$$\left(\frac{c_0}{\alpha c_1}\right)^{r(k-r)} \le \mathcal{M}(\mathbb{B}(V, \epsilon), \rho, \alpha\epsilon) \le \left(\frac{2c_1}{\alpha c_0}\right)^{r(k-r)}.$$

Proof We only prove the entropy upper bound

$$\mathcal{M}(\mathbb{B}(V, \epsilon), d, \alpha\epsilon) \le \left(\frac{c_0}{\alpha c_1}\right)^{r(k-r)}, \tag{54}$$

as the other results has been proved in Lemma 1 of Cai et al. (2013). Specifically, Let G_ϵ be the ϵ-packing set of $O(k, r)$. It then holds that

$$\begin{aligned}\mathcal{M}(O(k,r), d, \alpha\epsilon) &\ge \sum_{V \in G_\epsilon} \mathcal{M}(\mathbb{B}(V, \epsilon), d, \alpha\epsilon) \ge |G_\epsilon| \mathcal{M}(\mathbb{B}(V^*, \epsilon), d, \alpha\epsilon) \\ &= \mathcal{M}(O(k,r), d, \epsilon)\mathcal{M}(\mathbb{B}(V^*, \epsilon), d, \alpha\epsilon)\end{aligned}$$

for some $V^* \in O(k, r)$. Hence,

$$\mathcal{M}(\mathbb{B}(V^*, \epsilon), d, \alpha\epsilon) \le \frac{\mathcal{M}(O(k,r)), d, \alpha\epsilon)}{\mathcal{M}(O(k,r)), d, \epsilon)}.$$

By the equivalence between the packing and the covering numbers, it holds that

$$\mathcal{M}(\mathbb{B}(V^*, \epsilon), d, \alpha\epsilon) \le \frac{\mathcal{N}(O(k,r)), d, \alpha\epsilon/2)}{\mathcal{N}(O(k,r)), d, \epsilon)} \le \left(\frac{2c_1}{\alpha c_0}\right)^{r(k-r)},$$

where the last inequality follows from the first statement of the lemma. Then (54) holds since the metric d is unitarily invariant. ■

The following lemma is an estimate of the Dudley's entropy integral for the orthogonal group $O(p,r)$.

Lemma 36 *For any given $\mathbf{U} \in O(p,r)$, there exists some constant $C > 0$ such that $\int_0^\infty \sqrt{\log \mathcal{N}(\mathcal{T}(O(p,r),\mathbf{U}), d_2, \epsilon)}d\epsilon \le C\sqrt{pr}$. Therefore, we have $\Delta^2(O(p,r)) \le Cpr$.*

Proof By definition, for any $\mathbf{G} \in \mathcal{T}(O(p,r),\mathbf{U})$, it is at most rank $2r$, and suppose its SVD is $\mathbf{G} = \bar{\mathbf{U}}\bar{\mathbf{\Gamma}}\bar{\mathbf{V}}^\top$, then $\bar{\mathbf{\Gamma}}$ is a diagonal matrix with nonnegative diagonal entries and Frobenius norm equal to one. Thus, if we define $\mathcal{P}_0 = \{\bar{\mathbf{U}}\bar{\mathbf{\Gamma}}\bar{\mathbf{V}}^\top : \bar{\mathbf{U}}, \bar{\mathbf{V}} \in O(p,2r), \|(\bar{\mathbf{\Gamma}}_{ii})_{1\le i\le 2r}\|_2 = 1\}$, then by Lemma 24,

$$\mathcal{N}(\mathcal{T}(O(p,r),\mathbf{U}), d_2, \epsilon) \le \mathcal{N}(\mathcal{P}_0, d_2, \epsilon).$$

By Lemma 34, we can calculate that

$$\begin{aligned}\int_0^\infty \sqrt{\log \mathcal{N}(\mathcal{T}(O(p,r),\mathbf{U}), d_2, \epsilon)}d\epsilon &\le \int_0^\infty \sqrt{\log \mathcal{N}(\mathcal{P}_0, d_2, \epsilon)}d\epsilon \\ &\le C\sqrt{pr}\int_0^{\sqrt{2}} \sqrt{\log(c/\epsilon)}d\epsilon \le C\sqrt{pr}. \end{aligned} \tag{55}$$

The second statement follows directly from the definition of $\Delta^2(O(p,r))$. ■

B.1 Sparse PCA/SVD: Proof of Proposition 11 and Theorem 12

Matrix denoising model with $\mathcal{C}_S(p_1,r,k)$, or sparse SVD. By Lemma 33, we can construct a subset $\Theta_\epsilon(k) \subset \mathcal{C}_S(p_1,r,k)$ as follows. Let $\Omega_M = \{\omega^{(1)},...,\omega^{(M)}\} \subset \{0,1\}^{p_1-r-1}$ be the set obtained from Lemma 33 where $n = p_1 - r - 1$, $d = k/e < (p_1-r-1)/4$ and M is the smallest integer such that $\log M \ge cd\log n/d$, i.e., $M = \lceil \exp(ck\log\frac{e(p_1-r-1)}{k})\rceil$. We define

$$\Theta_\epsilon = \left\{ \begin{bmatrix} \mathbf{v} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{r-1} \end{bmatrix} : \mathbf{v} = (\sqrt{1-\epsilon^2}, \epsilon\omega/\sqrt{d}) \in \mathbb{S}^{p_1-r-1}, \omega \in \Omega_M \right\}, \qquad \epsilon \in (0,1).$$

Then Θ_ϵ is a $\frac{\epsilon}{2}$-packing set of $\mathbb{B}(\mathbf{U}_0, \sqrt{2}\epsilon) \cap \mathcal{C}_S(p_1,r,k)$ with $\mathbf{U}_0 = \begin{bmatrix} \mathbf{v}_0 & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{r-1} \end{bmatrix}$ where $\mathbf{v}_0 = (1,0,...,0)^\top$, $|\Theta_\epsilon| = M$. Now we set

$$\epsilon^2 = \frac{c_1(t^2+\sigma^2p_2)\sigma^2 k\log(e(p_1-r-1)/k)}{t^4} \wedge 1,$$

for some sufficiently small $c_1 > 0$. It follows that

$$\left(\frac{c_2\sigma^2(t^2+\sigma^2p_2)}{t^4}\log|\Theta_\epsilon| \wedge 1\right) \le \epsilon^2 \le \left(\frac{\sigma^2(t^2+\sigma^2p_2)}{640t^4}\log|\Theta_\epsilon| \wedge 1\right)$$

for some $c_2 \in (0, 1/640)$. So the condition of Theorem 5 holds with $\epsilon_0 = \sqrt{2}\epsilon$, $\alpha = 1/(2\sqrt{2})$ and $\log|\Theta_\epsilon| \asymp k\log(ep_1/k)$. Moreover, for any $\mathbf{U}' \in O(k,r)$, suppose $M_\epsilon \subset O(k,r)$ is an $\alpha\epsilon$-packing set of $\mathbb{B}(\mathbf{U}', \epsilon)$ constructed as in Lemma 35, then the set

$$\Theta'_\epsilon = \left\{ \mathbf{U} = \begin{bmatrix} \mathbf{W} \\ \mathbf{0} \end{bmatrix}, \mathbf{W} \in M_\epsilon \right\} \subset \mathcal{C}_S(p_1, r, k), \tag{56}$$

is an $\alpha\epsilon$-packing set of $\mathcal{C}_S(p_1,r,k) \cap \mathbb{B}(\mathbf{U}_0, \epsilon)$ where $\mathbf{U}_0 = \begin{bmatrix} \mathbf{U}' \\ \mathbf{0} \end{bmatrix}$, and $|\Theta'_\epsilon| \geq (c/\alpha)^{r(k-r)}$. Now we set

$$\epsilon^2 = \frac{c_1(t^2 + \sigma^2 p_2)\sigma^2 r(k-r)}{t^4} \wedge r^2,$$

for some sufficiently small $c_1 > 0$. It follows that

$$\left(\frac{c_2\sigma^2(t^2+\sigma^2p_2)}{t^4} \log|\Theta'_\epsilon| \wedge r \right) \leq \epsilon^2 \leq \left(\frac{\sigma^2(t^2+\sigma^2p_2)}{640t^4} \log|\Theta'_\epsilon| \wedge r \right)$$

for some $c_2 \in (0, 1/640)$. Thus, the condition of Theorem 5 holds with $\log|\Theta'_\epsilon| \asymp r(k-r)$.

To obtain an upper bound for $\Delta(\mathcal{C}_S(p_1,r,k))$, we notice that any element $\mathbf{H} \in \mathcal{T}(\mathcal{C}_S(p_1,r,k), \mathbf{U})$ satisfies $\mathbf{H} = \mathbf{H}^\top$ and

$$\max_{1\leq i\leq p_1} \|\mathbf{H}_{i\cdot}\|_0 \leq k, \quad \max_{1\leq i\leq p_1} \|\mathbf{H}_{\cdot i}\|_0 \leq k.$$

Then $\mathcal{T}(\mathcal{C}_S(p_1,r,k), \mathbf{U})$ can be covered by the union of its $\binom{p_1}{k}$ disjoint subsets, with each subset corresponding to a fixed sparsity configuration. Each of the above subsets can be identified with $\mathcal{T}(O(k,r), \mathbf{U}')$ for some $\mathbf{U}' \in O(k,r)$, and by Lemma 34 and the proof of Lemma 36,

$$\mathcal{N}(\mathcal{T}(O(k,r), \mathbf{U}'), d_2, \epsilon) \leq (c_1/\epsilon)^{2r(2k+1)}.$$

for any $\epsilon \in (0, \sqrt{2})$. Then by taking a union of the covering sets, we have

$$\mathcal{N}(\mathcal{T}(\mathcal{C}_S(p_1,r,k), \mathbf{U}), d_2, \epsilon) \leq \binom{p_1}{k} (c_1/\epsilon)^{2r(2k+1)} \leq (ep_1/k)^k (c_1/\epsilon)^{2r(2k+1)}.$$

As a result,

$$\begin{aligned} \int_0^\infty \sqrt{\log \mathcal{N}(\mathcal{T}(\mathcal{C}_S(p_1,r,k), \mathbf{U}), d_2, \epsilon)} d\epsilon &\leq \sqrt{2k\log(ep_1/k)} + \sqrt{2r(2k+1)} \int_0^{\sqrt{2}} \sqrt{\log \frac{c_1}{\epsilon}} d\epsilon \\ &\leq C(\sqrt{k\log(ep_1/k)} + \sqrt{rk}). \end{aligned}$$

In addition, we also have

$$\int_0^\infty \log \mathcal{N}(\mathcal{T}(\mathcal{C}_S(p_1,r,k), \mathbf{U}), d_2, \epsilon) d\epsilon \leq C(k\log(ep_1/k) + rk).$$

By Part III of the proof of Theorem 8, the upper bound result follows whenever $\frac{t}{\sigma} \gtrsim \sqrt{k\log(ep_1/k)} + \sqrt{rk}$. In particular, in light of the minimax lower bound (from Theorem 5), if $r = O(1)$, then, whenever consistent estimation is possible, or

$$\frac{\sigma\sqrt{t^2+\sigma^2p_2}}{t^2} \left(\sqrt{k\log\frac{ep_1}{k}} + \sqrt{k} \right) \lesssim 1,$$

the condition $\frac{t}{\sigma} \gtrsim \sqrt{k\log(ep_1/k)} + \sqrt{rk}$ is satisfied and the proposed estimator is minimax optimal. The final results follows by combining Theorems 5 and 8.

Spiked Wishart model with $\mathcal{C}_S(p,r,k)$, or sparse PCA. We omitted the proof of this case as it is similar to the proof of the sparse SVD.

B.2 Non-Negative PCA/SVD: Proof of Proposition 13 and Theorem 15

Matrix denoising model with $\mathcal{C}_N(p_1,r)$, or non-negative SVD. On the one hand, with Lemma 33, we can construct a subset $\Theta_\epsilon \subset O(p_1,r)$ as follows. Let $\Omega_M = \{\omega^{(1)},...,\omega^{(M)}\} \subset \{0,1\}^n$ be the set obtained from Lemma 33 where $n = p_1-r-1$, $d=(p_1-r-1)/4$ and M is the smallest integer such that $\log M \geq cd\log n/d$, i.e., $M = \lceil \exp(\frac{c(p_1-r-1)\log 2}{2})\rceil$. Following the idea of Vu and Lei (2012) and Cai et al. (2013), we define

$$\Theta_\epsilon = \left\{ \begin{bmatrix} \mathbf{v} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{r-1} \end{bmatrix} : \mathbf{v} = (\sqrt{1-\epsilon^2}, \epsilon\omega/\sqrt{d}) \in \mathbb{S}^{p_1-r-1}, \omega \in \Omega_M \right\}, \qquad \epsilon \in (0,1).$$

Then it holds that $\Theta_\epsilon \subset \mathbb{B}(\mathbf{U}_0, \sqrt{2}\epsilon)$ for $\mathbf{U}_0 = \begin{bmatrix} \mathbf{v}_0 & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{r-1} \end{bmatrix}$ where $\mathbf{v}_0 = (1,0,...,0)^\top$, $|\Theta_\epsilon| = M$, and that for any $\mathbf{U} \neq \mathbf{U}' \in \Theta_\epsilon$,

$$d(\mathbf{U},\mathbf{U}') \geq \sqrt{2}\cdot\sqrt{1-(1-\epsilon^2/8)^2} \geq \frac{\epsilon}{2}.$$

In other words, Θ_ϵ is a $\frac{\epsilon}{2}$-packing set of $\mathbb{B}(\mathbf{U}_0,\sqrt{2}\epsilon)\cap\mathcal{C}_{NN}(p_1,r)$. Now we set

$$\epsilon^2 = \frac{c_1(t^2+\sigma^2p_2)\sigma^2(p_1-r-1)}{t^4} \wedge 1,$$

for some sufficiently small $c_1 > 0$. It follows that

$$\left(\frac{c_2\sigma^2(t^2+\sigma^2p_2)}{t^4}\log|\Theta_\epsilon| \wedge 1\right) \leq \epsilon^2 \leq \left(\frac{\sigma^2(t^2+\sigma^2p_2)}{640t^4}\log|\Theta_\epsilon|\wedge 1\right)$$

for some $c_2 \in (0, 1/640)$. So the condition of Theorem 5 holds with $\epsilon_0 = \sqrt{2}\epsilon$, $\alpha = 1/(2\sqrt{2})$ and $\log|\Theta_\epsilon| \asymp p_1$.

On the other hand, we need to obtain an upper bound for $\Delta(\mathcal{C}_N(p_1,r))$. To bound the Dudley's entropy integral $\int_0^\infty \sqrt{\log\mathcal{N}(\mathcal{T}(\mathcal{C}_N(p_1,r),\mathbf{U}),d_2,\epsilon)}d\epsilon$, we simply use the fact that $\mathcal{C}_N(p_1,r)\subset O(p_1,r)$ and

$$\mathcal{N}(\mathcal{T}(\mathcal{C}_N(p_1,r),\mathbf{U}),d_2,\epsilon) \leq \mathcal{N}(\mathcal{T}(O(p_1,r),\mathbf{U}),d_2,\epsilon).$$

Then by Lemma 36, we have $\Delta^2(\mathcal{C}_{NN}(p_1,r)) \lesssim p_1 r$. Combining Theorems 5 and 8, we have $\Delta^2(\mathcal{C}_{NN}(p_1,r)) \gtrsim \log|\Theta_\epsilon|$, which implies $\Delta^2(\mathcal{C}_{NN}(p_1,r)) \asymp \log|\Theta_\epsilon| \asymp p_1$ if $r = O(1)$. Again, by Part III of the proof of Theorem 8, the upper bound follows whenever $\frac{t}{\sigma} \gtrsim \sqrt{rp_1}$. In particular, when $r = O(1)$, this condition is satisfied whenever

$$\frac{\sigma\sqrt{p_1(t^2+\sigma^2p_2)}}{t^2} \lesssim 1.$$

In other words, in light of the minimax lower bound (from Theorem 5), whenever consistent estimation is possible, the condition $\frac{t}{\sigma} \gtrsim \sqrt{p_1}$ is satisfied and the proposed estimator is minimax optimal.

Spiked Wishart model with $\mathcal{C}_N(p,r)$, or non-negative PCA. Similarly, let $\Omega_M = \{\omega^{(1)}, ..., \omega^{(M)}\} \subset \{0,1\}^{p-r-1}$ be the set obtained from Lemma 33 where $d = (p-r-1)/4$ and M is the smallest integer such that $\log M \geq cd\log(p-r-1)/d$, i.e., $M = \lceil \exp(\frac{c(p-r-1)\log 2}{2}) \rceil$. We define

$$\Theta_\epsilon = \left\{ \begin{bmatrix} \mathbf{v} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{r-1} \end{bmatrix} : \mathbf{v} = (\sqrt{1-\epsilon^2}, \epsilon\omega/\sqrt{d}) \in \mathbb{S}^{p-r-1}, \omega \in \Omega_M \right\}, \qquad \epsilon \in (0,1).$$

Then it holds that $\Theta_\epsilon \subset \mathbb{B}(\mathbf{U}_0, \sqrt{2}\epsilon)$ for $\mathbf{U}_0 = \begin{bmatrix} \mathbf{v}_0 & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{r-1} \end{bmatrix}$ where $\mathbf{v}_0 = (1,0,...,0)^\top$, $|\Theta_\epsilon| = M$, and that for any $\mathbf{U} \neq \mathbf{U}' \in \Theta_\epsilon$,

$$d(\mathbf{U}, \mathbf{U}') \geq \sqrt{2} \cdot \sqrt{1 - (1 - \epsilon^2/8)^2} \geq \frac{\epsilon}{2}.$$

In other words, Θ_ϵ is a $\frac{\epsilon}{2}$-packing set of $\mathbb{B}(\mathbf{U}_0, \sqrt{2}\epsilon) \cap \mathcal{C}_{NN}(p,r)$. Now we set

$$\epsilon^2 = \frac{c_1\sigma^2(\sigma^2+t)(p-r-1)}{nt^2} \wedge 1,$$

for some sufficiently small $c_1 > 0$. It follows that

$$\left(\frac{c_2\sigma^2(\sigma^2+t)}{nt^2}\log|\Theta_\epsilon| \wedge 1\right) \leq \epsilon^2 \leq \left(\frac{\sigma^2(\sigma^2+t)}{nt^2}\frac{(p-r-1)\log 2}{10} \wedge 1\right) \leq \left(\frac{\sigma^2(\sigma^2+t)}{32nt^2}\log|\Theta_\epsilon| \wedge 1\right)$$

for some $c_2 \in (0, 1/32)$, so that condition of Theorem 9 holds and $\log|\Theta_\epsilon| \asymp p$. The rest of the arguments such as the calculation of Dudley's entropy integral are the same as the above proof of the non-negative SVD.

B.3 Subspace PCA/SVD: Proof of Proposition 17 and Theorem 18

To prove this proposition, in light of Lemmas 34, 35 and 36, it suffices to establish the isometry between $(\mathcal{C}_A(p,r,k), d)$ and $(O(k,r), d)$. Let $\mathbf{Q} \in O(p,k)$ has its columns being the basis of the null space of A. We consider the map $F : O(k,r) \to \mathcal{C}_A(p,r,k)$ where $F(\mathbf{W}) = \mathbf{QW}$. To show that F is a bijection, we notice that

1. For any $\mathbf{G} \in \mathcal{C}_A(p,r,k)$, for each of its columns $\mathbf{Q}_{\cdot i}$, there exists some $\mathbf{v}_i \in \mathbb{S}^{k-1}$ such that $\mathbf{G}_{\cdot i} = \mathbf{Q}\mathbf{v}_i$ and $\mathbf{v}_i^\top \mathbf{v}_j = \mathbf{v}_i^\top \mathbf{Q}^\top \mathbf{Q}\mathbf{v}_j = \mathbf{G}_{\cdot i}^\top \mathbf{G}_{\cdot j} = 0$. Then let $\mathbf{W} = [\mathbf{v}_1, ..., \mathbf{v}_r] \in O(k,r)$, apparently, we have $F(\mathbf{W}) = \mathbf{G}$. This proves that the map is onto.

2. For any $\mathbf{W}_1 \neq \mathbf{W}_2 \in O(k,r)$, it follows that $F(\mathbf{W}_1) \neq F(\mathbf{W}_2)$. This proves the injection.

To show the map F is isometric, we notice that

1. For any $\mathbf{G}_1 = F(\mathbf{W}_1), \mathbf{G}_2 = F(\mathbf{W}_2) \in \mathcal{C}_A(p,r,k)$,

$$\begin{aligned} d(F(\mathbf{W}_1), F(\mathbf{W}_2)) &= \|\mathbf{Q}\mathbf{W}_1\mathbf{W}_1^\top\mathbf{Q}^\top - \mathbf{Q}\mathbf{W}_2\mathbf{W}_2^\top\mathbf{Q}^\top\|_F \\ &\leq \|\mathbf{Q}\|^2\|\mathbf{W}_1\mathbf{W}_1^\top - \mathbf{W}_2\mathbf{W}_2^\top\|_F \\ &\leq d(\mathbf{W}_1, \mathbf{W}_2). \end{aligned}$$

2. For any $\mathbf{W}_1, \mathbf{W}_2 \in O(k,r)$,
$$d(\mathbf{W}_1, \mathbf{W}_2) = \|\mathbf{Q}^\top \mathbf{Q}\mathbf{W}_1\mathbf{W}_1^\top \mathbf{Q}^\top - \mathbf{Q}\mathbf{W}_2\mathbf{W}_2^\top \mathbf{Q}^\top \mathbf{Q}\|_F \le d(F(\mathbf{W}_1, \mathbf{W}_2)).$$

Thus $d(F(\mathbf{W}_1), F(\mathbf{W}_2)) = d(\mathbf{W}_1, \mathbf{W}_2)$.

B.4 Spectral Clustering: Proof of Proposition 19 and Theorem 20

The upper bound $\Delta^2(\mathcal{C}_\pm^n) \lesssim n$ follows from the same argument as in the proof of Proposition 15. For the second statement, by Lemma 33, we can construct a subset $\Theta(d) \subset \mathbb{S}^{n-1}$ as follows. Let $\Omega_M = \{\omega^{(1)}, ..., \omega^{(M)}\} \subset \{0,1\}^n$ be the set obtained from Lemma 33 where $\|\omega^{(j)}\|_0 = d \le n/4$ for all $1 \le j \le n$ and M is the smallest integer such that $\log M \ge cd$, i.e., $M = \lceil \exp(cd \log \frac{n}{d}) \rceil$. We define

$$\Theta(d) = \left\{ \frac{2|\omega - 0.5 \cdot \mathbf{1}|}{\sqrt{n}} \in \mathcal{C}_\pm^n : \omega \in \Omega_M \cup \{(0, ..., 0)\} \right\},$$

where $\mathbf{1} = (1, ..., 1)^\top \in \mathbb{R}^n$. Then since for $\mathbf{u}_0 = (-1/\sqrt{n}, ..., -1/\sqrt{n})^\top$ and any $\mathbf{u} \in \Theta(d)$,

$$d(\mathbf{u}_0, \mathbf{u}) \le \|\mathbf{u}_0 - \mathbf{u}\|_2 \le 2\sqrt{\frac{d}{n}},$$

it holds that $\Theta(d) \subset \mathbb{B}(\mathbf{u}_0, 2\sqrt{d/n})$ with and that for any $\mathbf{u} \ne \mathbf{u}' \in \Theta(d)$,

$$d(\mathbf{u}, \mathbf{u}') \ge \frac{1}{\sqrt{2}} \|\mathbf{u} - \mathbf{u}'\|_2 \ge \sqrt{\frac{d}{n}}$$

so that $\Theta(d)$ is a $\sqrt{\frac{d}{n}}$-packing set of $\mathbb{B}(\mathbf{u}_0, 2\sqrt{d/n}) \cap \mathcal{C}_\pm^n$. Now since $t^2 = C\sigma^2(n + \sqrt{np})$, we can set

$$\epsilon_0 = \sqrt{\frac{d}{n}}, \quad \text{where} \quad d = c_1 n,$$

for some sufficiently small $c_1 > 0$, and thus it follows that

$$\left(\frac{c_2 \sigma^2 (t^2 + \sigma^2 p)}{t^4} \log |\Theta(d)| \wedge 1 \right) \le \epsilon_0^2 \le \left(\frac{\sigma^2 (t^2 + \sigma^2 p)}{128 t^4} \log |\Theta(d)| \wedge 1 \right)$$

for some $c_2 \in (0, 1/128)$. So the condition of Theorem 5 holds with $\alpha = 1/2$ and $\log |\Theta(d)| \asymp n$.

Appendix C. Proof of Technical Lemmas

Proof of Lemma 22. For the first statement, the first inequality can be proved by

$$\begin{aligned}
\langle \mathbf{U}\mathbf{\Gamma}^2\mathbf{U}, \mathbf{U}\mathbf{U}^\top - \mathbf{W}\mathbf{W}^\top \rangle &= \operatorname{tr}(\mathbf{U}\mathbf{\Gamma}^2\mathbf{U}^\top) - \operatorname{tr}(\mathbf{W}^\top \mathbf{U}\mathbf{\Gamma}^2\mathbf{U}^\top \mathbf{W}) \\
&= \operatorname{tr}(\mathbf{\Gamma}^2) - \operatorname{tr}(\mathbf{\Gamma}^2 \mathbf{U}^\top \mathbf{W}\mathbf{W}^\top \mathbf{U}) \\
&= \sum_{i=1}^r \lambda_i^2 (1 - (\mathbf{U}^\top \mathbf{W}\mathbf{W}^\top \mathbf{U})_{ii}) \\
&\ge \lambda_r^2 (r - \operatorname{tr}(\mathbf{U}^\top \mathbf{W}\mathbf{W}^\top \mathbf{U})) \\
&= \frac{\lambda_r^2}{2} \|\mathbf{U}\mathbf{U}^\top - \mathbf{W}\mathbf{W}^\top\|_F^2.
\end{aligned}$$

The second inequality follows from the same rationale. The second statement has been proved in Lemma 3 of Cai et al. (2013).

Proof of Lemma 29. Throughout the proof, for simplicity, we write $\mathcal{P} = \mathcal{P}(\mathcal{C}, \mathbf{U})$ and $\mathcal{T} = \mathcal{T}(\mathcal{C}, \mathbf{U})$. By Corollary 2.3.2 of Talagrand (2014), for any metric space (T, d), if we define

$$e_n(T) = \inf\{\epsilon : \mathcal{N}(T, d, \epsilon) \le N_n\}, \quad \text{where } N_0 = 1; N_n = 2^{2^n} \text{ for } n \ge 1, \tag{57}$$

then there exists some constant $K(\alpha)$ only depending on α such that

$$\gamma_\alpha(T, d) \le K(\alpha) \sum_{n \ge 0} 2^{n/\alpha} e_n(T). \tag{58}$$

The following inequalities establish the correspondence between e_n and the Dudley's entropy integral,

$$\begin{aligned} \sum_{n \ge 0} 2^{n/2} e_n(T) &\le C \int_0^\infty \sqrt{\log \mathcal{N}(T, d, \epsilon)} d\epsilon, \\ \sum_{n \ge 0} 2^{n} e_n(T) &\le C \int_0^\infty \log \mathcal{N}(T, d, \epsilon) d\epsilon, \end{aligned} \tag{59}$$

whose derivation is delayed to the end of this proof. Combining (58) and (59), it follows that

$$\gamma_\alpha(T, d) \le K(\alpha) \int_0^\infty \log^{1/\alpha} \mathcal{N}(T, d, \epsilon) d\epsilon. \tag{60}$$

By (60), it suffices to obtain estimates of the metric entropies $\log \mathcal{N}(\mathcal{P}, d_\infty, \epsilon)$ and $\sqrt{\log \mathcal{N}(\mathcal{P}, d_2, \epsilon)}$. By definition of $\mathcal{T}$, apparently $(\mathcal{P}, d_\infty)$ is isomorphic to $(\mathcal{T}, d_\infty)$, then by Lemma 24, it holds that

$$\mathcal{N}(\mathcal{P}, d_\infty, \epsilon) = \mathcal{N}(\mathcal{T}, d_\infty, \epsilon).$$

Along with the fact that, for any $\mathbf{G}_1, \mathbf{G}_2 \in \mathcal{T}$, $d_\infty(\mathbf{G}_1, \mathbf{G}_2) \le d_2(\mathbf{G}_1, \mathbf{G}_2)$ and therefore

$$\mathcal{N}(\mathcal{T}, d_\infty, \epsilon) \le \mathcal{N}(\mathcal{T}, d_2, \epsilon),$$

we prove the first statement of the lemma. On the other hand, consider the map $F : (\mathcal{P}, d_2) \to (\mathcal{T}, d_2)$ where for any $\mathbf{D} \in \mathcal{P}$, $F(\mathbf{D}) \in \mathbb{R}^{p_1 \times p_1}$ is the submatrix of $\mathbf{D}$ by extracting its entries in the first p_1 columns and rows. Then, for any $\mathbf{D}_1, \mathbf{D}_2 \in \mathcal{P}$, it holds that

$$d_2(F(\mathbf{D}_1), F(\mathbf{D}_2)) = \|F(\mathbf{D}_1) - F(\mathbf{D}_2)\|_F = \frac{1}{\sqrt{p_2}} d_2(\mathbf{D}_1, \mathbf{D}_2).$$

Again, applying Lemma 6, we have

$$\mathcal{N}(\mathcal{P}, d_2, \epsilon) = \mathcal{N}(\mathcal{T}, d_2, \epsilon/\sqrt{p_2}).$$

The second statement of the lemma then follows simply from the change of variable

$$\gamma_2(\mathcal{P}, d_2) \le C_2 \int_0^\infty \sqrt{\log \mathcal{N}(\mathcal{T}, d_2, \epsilon/\sqrt{p_2})} d\epsilon = C_2 \sqrt{p_2} \int_0^\infty \sqrt{\log \mathcal{N}(\mathcal{T}, d_2, \epsilon)} d\epsilon.$$

Proof of (59). The proof of the first inequality can be found, for example, on page 22 of Talagrand (2014). Nevertheless, we provide a detailed proof for completeness. By definition of e_n, if $\epsilon < e_n(T)$, we have $\mathcal{N}(T,d,\epsilon) > N_n$ and $\mathcal{N}(T,d,\epsilon) \geq N_n + 1$. Then

$$\sqrt{\log(1+N_n)}(e_n(T) - e_{n+1}(T)) \leq \int_{e_{n+1}(T)}^{e_n(T)} \sqrt{\log \mathcal{N}(T,d,\epsilon)}.$$

Since $\log(1+N_n) \geq 2^n \log 2$ for $n \geq 0$, summation over $n \geq 0$ yields

$$\sqrt{\log 2}\sum_{n\geq 0} 2^{n/2}(e_n - e_{n+1}(T)) \leq \int_0^{e_0(T)} \sqrt{\log \mathcal{N}(T,d,\epsilon)}.$$

Then the final inequality (59) follows by noting that

$$\begin{aligned}\sum_{n\geq 0} 2^{n/2}(e_n - e_{n+1}(T)) &= \sum_{n\geq 0} 2^{n/2} e_n(T) - \sum_{n\geq 1} 2^{(n-1)/2} e_n(T) \\ &\geq (1 - 1/\sqrt{2}) \sum_{n\geq 0} 2^{n/2} e_n(T).\end{aligned}$$

The second inequality can be obtained similarly by working with the inequality

$$\log(1+N_n)(e_n(T) - e_{n+1}(T)) \leq \int_{e_{n+1}(T)}^{e_n(T)} \log \mathcal{N}(T,d,\epsilon).$$

Proof of Lemma 32. The proof of this lemma generalizes the ideas in Cai and Zhang (2018) and Ma et al. (2019). In general, direct calculation of $D(P_i, P_j)$ is difficult. We detour by introducing an approximate density of P_i as

$$\tilde{P}_i(\mathbf{Y}) = \frac{\sigma^{-p_1 p_2}}{(2\pi)^{p_1 p_2/2}} \int \exp(-\|\mathbf{Y} - t\mathbf{U}_i\mathbf{W}^\top\|_F^2/(2\sigma^2)) \left(\frac{p_2}{2\pi}\right)^{rp_2/2} \exp(-p_2\|\mathbf{W}\|_F^2/2) d\mathbf{W}.$$

Now for $\mathbf{Y} \sim \tilde{P}_i$, if Y_k is the k-th column of $\mathbf{Y}$, we have

$$Y_k|\mathbf{U}_i \sim_{i.i.d.} N\left(0, \sigma^2\left(I_n - \frac{4t^2}{4t^2 + \sigma^2 p_2}\mathbf{U}_i\mathbf{U}_i^\top\right)^{-1}\right) = N\left(0, \sigma^2 I_n + \frac{4t^2}{p_2}\mathbf{U}_i\mathbf{U}_i^\top\right), \tag{61}$$

for $k = 1, ..., p_2$. It is well-known that the KL-divergence between two p-dimensional multivariate Gaussian distribution is

$$D(N(\mu_0, \mathbf{\Sigma}_0)\|N(\mu_1, \mathbf{\Sigma}_1)) = \frac{1}{2}\left(\mathrm{tr}(\mathbf{\Sigma}_0^{-1}\mathbf{\Sigma}_1) + (\mu_1 - \mu_0)^\top \mathbf{\Sigma}_1^{-1}(\mu_1 - \mu_0) - p + \log\left(\frac{\det \mathbf{\Sigma}_1}{\det \mathbf{\Sigma}_0}\right)\right).$$

As a result, we can calculate that for any $\tilde{P}_i$ and $\tilde{P}_j$,

$$\begin{aligned} D(\tilde{P}_i, \tilde{P}_j) &= \frac{p_2}{2}\left\{\mathrm{tr}\left(\left(I_{p_1} - \frac{4t^2}{4t^2 + \sigma^2 p_2}\mathbf{U}_i\mathbf{U}_i^\top\right)\left(I_{p_1} + \frac{4t^2}{\sigma^2 p_2}\mathbf{U}_j\mathbf{U}_j^\top\right)\right) - p_1\right\} \\ &\leq \frac{Ct^4}{4t^2 + \sigma^2 p_2}(r - \|\mathbf{U}_i^\top\mathbf{U}_j\|_F^2) \\ &= \frac{Ct^4 d(\mathbf{U}_i, \mathbf{U}_j)}{4t^2 + \sigma^2 p_2} \end{aligned} \tag{62}$$

where the last inequality follows from Lemma 21. Hence, the proof of this proposition is complete if we can show that there exist some constant $C > 0$ such that

$$D(P_i, P_j) \le D(\tilde{P}_i, \tilde{P}_j) + C. \tag{63}$$

The rest of the proof is devoted to the proof of (63).

Proof of (63). Define the event $\mathcal{G} = \{\mathbf{W} \in \mathbb{R}^{r\times p_2} : 1/2 \le \lambda_{\min}(\mathbf{W}) \le \lambda_{\max}(\mathbf{W}) \le 2\}$. For any given u,

$$\begin{aligned}
\frac{P_i}{\tilde{P}_i} &= \frac{1}{(2\pi)^{\frac{rp_2}{2}}(\frac{\sigma^2}{4t^2+\sigma^2p_2})^{\frac{rp_2}{2}}} \exp\left(\frac{1}{2\sigma^2}\sum_{k=1}^{p_2} Y_k^\top (I_{p_1} - \frac{4t^2}{4t^2+\sigma^2p_2}\mathbf{U}_i\mathbf{U}_i^\top)Y_k\right) \\
&\quad \times C_{\mathbf{U}_i,t}\int_{\mathcal{G}} \exp(-\|\mathbf{Y} - t\mathbf{U}_i\mathbf{W}^\top\|_F^2/(2\sigma^2) - p_2\|\mathbf{W}\|_F^2/2)d\mathbf{W} \\
&= \left(\frac{4t^2+\sigma^2p_2}{2\pi\sigma^2}\right)^{p_2r/2} \exp\left(-(4t^2+\sigma^2p_2)\left\|\mathbf{W} - \frac{2t}{4t^2+\sigma^2p_2}\mathbf{U}_i^\top\mathbf{Y}\right\|_F^2/2\right)d\mathbf{W} \\
&= C_{\mathbf{U}_i,t} P\left(\mathbf{W}' \in \mathcal{G}\middle| \mathbf{W}' \sim N\left(\frac{2t}{4t^2+\sigma^2p_2}\mathbf{U}_i^\top\mathbf{Y}, \frac{\sigma^2}{4t^2+\sigma^2p_2}\mathbf{I}_{p_1}\right)\right) \\
&\le C_{\mathbf{U}_i,t}.
\end{aligned} \tag{64}$$

Recall that

$$C_{\mathbf{U}_i,t}^{-1} = P\big(\mathbf{W} = (w_{jk}) \in \mathcal{G} | w_{jk} \sim N(0, 1/p_2)\big).$$

By concentration of measure inequalities for Gaussian random matrices (see, for example, Corollary 5.35 of Vershynin (2010)), we have, for sufficiently large (p_2, r),

$$P(\mathbf{W} \in \mathcal{G}) \ge 1 - 2\exp(-cp_2), \tag{65}$$

for some constant $c > 0$. In other words, we have

$$C_{\mathbf{U}_i,t}^{-1} \ge 1 - p_2^{-c} \tag{66}$$

and

$$\frac{P_i}{\tilde{P}_i} \le 1 + p_2^{-c} \tag{67}$$

uniformly for some constant $c > 0$. Thus, for some constant $\delta > 0$, we have

$$\begin{aligned}
D(P_i, P_j) &= \int P_i\left[\log\left(\frac{P_i}{\tilde{P}_i}\right) + \log\left(\frac{\tilde{P}_i}{\tilde{P}_j}\right) + \log\left(\frac{\tilde{P}_j}{P_j}\right)\right]d\mathbf{Y} \\
&\le \log(1+\delta) + D(\tilde{P}_i, \tilde{P}_j) + \int (P_i - \tilde{P}_i)\log\left(\frac{\tilde{P}_i}{\tilde{P}_j}\right)d\mathbf{Y} + \int P_i \log\left(\frac{\tilde{P}_i}{P_j}\right)d\mathbf{Y} \\
&\le \log(1+\delta) + D(\tilde{P}_i, \tilde{P}_j) + \int \tilde{P}_i\left(\frac{P_i}{\tilde{P}_i} - 1\right)\log\left(\frac{\tilde{P}_i}{\tilde{P}_j}\right)d\mathbf{Y} \\
&\quad + (1+\delta)\int \tilde{P}_i\left|\log\left(\frac{\tilde{P}_j}{P_j}\right)\right|d\mathbf{Y} \\
&\le \log(1+\delta) + D(\tilde{P}_i, \tilde{P}_j) + p_2^{-c}\int \tilde{P}_i\left|\log\left(\frac{\tilde{P}_i}{\tilde{P}_j}\right)\right|d\mathbf{Y} + (1+\delta)\int \tilde{P}_i\left|\log\left(\frac{\tilde{P}_j}{P_j}\right)\right|d\mathbf{Y}.
\end{aligned} \tag{68}$$

Now since

$$\begin{aligned}\int \tilde{P}_i\left|\log\left(\frac{\tilde{P}_i}{\tilde{P}_j}\right)\right|d\mathbf{Y} &= \frac{1}{2\sigma^2}\int \tilde{P}_i\left|\frac{4t^2}{4t^2+\sigma^2 p_2}\sum_{k=1}^{p_2} Y_k^\top(\mathbf{U}_i\mathbf{U}_i^\top - \mathbf{U}_j\mathbf{U}_j^\top)Y_k\right|d\mathbf{Y}\\ &\le \frac{1}{2\sigma^2}\mathbb{E}\bigg[\frac{4t^2}{4t^2+\sigma^2 p_2}\sum_{k=1}^{p_2} Y_k^\top(\mathbf{U}_i\mathbf{U}_i^\top + \mathbf{U}_j\mathbf{U}_j^\top)Y_k\bigg]\\ &= \frac{4t^2 p_2}{2\sigma^2(4t^2+\sigma^2 p_2)}\mathrm{tr}\bigg((\mathbf{U}_i\mathbf{U}_i^\top + \mathbf{U}_j\mathbf{U}_j^\top)(\sigma^2 I_{p_1} + \frac{4t^2}{p_2}\mathbf{U}_i\mathbf{U}_i^\top)\bigg)\\ &\le \frac{4t^2 p_2}{4t^2+\sigma^2 p_2}\mathrm{tr}\bigg(\mathbf{U}_i^\top(I_{p_1} + \frac{4t^2}{\sigma^2 p_2})\mathbf{U}_i\bigg)\\ &= \frac{4rt^2}{\sigma^2} \le rp_2,\end{aligned}$$

where in the second row the expectation is with respect to $Y_k \sim N\big(0, \sigma^2 I_{p_1} + \frac{4t^2\sigma^2}{\sigma^2 p_2}\mathbf{U}_i\mathbf{U}_i^\top\big)$. we know that the third term in (68) can be bounded by

$$p_2^{-c}\int \tilde{P}_i\left|\log\left(\frac{\tilde{P}_i}{\tilde{P}_j}\right)\right|d\mathbf{Y} \le rp_2\cdot p_2^{-c} \le C$$

for some constants $C, c > 0$. Finally, by (64), we have

$$\int \tilde{P}_i\left|\log\left(\frac{\tilde{P}_j}{P_j}\right)\right|dY \le \int \tilde{P}_i\left|\log\frac{1}{C_{\mathbf{U}_j,t}}\right|dY + \int \tilde{P}_i\left|\log\frac{1}{P(\mathbf{W}'\in\mathcal{G}|E)}\right|d\mathbf{Y},$$

where we denoted

$$E = \bigg\{\mathbf{W}' \sim N\bigg(\frac{2t}{4t^2+\sigma^2 p_2}\mathbf{U}_i^\top\mathbf{Y}, \frac{\sigma^2}{4t^2+\sigma^2 p_2}\mathbf{I}_{p_1}\bigg)\bigg\}.$$

Now on the one hand,

$$\int \tilde{P}_i\left|\log\frac{1}{C_{\mathbf{U}_i,t}}\right|d\mathbf{Y} \le \big(\log(1+\delta)\vee|\log(1-\delta)^{-1}|\big).$$

On the other hand, for fixed $\mathbf{Y}$ and $\mathbf{U}_i^\top\mathbf{Y}\in\mathbb{R}^{r\times p_2}$, we can find $\mathbf{Q}\in O(p_2, p_2-r)$ which is orthogonal to $\mathbf{U}_i^\top\mathbf{Y}$, i.e., $\mathbf{U}_i^\top\mathbf{Y}\mathbf{Q}=0$. Then $\mathbf{W}'\mathbf{Q}\in\mathbb{R}^{r\times(p_2-r)}$ are i.i.d. normal distributed with mean 0 and variance $\frac{\sigma^2}{4t^2+\sigma^2 p_2}$. Then again by standard result in random matrix (e.g. Corollary 5.35 in Vershynin (2010)), we have

$$\lambda_{\min}(\mathbf{W}') = \lambda_r(\mathbf{W}') \ge \lambda_r(\mathbf{W}'\mathbf{Q}) \ge \frac{\sigma}{\sqrt{4t^2+\sigma^2 p_2}}(\sqrt{p_2-r}-\sqrt{r}-x)$$

with probability at least $1-2\exp(-x^2/2)$. Since $t^2 < \sigma^2 p_2/4$, for p_2 sufficiently large, we can find c such that by setting $x = c\sqrt{p_2}$,

$$P(\lambda_{\min}(\mathbf{W}') \ge 1/2) \ge 1 - e^{-cp_2}. \tag{69}$$

Analogous to the argument on $\lambda_{\min}(\mathbf{W}')$, we also have

$$P(\lambda_{\max}(\mathbf{W}') \leq 2) \geq 1 - e^{-cp_2}. \tag{70}$$

Thus, by the union bound inequality, we have

$$P(\mathbf{W}' \in \mathcal{G}) \geq 1 - 2e^{-cp_2},$$

and consequently,

$$\int \tilde{P}_i \left| \log \frac{1}{P(\mathbf{W}' \in \mathcal{G}|E)} \right| d\mathbf{Y} \leq \left| \log \frac{1}{1 - p_2^{-c}} \right| \leq p_2^{-c}.$$

This helps us to bound the last term of (68). Combining the above results, we have proven the inequality (63) and therefore completed the proof.

References

Emmanuel Abbe, Jianqing Fan, and Kaizheng Wang. An ℓ_p theory of pca and spectral clustering. *arXiv preprint arXiv:2006.14062*, 2020.

Miguel A Arcones and Evarist Giné. On decoupling, series expansions, and tail behavior of chaos processes. *J. Theor. Probab.*, 6(1):101–122, 1993.

Martin Azizyan, Aarti Singh, and Larry Wasserman. Minimax theory for high-dimensional gaussian mixtures with sparse mean separation. In *NIPS*, pages 2139–2147, 2013.

Zhidong Bai and Jian-feng Yao. Central limit theorems for eigenvalues in a spiked population model. In *Annales de l'IHP Probabilités et Statistiques*, volume 44, pages 447–474, 2008.

Jinho Baik and Jack W Silverstein. Eigenvalues of large sample covariance matrices of spiked population models. *J. Multiv. Anal.*, 97(6):1382–1408, 2006.

Afonso S Bandeira, Nicolas Boumal, and Amit Singer. Tightness of the maximum likelihood semidefinite relaxation for angular synchronization. *Mathematical Programming*, 163(1-2):145–167, 2017.

Zhigang Bao, Xiucai Ding, and Ke Wang. Singular vector and singular subspace distribution for the matrix denoising model. *arXiv preprint arXiv:1809.10476*, 2018.

Peter L Bartlett and Shahar Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *J. Mach. Learn. Res.*, 3(Nov):463–482, 2002.

Aharon Birnbaum, Iain M Johnstone, Boaz Nadler, and Debashis Paul. Minimax bounds for sparse PCA with noisy high-dimensional data. *Ann. Statist.*, 41(3):1055–1084, 2013.

Nicolas Boumal. Nonconvex phase synchronization. *SIAM J. Optimiz.*, 26(4):2355–2377, 2016.

Olivier Bousquet, Vladimir Koltchinskii, and Dmitriy Panchenko. Some local measures of complexity of convex hulls and generalization bounds. In *International Conference on Computational Learning Theory*, pages 59–73. Springer, 2002.

T Tony Cai and Anru Zhang. Rate-optimal perturbation bounds for singular subspaces with applications to high-dimensional statistics. *Ann. Statist.*, 46(1):60–89, 2018.

T Tony Cai, Zongming Ma, and Yihong Wu. Sparse PCA: Optimal rates and adaptive estimation. *Ann. Statist.*, 41(6):3074–3110, 2013.

T. Tony Cai, Zongming Ma, and Yihong Wu. Optimal estimation and rank detection for sparse spiked covariance matrices. *Probab. Theory Related Fields*, 161:781–815, 2015.

T Tony Cai, Tengyuan Liang, and Alexander Rakhlin. Geometric inference for general high-dimensional linear inverse problems. *Ann. Statist.*, 44(4):1536–1563, 2016.

Emmanuel J Candes and Yaniv Plan. Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *IEEE Trans. Inform. Theory*, 57(4):2342–2359, 2011.

Yuxin Chen and Emmanuel J Candès. The projected power method: An efficient algorithm for joint alignment from pairwise differences. *Comm. Pure Appl. Math.*, 71(8):1648–1714, 2018.

Yunjin Choi, Jonathan Taylor, and Robert Tibshirani. Selecting the number of principal components: Estimation of the true rank of a noisy matrix. *Ann. Statist.*, 45(6):2590–2617, 2017.

Alexandre d'Aspremont, Laurent E Ghaoui, Michael I Jordan, and Gert R Lanckriet. A direct formulation for sparse PCA using semidefinite programming. In *NIPS*, pages 41–48, 2005.

Yash Deshpande and Andrea Montanari. Information-theoretically optimal sparse PCA. In *2014 IEEE Int. Symp. Info.*, pages 2197–2201. IEEE, 2014.

Yash Deshpande, Andrea Montanari, and Emile Richard. Cone-constrained principal component analysis. In *NIPS*, pages 2717–2725, 2014.

David Donoho and Matan Gavish. Minimax risk of matrix denoising by singular value thresholding. *Ann. Statist.*, 42(6):2413–2440, 2014.

David L Donoho, Matan Gavish, and Iain M Johnstone. Optimal shrinkage of eigenvalues in the spiked covariance model. *Ann. Statist.*, 46(4):1742, 2018.

Richard M Dudley. The sizes of compact subsets of Hilbert space and continuity of Gaussian processes. In *Selected Works of RM Dudley*, pages 125–165. Springer, 2010.

Alexandre d'Aspremont, Francis Bach, and Laurent El Ghaoui. Optimal solutions for sparse principal component analysis. *J. Mach. Learn. Res.*, 9(Jul):1269–1294, 2008.

Orizon Pereira Ferreira, Alfredo N Iusem, and Sandor Z Németh. Projections onto convex sets on the sphere. *Journal of Global Optimization*, 57(3):663–676, 2013.

Christophe Giraud and Nicolas Verzelen. Partial recovery bounds for clustering with the relaxed k means. *arXiv preprint arXiv:1807.07547*, 2018.

Gene H Golub and Charles F Van Loan. *Matrix Computations*, volume 3. JHU Press, 2012.

David Haussler and Manfred Opper. Metric entropy and minimax risk in classification. In *Structures in Logic and Computer Science*, pages 212–235. Springer, 1997a.

David Haussler and Manfred Opper. Mutual information, metric entropy and cumulative relative entropy risk. *Ann. Statist.*, 25(6):2451–2492, 1997b.

Adel Javanmard, Andrea Montanari, and Federico Ricci-Tersenghi. Phase transitions in semidefinite relaxations. *P. Natl. Acad. Sci.*, 113(16):E2218–E2223, 2016.

Jiashun Jin and Wanjie Wang. Influential features PCA for high dimensional clustering. *Ann. Statist.*, 44(6):2323–2359, 2016.

Jiashun Jin, Zheng Tracy Ke, and Wanjie Wang. Phase transitions for high dimensional clustering and related problems. *Ann. Statist.*, 45(5):2151–2189, 2017.

Iain M Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *Ann. Statist.*, 29(2):295–327, 2001.

Michel Journée, Yurii Nesterov, Peter Richtárik, and Rodolphe Sepulchre. Generalized power method for sparse principal component analysis. *J. Mach. Learn. Res.*, 11(Feb): 517–553, 2010.

Jaya Kawale and Daniel Boley. Constrained spectral clustering using l1 regularization. In *Proceedings of the 2013 SIAM International Conference on Data Mining*, pages 103–111. SIAM, 2013.

Matthäus Kleindessner, Samira Samadi, Pranjal Awasthi, and Jamie Morgenstern. Guarantees for spectral clustering with fairness constraints. *arXiv preprint arXiv:1901.08668*, 2019.

Vladimir Koltchinskii. Local rademacher complexities and oracle inequalities in risk minimization. *Ann. Statist.*, 34(6):2593–2656, 2006.

Felix Krahmer, Shahar Mendelson, and Holger Rauhut. Suprema of chaos processes and the restricted isometry property. *Comm. Pure Appl. Math.*, 67(11):1877–1904, 2014.

Guillaume Lecué and Shahar Mendelson. Aggregation via empirical risk minimization. *Probab. Theory Related Fields*, 145(3-4):591–613, 2009.

Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: isoperimetry and processes*. Springer Science & Business Media, 2013.

Matthias Löffler, Anderson Y Zhang, and Harrison H Zhou. Optimality of spectral clustering for gaussian mixture model. *arXiv preprint arXiv:1911.00538*, 2019.

Yu Lu and Harrison H Zhou. Statistical and computational guarantees of lloyd's algorithm and its variants. *arXiv preprint arXiv:1612.02099*, 2016.

Gábor Lugosi and Andrew B Nobel. Adaptive model selection using empirical complexities. *Ann. Statist.*, 27(6):1830–1864, 1999.

Rong Ma, T Tony Cai, and Hongzhe Li. Optimal estimation of bacterial growth rates based on permuted monotone matrix. *arXiv preprint arXiv:1911.12516*, 2019.

Rong Ma, T Tony Cai, and Hongzhe Li. Optimal permutation recovery in permuted monotone matrix model. *J. Amer. Statist. Assoc.*, pages 1–15, 2020.

Zongming Ma. Sparse principal component analysis and iterative thresholding. *Ann. Statist.*, 41(2):772–801, 2013.

Pascal Massart. *Concentration Inequalities and Model Selection: Ecole d'Eté de Probabilités de Saint-Flour XXXIII-2003*. Springer, 2007.

Andrea Montanari and Emile Richard. Non-negative principal component analysis: Message passing algorithms and sharp asymptotics. *IEEE Trans. Inform. Theory*, 62(3):1458–1484, 2015.

Mohamed Ndaoud. Sharp optimal recovery in the two component gaussian mixture model. *arXiv preprint arXiv:1812.08078*, 2018.

Efe Onaran and Soledad Villar. Projected power iteration for network alignment. In *Wavelets and Sparsity XVII*, volume 10394, page 103941C. International Society for Optics and Photonics, 2017.

Debashis Paul. Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statist. Sin.*, 17:1617–1642, 2007.

Amelia Perry, Alexander S Wein, Afonso S Bandeira, and Ankur Moitra. Optimality and sub-optimality of PCA i: Spiked random matrix models. *Ann. Statist.*, 46(5):2416–2451, 2018.

Alexander Rakhlin, Karthik Sridharan, and Alexandre B Tsybakov. Empirical entropy, minimax regret and minimax risk. *Bernoulli*, 23(2):789–824, 2017.

Sundeep Rangan and Alyson K Fletcher. Iterative estimation of constrained rank-one matrices in noise. In *2012 IEEE Int. Symp. Info. Proceedings*, pages 1246–1250. IEEE, 2012.

Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Trans. Inform. Theory*, 57(10): 6976–6994, 2011.

Haipeng Shen and Jianhua Z Huang. Sparse principal component analysis via regularized low rank matrix approximation. *J. Multiv. Anal.*, 99(6):1015–1034, 2008.

Amit Singer. Angular synchronization by eigenvectors and semidefinite programming. *Applied and Computational Harmonic Analysis*, 30(1):20–36, 2011.

Stanisław Szarek. Metric entropy of homogeneous spaces. *Banach Center Publications*, 43 (1):395–410, 1998.

Michel Talagrand. *Upper and lower bounds for stochastic processes: modern methods and classical problems*, volume 60. Springer Science & Business Media, 2014.

Alexandre B Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer, New York, 2009.

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

Roman Vershynin. *High-dimensional Probability: An Introduction with Applications in Data Science*, volume 47. Cambridge University Press, 2018.

Nicolas Verzelen. Minimax risks for sparse regressions: Ultra-high dimensional phenomenons. *Electronic Journal of Statistics*, 6:38–90, 2012.

Vincent Vu and Jing Lei. Minimax rates of estimation for sparse PCA in high dimensions. In *Artificial Intelligence and Statistics*, pages 1278–1286, 2012.

Vincent Q Vu and Jing Lei. Minimax sparse principal subspace estimation in high dimensions. *Ann. Statist.*, 41(6):2905–2947, 2013.

Vincent Q Vu, Juhee Cho, Jing Lei, and Karl Rohe. Fantope projection and selection: A near-optimal convex relaxation of sparse PCA. In *NIPS*, pages 2670–2678, 2013.

Weichen Wang and Jianqing Fan. Asymptotics of empirical eigenstructure for high dimensional spiked covariance. *Ann. Statist.*, 45(3):1342, 2017.

Xiang Wang and Ian Davidson. Flexible constrained spectral clustering. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 563–572. ACM, 2010.

Daniela M Witten, Robert Tibshirani, and Trevor Hastie. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, 10(3):515–534, 2009.

Yihong Wu and Pengkun Yang. Minimax rates of entropy estimation on large alphabets via best polynomial approximation. *IEEE Trans. Inform. Theory*, 62(6):3702–3720, 2016.

Dan Yang, Zongming Ma, and Andreas Buja. A sparse SVD method for high-dimensional data. *arXiv preprint arXiv:1112.2433*, 2011.

Yuhong Yang. Minimax nonparametric classification. i. rates of convergence. *IEEE Trans. Inform. Theory*, 45(7):2271–2284, 1999.

Yuhong Yang and Andrew Barron. Information-theoretic determination of minimax rates of convergence. *Ann. Statist.*, 27(5):1564–1599, 1999.

Yannis G Yatracos. A lower bound on the error in nonparametric regression type problems. *Ann. Statist.*, 16(3):1180–1187, 1988.

Xiao-Tong Yuan and Tong Zhang. Truncated power method for sparse eigenvalue problems. *J. Mach. Learn. Res.*, 14(Apr):899–925, 2013.

Anru Zhang, T Tony Cai, and Yihong Wu. Heteroskedastic PCA: Algorithm, optimality, and applications. *arXiv preprint arXiv:1810.08316*, 2018.

Hui Zou, Trevor Hastie, and Robert Tibshirani. Sparse principal component analysis. *J. Comput. Graph. Stat.*, 15(2):265–286, 2006.