On the Sample Complexity of Estimating Small Singular Modes

Xiangxiang Xu
Dept. of Electronic Engineering
Tsinghua University
Beijing, China 100084
Email: xuxx14@mails.tsinghua.edu.cn

Weida Wang, Shao-Lun Huang
DSIT Research Center
Tsinghua-Berkeley Shenzhen Institute
Shenzhen, China 518055
Email: {wangwd19@mails, shaolun.huang@sz}.tsinghua.edu.cn

***Abstract*—While it is commonly believed that estimating the small singular modes for a nearly low-rank matrix requires more samples, the sample size needed is generally unclear. In this paper, we investigate this sample complexity by considering the difference between the estimation errors of estimating a matrix with or without estimating these small singular modes. Specifically, we develop a mathematical framework based on the matrix perturbation analysis to characterize the noise level of estimating small singular modes by n samples. In particular, we show that under mild assumptions on the sample noise, it requires at least $n = O(\eta^{-2})$ samples to well estimate the singular modes with the singular value in the order of some small η. More importantly, our results are applied to the channel state estimation and Hirschfeld-Gebelein-Rényi (HGR) maximal correlation problems, from which we characterize that for how many samples, utilizing the low-rank approximation in these problems are beneficial. Finally, numerical simulations are provided to verify our results.**

I. Introduction

The low-rank approximation method has been widely used in information theory [1], machine learning [2], [3] and related applications, with benefits from two main aspects. Firstly, the implementation of low-rank approximation requires much less storage and computation resources than dealing with the full matrix and thus can be applied in the scenarios with limited resources [4]. Secondly—and perhaps more importantly—the low-rank approximation can be used as an approach for noise reduction, with applications in image denoising [5] and channel state estimation for MIMO (multiple-input multiple-output) system [6]. In particular, since the low-rank approximation method truncates all the small singular modes in the data, the noise in these modes is also eliminated. Specifically, when the estimation is based on data samples, these small singular modes can be extremely sensitive to the sample noise, and thus cannot be well-estimated, especially when the sample size is insufficient. Therefore, the performance gain from low-rank approximation can be used to equivalently characterize the sample complexity in estimating the small singular modes. As a motivating example, in the Hirschfeld-Gebelein-Rényi (HGR) maximal correlation problem [7]–[9], with given data samples, the goal is to estimate a normalized joint distribution matrix, called the canonical dependence matrix (CDM) [1]. With infinite number of samples, the empirical distribution of samples coincides with the true joint distribution, and the CDM can be accurately estimated. However, when the sample size is relatively small, the empirical distribution-based estimation is typically outperformed by its low-rank approximation, and this performance difference essentially characterizes the number of samples required to estimate those small singular modes.

In this paper, we characterizes this sample complexity problem, via considering the difference between the estimation errors of estimating a matrix with or without estimating these small singular modes. To this end, based on the matrix perturbation theory, we first develop a mathematical framework to characterize how much the small singular modes are polluted by the sample noise. In particular, for nearly low-rank matrices with small singular values of the order η, we provide non-asymptotic characterizations on the regimes of the number n of samples such that the small singular modes can be well-estimated with high probability. Specifically, we demonstrate that under mild assumptions on the sample noise, $n = O(\eta^{-2})$ samples are required to estimate the small singular modes of the order η well. Moreover, we apply our results in channel state estimation, and the Hirschfeld-Gebelein-Rényi (HGR) maximal correlation problem, which illustrates the sample complexity for estimating small singular modes, and also provides the criteria in deciding whether or not to adopt the low-rank approximation method in practice. Finally, numerical simulations are provided to verify our results.

II. Preliminaries and Definitions

Let $\mathbf{A} \in \mathbb{R}^{m_1 \times m_2}$ denote the nearly low-rank matrix to be estimated, and suppose it is generated by some matrix $\mathbf{Z} \in \mathbb{R}^{m_1 \times m_2}$ that can be observed via $\mathbf{A} = \tau(\mathbf{Z})$, with $\tau(\cdot)$ representing a known mapping, e.g., the mapping from joint distribution to the CDM in the HGR maximal correlation problem. Then, with n i.i.d. samples $\mathbf{Z}^{(1)}, \ldots, \mathbf{Z}^{(n)}$, the estimation $\hat{\mathbf{A}}$ is obtained via $\hat{\mathbf{A}} = \tau(\hat{\mathbf{Z}})$, where $\hat{\mathbf{Z}} \triangleq \frac{1}{n}\sum_{\ell=1}^{n} \mathbf{Z}^{(\ell)}$ represents the sample mean. In particular, we assume the samples $\mathbf{Z}^{(\ell)}$ satisfy $\mathbb{E}\left[\mathbf{Z}^{(\ell)}\right] = \mathbf{Z}$ and[1] $\mathrm{cov}(\mathrm{vec}(\mathbf{Z}^{(\ell)} - \mathbf{Z})) = \mathbf{R}$. In addition, we characterize the mapping τ by its gradient via

[1]Here, we apply the vectorization operation $\mathrm{vec}(\cdot)$ to stack all columns of a matrix into a vector. Specifically, for $\mathbf{W} = [w_{ij}] \in \mathbb{R}^{p \times q}$, $\mathrm{vec}(\mathbf{W})$ is a pq-dimensional column vector with the $[p(j-1)+i]$-th entry being w_{ij}.

defining the $m_1m_2 \times m_1m_2$ matrix $\mathbf{L}$ with the entry at $[(j-1)m_2+i]$-th row and $[(j'-1)m_2+i']$-th column being

$$L_{ij,i'j'} \triangleq \frac{\partial\,[\tau(\mathbf{Z})]_{ij}}{\partial\, Z_{i'j'}}. \tag{1}$$

Moreover, suppose that $\mathbf{A}$ has k dominant singular modes, with the singular value decomposition (SVD)

$$\mathbf{A} = \begin{bmatrix}\mathbf{U}_k & \mathbf{U}_\perp\end{bmatrix}\begin{bmatrix}\mathbf{\Sigma}_k & \mathbf{O}\\ \mathbf{O} & \mathbf{\Sigma}_\perp\end{bmatrix}\begin{bmatrix}\mathbf{V}_k^{\mathrm{T}}\\ \mathbf{V}_\perp^{\mathrm{T}}\end{bmatrix} \tag{2}$$

where $\mathbf{U}_k$ and $\mathbf{V}_k$ are the left and right singular vectors of the dominant k modes, $\mathbf{U}_\perp$ and $\mathbf{V}_\perp$ denote the singular vectors for the small singular modes, with $\mathbf{\Sigma}_k, \mathbf{\Sigma}_\perp$ being the corresponding singular values. Specifically, we assume that $\mathbf{\Sigma}_\perp = \eta\bar{\mathbf{\Sigma}}_\perp$ where $\eta > 0$ is a small quantity representing the order of the small singular values. In addition, we denote $\sigma_0 \triangleq \|\bar{\mathbf{\Sigma}}_\perp\|_{\mathrm{F}} > 0$, where $\|\cdot\|_{\mathrm{F}}$ denotes the Frobenius norm.

Similarly, the SVD of $\hat{\mathbf{A}}$ is denoted by

$$\hat{\mathbf{A}} = \begin{bmatrix}\hat{\mathbf{U}}_k & \hat{\mathbf{U}}_\perp\end{bmatrix}\begin{bmatrix}\hat{\mathbf{\Sigma}}_k & \mathbf{O}\\ \mathbf{O} & \hat{\mathbf{\Sigma}}_\perp\end{bmatrix}\begin{bmatrix}\hat{\mathbf{V}}_k^{\mathrm{T}}\\ \hat{\mathbf{V}}_\perp^{\mathrm{T}}\end{bmatrix}, \tag{3}$$

then the rank-k approximation $\hat{\mathbf{A}}_k$ of $\hat{\mathbf{A}}$ can be expressed as

$$\hat{\mathbf{A}}_k = \hat{\mathbf{U}}_k\hat{\mathbf{\Sigma}}_k\hat{\mathbf{V}}_k^{\mathrm{T}}. \tag{4}$$

To measure the performance of a given estimation $\mathbf{A}'$ of $\mathbf{A}$ in approximating $\mathbf{A}$, we use $d(\mathbf{A}', \mathbf{A})$ as the metric with $d\colon \mathbb{R}^{m_1\times m_2}\times\mathbb{R}^{m_1\times m_2}\mapsto\mathbb{R}$. Without loss of generality, we assume that $d(\mathbf{A}', \mathbf{A})$ is twice differentiable with respect to $\mathbf{A}'$, and $d(\mathbf{A}', \mathbf{A}) \geq 0$ with equality if and only if $\mathbf{A}' = \mathbf{A}$. Then, for given $\mathbf{\Xi}$ and a small ϵ, with

$$\mathbf{A}' = \mathbf{A} + \epsilon\mathbf{\Xi} + o(\epsilon), \tag{5}$$

the second-order Taylor series expansion of $d(\mathbf{A}', \mathbf{A})$ with respect to $\mathbf{A}'$ at $\mathbf{A}' = \mathbf{A}$ can be expressed as

$$d(\mathbf{A}', \mathbf{A}) = \frac{\epsilon^2}{2}\operatorname{vec}^{\mathrm{T}}(\mathbf{\Xi})\operatorname{Hess}_d\operatorname{vec}(\mathbf{\Xi}) + o(\epsilon^2), \tag{6}$$

where $\operatorname{Hess}_d \in \mathbb{R}^{m_1m_2\times m_1m_2}$ represents the Hessian matrix.

Then, the sample complexity of estimating the small singular modes $\mathbf{U}_\perp\mathbf{\Sigma}_\perp\mathbf{V}_\perp^{\mathrm{T}}$ of $\mathbf{A}$ can be equivalently characterized by the difference of estimation errors $d(\hat{\mathbf{A}}, \mathbf{A}) - d(\hat{\mathbf{A}}_k, \mathbf{A})$, since $\hat{\mathbf{A}} = \hat{\mathbf{A}}_k + \hat{\mathbf{U}}_\perp\hat{\mathbf{\Sigma}}_\perp\hat{\mathbf{V}}_\perp^{\mathrm{T}}$ differs from $\hat{\mathbf{A}}_k$ only in whether or not the small singular modes $\hat{\mathbf{U}}_\perp\hat{\mathbf{\Sigma}}_\perp\hat{\mathbf{V}}_\perp^{\mathrm{T}}$ are estimated. In this paper, we are interested in characterizing

$$p_n \triangleq \mathbb{P}_n\left\{d(\hat{\mathbf{A}}, \mathbf{A}) > d(\hat{\mathbf{A}}_k, \mathbf{A})\right\}, \tag{7}$$

and particularly the regime of n such that p_n is close to either 0 or 1, where the probability $\mathbb{P}_n\{\cdot\}$ is measured by n i.i.d samples $\mathbf{Z}^{(1)}, \ldots, \mathbf{Z}^{(n)}$. For this purpose, in the following we develop a mathematical framework for computing the difference $d(\hat{\mathbf{A}}, \mathbf{A}) - d(\hat{\mathbf{A}}_k, \mathbf{A})$ for some small perturbation $\hat{\mathbf{A}} - \mathbf{A}$, which corresponds to the practical scenario of large sample size n.

III. The Perturbation Analysis

For a given matrix $\mathbf{A}'$ as defined in (5), let $\mathbf{A}'_k$ denote the rank-k approximation of $\mathbf{A}'$ defined similar to (4), and we have the following characterization of $d(\mathbf{A}', \mathbf{A}) - d(\mathbf{A}'_k, \mathbf{A})$ for $\mathbf{A}'$ close to $\mathbf{A}$.

Lemma 1. *We have*

$$\begin{aligned} &d(\mathbf{A}', \mathbf{A}) - d(\mathbf{A}'_k, \mathbf{A})\\ &= \frac{\epsilon^2}{2}\operatorname{vec}^{\mathrm{T}}(\mathbf{\Xi})\mathbf{J}_d\operatorname{vec}(\mathbf{\Xi})\\ &\quad - \frac{\eta^2}{2}\operatorname{vec}^{\mathrm{T}}(\mathbf{U}_\perp\bar{\mathbf{\Sigma}}_\perp\mathbf{V}_\perp^{\mathrm{T}})\operatorname{Hess}_d\operatorname{vec}(\mathbf{U}_\perp\bar{\mathbf{\Sigma}}_\perp\mathbf{V}_\perp^{\mathrm{T}})\\ &\quad + \epsilon\eta\operatorname{vec}^{\mathrm{T}}(\mathbf{U}_\perp\bar{\mathbf{\Sigma}}_\perp\mathbf{V}_\perp^{\mathrm{T}})\operatorname{Hess}_d(\mathbf{I}-\mathbf{G})\operatorname{vec}(\mathbf{\Xi})\\ &\quad + o(\epsilon^2+\eta^2), \end{aligned} \tag{8}$$

where $\mathbf{J}_d \triangleq \operatorname{Hess}_d - (\mathbf{I}-\mathbf{G})\operatorname{Hess}_d(\mathbf{I}-\mathbf{G})$, *and where we have defined the projection matrix* $\mathbf{G} \triangleq (\mathbf{V}_\perp\mathbf{V}_\perp^{\mathrm{T}})\otimes(\mathbf{U}_\perp\mathbf{U}_\perp^{\mathrm{T}})$ *with "$\otimes$" denoting the Kronecker product.*

Proof. From (5)–(6), we have

$$\begin{aligned} &d(\mathbf{A}', \mathbf{A}) - d(\mathbf{A}'_k, \mathbf{A})\\ &= \frac{\epsilon^2}{2}\operatorname{vec}^{\mathrm{T}}(\mathbf{\Xi})\operatorname{Hess}_d\operatorname{vec}(\mathbf{\Xi})\\ &\quad - \frac{1}{2}\operatorname{vec}^{\mathrm{T}}(\mathbf{A}'_k-\mathbf{A})\operatorname{Hess}_d\operatorname{vec}(\mathbf{A}'_k-\mathbf{A}) + o(\epsilon^2+\eta^2). \end{aligned} \tag{9}$$

To reduce (9), we first establish the rank-k matrix $\mathbf{A}'_k$ using the matrix perturbation analysis. To this end, note that $\mathbf{A}'$ can be expressed as $\mathbf{A}' = \mathbf{A}_k + \mathbf{N}$, with $\mathbf{N} \triangleq \mathbf{A}' - \mathbf{A}_k = \eta\mathbf{U}_\perp\bar{\mathbf{\Sigma}}_\perp\mathbf{V}_\perp^{\mathrm{T}} + \epsilon\mathbf{\Xi} + o(\epsilon)$ denoting the perturbation term. Let $\mathbf{U}', \mathbf{V}'$ denote the matrices composed by the first k left and right singular vectors of $\mathbf{A}'$, respectively, then it follows from [10] that

$$\begin{aligned} \mathbf{U}'_k\mathbf{U}'^{\mathrm{T}}_k &= \mathbf{U}_k\mathbf{U}_k^{\mathrm{T}} - \mathbf{U}_k\mathbf{Q}_1\mathbf{Q}_1^{\mathrm{T}}\mathbf{U}_k^{\mathrm{T}} - \mathbf{U}_k\mathbf{Q}_2\mathbf{U}_\perp^{\mathrm{T}}\\ &\quad - \mathbf{U}_\perp\mathbf{Q}_2^{\mathrm{T}}\mathbf{U}_k^{\mathrm{T}} + \mathbf{U}_\perp\mathbf{Q}_1^{\mathrm{T}}\mathbf{Q}_1\mathbf{U}_\perp^{\mathrm{T}} + o(\epsilon^2)\\ &= \mathbf{U}_k\mathbf{U}_k^{\mathrm{T}} + \epsilon\operatorname{sym}(\mathbf{U}_k\mathbf{\Sigma}_k^{-1}\mathbf{V}_k^{\mathrm{T}}\mathbf{\Xi}^{\mathrm{T}}\mathbf{U}_\perp\mathbf{U}_\perp^{\mathrm{T}}) + o(\epsilon), \end{aligned}$$

where we have defined $\mathbf{Q}_1 \triangleq -\mathbf{\Sigma}_k^{-1}\mathbf{V}_k^{\mathrm{T}}\mathbf{N}^{\mathrm{T}}\mathbf{U}_\perp$, $\mathbf{Q}_2 \triangleq \mathbf{Q}_1 + (\mathbf{\Sigma}_k^{-1}\mathbf{V}_k^{\mathrm{T}}\mathbf{N}^{\mathrm{T}}\mathbf{U}_k\mathbf{\Sigma}_k^{-1}\mathbf{V}_k^{\mathrm{T}} - \mathbf{\Sigma}_k^{-2}\mathbf{U}_k^{\mathrm{T}}\mathbf{N}\mathbf{V}_\perp\mathbf{V}_\perp^{\mathrm{T}})\mathbf{N}^{\mathrm{T}}\mathbf{U}_\perp$, and for square matrices $\mathbf{M}$, $\operatorname{sym}(\mathbf{M}) \triangleq \mathbf{M} + \mathbf{M}^{\mathrm{T}}$. Similarly, we have

$$\mathbf{V}'_k\mathbf{V}'^{\mathrm{T}}_k = \mathbf{V}_k\mathbf{V}_k^{\mathrm{T}} + \epsilon\operatorname{sym}(\mathbf{V}_k\mathbf{\Sigma}_k^{-1}\mathbf{U}_k^{\mathrm{T}}\mathbf{\Xi}\mathbf{V}_\perp\mathbf{V}_\perp^{\mathrm{T}}) + o(\epsilon).$$

Therefore,

$$\begin{aligned} &(\mathbf{V}'_k\mathbf{V}'^{\mathrm{T}}_k)\otimes(\mathbf{U}'_k\mathbf{U}'^{\mathrm{T}}_k)\\ &= (\mathbf{V}_k\mathbf{V}_k^{\mathrm{T}})\otimes(\mathbf{U}_k\mathbf{U}_k^{\mathrm{T}})\\ &\quad + \epsilon\big[(\mathbf{V}_k\mathbf{V}_k^{\mathrm{T}})\otimes\operatorname{sym}(\mathbf{U}_k\mathbf{\Sigma}_k^{-1}\mathbf{V}_k^{\mathrm{T}}\mathbf{\Xi}^{\mathrm{T}}\mathbf{U}_\perp\mathbf{U}_\perp^{\mathrm{T}})\\ &\qquad + \operatorname{sym}(\mathbf{V}_k\mathbf{\Sigma}_k^{-1}\mathbf{U}_k^{\mathrm{T}}\mathbf{\Xi}\mathbf{V}_\perp\mathbf{V}_\perp^{\mathrm{T}})\otimes(\mathbf{U}_k\mathbf{U}_k^{\mathrm{T}})\big] + o(\epsilon)\\ &= \mathbf{G}_k + \epsilon\mathbf{\Gamma} + o(\epsilon), \end{aligned} \tag{10}$$

where we have defined $\mathbf{G}_k \triangleq (\mathbf{V}_k\mathbf{V}_k^{\mathrm{T}})\otimes(\mathbf{U}_k\mathbf{U}_k^{\mathrm{T}})$ and

$$\mathbf{\Gamma} \triangleq (\mathbf{V}_k\mathbf{V}_k^{\mathrm{T}})\otimes\operatorname{sym}(\mathbf{U}_k\mathbf{\Sigma}_k^{-1}\mathbf{V}_k^{\mathrm{T}}\mathbf{\Xi}^{\mathrm{T}}\mathbf{U}_\perp\mathbf{U}_\perp^{\mathrm{T}})$$

$$+\operatorname{sym}(\mathbf{V}_k\mathbf{\Sigma}_k^{-1}\mathbf{U}_k^{\mathrm{T}}\mathbf{\Xi}\mathbf{V}_\perp\mathbf{V}_\perp^{\mathrm{T}})\otimes(\mathbf{U}_k\mathbf{U}_k^{\mathrm{T}}).$$

Also, note that

$$\mathbf{G}_k\operatorname{vec}(\mathbf{U}_k\mathbf{\Sigma}_k\mathbf{V}_k^{\mathrm{T}})=\operatorname{vec}(\mathbf{U}_k\mathbf{\Sigma}_k\mathbf{V}_k^{\mathrm{T}}),$$
$$\mathbf{G}_k\operatorname{vec}(\mathbf{U}_\perp\bar{\mathbf{\Sigma}}_\perp\mathbf{V}_\perp^{\mathrm{T}})=\mathbf{\Gamma}\operatorname{vec}(\mathbf{U}_\perp\bar{\mathbf{\Sigma}}_\perp\mathbf{V}_\perp^{\mathrm{T}})=\mathbf{0},$$

and

$$\mathbf{\Gamma}\operatorname{vec}(\mathbf{U}_k\mathbf{\Sigma}_k\mathbf{V}_k^{\mathrm{T}})$$
$$=\operatorname{vec}(\mathbf{U}_\perp\mathbf{U}_\perp^{\mathrm{T}}\mathbf{\Xi}\mathbf{V}_k\mathbf{V}_k^{\mathrm{T}})+\operatorname{vec}(\mathbf{U}_k\mathbf{U}_k^{\mathrm{T}}\mathbf{\Xi}\mathbf{V}_\perp\mathbf{V}_\perp^{\mathrm{T}}),$$

where we have used the fact that $(\mathbf{M}_1\otimes\mathbf{M}_2)\operatorname{vec}(\mathbf{M}_3)=\operatorname{vec}(\mathbf{M}_2^{\mathrm{T}}\mathbf{M}_3\mathbf{M}_1)$ for matrices $\mathbf{M}_1,\mathbf{M}_2,\mathbf{M}_3$ with compatible dimensions.

Then, we have

$$\begin{aligned}
&\operatorname{vec}(\mathbf{A}_k')-\operatorname{vec}(\mathbf{A})\\
&=\operatorname{vec}(\hat{\mathbf{U}}_k\hat{\mathbf{U}}_k^{\mathrm{T}}\mathbf{A}'\hat{\mathbf{V}}_k\hat{\mathbf{V}}_k^{\mathrm{T}})-\operatorname{vec}(\mathbf{A})\\
&=(\hat{\mathbf{V}}_k\hat{\mathbf{V}}_k^{\mathrm{T}})\otimes(\hat{\mathbf{U}}_k\hat{\mathbf{U}}_k^{\mathrm{T}})\operatorname{vec}(\mathbf{A}')-\operatorname{vec}(\mathbf{A})\\
&=(\hat{\mathbf{V}}_k\hat{\mathbf{V}}_k^{\mathrm{T}})\otimes(\hat{\mathbf{U}}_k\hat{\mathbf{U}}_k^{\mathrm{T}})(\operatorname{vec}(\mathbf{A})+\epsilon\operatorname{vec}(\mathbf{\Xi}))-\operatorname{vec}(\mathbf{A})+o(\epsilon)\\
&=[(\hat{\mathbf{V}}_k\hat{\mathbf{V}}_k^{\mathrm{T}})\otimes(\hat{\mathbf{U}}_k\hat{\mathbf{U}}_k^{\mathrm{T}})-\mathbf{I}]\operatorname{vec}(\mathbf{U}_k\mathbf{\Sigma}_k\mathbf{V}_k^{\mathrm{T}})\\
&\quad+\eta[(\hat{\mathbf{V}}_k\hat{\mathbf{V}}_k^{\mathrm{T}})\otimes(\hat{\mathbf{U}}_k\hat{\mathbf{U}}_k^{\mathrm{T}})-\mathbf{I}]\operatorname{vec}(\mathbf{U}_\perp\bar{\mathbf{\Sigma}}_\perp\mathbf{V}_\perp^{\mathrm{T}})\\
&\quad+\epsilon(\hat{\mathbf{V}}_k\hat{\mathbf{V}}_k^{\mathrm{T}})\otimes(\hat{\mathbf{U}}_k\hat{\mathbf{U}}_k^{\mathrm{T}})\operatorname{vec}(\mathbf{\Xi})+o(\epsilon)\\
&=(\mathbf{G}_k+\epsilon\mathbf{\Gamma}-\mathbf{I})\operatorname{vec}(\mathbf{U}_k\mathbf{\Sigma}_k\mathbf{V}_k^{\mathrm{T}})\\
&\quad+\eta(\mathbf{G}_k+\epsilon\mathbf{\Gamma}-\mathbf{I})\operatorname{vec}(\mathbf{U}_\perp\bar{\mathbf{\Sigma}}_\perp\mathbf{V}_\perp^{\mathrm{T}})\\
&\quad+(\mathbf{G}_k+\epsilon\mathbf{\Gamma})\operatorname{vec}(\mathbf{\Xi})\epsilon+o(\epsilon)\\
&=-\eta\operatorname{vec}(\mathbf{U}_\perp\bar{\mathbf{\Sigma}}_\perp\mathbf{V}_\perp^{\mathrm{T}})+\epsilon(\mathbf{I}-\mathbf{G})\operatorname{vec}(\mathbf{\Xi})+o(\epsilon). \qquad (11)
\end{aligned}$$

Finally, from (9) and (11), we obtain

$$\begin{aligned}
&d(\mathbf{A}',\mathbf{A})-d(\mathbf{A}_k',\mathbf{A})\\
&=\frac{\epsilon^2}{2}\operatorname{vec}^{\mathrm{T}}(\mathbf{\Xi})\mathbf{J}_d\operatorname{vec}(\mathbf{\Xi})\\
&\quad+\epsilon\eta\operatorname{vec}^{\mathrm{T}}(\mathbf{U}_\perp\bar{\mathbf{\Sigma}}_\perp\mathbf{V}_\perp^{\mathrm{T}})\operatorname{Hess}_d(\mathbf{I}-\mathbf{G})\operatorname{vec}(\mathbf{\Xi})\\
&\quad-\frac{\eta^2}{2}\operatorname{vec}^{\mathrm{T}}(\mathbf{U}_\perp\bar{\mathbf{\Sigma}}_\perp\mathbf{V}_\perp^{\mathrm{T}})\operatorname{Hess}_d\operatorname{vec}(\mathbf{U}_\perp\bar{\mathbf{\Sigma}}_\perp\mathbf{V}_\perp^{\mathrm{T}})\\
&\quad+o(\epsilon^2+\eta^2). \qquad (12)
\end{aligned}$$

□

In particular, it is natural to require that the low-rank approximation $\hat{\mathbf{A}}_k$ outperforms the original estimation $\hat{\mathbf{A}}$ when the small singular values of $\mathbf{A}$ is much smaller than the noise level, i.e., $d(\mathbf{A}',\mathbf{A})-d(\mathbf{A}_k',\mathbf{A})>0$ if $\eta=o(\epsilon)$. In fact, from Lemma 1, if $\eta=o(\epsilon)$ we have

$$d(\mathbf{A}',\mathbf{A})-d(\mathbf{A}_k',\mathbf{A})=\frac{\epsilon^2}{2}\operatorname{vec}^{\mathrm{T}}(\mathbf{\Xi})\mathbf{J}_d\operatorname{vec}(\mathbf{\Xi})+o(\epsilon^2).$$

Therefore, $\mathbf{J}_d$ is required to be positive semidefinite, and we have the following result.

Proposition 2. *The matrix $\mathbf{J}_d$ is positive semidefinite if and only if*

$$\operatorname{Hess}_d\mathbf{G}=\mathbf{G}\operatorname{Hess}_d, \qquad (13)$$

and in this case we have

$$\begin{aligned}
&d(\mathbf{A}',\mathbf{A})-d(\mathbf{A}_k',\mathbf{A})\\
&=\frac{\epsilon^2}{2}\operatorname{vec}^{\mathrm{T}}(\mathbf{\Xi})\mathbf{G}\operatorname{Hess}_d\mathbf{G}\operatorname{vec}(\mathbf{\Xi})\\
&\quad-\frac{\eta^2}{2}\operatorname{vec}^{\mathrm{T}}(\mathbf{U}_\perp\bar{\mathbf{\Sigma}}_\perp\mathbf{V}_\perp^{\mathrm{T}})\operatorname{Hess}_d\operatorname{vec}(\mathbf{U}_\perp\bar{\mathbf{\Sigma}}_\perp\mathbf{V}_\perp^{\mathrm{T}})\\
&\quad+o(\epsilon^2+\eta^2). \qquad (14)
\end{aligned}$$

Proof. The "if" part of the claim is immediate. Indeed, note that since $\mathbf{G}=\mathbf{G}^2$, it follows from (13) that

$$\begin{aligned}
\mathbf{G}\operatorname{Hess}_d&=\mathbf{G}^2\operatorname{Hess}_d\\
&=\mathbf{G}(\mathbf{G}\operatorname{Hess}_d)=\mathbf{G}(\operatorname{Hess}_d\mathbf{G})=\mathbf{G}\operatorname{Hess}_d\mathbf{G},
\end{aligned}$$

and, similarly,

$$\operatorname{Hess}_d\mathbf{G}=\mathbf{G}\operatorname{Hess}_d\mathbf{G}.$$

Therefore,

$$\begin{aligned}
\mathbf{J}_d&=\operatorname{Hess}_d-(\mathbf{I}-\mathbf{G})\operatorname{Hess}_d(\mathbf{I}-\mathbf{G})\\
&=\mathbf{G}\operatorname{Hess}_d+\operatorname{Hess}_d\mathbf{G}-\mathbf{G}\operatorname{Hess}_d\mathbf{G}\\
&=\mathbf{G}\operatorname{Hess}_d\mathbf{G}, \qquad (15)
\end{aligned}$$

and thus is positive semidefinite.

For the "only if" part of the claim, for all $\boldsymbol{x}\in\mathbb{R}^{m_1m_2}$ and $t>0$, let $\boldsymbol{w}\triangleq\mathbf{G}\boldsymbol{x}-t(\mathbf{I}-\mathbf{G})\operatorname{Hess}_d\mathbf{G}\boldsymbol{x}$. Then, since $\mathbf{J}_d$ is positive semidefinite, we have

$$\begin{aligned}
0&\le\boldsymbol{w}^{\mathrm{T}}\mathbf{J}_d\boldsymbol{w}\\
&=\boldsymbol{w}^{\mathrm{T}}\left[\operatorname{Hess}_d-(\mathbf{I}-\mathbf{G})\operatorname{Hess}_d(\mathbf{I}-\mathbf{G})\right]\boldsymbol{w}\\
&=\boldsymbol{w}^{\mathrm{T}}\left[\mathbf{G}\operatorname{Hess}_d(\mathbf{I}-\mathbf{G})+\operatorname{Hess}_d\mathbf{G}\right]\boldsymbol{w}\\
&=\boldsymbol{w}^{\mathrm{T}}[\operatorname{Hess}_d\mathbf{G}\boldsymbol{x}-t\mathbf{G}\operatorname{Hess}_d(\mathbf{I}-\mathbf{G})\operatorname{Hess}_d\mathbf{G}\boldsymbol{x}]\\
&=\boldsymbol{x}^{\mathrm{T}}\mathbf{G}\operatorname{Hess}_d\mathbf{G}\boldsymbol{x}-2t\boldsymbol{x}^{\mathrm{T}}\mathbf{G}\operatorname{Hess}_d(\mathbf{I}-\mathbf{G})\operatorname{Hess}_d\mathbf{G}\boldsymbol{x}\\
&=\boldsymbol{x}^{\mathrm{T}}\mathbf{G}\operatorname{Hess}_d\mathbf{G}\boldsymbol{x}-2t\|(\mathbf{I}-\mathbf{G})\operatorname{Hess}_d\mathbf{G}\boldsymbol{x}\|^2.
\end{aligned}$$

Note that since t can be arbitrarily large, we must have $\|(\mathbf{I}-\mathbf{G})\operatorname{Hess}_d\mathbf{G}\boldsymbol{x}\|=0$ for all $\boldsymbol{x}$, which implies that $(\mathbf{I}-\mathbf{G})\operatorname{Hess}_d\mathbf{G}=\mathbf{O}$. As a result, we obtain

$$\begin{aligned}
\operatorname{Hess}_d\mathbf{G}=\mathbf{G}\operatorname{Hess}_d\mathbf{G}&=(\mathbf{G}\operatorname{Hess}_d\mathbf{G})^{\mathrm{T}}\\
&=(\operatorname{Hess}_d\mathbf{G})^{\mathrm{T}}=\mathbf{G}\operatorname{Hess}_d.
\end{aligned}$$

Moreover, if $\operatorname{Hess}_d\mathbf{G}=\mathbf{G}\operatorname{Hess}_d$, we have

$$\begin{aligned}
(\mathbf{I}-\mathbf{G})\operatorname{Hess}_d&=\operatorname{Hess}_d-\mathbf{G}\operatorname{Hess}_d\\
&=\operatorname{Hess}_d-\operatorname{Hess}_d\mathbf{G}=\operatorname{Hess}_d(\mathbf{I}-\mathbf{G}). \qquad (16)
\end{aligned}$$

Furthermore, note that since

$$\begin{aligned}
\mathbf{G}(\mathbf{V}_\perp\otimes\mathbf{U}_\perp)&=(\mathbf{V}_\perp\mathbf{V}_\perp^{\mathrm{T}})\otimes(\mathbf{U}_\perp\mathbf{U}_\perp^{\mathrm{T}})(\mathbf{V}_\perp\otimes\mathbf{U}_\perp)\\
&=\mathbf{V}_\perp\otimes\mathbf{U}_\perp,
\end{aligned}$$

we have

$$(\mathbf{I}-\mathbf{G})(\mathbf{V}_\perp\otimes\mathbf{U}_\perp)=(\mathbf{V}_\perp\otimes\mathbf{U}_\perp)-(\mathbf{V}_\perp\otimes\mathbf{U}_\perp)=\mathbf{O}. \qquad (17)$$

Therefore, we obtain

$$\begin{aligned}
&\operatorname{vec}^{\mathrm{T}}(\mathbf{U}_\perp\bar{\mathbf{\Sigma}}_\perp\mathbf{V}_\perp^{\mathrm{T}})\operatorname{Hess}_d(\mathbf{I}-\mathbf{G})\operatorname{vec}(\mathbf{\Xi}) && (18)\\
&=\operatorname{vec}^{\mathrm{T}}(\mathbf{\Xi})(\mathbf{I}-\mathbf{G})\operatorname{Hess}_d\operatorname{vec}(\mathbf{U}_\perp\bar{\mathbf{\Sigma}}_\perp\mathbf{V}_\perp^{\mathrm{T}}) && (19)\\
&=\operatorname{vec}^{\mathrm{T}}(\mathbf{\Xi})\operatorname{Hess}_d(\mathbf{I}-\mathbf{G})(\mathbf{V}_\perp\otimes\mathbf{U}_\perp)\operatorname{vec}(\bar{\mathbf{\Sigma}}) && (20)\\
&=\mathbf{0}, && (21)
\end{aligned}$$

where to obtain the second equality we have used (16), and to obtain the last equality we have used (17).

Finally, from (8), (15) and (21) we obtain (14) as desired. □

As a special case, note that the square error $d(\mathbf{A}',\mathbf{A})=\|\mathbf{A}'-\mathbf{A}\|_{\mathrm{F}}^2$ with $\operatorname{Hess}_d=2\mathbf{I}$ satisfies the condition (13). In general, the metrics $d(\cdot,\cdot)$ satisfying (13) would depend on $\mathbf{A}$ since $\mathbf{G}=(\mathbf{V}_\perp\mathbf{V}_\perp^{\mathrm{T}})\otimes(\mathbf{U}_\perp\mathbf{U}_\perp^{\mathrm{T}})$ depends on $\mathbf{A}$. For the simplicity of exposition, in the remainder of this paper we will restrict our attention to the square error, and the results for general metrics with (13) can be similarly obtained. In particular, we have the following characterization for the square error.

Corollary 3. *We have*

$$\begin{aligned}
&\|\mathbf{A}'-\mathbf{A}\|_{\mathrm{F}}^2-\|\mathbf{A}_k'-\mathbf{A}\|_{\mathrm{F}}^2\\
&=\epsilon^2\|\mathbf{U}_\perp^{\mathrm{T}}\mathbf{\Xi}\mathbf{V}_\perp\|_{\mathrm{F}}^2-\eta^2\sigma_0^2+o(\epsilon^2+\eta^2). && (22)
\end{aligned}$$

In addition, if $\mathbf{A}'$ satisfies $\mathbf{A}'=\tau(\mathbf{Z}')$ for some $\mathbf{Z}'$ with $\|\mathbf{Z}'-\mathbf{Z}\|_{\mathrm{F}}<C\epsilon$, where $C>0$ is a constant, then we have

$$\begin{aligned}
&\|\mathbf{A}'-\mathbf{A}\|_{\mathrm{F}}^2-\|\mathbf{A}_k'-\mathbf{A}\|_{\mathrm{F}}^2\\
&=\left\|\bar{\mathbf{L}}\operatorname{vec}(\mathbf{Z}'-\mathbf{Z})\right\|^2-\eta^2\sigma_0^2+o(\epsilon^2+\eta^2), && (23)
\end{aligned}$$

where $\bar{\mathbf{L}}$ is defined as

$$\bar{\mathbf{L}}\triangleq(\mathbf{V}_\perp^{\mathrm{T}}\otimes\mathbf{U}_\perp^{\mathrm{T}})\mathbf{L}. \quad (24)$$

Proof. When $d(\mathbf{A}',\mathbf{A})=\|\mathbf{A}'-\mathbf{A}\|_{\mathrm{F}}^2$, we have $\operatorname{Hess}_d=2\mathbf{I}$, and it follows from (15) that

$$\mathbf{J}_d=\mathbf{G}\operatorname{Hess}_d\mathbf{G}=2\mathbf{G}^2=2\mathbf{G}.$$

Therefore, from (14) we have

$$\begin{aligned}
&\|\mathbf{A}'-\mathbf{A}\|_{\mathrm{F}}^2-\|\mathbf{A}_k'-\mathbf{A}\|_{\mathrm{F}}^2\\
&=\frac{\epsilon^2}{2}\operatorname{vec}^{\mathrm{T}}(\mathbf{\Xi})\mathbf{G}\operatorname{Hess}_d\mathbf{G}\operatorname{vec}(\mathbf{\Xi})\\
&\quad-\frac{\eta^2}{2}\operatorname{vec}^{\mathrm{T}}(\mathbf{U}_\perp\bar{\mathbf{\Sigma}}_\perp\mathbf{V}_\perp^{\mathrm{T}})\operatorname{Hess}_d\operatorname{vec}(\mathbf{U}_\perp\bar{\mathbf{\Sigma}}_\perp\mathbf{V}_\perp^{\mathrm{T}})\\
&\quad+o(\epsilon^2+\eta^2)\\
&=\epsilon^2\operatorname{vec}^{\mathrm{T}}(\mathbf{\Xi})\mathbf{G}\operatorname{vec}(\mathbf{\Xi})-\eta^2\|\mathbf{U}_\perp\bar{\mathbf{\Sigma}}_\perp\mathbf{V}_\perp^{\mathrm{T}}\|_{\mathrm{F}}^2+o(\epsilon^2+\eta^2)\\
&=\epsilon^2\left\|\mathbf{U}_\perp^{\mathrm{T}}\mathbf{\Xi}\mathbf{V}_\perp\right\|_{\mathrm{F}}^2-\eta^2\sigma_0^2+o(\epsilon^2+\eta^2),
\end{aligned}$$

where to obtain the last equality we have used the facts that

$$\begin{aligned}
\operatorname{vec}^{\mathrm{T}}(\mathbf{\Xi})\mathbf{G}\operatorname{vec}(\mathbf{\Xi})&=\operatorname{vec}^{\mathrm{T}}(\mathbf{\Xi})(\mathbf{V}_\perp\mathbf{V}_\perp^{\mathrm{T}})\otimes(\mathbf{U}_\perp\mathbf{U}_\perp^{\mathrm{T}})\operatorname{vec}(\mathbf{\Xi})\\
&=\operatorname{vec}^{\mathrm{T}}(\mathbf{\Xi})(\mathbf{V}_\perp\otimes\mathbf{U}_\perp)(\mathbf{V}_\perp^{\mathrm{T}}\otimes\mathbf{U}_\perp^{\mathrm{T}})\operatorname{vec}(\mathbf{\Xi})\\
&=\left\|(\mathbf{V}_\perp^{\mathrm{T}}\otimes\mathbf{U}_\perp^{\mathrm{T}})\operatorname{vec}(\mathbf{\Xi})\right\|^2 && (25)\\
&=\left\|\operatorname{vec}\left(\mathbf{U}_\perp\mathbf{\Xi}\mathbf{V}_\perp^{\mathrm{T}}\right)\right\|^2 && (26)\\
&=\left\|\mathbf{U}_\perp^{\mathrm{T}}\mathbf{\Xi}\mathbf{V}_\perp\right\|_{\mathrm{F}}^2 && (27)
\end{aligned}$$

and

$$\|\mathbf{U}_\perp\bar{\mathbf{\Sigma}}_\perp\mathbf{V}_\perp^{\mathrm{T}}\|_{\mathrm{F}}^2=\|\bar{\mathbf{\Sigma}}_\perp\|_{\mathrm{F}}^2=\sigma_0^2.$$

In addition, when $\mathbf{A}'$ satisfies $\mathbf{A}'=\tau(\mathbf{Z}')$ for some $\mathbf{Z}'$ with $\|\mathbf{Z}'-\mathbf{Z}\|_{\mathrm{F}}<C\epsilon$ and $C>0$ being a constant, we have

$$\begin{aligned}
\operatorname{vec}(\mathbf{A}')&=\operatorname{vec}(\tau(\mathbf{Z}+\mathbf{Z}'-\mathbf{Z}))\\
&=\operatorname{vec}(\tau(\mathbf{Z}))+\mathbf{L}\operatorname{vec}(\mathbf{Z}'-\mathbf{Z})+o(\epsilon)\\
&=\operatorname{vec}(\mathbf{A})+\mathbf{L}\operatorname{vec}(\mathbf{Z}'-\mathbf{Z})+o(\epsilon).
\end{aligned}$$

Then, from (5) we have

$$\epsilon\operatorname{vec}(\mathbf{\Xi})=\mathbf{L}\operatorname{vec}(\mathbf{Z}'-\mathbf{Z})+o(\epsilon).$$

Therefore, from (25) and (27) we have

$$\begin{aligned}
\epsilon^2\left\|\mathbf{U}_\perp^{\mathrm{T}}\mathbf{\Xi}\mathbf{V}_\perp\right\|_{\mathrm{F}}^2&=\left\|(\mathbf{V}_\perp^{\mathrm{T}}\otimes\mathbf{U}_\perp^{\mathrm{T}})(\epsilon\operatorname{vec}(\mathbf{\Xi}))\right\|^2\\
&=\left\|(\mathbf{V}_\perp^{\mathrm{T}}\otimes\mathbf{U}_\perp^{\mathrm{T}})\mathbf{L}\operatorname{vec}(\mathbf{Z}'-\mathbf{Z})\right\|^2+o(\epsilon^2)\\
&=\left\|\bar{\mathbf{L}}\operatorname{vec}(\mathbf{Z}'-\mathbf{Z})\right\|^2+o(\epsilon^2),
\end{aligned}$$

and thus (23) follows. □

As an immediate consequence of Corollary 3, there exist constants $\epsilon_0>0$ and $\eta_0>0$, such that for all $\mathbf{Z}'$ with $\|\mathbf{Z}'-\mathbf{Z}\|_{\mathrm{F}}<\epsilon_0$, we have

$$\begin{aligned}
&\frac{2}{3}\left\|\bar{\mathbf{L}}\operatorname{vec}(\mathbf{Z}'-\mathbf{Z})\right\|^2-\frac{4\eta^2}{3}\sigma_0^2\\
&<\|\mathbf{A}'-\mathbf{A}\|_{\mathrm{F}}^2-\|\mathbf{A}_k'-\mathbf{A}\|_{\mathrm{F}}^2\\
&<\frac{4}{3}\left\|\bar{\mathbf{L}}\operatorname{vec}(\mathbf{Z}'-\mathbf{Z})\right\|^2-\frac{2\eta^2}{3}\sigma_0^2, && (28)
\end{aligned}$$

which will be useful later in the characterization of (7).

IV. Main Results

In this section, we investigate the sample complexity of estimating small singular modes of $\mathbf{A}$, via the characterization for the probability p_n as defined in (7). In particular, we are interested in the regime of n such that $p_n>1-\delta$ or $p_n<\delta$ for some small δ, corresponding to the case where the small singular modes dominate the sample noise, or conversely.

To begin, note that for η_0 and ϵ_0 as defined in (28), and $\eta\in(0,\eta_0)$, we have

$$\begin{aligned}
p_n&=\mathbb{P}_n\left\{\|\hat{\mathbf{A}}-\mathbf{A}\|_{\mathrm{F}}^2>\|\hat{\mathbf{A}}_k-\mathbf{A}\|_{\mathrm{F}}^2\right\}\\
&\geq\mathbb{P}_n\left\{\|\hat{\mathbf{A}}-\mathbf{A}\|_{\mathrm{F}}^2>\|\hat{\mathbf{A}}_k-\mathbf{A}\|_{\mathrm{F}}^2,\ \|\hat{\mathbf{Z}}-\mathbf{Z}\|_{\mathrm{F}}^2\leq\epsilon_0^2\right\}\\
&\geq\mathbb{P}_n\left\{\left\|\bar{\mathbf{L}}\operatorname{vec}(\hat{\mathbf{Z}}-\mathbf{Z})\right\|^2\geq2\eta^2\sigma_0^2,\ \|\hat{\mathbf{Z}}-\mathbf{Z}\|_{\mathrm{F}}^2\leq\epsilon_0^2\right\}\\
&\geq1-\mathbb{P}_n\left\{\left\|\bar{\mathbf{L}}\operatorname{vec}(\hat{\mathbf{Z}}-\mathbf{Z})\right\|^2<2\eta^2\sigma_0^2\right\}\\
&\quad-\mathbb{P}_n\left\{\|\hat{\mathbf{Z}}-\mathbf{Z}\|_{\mathrm{F}}^2>\epsilon_0^2\right\}, && (29)
\end{aligned}$$

where $\bar{\mathbf{L}}$ is as defined in (24), where to obtain the second equality we have used (28), and where to obtain the last equality

we have used the fact that for two events E_1 and E_2, we have $\mathbb{P}\{E_1 \cap E_2\} = 1 - \mathbb{P}\{E_1' \cup E_2'\} \geq 1 - \mathbb{P}\{E_1'\} - \mathbb{P}\{E_2'\}$, with E_1' and E_2' denoting the complementary events of E_1 and E_2, respectively.

Similarly, we have, for all $\eta \in (0, \eta_0)$,

$$\begin{aligned}
p_n &= \mathbb{P}_n\left\{\|\hat{\mathbf{A}} - \mathbf{A}\|_{\mathrm{F}}^2 > \|\hat{\mathbf{A}}_k - \mathbf{A}\|_{\mathrm{F}}^2\right\} \\
&= \mathbb{P}_n\left\{\left\|\hat{\mathbf{A}} - \mathbf{A}\right\|_{\mathrm{F}}^2 > \left\|\hat{\mathbf{A}}_k - \mathbf{A}\right\|_{\mathrm{F}}^2, \left\|\hat{\mathbf{Z}} - \mathbf{Z}\right\|_{\mathrm{F}}^2 \leq \epsilon_0^2\right\} \\
&\quad + \mathbb{P}_n\left\{\|\hat{\mathbf{A}} - \mathbf{A}\|_{\mathrm{F}}^2 > \|\hat{\mathbf{A}}_k - \mathbf{A}\|_{\mathrm{F}}^2, \left\|\hat{\mathbf{Z}} - \mathbf{Z}\right\|_{\mathrm{F}}^2 > \epsilon_0^2\right\} \\
&\leq \mathbb{P}_n\left\{\left\|\bar{\mathbf{L}}\operatorname{vec}(\hat{\mathbf{Z}} - \mathbf{Z})\right\|^2 > \frac{1}{2}\eta^2\sigma_0^2, \left\|\hat{\mathbf{Z}} - \mathbf{Z}\right\|_{\mathrm{F}}^2 \leq \epsilon_0^2\right\} \\
&\quad + \mathbb{P}_n\left\{\left\|\hat{\mathbf{Z}} - \mathbf{Z}\right\|_{\mathrm{F}}^2 > \epsilon_0^2\right\} \\
&\leq \mathbb{P}_n\left\{\left\|\bar{\mathbf{L}}\operatorname{vec}(\hat{\mathbf{Z}} - \mathbf{Z})\right\|^2 > \frac{1}{2}\eta^2\sigma_0^2\right\} \\
&\quad + \mathbb{P}_n\left\{\left\|\hat{\mathbf{Z}} - \mathbf{Z}\right\|_{\mathrm{F}}^2 > \epsilon_0^2\right\}, \qquad (30)
\end{aligned}$$

where to obtain the first inequality we have again used (28).

Therefore, to obtain $p_n > 1 - \delta$ it suffices to choose n such that

$$\mathbb{P}_n\left\{\left\|\bar{\mathbf{L}}\operatorname{vec}(\hat{\mathbf{Z}} - \mathbf{Z})\right\|^2 < 2\eta^2\sigma_0^2\right\} < \frac{\delta}{2}, \qquad (31)$$

and

$$\mathbb{P}_n\left\{\left\|\hat{\mathbf{Z}} - \mathbf{Z}\right\|_{\mathrm{F}}^2 > \epsilon_0^2\right\} < \frac{\delta}{2}. \qquad (32)$$

Similarly, to obtain $p_n < \delta$ it suffices to choose n to satisfy (32) and

$$\mathbb{P}_n\left\{\left\|\bar{\mathbf{L}}\operatorname{vec}(\hat{\mathbf{Z}} - \mathbf{Z})\right\|^2 > \frac{1}{2}\eta^2\sigma_0^2\right\} < \frac{\delta}{2}. \qquad (33)$$

As a result, it suffices to establish (31)–(33), which characterize the concentration and anti-concentration properties of the sample mean $\hat{\mathbf{Z}} = \frac{1}{n}\sum_{\ell=1}^n \mathbf{Z}^{(\ell)}$. To this end, we introduce the following useful lemma.

Lemma 4. *Suppose $\boldsymbol{\zeta}^{(1)}, \ldots, \boldsymbol{\zeta}^{(n)} \in \mathbb{R}^m$ are i.i.d random vectors, with zero-mean, covariance matrix $\boldsymbol{\Lambda}$ and $\mathbb{E}\left[\left\|\boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{\zeta}^{(1)}\right\|^3\right] < \infty$. Let $\hat{\boldsymbol{\zeta}} \triangleq \frac{1}{n}\sum_{\ell=1}^n \boldsymbol{\zeta}^{(\ell)}$, then for all $t > 0$, we have*

$$\mathbb{P}_n\left\{\hat{\boldsymbol{\zeta}}^{\mathrm{T}}\hat{\boldsymbol{\zeta}} < t\right\} \geq 1 - \frac{\operatorname{tr}(\boldsymbol{\Lambda})}{nt}, \qquad (34)$$

and

$$\mathbb{P}_n\left\{\hat{\boldsymbol{\zeta}}^{\mathrm{T}}\hat{\boldsymbol{\zeta}} > t\right\} \geq 1 - \sqrt{\frac{nt}{\lambda}} - \frac{C_0}{\sqrt{n}}, \qquad (35)$$

where $C_0 > 0$ is a constant independent of $\boldsymbol{\Lambda}$ and t, and λ denotes the minimum positive diagonal entry of $\boldsymbol{\Lambda}$. In particular, if $\left\|\boldsymbol{\zeta}^{(\ell)}\right\| \leq C$ almost surely for some constant $C > 0$, then for all $t > 0$, we have

$$\mathbb{P}_n\left\{\hat{\boldsymbol{\zeta}}^{\mathrm{T}}\hat{\boldsymbol{\zeta}} < t\right\} \geq 1 - (m+1)\exp\left(-\frac{nt}{4C^2}\right). \qquad (36)$$

Proof. For the first part, note that since $\operatorname{tr}\left(\hat{\boldsymbol{\zeta}}^{\mathrm{T}}\hat{\boldsymbol{\zeta}}\right) = \operatorname{tr}\left(\hat{\boldsymbol{\zeta}}\hat{\boldsymbol{\zeta}}^{\mathrm{T}}\right)$ and $\operatorname{cov}\left(\hat{\boldsymbol{\zeta}}\right) = \frac{1}{n}\boldsymbol{\Lambda}$, we have

$$\begin{aligned}
\mathbb{E}\left[\hat{\boldsymbol{\zeta}}^{\mathrm{T}}\hat{\boldsymbol{\zeta}}\right] = \mathbb{E}\left[\operatorname{tr}(\hat{\boldsymbol{\zeta}}^{\mathrm{T}}\hat{\boldsymbol{\zeta}})\right] &= \mathbb{E}\left[\operatorname{tr}(\hat{\boldsymbol{\zeta}}\hat{\boldsymbol{\zeta}}^{\mathrm{T}})\right] \\
&= \operatorname{tr}\left(\mathbb{E}\left[\hat{\boldsymbol{\zeta}}\hat{\boldsymbol{\zeta}}^{\mathrm{T}}\right]\right) \\
&= \operatorname{tr}\left(\operatorname{cov}(\hat{\boldsymbol{\zeta}})\right) = \frac{\operatorname{tr}(\boldsymbol{\Lambda})}{n}.
\end{aligned}$$

Then, it follows from Markov's inequality that

$$\mathbb{P}_n\left\{\hat{\boldsymbol{\zeta}}^{\mathrm{T}}\hat{\boldsymbol{\zeta}} \geq t\right\} \leq \frac{\mathbb{E}\left[\hat{\boldsymbol{\zeta}}^{\mathrm{T}}\hat{\boldsymbol{\zeta}}\right]}{t} = \frac{\operatorname{tr}(\boldsymbol{\Lambda})}{nt},$$

which implies (34).

For the second part, to obtain (35), note that from Berry-Esseen inequality [11] we have

$$\left|\mathbb{P}_n\left\{\hat{\boldsymbol{\zeta}}^{\mathrm{T}}\hat{\boldsymbol{\zeta}} > t\right\} - \mathbb{P}_n\left\{\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{\alpha} > t\right\}\right| \leq C_0 n^{-\frac{1}{2}}$$

where we have defined the random vector $\boldsymbol{\alpha} \sim \mathcal{N}(\mathbf{0}, \frac{1}{n}\boldsymbol{\Lambda})$, and where $C_0 > 0$ is a constant independent of $n, t, \hat{\boldsymbol{\zeta}}$.

Note that

$$\begin{aligned}
\mathbb{P}_n\left\{\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{\alpha} > t\right\} &= \mathbb{P}_n\left\{\bar{\boldsymbol{\alpha}}^{\mathrm{T}}\bar{\boldsymbol{\alpha}} > nt\right\} \\
&\geq \mathbb{P}_n\left\{\bar{\alpha}_i^2 > nt\right\} \qquad (37) \\
&= 2Q\left(\sqrt{\frac{nt}{\lambda}}\right) \qquad (38) \\
&> 1 - \sqrt{\frac{nt}{\lambda}}, \qquad (39)
\end{aligned}$$

where $\bar{\alpha}_i$ is the i-th entry of $\bar{\boldsymbol{\alpha}} \triangleq \sqrt{n}\boldsymbol{\alpha}$ with its variance being the minimum positive diagonal entry of $\boldsymbol{\Lambda}$, and where the Q-function $Q(x) \triangleq \frac{1}{\sqrt{2\pi}}\int_x^\infty \exp(-\frac{1}{2}u^2)du$ represents the tail distribution of the standard normal distribution. We obtain (37) from the fact that $n\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{\alpha} \geq \bar{\alpha}_i^2$, and we obtain (39) from the fact that $Q(x) \geq \frac{1}{2} - \frac{x}{\sqrt{2\pi}} \geq \frac{1}{2} - \frac{x}{2}$.

Therefore, using (39), from Berry-Essen inequality we have

$$\begin{aligned}
\mathbb{P}_n\left\{\hat{\boldsymbol{\zeta}}^{\mathrm{T}}\hat{\boldsymbol{\zeta}} > t\right\} &\geq \mathbb{P}_n\left\{\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{\alpha} > t\right\} - C_0 n^{-\frac{1}{2}} \\
&> 1 - \sqrt{\frac{nt}{\lambda}} - C_0 n^{-\frac{1}{2}}.
\end{aligned}$$

As for the last part, it follows from Chernoff bound for random matrices [12] that

$$\begin{aligned}
\mathbb{P}_n\left\{\hat{\boldsymbol{\zeta}}^{\mathrm{T}}\hat{\boldsymbol{\zeta}} \geq t\right\} &= \mathbb{P}_n\left\{\left\|\frac{1}{n}\sum_{\ell=1}^n \boldsymbol{\zeta}^{(\ell)}\right\|^2 \geq t\right\} \\
&\leq (m+1)\exp\left(-\frac{3tn}{8C^2}\right) \\
&\leq (m+1)\exp\left(-\frac{tn}{4C^2}\right),
\end{aligned}$$

which implies (36). □

Using Lemma 4, we have the following characterizations for the probability p_n, where we have taken the mild assumption that the noise $\mathbf{Z}^{(\ell)}$ has a finite third moment.

Theorem 5. *Suppose that* $\mathbb{E}\left[\|\mathbf{R}^{-\frac{1}{2}}\operatorname{vec}(\mathbf{Z}^{(1)}-\mathbf{Z})\|_{\mathrm{F}}^{3}\right]<\infty$, *then for all* $\delta\in(0,1)$, *there exists* $\eta'>0$ *such that for all* $\eta\in(0,\eta')$, *we have*

1) $p_n>1-\delta$, *if* $\frac{\delta^2 R_{\mathrm{m}}}{128\eta^2\sigma_0^2}\le n\le\frac{\delta^2 R_{\mathrm{m}}}{32\eta^2\sigma_0^2}$,

2) $p_n<\delta$, *if* $n\ge\frac{4\operatorname{tr}(\bar{\mathbf{L}}\mathbf{R}\bar{\mathbf{L}}^{\mathrm{T}})}{\eta^2\sigma_0^2\delta}$,

where R_{m} *is the minimum non-zero diagonal entry of* $\bar{\mathbf{L}}\mathbf{R}\bar{\mathbf{L}}^{\mathrm{T}}$, *and* $\bar{\mathbf{L}}$ *is as defined in* (24). *In particular, if* $\|\mathbf{Z}^{(\ell)}-\mathbf{Z}\|_{\mathrm{F}}\le C$ *almost surely for some* $C>0$, *we have* $p_n<\delta$ *for* $n>\frac{8C^2\|\mathbf{L}\|_{\mathrm{s}}^2}{\eta^2\sigma_0^2}\ln\left(\frac{4m_1m_2}{\delta}\right)$ *with* $\|\cdot\|_{\mathrm{s}}$ *denoting the spectral norm.*

Proof. To prove that $p_n>1-\delta$, for $\eta\in(0,\eta_0)$, it suffices to choose n satisfying (31)–(32). To this end, note that from (35), we have

$$\mathbb{P}_n\left\{\left\|\bar{\mathbf{L}}\operatorname{vec}(\hat{\mathbf{Z}}-\mathbf{Z})\right\|^2<2\eta^2\sigma_0^2\right\}\le\eta\sigma_0\sqrt{\frac{2n}{R_{\mathrm{m}}}}+\frac{C_0}{\sqrt{n}}.$$

Therefore, from (31) it suffices to have

$$\eta\sigma_0\sqrt{\frac{2n}{R_{\mathrm{m}}}}+\frac{C_0}{\sqrt{n}}\le\frac{\delta}{2},\tag{40}$$

which is equivalent to

$$\gamma(\eta,\delta)\triangleq\delta^2-16\eta\sigma_0C_0\sqrt{\frac{2}{R_{\mathrm{m}}}}\ge0,$$

and

$$\frac{R_{\mathrm{m}}\left(\delta-\sqrt{\gamma(\eta,\delta)}\right)^2}{32\eta^2\sigma_0^2}\le n\le\frac{R_{\mathrm{m}}\left(\delta+\sqrt{\gamma(\eta,\delta)}\right)^2}{32\eta^2\sigma_0^2},$$

where $\gamma(\eta,\delta)$ is the discriminant for (40).

Note that when $\eta<\eta_1\triangleq\frac{\delta^2\sqrt{R_{\mathrm{m}}}}{64\sigma_0C_0}$, we have $\gamma(\eta,\delta)>\frac{\delta^2}{4}$, which implies that

$$\left(\delta-\sqrt{\gamma(\eta,\delta)}\right)^2<\frac{\delta^2}{4}\quad\text{and}\quad\left(\delta+\sqrt{\gamma(\eta,\delta)}\right)^2>\delta^2.$$

Thus, when $\eta<\min\{\eta_0,\eta_1\}$, to obtain (31) it suffices to choose n such that $n_0<n<4n_0$ with $n_0=\frac{\delta^2R_{\mathrm{m}}}{128\eta^2\sigma_0^2}$.

Also, it follows from (34) that if $n>\frac{2\operatorname{tr}(\mathbf{R})}{\delta\epsilon_0^2}$, we have

$$\begin{aligned}\mathbb{P}_n\left\{\|\hat{\mathbf{Z}}-\mathbf{Z}\|_{\mathrm{F}}^2>\epsilon_0^2\right\}&=\mathbb{P}_n\left\{\left\|\operatorname{vec}(\hat{\mathbf{Z}}-\mathbf{Z})\right\|^2>\epsilon_0^2\right\}\\&\le\frac{\operatorname{tr}(\mathbf{R})}{n\epsilon_0^2}\le\frac{\delta}{2},\end{aligned}\tag{41}$$

which implies (33).

In particular, when $\eta<\eta_2\triangleq\frac{\delta\epsilon_0}{16\sigma_0}\sqrt{\frac{\delta R_{\mathrm{m}}}{\operatorname{tr}(\mathbf{R})}}$, we have $\frac{2\operatorname{tr}(\mathbf{R})}{\delta\epsilon_0^2}<n_0$. Therefore, when $\eta<\min\{\eta_0,\eta_1,\eta_2\}$, if $n>n_0$ we also have $n>\frac{2\operatorname{tr}(\mathbf{R})}{\delta\epsilon_0^2}$ and thus (41) holds.

Similarly, to prove that $p_n<\delta$, for $\eta\in(0,\eta_0)$, it suffices to choose n satisfying (32) and (33). For this purpose, note that from (34), we have

$$\mathbb{P}_n\left\{\left\|\bar{\mathbf{L}}\operatorname{vec}(\hat{\mathbf{Z}}-\mathbf{Z})\right\|^2>\frac{1}{2}\eta^2\sigma_0^2\right\}\le\frac{2\operatorname{tr}(\bar{\mathbf{L}}\mathbf{R}\bar{\mathbf{L}}^{\mathrm{T}})}{n\eta^2\sigma_0^2}.$$

Therefore, from (33) it suffices to have

$$\frac{2\operatorname{tr}(\bar{\mathbf{L}}\mathbf{R}\bar{\mathbf{L}}^{\mathrm{T}})}{n\eta^2\sigma_0^2}<\frac{\delta}{2},$$

which is equivalent to $n>n_2\triangleq\frac{4\operatorname{tr}(\bar{\mathbf{L}}\mathbf{R}\bar{\mathbf{L}}^{\mathrm{T}})}{\eta^2\sigma_0^2\delta}$.

Moreover, from (41) we know that it suffices to choose $n>\frac{2\operatorname{tr}(\mathbf{R})}{\delta\epsilon_0^2}$ to obtain (32). In fact, when $\eta<\eta_3\triangleq\frac{2\operatorname{tr}(\bar{\mathbf{L}}\mathbf{R}\bar{\mathbf{L}}^{\mathrm{T}})\epsilon_0^2}{\operatorname{tr}(\mathbf{R})\sigma_0^2}$, it can be verified that $n_2>\frac{2\operatorname{tr}(\mathbf{R})}{\delta\epsilon_0^2}$, and thus it suffices to have $n>n_2$. Therefore, η' can be chosen as $\eta'=\min\{\eta_0,\eta_1,\eta_2,\eta_3\}$.

In addition, when $\|\mathbf{Z}^{(\ell)}-\mathbf{Z}\|_{\mathrm{F}}<C$ for some constant $C>0$ almost surely, we have

$$\begin{aligned}\left\|\bar{\mathbf{L}}\operatorname{vec}(\hat{\mathbf{Z}}-\mathbf{Z})\right\|&\le\|\bar{\mathbf{L}}\|_{\mathrm{s}}\|\operatorname{vec}(\mathbf{Z}^{(\ell)}-\mathbf{Z})\|\\&\le\|\mathbf{L}\|_{\mathrm{s}}\|\mathbf{Z}^{(\ell)}-\mathbf{Z}\|_{\mathrm{F}}<C\|\mathbf{L}\|_{\mathrm{s}}\end{aligned}$$

almost surely. Therefore, it follows from (36) that, if $n\ge n_2'\triangleq\frac{8C^2\|\mathbf{L}\|_{\mathrm{s}}^2}{\eta^2\sigma_0^2}\ln\left(\frac{4m_1m_2}{\delta}\right)$ we have

$$\begin{aligned}&\mathbb{P}_n\left\{\left\|\bar{\mathbf{L}}\operatorname{vec}(\hat{\mathbf{Z}}-\mathbf{Z})\right\|^2>\frac{1}{2}\eta^2\sigma_0^2\right\}\\&\le(m_1m_2+1)\exp\left(-\frac{n\eta^2\sigma_0^2}{8C^2\|\bar{\mathbf{L}}\|_{\mathrm{s}}^2}\right)\le\frac{\delta}{2},\end{aligned}\tag{42}$$

and, if $n\ge n_0'\triangleq\frac{4C^2}{\epsilon_0^2}\ln\left(\frac{4m_1m_2}{\delta}\right)$, we have

$$\mathbb{P}_n\left\{\left\|\hat{\mathbf{Z}}-\mathbf{Z}\right\|_{\mathrm{F}}^2>\epsilon_0^2\right\}\le(m_1m_2+1)\exp\left(-\frac{n\epsilon_0^2}{4C^2}\right)\le\frac{\delta}{2}.$$

In particular, when $\eta<\frac{\epsilon_0\|\bar{\mathbf{L}}\|_{\mathrm{s}}}{\sigma_0}$, it can be verified that $n_2'>n_0'$. Therefore, when $\eta<\min\left\{\eta_0,\frac{\epsilon_0\|\bar{\mathbf{L}}\|_{\mathrm{s}}}{\sigma_0}\right\}$, to obtain $p_n<\delta$ it suffices to choose n such that $n>n_2'=\frac{8C^2\|\mathbf{L}\|_{\mathrm{s}}^2}{\eta^2\sigma_0^2}\ln\left(\frac{4m_1m_2}{\delta}\right)$. □

Remark 6. *From Theorem 5, for given* $\delta\in(0,1)$, *the regimes of* n *for both* $p_n>1-\delta$ *and* $p_n<\delta$ *satisfy* $n=O(\eta^{-2})$, *when* η *is sufficiently small. Therefore, the number of samples required to estimate the small singular modes, which lies in between these two regimes, is also* $O(\eta^{-2})$.

In particular, when the noise $\mathbf{Z}^{(\ell)}-\mathbf{Z}$ follows a multivariate Gaussian distribution, we have the following refined characterization for n.

Theorem 7. *Suppose that* $\operatorname{vec}(\mathbf{Z}^{(\ell)}-\mathbf{Z})\sim\mathcal{N}(\mathbf{0},\mathbf{R})$, *for all* $\delta\in(0,1)$, *there exists* $\bar{\eta}_0>0$ *such that for all* $\eta\in(0,\bar{\eta}_0)$, *we have*

1) $p_n>1-\delta$, *if* $\frac{2m_1m_2\|\mathbf{R}\|_{\mathrm{s}}}{\epsilon_0^2}\ln\left(\frac{2m_1m_2}{\delta}\right)\le n\le\frac{\delta^2R_{\mathrm{m}}}{8\eta^2\sigma_0^2}$,

2) $p_n<\delta$, *if* $n\ge\frac{4m_1m_2\|\bar{\mathbf{L}}\mathbf{R}^{\frac{1}{2}}\|_{\mathrm{s}}^2}{\eta^2\sigma_0^2}\ln\left(\frac{2m_1m_2}{\delta}\right)$,

where R_{m} *is as defined in Theorem 5.*

Proof. To prove that $p_n < \delta$, for $\eta \in (0, \eta_0)$, it suffices to choose n satisfying (32) and (33). Note that

$$\begin{aligned}
&\mathbb{P}_n\left\{\left\|\bar{\mathbf{L}}\operatorname{vec}(\hat{\mathbf{Z}}-\mathbf{Z})\right\|^2 > t\right\} \\
&\le \mathbb{P}\left\{\left\|\bar{\mathbf{L}}\mathbf{R}^{\frac{1}{2}}\right\|_{\mathrm{s}}^2 \left\|\mathbf{R}^{-\frac{1}{2}}\operatorname{vec}(\hat{\mathbf{Z}}-\mathbf{Z})\right\|^2 > t\right\} && (43)\\
&= \mathbb{P}\left\{\left\|\sqrt{n}\mathbf{R}^{-\frac{1}{2}}\operatorname{vec}(\hat{\mathbf{Z}}-\mathbf{Z})\right\|^2 > \frac{nt}{\|\bar{\mathbf{L}}\mathbf{R}^{\frac{1}{2}}\|_{\mathrm{s}}^2}\right\} && (44)\\
&\le m_1 m_2 \exp\left(-\frac{nt}{2m_1 m_2\|\bar{\mathbf{L}}\mathbf{R}^{\frac{1}{2}}\|_{\mathrm{s}}^2}\right), && (45)
\end{aligned}$$

where to obtain the first inequality we have used the submultiplicativity of the spectral norm, and to obtain the second inequality we have used the fact that for an m-dimensional standard normal vector $\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_m)$ we have

$$\mathbb{P}\left\{\boldsymbol{z}^{\mathrm{T}}\boldsymbol{z} > t\right\} < m\exp\left(-\frac{t}{2m}\right), \quad t > 0. \tag{46}$$

In fact, note that

$$\begin{aligned}
\mathbb{P}\left\{\boldsymbol{z}^{\mathrm{T}}\boldsymbol{z} \le t\right\} &\ge \mathbb{P}\left\{\max_{1\le i\le m} z_i^2 < \frac{t}{m}\right\} \\
&= \prod_{i=1}^{m}\mathbb{P}\left\{|z_i| < \sqrt{\frac{t}{m}}\right\} && (47)\\
&= \prod_{i=1}^{m}\left[1 - 2Q\left(\sqrt{\frac{t}{m}}\right)\right] && (48)\\
&\ge \prod_{i=1}^{m}\left[1 - \exp\left(-\frac{t}{2m}\right)\right] && (49)\\
&\ge 1 - m\exp\left(-\frac{t}{2m}\right), && (50)
\end{aligned}$$

where to obtain (47) we have used the fact that $z_1, \ldots, z_m$ are independent, to obtain (49) we have used the upper bound for Q-function that $Q(x) \le \frac{1}{2}\exp\left(-\frac{x^2}{2}\right)$ for $x > 0$ (see, e.g., [13, Eq. (2), Eq. (5)]), and where to obtain (50) we have used Bernoulli inequality.

Then, it follows from (45) that

$$\begin{aligned}
&\mathbb{P}_n\left\{\left\|\bar{\mathbf{L}}\operatorname{vec}(\hat{\mathbf{Z}}-\mathbf{Z})\right\|^2 > \frac{1}{2}\eta^2\sigma_0^2\right\} \\
&\le m_1 m_2 \exp\left(-\frac{n\eta^2\sigma_0^2}{4m_1 m_2\|\bar{\mathbf{L}}\mathbf{R}^{\frac{1}{2}}\|_{\mathrm{s}}^2}\right). && (51)
\end{aligned}$$

Therefore, from (33), it suffices to have

$$m_1 m_2 \exp\left(-\frac{n\eta^2\sigma_0^2}{4m_1 m_2\|\bar{\mathbf{L}}\mathbf{R}^{\frac{1}{2}}\|_{\mathrm{s}}^2}\right) < \frac{\delta}{2}, \tag{52}$$

which is equivalent to

$$n > \bar{n}_2 \triangleq \frac{4m_1 m_2\|\bar{\mathbf{L}}\mathbf{R}^{\frac{1}{2}}\|_{\mathrm{s}}^2}{\eta^2\sigma_0^2}\ln\left(\frac{2m_1 m_2}{\delta}\right).$$

Similarly, from (45), if $n \ge \frac{2m_1 m_2\|\mathbf{R}\|_{\mathrm{s}}}{\epsilon_0^2}\ln\left(\frac{2m_1 m_2}{\delta}\right)$, we have

$$\begin{aligned}
&\mathbb{P}_n\left\{\|\hat{\mathbf{Z}}-\mathbf{Z}\|_{\mathrm{F}}^2 > \epsilon_0^2\right\} \\
&= \mathbb{P}_n\left\{\left\|\operatorname{vec}(\hat{\mathbf{Z}}-\mathbf{Z})\right\|^2 > \epsilon_0^2\right\} \\
&\le m_1 m_2\exp\left(-\frac{n\epsilon_0^2}{2m_1 m_2\|\mathbf{R}\|_{\mathrm{s}}}\right) < \frac{\delta}{2}, && (53)
\end{aligned}$$

and thus (32) holds.

Note that when $\eta < \bar{\eta}_1 \triangleq \frac{\epsilon_0\|\bar{\mathbf{L}}\mathbf{R}^{\frac{1}{2}}\|_{\mathrm{s}}}{\sigma_0\sqrt{\|\mathbf{R}\|_{\mathrm{s}}}}$, we have

$$\bar{n}_2 > \frac{2m_1 m_2\|\mathbf{R}\|_{\mathrm{s}}}{\epsilon_0^2}\ln\left(\frac{2m_1 m_2}{\delta}\right). \tag{54}$$

Therefore, it follows from (51)–(54) that when $\eta < \min\{\eta_0, \bar{\eta}_1\}$, to obtain $p_n < \delta$ it suffices to have $n > \bar{n}_2$.

In addition, to prove that $p_n > 1 - \delta$, for $\eta \in (0, \eta_0)$, it suffices to choose n satisfying (31)–(32). Similarly as (39), we have

$$\mathbb{P}_n\left\{\left\|\bar{\mathbf{L}}\operatorname{vec}(\hat{\mathbf{Z}}-\mathbf{Z})\right\|^2 < 2\eta^2\sigma_0^2\right\} \le \eta\sigma_0\sqrt{\frac{2n}{R_{\mathrm{m}}}}.$$

Therefore, from (32), it suffices to let

$$\eta\sigma_0\sqrt{\frac{2n}{R_{\mathrm{m}}}} \le \frac{\delta}{2}, \tag{55}$$

which is equivalent to $n < \bar{n}_1 \triangleq \frac{\delta^2 R_{\mathrm{m}}}{8\eta^2\sigma_0^2}$.

Moreover, from (45), if $n > \bar{n}_0 \triangleq \frac{2m_1 m_2\|\mathbf{R}\|_{\mathrm{s}}}{\epsilon_0^2}\ln\left(\frac{2m_1 m_2}{\delta}\right)$ we have

$$\begin{aligned}
&\mathbb{P}_n\left\{\left\|\operatorname{vec}(\hat{\mathbf{Z}}-\mathbf{Z})\right\|^2 > \epsilon_0^2\right\} \\
&\le m_1 m_2\exp\left(-\frac{n\epsilon_0^2}{2m_1 m_2\|\mathbf{R}\|_{\mathrm{s}}}\right) \le \frac{\delta}{2}, && (56)
\end{aligned}$$

and thus (31) holds.

Also, if $\eta < \bar{\eta}_2 \triangleq \frac{\delta\epsilon_0}{4\sigma_0}\left(\frac{m_1 m_2\|\mathbf{R}\|_{\mathrm{s}}}{R_{\mathrm{m}}}\ln\left(\frac{2m_1 m_2}{\delta}\right)\right)^{-\frac{1}{2}}$, it can be verified that $\bar{n}_1 > \bar{n}_0$. Therefore, when $\eta < \min\{\eta_0, \bar{\eta}_2\}$, to obtain $p_n > 1-\delta$ it suffices to choose n such that $\bar{n}_0 < n < \bar{n}_1$. Finally, the $\bar{\eta}_0$ can be chosen as $\bar{\eta}_0 = \min\{\eta_0, \bar{\eta}_1, \bar{\eta}_2\}$.

□

V. Applications

The above theoretical characterizations can be applied to the channel state estimation problem and the HGR maximal correlation problem, from which we can establish the criteria in whether or not to adopt the low-rank approximation method given the sample size.

A. Channel State Estimation

In the MIMO system with an array of m_2 transmit antennas and an array of m_1 receive antennas, each transmission can be modeled as

$$\boldsymbol{y} = \mathbf{H}\boldsymbol{x} + \boldsymbol{n},$$

where $\mathbf{H} \in \mathbb{C}^{m_1 \times m_2}$ represents the channel matrix, and $\boldsymbol{y}$, $\boldsymbol{x}$, $\boldsymbol{n}$ represent the receive, transmit, and noise vector, respectively. In particular, due to the large-scale of the transmit antenna array with small spacing between antennas, the signals received from massive antennas would have very high correlation [6] and thus $\mathbf{H}$ exhibits a low-rank structure.

In practice, the true channel state $\mathbf{H}$ is unknown and estimated by n i.i.d. samples, $\mathbf{H}^{(1)}, \ldots, \mathbf{H}^{(n)}$, with $\operatorname{vec}(\mathbf{H}^{(i)}) \sim \mathcal{CN}(\operatorname{vec}(\mathbf{H}), \mathbf{R})$, where the effect from noise and interference is modeled by circularly symmetric complex Gaussian random variables [14] with $\mathbf{R}$ being the covariance matrix. Then, based on these samples, the channel state $\mathbf{H}$ can be estimated by computing the sample mean

$$\hat{\mathbf{H}} = \frac{1}{n}\sum_{\ell=1}^{n} \mathbf{H}^{(\ell)}$$

or its rank-k approximation $\hat{\mathbf{H}}_k$. Note that this is a special case of our model, with the correspondences $\mathbf{A} = \mathbf{H}$, $\hat{\mathbf{A}} = \hat{\mathbf{H}} = \hat{\mathbf{Z}}$, and τ being the identity map. Then, we have the following result comparing the estimation errors of $\hat{\mathbf{H}}$ and $\hat{\mathbf{H}}_k$.

Proposition 8. *Suppose that* $\operatorname{vec}(\mathbf{H}^{(\ell)} - \mathbf{H}) \sim \mathcal{CN}(\mathbf{0}, \mathbf{R})$, *for all* $\delta \in (0,1)$ *there exists* $\bar{\eta}_0' > 0$ *such that for all* $\eta \in (0, \bar{\eta}_0')$, *we have*

1) $p_n > 1-\delta$, *if* $\frac{2m_1m_2\|\mathbf{R}\|_{\mathrm{s}}}{\epsilon_0^2}\ln\left(\frac{4m_1m_2}{\delta}\right) \le n \le \frac{\delta^2 R'_{\mathrm{m}}}{8\eta^2\sigma_0^2}$,

2) $p_n < \delta$, *if* $n \ge \frac{8m_1m_2\|\mathbf{R}'\|_{\mathrm{s}}}{\eta^2\sigma_0^2}\ln\left(\frac{4m_1m_2}{\delta}\right)$,

where $\mathbf{R}' \triangleq \frac{1}{2}\operatorname{Re}(\bar{\mathbf{L}})\mathbf{R}\operatorname{Re}(\bar{\mathbf{L}})^{\mathrm{T}} + \frac{1}{2}\operatorname{Im}(\bar{\mathbf{L}})\mathbf{R}\operatorname{Im}(\bar{\mathbf{L}})^{\mathrm{T}}$, *and* R'_{m} *represents the minimal positive diagonal entry of* $\mathbf{R}'$.

Proof. First, from the fact that for any complex vector $\boldsymbol{w}$ the norm square $\|\boldsymbol{w}\|^2 = \|\operatorname{Re}(\boldsymbol{w})\|^2 + \|\operatorname{Im}(\boldsymbol{w})\|^2$, we have

$$\begin{aligned}
&\left\|\bar{\mathbf{L}}\operatorname{vec}(\hat{\mathbf{H}}-\mathbf{H})\right\|^2 \\
&= \left\|\operatorname{Re}(\bar{\mathbf{L}})\operatorname{vec}(\operatorname{Re}(\hat{\mathbf{H}}-\mathbf{H})) - \operatorname{Im}(\bar{\mathbf{L}})\operatorname{vec}(\operatorname{Im}(\hat{\mathbf{H}}-\mathbf{H}))\right\|^2 \\
&\quad + \left\|\operatorname{Re}(\bar{\mathbf{L}})\operatorname{vec}(\operatorname{Im}(\hat{\mathbf{H}}-\mathbf{H})) + \operatorname{Im}(\bar{\mathbf{L}})\operatorname{vec}(\operatorname{Re}(\hat{\mathbf{H}}-\mathbf{H}))\right\|^2 \\
&= \left\|\boldsymbol{w}'\right\|^2 + \left\|\boldsymbol{w}''\right\|^2,
\end{aligned}$$

where $\boldsymbol{w}' \triangleq \operatorname{Re}(\bar{\mathbf{L}})\operatorname{vec}(\operatorname{Re}(\hat{\mathbf{H}}-\mathbf{H})) - \operatorname{Im}(\bar{\mathbf{L}})\operatorname{vec}(\operatorname{Im}(\hat{\mathbf{H}}-\mathbf{H}))$, $\boldsymbol{w}'' \triangleq \operatorname{Re}(\bar{\mathbf{L}})\operatorname{vec}(\operatorname{Im}(\hat{\mathbf{H}}-\mathbf{H})) + \operatorname{Im}(\bar{\mathbf{L}})\operatorname{vec}(\operatorname{Re}(\hat{\mathbf{H}}-\mathbf{H}))$. Note that $\operatorname{vec}(\operatorname{Re}(\hat{\mathbf{H}}-\mathbf{H})) \sim \mathcal{N}(\mathbf{0}, \frac{1}{2n}\mathbf{R})$, $\operatorname{vec}(\operatorname{Im}(\hat{\mathbf{H}}-\mathbf{H})) \sim \mathcal{N}(\mathbf{0}, \frac{1}{2n}\mathbf{R})$, thus $\boldsymbol{w}'$ and $\boldsymbol{w}''$ are both the sum of Gaussian distributions. Therefore we have $\boldsymbol{w}' \sim \mathcal{N}(\mathbf{0}, \frac{1}{n}\mathbf{R}')$ and $\boldsymbol{w}'' \sim \mathcal{N}(\mathbf{0}, \frac{1}{n}\mathbf{R}')$, where $\mathbf{R}' \triangleq \frac{1}{2}\operatorname{Re}(\bar{\mathbf{L}})\mathbf{R}\operatorname{Re}(\bar{\mathbf{L}})^{\mathrm{T}} + \frac{1}{2}\operatorname{Im}(\bar{\mathbf{L}})\mathbf{R}\operatorname{Im}(\bar{\mathbf{L}})^{\mathrm{T}}$.

To prove that $p_n < \delta$, for $\eta \in (0, \eta_0)$, it suffices to choose n satisfying (32) and (33). Note that for any constant $t > 0$,

$$\begin{aligned}
&\mathbb{P}_n\left\{\left\|\bar{\mathbf{L}}\operatorname{vec}(\hat{\mathbf{H}}-\mathbf{H})\right\|^2 > t\right\} \\
&= \mathbb{P}_n\left\{\left\|\boldsymbol{w}'\right\|^2 + \left\|\boldsymbol{w}''\right\|^2 > t\right\} \\
&\le \mathbb{P}\left\{\|\mathbf{R}'\|_{\mathrm{s}}\left\|(\mathbf{R}')^{-\frac{1}{2}}\boldsymbol{w}'\right\|^2 + \|\mathbf{R}'\|_{\mathrm{s}}\left\|(\mathbf{R}')^{-\frac{1}{2}}\boldsymbol{w}''\right\|^2 > t\right\} \\
&= \mathbb{P}\left\{\left\|n^{\frac{1}{2}}(\mathbf{R}')^{-\frac{1}{2}}\boldsymbol{w}'\right\|^2 + \left\|n^{\frac{1}{2}}(\mathbf{R}')^{-\frac{1}{2}}\boldsymbol{w}''\right\|^2 > \frac{nt}{\|\mathbf{R}'\|_{\mathrm{s}}}\right\} \\
&\le 2m_1m_2\exp\left(-\frac{nt}{4m_1m_2\|\mathbf{R}'\|_{\mathrm{s}}}\right), \qquad (57)
\end{aligned}$$

where to obtain the first inequality we have used the submultiplicativity of the spectral norm, and to obtain the second inequality we have used an immediate consequence of (46) that,

$$\mathbb{P}\left\{\|\boldsymbol{z}\|^2 + \left\|\boldsymbol{z}'\right\|^2 > t\right\} < 2m\exp\left(-\frac{t}{4m}\right), \quad t > 0, \qquad (58)$$

where $\boldsymbol{z}, \boldsymbol{z}' \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_m)$ are independent m-dimensional standard normal vectors.

Then, it follows from (57) that

$$\begin{aligned}
&\mathbb{P}_n\left\{\left\|\bar{\mathbf{L}}\operatorname{vec}(\hat{\mathbf{H}}-\mathbf{H})\right\|^2 > \frac{1}{2}\eta^2\sigma_0^2\right\} \\
&\le 2m_1m_2\exp\left(-\frac{n\eta^2\sigma_0^2}{8m_1m_2\|\mathbf{R}'\|_{\mathrm{s}}}\right). \qquad (59)
\end{aligned}$$

Therefore, from (33), it suffices to have

$$2m_1m_2\exp\left(-\frac{n\eta^2\sigma_0^2}{8m_1m_2\|\mathbf{R}'\|_{\mathrm{s}}}\right) < \frac{\delta}{2},$$

which is equivalent to

$$n > \bar{n}_2' \triangleq \frac{8m_1m_2\|\mathbf{R}'\|_{\mathrm{s}}}{\eta^2\sigma_0^2}\ln\left(\frac{4m_1m_2}{\delta}\right).$$

Similarly, from (57), if $n \ge \frac{2m_1m_2\|\mathbf{R}\|_{\mathrm{s}}}{\epsilon_0^2}\ln\left(\frac{4m_1m_2}{\delta}\right)$, we have

$$\begin{aligned}
&\mathbb{P}_n\left\{\|\hat{\mathbf{H}}-\mathbf{H}\|_{\mathrm{F}}^2 > \epsilon_0^2\right\} \\
&= \mathbb{P}_n\left\{\left\|\operatorname{vec}(\hat{\mathbf{H}}-\mathbf{H})\right\|^2 > \epsilon_0^2\right\} \\
&\le 2m_1m_2\exp\left(-\frac{n\epsilon_0^2}{4m_1m_2\|\frac{1}{2}\mathbf{R}\|_{\mathrm{s}}}\right) < \frac{\delta}{2}. \qquad (60)
\end{aligned}$$

Note that when $\eta < \bar{\eta}_1' \triangleq \frac{\epsilon_0\sqrt{2\|\mathbf{R}'\|_{\mathrm{s}}}}{\sigma_0\sqrt{\|\mathbf{R}\|_{\mathrm{s}}}}$, we have

$$\bar{n}_2' > \frac{2m_1m_2\|\mathbf{R}\|_{\mathrm{s}}}{\epsilon_0^2}\ln\left(\frac{4m_1m_2}{\delta}\right). \qquad (61)$$

Therefore, from (59)–(61), we know that when $\eta < \min\{\eta_0, \bar{\eta}_1'\}$, to obtain $p_n < \delta$ it suffices to choose $n > \bar{n}_2'$.

To prove that $p_n > 1-\delta$, for $\eta \in (0, \eta_0)$, it suffices to choose n satisfying (31)–(32).

Similar to (39), we have

$$\mathbb{P}_n\left\{\left\|\bar{\mathbf{L}}\,\mathrm{vec}(\hat{\mathbf{H}}-\mathbf{H})\right\|^2<2\eta^2\sigma_0^2\right\}\le\eta\sigma_0\sqrt{\frac{2n}{R'_{\mathrm{m}}}}.$$

Hence, from (32), it suffices to have

$$\eta\sigma_0\sqrt{\frac{2n}{R'_{\mathrm{m}}}}\le\frac{\delta}{2},$$

which is equivalent to $n<\bar{n}'_1=\frac{\delta^2R'_{\mathrm{m}}}{8\eta^2\sigma_0^2}$.

In addition, from (57), if $n>\bar{n}'_0\triangleq\frac{2m_1m_2\|\mathbf{R}\|_{\mathrm{s}}}{\epsilon_0^2}\ln\left(\frac{4m_1m_2}{\delta}\right)$ we have

$$\begin{aligned}&\mathbb{P}_n\left\{\left\|\mathrm{vec}(\hat{\mathbf{H}}-\mathbf{H})\right\|^2>\epsilon_0^2\right\}\\&\le 2m_1m_2\exp\left(-\frac{n\epsilon_0^2}{4m_1m_2\|\frac{1}{2}\mathbf{R}\|_{\mathrm{s}}}\right)\le\frac{\delta}{2},\end{aligned}$$

and thus (31) holds. Note that when $\eta<\bar{\eta}'_2\triangleq\frac{\delta\epsilon_0}{4\sigma_0}\left(\frac{m_1m_2\|\mathbf{R}\|_{\mathrm{s}}}{R'_{\mathrm{m}}}\ln\left(\frac{4m_1m_2}{\delta}\right)\right)^{-\frac{1}{2}}$, it can be verified that $\bar{n}'_1>\bar{n}'_0$. Finally, $\bar{\eta}'_0$ can be chosen as $\bar{\eta}'_0=\min\{\eta_0,\bar{\eta}'_1,\bar{\eta}'_2\}$. □

B. HGR Maximal Correlation

Given jointly distributed discrete random variables X and Y with ranges $\mathcal{X}$ and $\mathcal{Y}$ respectively, the (generalized) Hirschfeld-Gebelein-Rényi (HGR) maximal correlation [7]–[9], defined as

$$\max_{\substack{f\colon\mathcal{X}\mapsto\mathbb{R}^m,\ g\colon\mathcal{Y}\mapsto\mathbb{R}^m\\ \mathbb{E}[f(X)]=\mathbb{E}[g(Y)]=\mathbf{0}\\ \mathbb{E}[f(X)f^{\mathrm{T}}(X)]=\mathbb{E}[g(Y)g^{\mathrm{T}}(Y)]=\mathbf{I}_m}}\mathbb{E}\left[f^{\mathrm{T}}(X)g(Y)\right],$$

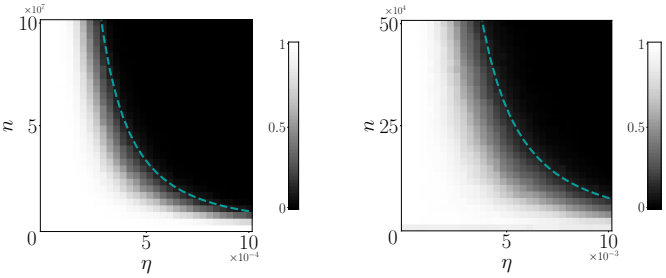
is a normalized measure of the dependence between X and Y, and the functions f and g that achieve this maximal value, called maximal correlation functions, have been shown useful in information theory [1] and machine learning [3]. It can be shown that the maximal correlation functions correspond to the singular vectors of the canonical dependence matrix (CDM) [1] $\tilde{\mathbf{B}}\in\mathbb{R}^{|\mathcal{X}|\times|\mathcal{Y}|}$ of X and Y, defined as

$$\tilde{B}(x,y)\triangleq\frac{P_{XY}(x,y)-P_X(x)P_Y(y)}{\sqrt{P_X(x)P_Y(y)}},\tag{62}$$

where P_{XY} denotes the joint distribution of X and Y. Therefore, the HGR maximal correlation problem can be effectively solved via estimating $\tilde{\mathbf{B}}$. In practice, given n i.i.d. samples $(x^{(1)},y^{(1)}),\ldots,(x^{(n)},y^{(n)})$ with the empirical distribution $\hat{P}_{XY}$, the plug-in estimator for $\tilde{\mathbf{B}}$, which we denote by $\hat{\mathbf{B}}$, is

$$\hat{B}(x,y)\triangleq\frac{\hat{P}_{XY}(x,y)-\hat{P}_X(x)\hat{P}_Y(y)}{\sqrt{\hat{P}_X(x)\hat{P}_Y(y)}},\tag{63}$$

from which the maximal correlation function can also be effectively estimated [15].



(a) Channel state estimation (b) HGR maximal correlation

Fig. 1: Empirical probabilities for the event that low-rank approximation has better performance, where the curves take the form $n\eta^2=c$ for some constants $c>0$.

Then, given a statistical model that exhibits a low-rank structure, the CDM $\tilde{\mathbf{B}}$ will also be a nearly low-rank matrix. In particular, suppose that $\tilde{\mathbf{B}}$ has only k significant singular values, an estimation of $\tilde{\mathbf{B}}$ can be obtained by using either $\hat{\mathbf{B}}$ or its rank-k approximation $\hat{\mathbf{B}}_k$. Note that this is a special case of our model, with the correspondences $\mathbf{A}=\tilde{\mathbf{B}}$, $\hat{\mathbf{A}}=\hat{\mathbf{B}}$, and[2] $Z_{xy}^{(\ell)}=\delta_{x^{(\ell)}x}\delta_{y^{(\ell)}y}$, for all $x\in\mathcal{X},y\in\mathcal{Y}$, and with τ being the mapping from the joint probability matrix $[P_{XY}(x,y)]_{(x,y)\in|\mathcal{X}|\times|\mathcal{Y}|}$ to $\tilde{\mathbf{B}}$. Since $Z_{xy}^{(\ell)}\in[0,1]$, we obtain the following corollary, as an immediate consequence of Theorem 5.

Corollary 9. *For given $\tilde{\mathbf{B}}$, we have that for all $\delta\in(0,1)$, there exists $\eta''>0$ such that for all $\eta\in(0,\eta'')$, we have*

1) $p_n>1-\delta$, if $\frac{\delta^2R_{\mathrm{m}}}{128\eta^2\sigma_0^2}\le n\le\frac{\delta^2R_{\mathrm{m}}}{32\eta^2\sigma_0^2}$,

2) $p_n<\delta$, if $n\ge\frac{32\|\bar{\mathbf{L}}\|_{\mathrm{s}}^2}{\eta^2\sigma_0^2}\ln\left(\frac{4|\mathcal{X}||\mathcal{Y}|}{\delta}\right)$,

where R_{m} is as defined in Theorem 5, where $\bar{\mathbf{L}}=(\mathbf{V}_\perp^{\mathrm{T}}\otimes\mathbf{U}_\perp^{\mathrm{T}})\mathbf{L}$ is as defined in (24), and where $\mathbf{L}$ has entries

$$\begin{aligned}L_{xy,x'y'}=\frac{1}{\sqrt{P_X(x)P_Y(y)}}\Bigg[&\delta_{xx'}\delta_{yy'}-\frac{1}{2}\left(\frac{\delta_{xx'}}{P_X(x)}+\frac{\delta_{yy'}}{P_Y(y)}\right)\\&\cdot\left(P_{XY}(x,y)+P_X(x)P_Y(y)\right)\Bigg].\end{aligned}$$

VI. Numerical Simulations

To verify the relationships between the probability p_n and the choices of (η,n), we conduct numerical simulations for both channel state estimation and HGR maximal correlation.

A. Channel State Estimation

In the channel state estimation problem, we set $m_1=m_2=4$, $k=2$ and $\mathrm{vec}(\mathbf{H}^{(i)})\sim\mathcal{CN}(\mathrm{vec}(\mathrm{diag}(1,1,\eta,\eta)),\mathbf{I})$. For each pair of (η,n), we randomly generate the samples $\mathbf{H}^{(1)},\ldots,\mathbf{H}^{(n)}$, and compute the difference of estimation errors $\|\hat{\mathbf{H}}-\mathbf{H}\|_{\mathrm{F}}^2-\|\hat{\mathbf{H}}_k-\mathbf{H}\|_{\mathrm{F}}^2$ from $\hat{\mathbf{H}}=$

[2] We use the Kronecker delta

$$\delta_{ij}=\begin{cases}1 & \text{if } i=j\\ 0 & \text{otherwise}\end{cases}.$$

$\frac{1}{n}\sum_{\ell=1}^{n}\mathbf{H}^{(\ell)}$. This computation process is repeated for 1000 times, and the corresponding empirical probabilities of the event $\left\{\|\hat{\mathbf{H}}-\mathbf{H}\|_{\mathrm{F}}^2-\|\hat{\mathbf{H}}_k-\mathbf{H}\|_{\mathrm{F}}^2>0\right\}$ are shown in Fig. 1a, where the grayscale for each pixel represents the empirical probability for given (η, n).

B. HGR Maximal Correlation

In the HGR maximal correlation problem, we set $m_1 = m_2 = 4$, with the joint probability matrix $[P_{XY}(x,y)]_{(x,y)\in\mathcal{X}\times\mathcal{Y}}$ constructed as

$$\frac{1}{16}\begin{bmatrix}1+2\sigma & 1-2\sigma & 1 & 1\\ 1-2\sigma & 1+2\sigma & 1 & 1\\ 1 & 1 & 1+2\eta & 1-2\eta\\ 1 & 1 & 1-2\eta & 1+2\eta\end{bmatrix}.$$

Then, it can be verified that the corresponding CDM $\tilde{\mathbf{B}}$ has the singular values $\mathbf{\Sigma}_k=\sigma$ with $k=1$, and $\bar{\mathbf{\Sigma}}_\perp=\mathrm{diag}(1,0,0)$. In particular, in our experiment we set $\sigma=0.05$, and for each pair of (η, n), we randomly generate n sample pairs of (x,y) from the corresponding P_{XY}. Then, the difference of estimation errors $\|\hat{\mathbf{B}}-\tilde{\mathbf{B}}\|_{\mathrm{F}}^2-\|\hat{\mathbf{B}}_k-\tilde{\mathbf{B}}\|_{\mathrm{F}}^2$ is computed from the empirical distribution $\hat{P}_{XY}$ of samples. This computation process is repeated for 1000 times, with the empirical probabilities of the event $\left\{\|\hat{\mathbf{B}}-\tilde{\mathbf{B}}\|_{\mathrm{F}}^2-\|\hat{\mathbf{B}}_k-\tilde{\mathbf{B}}\|_{\mathrm{F}}^2>0\right\}$ shown in Fig. 1b.

From both Fig. 1a and Fig. 1b, the contours of the empirical probabilities coincide with the curves in the form $n\eta^2=c$ $(c>0)$, which validates the theoretical results.

Acknowledgment

We thank Runpeng Yu for assistance with numerical simulations. The work of Shao-Lun Huang was supported in part by the Natural Science Foundation of China under Grant 61807021, in part by the Shenzhen Science and Technology Research and Development Funds under Grant JCYJ20170818094022586, and in part by the Innovation and Entrepreneurship Project for Overseas High-Level Talents of Shenzhen under Grant KQJSCX20180327144037831.

References

[1] S.-L. Huang, A. Makur, G. W. Wornell, and L. Zheng, "On universal features for high-dimensional learning and inference," *arXiv preprint arXiv:1911.09105*, 2019.

[2] O. Levy and Y. Goldberg, "Neural word embedding as implicit matrix factorization," in *Advances in neural information processing systems*, 2014, pp. 2177–2185.

[3] L. Wang, J. Wu, S.-L. Huang, L. Zheng, X. Xu, L. Zhang, and J. Huang, "An efficient approach to informative feature extraction from multimodal data," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 5281–5288.

[4] M. B. Cohen, S. Elder, C. Musco, C. Musco, and M. Persu, "Dimensionality reduction for k-means clustering and low rank approximation," in *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*. ACM, 2015, pp. 163–172.

[5] W. Dong, G. Shi, and X. Li, "Nonlocal image restoration with bilateral variance estimation: a low-rank approach," *IEEE transactions on image processing*, vol. 22, no. 2, pp. 700–711, 2012.

[6] H. Xie, F. Gao, and S. Jin, "An overview of low-rank channel estimation for massive mimo systems," *IEEE Access*, vol. 4, pp. 7313–7321, 2016.

[7] H. O. Hirschfeld, "A connection between correlation and contingency," *Proc. Cambridge Phil. Soc.*, vol. 31, pp. 520–524, 1935.

[8] H. Gebelein, "Das statistische problem der korrelation als variations-und eigenwertproblem und sein zusammenhang mit der ausgleichungsrechnung," *Z. für angewandte Math., Mech.*, vol. 21, pp. 364–379, 1941.

[9] A. Rényi, "On measures of dependence," *Acta Mathematica Academiae Scientiarum Hungarica*, vol. 10, no. 3–4, pp. 441–451, 1959.

[10] R. J. Vaccaro, "A second-order perturbation expansion for the svd," *SIAM Journal on Matrix Analysis and Applications*, vol. 15, no. 2, pp. 661–671, 1994.

[11] V. Bentkus, "A lyapunov-type bound in rd," *Theory of Probability & Its Applications*, vol. 49, no. 2, pp. 311–323, 2005.

[12] J. A. Tropp, "User-friendly tail bounds for sums of random matrices," *Foundations of computational mathematics*, vol. 12, no. 4, pp. 389–434, 2012.

[13] M. Chiani, D. Dardari, and M. K. Simon, "New exponential bounds and approximations for the computation of error probability in fading channels," *IEEE Transactions on Wireless Communications*, vol. 2, no. 4, pp. 840–845, 2003.

[14] E. Bjornson and B. Ottersten, "A framework for training-based estimation in arbitrarily correlated Rician MIMO channels with Rician disturbance," *IEEE Transactions on Signal Processing*, vol. 58, no. 3, pp. 1807–1820, 2009.

[15] S.-L. Huang and X. Xu, "On the sample complexity of HGR maximal correlation functions," *arXiv preprint arXiv:1907.00393*, 2019.