Uncertainty Quantification for Nonconvex Tensor Completion: Confidence Intervals, Heteroscedasticity and Optimality

Changxiao Cai, H. Vincent Poor, *Life Fellow, IEEE*, and Yuxin Chen, *Member, IEEE*

***Abstract*— We study the distribution and uncertainty of nonconvex optimization for noisy tensor completion—the problem of estimating a low-rank tensor given incomplete and corrupted observations of its entries. Focusing on a two-stage estimation algorithm proposed by Cai *et al.*, we characterize the distribution of this nonconvex estimator down to fine scales. This distributional theory in turn allows one to construct valid and short confidence intervals for both the unseen tensor entries and the unknown tensor factors. The proposed inferential procedure enjoys several important features: (1) it is fully adaptive to noise heteroscedasticity, and (2) it is data-driven and automatically adapts to unknown noise distributions. Furthermore, our findings unveil the statistical optimality of nonconvex tensor completion: it attains un-improvable ℓ_2 accuracy—including both the rates and the pre-constants—when estimating both the unknown tensor and the underlying tensor factors.**

***Index Terms*— Confidence intervals, uncertainty quantification, tensor completion, nonconvex optimization, heteroscedasticity.**

I. Introduction

A. Noisy Low-Rank Tensor Completion

TENSOR data are routinely employed in data and information sciences to model (structured) multi-dimensional objects [3], [4], [5], [6], [7], [8], [9]. In many practical scenarios of interest, however, we do not have full access to a large-dimensional tensor of interest, as only a sampling of its entries are revealed to us; yet we would still wish to reliably infer all missing data. This task, commonly referred to as *tensor completion*, finds applications in numerous domains including medical imaging [10], visual data analysis [11], seismic data reconstruction [12], to name just a few. In order to make meaningful inference about the unseen entries, additional information about the unknown tensor plays a pivotal role (otherwise one is in the position with fewer equations than unknowns). A common type of such prior information is low-rank structure, which hypothesizes that the unknown tensor is decomposable into the superposition of a few rank-one tensors. Substantial attempts have been made in the past few years to understand and tackle such low-rank tensor completion problems.

To set the stage for a formal discussion, we formulate the problem as follows. Imagine that we are interested in reconstructing a third-order tensor $\boldsymbol{T}^\star = [T_{i,j,k}]_{1\le i,j,k\le d} \in \mathbb{R}^{d\times d\times d}$, which is *a priori* known to have low canonical polyadic (CP) rank [3]. This means that $\boldsymbol{T}^\star$ admits the following CP decomposition[1] [2]

$$\boldsymbol{T}^\star = \sum_{l=1}^{r} \boldsymbol{u}_l^\star \otimes \boldsymbol{u}_l^\star \otimes \boldsymbol{u}_l^\star =: \sum_{l=1}^{r} (\boldsymbol{u}_l^\star)^{\otimes 3}, \tag{1}$$

where $\boldsymbol{u}_l \in \mathbb{R}^d$ $(1 \le l \le r)$ represents the unknown tensor factor, and the rank r is considerably smaller than the ambient dimension d. What we have obtained is a highly incomplete collection of noisy observations about the entries of $\boldsymbol{T}^\star \in \mathbb{R}^{d\times d\times d}$; more precisely, we observe

$$T^{\mathsf{obs}}_{i,j,k} = T^\star_{i,j,k} + E_{i,j,k}, \qquad (i,j,k) \in \Omega, \tag{2}$$

where $\Omega \subseteq [d]\times[d]\times[d]$ with $[d] := \{1,\cdots,d\}$ is a subset of entries, $T^{\mathsf{obs}}_{i,j,k}$ denotes the observed entry in the (i,j,k)-th position, and we use $E_{i,j,k}$ to represent the associated measurement noise, in an attempt to model more realistic scenarios. The presence of missing data and noise, as well as the "notorious" tensor structure (which is often not as

[1]For any vectors $\boldsymbol{u},\boldsymbol{v},\boldsymbol{w} \in \mathbb{R}^d$, we denote by $\boldsymbol{u}\otimes\boldsymbol{v}\otimes\boldsymbol{w} \in \mathbb{R}^{d\times d\times d}$ a three-way array whose (i,j,k)-th element is given by the product of the corresponding vector entries $u_i v_j w_k$.

[2]We focus on symmetric order-3 tensors for simplicity of presentation. The results in this work generalize to the asymmetric case of higher order (c.f. Remark11)

Manuscript received 17 May 2021; revised 10 June 2022; accepted 7 September 2022. Date of publication 12 September 2022; date of current version 22 December 2022. The work of Changxiao Cai was supported in part by the Gordon Y. S. Wu Fellowship in Engineering. The work of H. Vincent Poor was supported in part by the National Science Foundation under Grant CCF-1908308 and in part by the Princeton Schmidt Data-X Research Award. The work of Yuxin Chen was supported in part by the Alfred P. Sloan Research Fellowship; in part by the Air Force Office of Scientific Research (AFOSR) Young Investigator Program (YIP) under Award FA9550-19-1-0030; in part by the Office of Naval Research (ONR) under Grant N00014-19-1-2120; in part by the Army Research Office (ARO) YIP under Award W911NF-20-1-0097; in part by the NSF under Grant CCF-1907661, Grant IIS-1900140/2218713, Grant IIS-2100158/2218773, and Grant DMS-2014279; and in part by the Princeton School of Engineering and Applied Science (SEAS) Innovation Award. An earlier version of this paper was presented in part at the 2020 International Conference on Machine Learning (ICML) [1]. *(Corresponding author: Yuxin Chen.)*

Changxiao Cai is with the Department of Biostatistics, University of Pennsylvania, Philadelphia, PA 19104 USA (e-mail: changxiao.cai@pennmedicine.upenn.edu).

H. Vincent Poor is with the Department of Electrical and Computer Engineering, Princeton University, Princeton, NJ 08540 USA (e-mail: poor@princeton.edu).

Yuxin Chen is with the Department of Statistics and Data Science, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104 USA (e-mail: yuxinc@wharton.upenn.edu).

Communicated by R. Talmon, Associate Editor for Signal Processing and Source Coding.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TIT.2022.3205781.

Digital Object Identifier 10.1109/TIT.2022.3205781

computationally friendly as its matrix analog), poses severe computational and statistical challenges for reliable tensor reconstruction.

B. Review: A Nonconvex Optimization Approach

A natural reconstruction strategy based on the partial data in hand is to resort to the following least-squares problem:

$$\min_{\boldsymbol{U}\in\mathbb{R}^{d\times r}} \quad f(\boldsymbol{U}) := \sum_{(i,j,k)\in\Omega}\left[\left(\sum_{l=1}^{r}\boldsymbol{u}_l^{\otimes 3}\right)_{i,j,k} - T_{i,j,k}^{\mathsf{obs}}\right]^2. \quad (3)$$

Here and in the sequel, we use $\boldsymbol{U} := [\boldsymbol{u}_1, \cdots, \boldsymbol{u}_r]$ to concisely represent the set $\{\boldsymbol{u}_l\}_{1\le l\le r}$. Unfortunately, owing to its highly nonconvex nature, the optimization problem (3) is in general daunting to solve.

To alleviate computational intractability, a number of polynomial-time algorithms have been proposed; partial examples include convex relaxation [13], [14], [15], spectral methods [16], [17], [18], sum of squares hierarchy [19], [20], alternating minimization [21], [22], and so on. Nevertheless, most of these algorithms either are still computationally prohibitive for large-scale problems, or do not come with optimal statistical guarantees; see Section IV for detailed discussions. To address the computational and statistical challenges at once, the recent work [2] proposed a two-stage nonconvex paradigm that guarantees efficient yet reliable solutions. In a nutshell, this algorithm starts by computing a rough (but reasonable) initial guess $\boldsymbol{U}^0 = [\boldsymbol{u}_1^0, \cdots, \boldsymbol{u}_r^0]$ for all tensor factors, and iteratively refines the estimate by means of the gradient descent (GD) update rule:

$$\boldsymbol{U}^{t+1} = \boldsymbol{U}^t - \eta_t \nabla f(\boldsymbol{U}^t), \qquad t = 0, 1, \cdots \quad (4)$$

See Algorithm 1 (note that the initialization scheme is more complex to describe, and is hence postponed to Appendix A-A). Encouragingly, despite the nonconvex optimization landscape, theoretical guarantees have been developed for Algorithm 1 under a suitable random sampling and random noise model. Take the noiseless case for instance: this approach converges linearly to the ground truth under near-minimal sample complexity. Furthermore, the algorithm achieves intriguing ℓ_2 and ℓ_∞ statistical accuracy under a broad family of noise models.

Algorithm 1 A Nonconvex Algorithm for Tensor Completion

Initialize $\boldsymbol{U}^0 = [\boldsymbol{u}_1^0, \cdots, \boldsymbol{u}_r^0]$ via Algorithm 2.
Gradient updates: **for** $t = 0, 1, \ldots, t_0 - 1$ **do**

$$\boldsymbol{U}^{t+1} = \boldsymbol{U}^t - \eta_t \nabla f(\boldsymbol{U}^t). \quad (5)$$

Output $\boldsymbol{U} = [\boldsymbol{u}_1, \cdots, \boldsymbol{u}_r] := \boldsymbol{U}^{t_0}$.

C. Uncertainty Quantification for Nonconvex Tensor Completion

In various practical scenarios (e.g. medical imaging), in order to enable informative decision making and trustworthy prediction, it is crucial not only to provide the users with the reconstruction outcome, but also to inform them of the uncertainty or risk underlying the reconstruction. The latter task, often termed *uncertainty quantification,* can be accomplished by characterizing the (approximate) distributions of our reconstruction, which can be further employed to construct valid confidence intervals (namely, giving lower and upper bounds) for the unknowns. In particular, two questions are of fundamental importance: given an estimate returned by the above nonconvex algorithm, how to identify a confidence interval when predicting an unseen entry, and how to produce a confidence region that is likely to contain the tensor factors of interest?

Unfortunately, classical distributional theory available in the statistics literature, which typically operates in the large-sample regime (with a fixed number of unknowns and a sample size tending to infinity), is not applicable to assess the uncertainty of the above nonconvex algorithm in high dimension. In fact, due to the nonconvex nature of the algorithm, it becomes remarkably challenging to track the distribution of the solution returned by Algorithm 1 or other nonconvex alternatives. The absence of distributional characterization prevents us from communicating a trustworthy uncertainty estimate to the users. While the statistical performance of Algorithm 1 has been investigated in [2], existing statistical guarantees — which hide the (potentially huge) pre-constants — can only yield confidence intervals that are overly wide and, as a result, practically uninformative. In contrast, one should aim for valid confidence intervals that are as short as possible.

Furthermore, an ideal procedure for uncertainty quantification should automatically adapt to unknown noise distributions. Accomplishing this goal, however, becomes particularly challenging when the noise levels are not only unknown but also location-varying — a scenario commonly referred to as *heteroscedasticity*. In fact, there is no shortage of realistic scenarios in which the data heteroscedasticity makes it infeasible to pre-estimate local variability in a uniformly reliable manner. Addressing this challenge calls for the design of model-agnostic data-driven procedures that are fully adaptive to noise heteroscedasticity.

D. Main Contributions and Insights

We now give an informal overview of the main contributions and insights of this paper. To the best of our knowledge, results of this kind were previously unavailable in the tensor completion/estimation literature.

1) *A distributional theory for nonconvex tensor completion.* Despite its nonconvex nature, the distributional representation of the estimate returned by Algorithm 1 can be established down to quite fine scales. Under mild conditions, (1) the resulting estimates for both the tensor factors and the tensor entries are nearly unbiased, and (2) the associated uncertainty follows a zero-mean Gaussian distribution whose (co)-variance can be accurately determined in a data-driven manner.
2) *Construction of entrywise confidence intervals*. The above distributional characterization leads to construction of

entrywise confidence intervals for both the unknown tensor and the associated tensor factors. The proposed inferential procedure is fully data-driven: it does not require prior knowledge about the noise distributions, and it automatically adapts to local variability of noise levels.

3) *Optimality w.r.t. both inference and estimation*. We develop fundamental lower bounds under i.i.d. Gaussian noise, confirming that the proposed entrywise confidence intervals are in some sense the shortest possible. As a byproduct, our results also reveal that nonconvex optimization achieves un-improvable ℓ_2 statistical accuracy — including both the rates and the pre-constants — for estimating both the unknown tensor and its underlying tensor factors.

All in all, our results shed light on the effectiveness of nonconvex optimization in noisy tensor completion, which enables optimal estimation and uncertainty quantification all at once.

The rest of the paper is organized as follows. Section II formulates the problem settings. Section III presents our distributional theory, discusses construction of confidence intervals, and develops fundamental lower bounds. Section IV provides an overview of related prior work. The proof outline of our main theory is supplied in Section V, with the proofs of auxiliary lemmas provided in the appendix. We conclude the paper with a discussion of future directions in Section VI.

E. Notation

For any matrix $\boldsymbol{M}$, we use $\|\boldsymbol{M}\|$ and $\|\boldsymbol{M}\|_{\mathrm{F}}$ to denote the spectral norm (operator norm) and the Frobenius norm of $\boldsymbol{M}$, respectively, and let $\boldsymbol{M}_{i,:}$ and $\boldsymbol{M}_{:,i}$ stand for the i-th row and i-th colomn, respectively. We denote by $\|\boldsymbol{M}\|_{2,\infty} := \max_l \|\boldsymbol{M}_{l,:}\|_2$ (resp. $\|\boldsymbol{M}\|_{\infty} := \max_{i,j} |M_{i,j}|$) the $\ell_{2,\infty}$ norm (resp. entrywise ℓ_∞ norm) of $\boldsymbol{M}$. In addition, let $\lambda_1(\boldsymbol{M}) \geq \lambda_2(\boldsymbol{M}) \geq \cdots$ denote the eigenvalues of $\boldsymbol{M}$ and $\sigma_1(\boldsymbol{M}) \geq \sigma_2(\boldsymbol{M}) \geq \cdots$ denote the singular values of $\boldsymbol{M}$. For any matrices $\boldsymbol{M}, \boldsymbol{N}$ of compatible dimensions, we let $\boldsymbol{M} \odot \boldsymbol{N}$ stand for the Hadamard (entrywise) product.

For any tensor $\boldsymbol{T} \in \mathbb{R}^{d\times d\times d}$, denote by $\mathcal{P}_\Omega(\boldsymbol{T})$ the Euclidean projection of $\boldsymbol{T}$ onto the subset of tensors that vanish outside the index set Ω. With this notation in place, the observed data (2) can be succinctly described as

$$\mathcal{P}_\Omega\big(\boldsymbol{T}^{\mathsf{obs}}\big) = \mathcal{P}_\Omega\big(\boldsymbol{T}^\star + \boldsymbol{E}\big), \tag{6}$$

where $\boldsymbol{T}^{\mathsf{obs}} := [T^{\mathsf{obs}}_{i,j,k}]_{1\leq i,j,k\leq d}$ and $\boldsymbol{E} := [E_{i,j,k}]_{1\leq i,j,k\leq d}$. Here and throughout, we let $T^{\mathsf{obs}}_{i,j,k} = 0$ for any $(i,j,k) \notin \Omega$. In addition, we use $u_{l,i}$ (resp. $u^\star_{l,i}$) to denote the i-th entry of $\boldsymbol{u}_l \in \mathbb{R}^d$ (resp. $\boldsymbol{u}^\star_l \in \mathbb{R}^d$).

For any tensor $\boldsymbol{T} \in \mathbb{R}^{d\times d\times d}$, let $\boldsymbol{T}_{i,:,:} \in \mathbb{R}^{d\times d}$ denote the mode-1 i-th slice with entries $(\boldsymbol{T}_{i,:,:})_{j,k} = T_{i,j,k}$, and $\boldsymbol{T}_{:,i,:}$ and $\boldsymbol{T}_{:,:,i}$ are defined analogously. Let $\mathsf{unfold}(\boldsymbol{T})$ represent the mode-3 matricization of $\boldsymbol{T}$, namely, $\mathsf{unfold}\,(\boldsymbol{T})$ is a matrix in $\mathbb{R}^{d\times d^2}$ whose entries are given by

$$\big(\mathsf{unfold}(\boldsymbol{T})\big)_{k,d(i-1)+j} = T_{i,j,k}, \qquad 1 \leq i,j,k \leq d. \tag{7}$$

For any tensors $\boldsymbol{T} \in \mathbb{R}^{d\times d\times d}$, the Frobenius norm of $\boldsymbol{T}$ is defined accordingly as $\|\boldsymbol{T}\|_{\mathrm{F}} := \sqrt{\sum_{i,j,k} T^2_{i,j,k}}$. We use $\|\boldsymbol{T}\|_\infty := \max_{i,j,k} |T_{i,j,k}|$ to denote the entrywise ℓ_∞ norm. For any vectors $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^d$, we define the vector products $\boldsymbol{T} \times_3 \boldsymbol{u} \in \mathbb{R}^{d\times d}$ and $\boldsymbol{T} \times_1 \boldsymbol{u} \times_2 \boldsymbol{v} \in \mathbb{R}^d$ such that

$$[\boldsymbol{T} \times_3 \boldsymbol{u}]_{i,j} := \sum\nolimits_k T_{i,j,k} u_k, \qquad 1 \leq i,j \leq d; \tag{8a}$$

$$[\boldsymbol{T} \times_1 \boldsymbol{u} \times_2 \boldsymbol{v}]_k := \sum\nolimits_{i,j} T_{i,j,k} u_i v_j, \qquad 1 \leq k \leq d. \tag{8b}$$

The products $\boldsymbol{T} \times_2 \boldsymbol{u} \in \mathbb{R}^{d\times d}$, $\boldsymbol{T} \times_3 \boldsymbol{u} \in \mathbb{R}^{d\times d}$, $\boldsymbol{T} \times_1 \boldsymbol{u} \times_3 \boldsymbol{v} \in \mathbb{R}^d$, $\boldsymbol{T} \times_2 \boldsymbol{u} \times_3 \boldsymbol{v} \in \mathbb{R}^d$ are defined analogously. In addition, the spectral norm of $\boldsymbol{T}$ is defined as $\|\boldsymbol{T}\| := \sup_{\boldsymbol{u},\boldsymbol{v},\boldsymbol{w}\in\mathbb{S}^{d-1}} \langle \boldsymbol{T}, \boldsymbol{u}\otimes\boldsymbol{v}\otimes\boldsymbol{w}\rangle$, where we denote by $\mathbb{S}^{d-1} := \{\boldsymbol{u}\in\mathbb{R}^d \,|\, \|\boldsymbol{u}\|_2 = 1\}$ the unit sphere in $\mathbb{R}^d$.

We use $[a \pm b]$ to denote the interval $[a-b, a+b]$, and we shall often let (i,j) denote $(i-1)d+j$ whenever it is clear from the context. We denote by $[d] := \{1, 2, \cdots, d\}$. The notation $f(d) \lesssim g(d)$ or $f(d) = O(g(d))$ (resp. $f(d) \gtrsim g(d)$) means that there exists a constant $C_0 > 0$ such that $|f(d)| \leq C_0|g(d)|$ (resp. $|f(d)| \geq C_0|g(d)|$). The notation $f(d) \asymp g(d)$ means that $C_0|f(d)| \leq |g(d)| \leq C_1|f(d)|$ holds for some universal constants $C_0, C_1 > 0$. In addition, $f(d) = o(g(d))$ means that $\lim_{d\to\infty} f(d)/g(d) = 0$, $f(d) \ll g(d)$ means that $f(d) \leq c_0 g(d)$ for some small constant $c_0 > 0$ and $f(d) \gg g(d)$ means that $f(d) \geq c_0 g(d)$ for some large constant $c_0 > 0$.

II. Models and Assumptions

In this paper, we shall consider a setting with random sampling and independent random noise as follows.

Assumption 1 (Random Sampling): Suppose that Ω is a symmetric index set.[3] Assume that each (i,j,k) with $i \leq j \leq k$ is included in Ω independently with probability p. Throughout this paper, we shall define

$$\chi_{i,j,k} := \mathbb{1}\{(i,j,k) \in \Omega\}, \qquad 1 \leq i,j,k \leq d. \tag{9}$$

Assumption 2 (Random Noise): Suppose that $\boldsymbol{E} = [E_{i,j,k}]_{1\leq i,j,k\leq d}$ is a symmetric tensor.[4] Assume that the noise components $\{E_{i,j,k}\}_{1\leq i\leq j\leq k\leq d}$ are independent sub-Gaussian random variables satisfying $\mathbb{E}[E_{i,j,k}] = 0$ and $\mathsf{Var}(E_{i,j,k}) = \sigma^2_{i,j,k}$. Denoting $\sigma_{\min} := \min_{i,j,k} \sigma_{i,j,k}$ and $\sigma_{\max} := \max_{i,j,k} \sigma_{i,j,k}$, we assume throughout that $\sigma_{\max}/\sigma_{\min} = o\,(1)$.

Next, we introduce additional parameters about the unknown tensor of interest. Recall that

$$\boldsymbol{T}^\star = \sum_{l=1}^r \boldsymbol{u}^\star_l \otimes \boldsymbol{u}^\star_l \otimes \boldsymbol{u}^\star_l = \sum_{l=1}^r \boldsymbol{u}^{\star\otimes 3}_l \in \mathbb{R}^{d\times d\times d}.$$

To begin with, we define the strength of each rank-one tensor component as follows

$$\lambda^\star_{\min} := \min_{1\leq l\leq r} \|\boldsymbol{u}^\star_l\|_2^3 \quad \text{and} \quad \lambda^\star_{\max} := \max_{1\leq l\leq r} \|\boldsymbol{u}^\star_l\|_2^3, \tag{10}$$

allowing us to define the condition number by

$$\kappa := \lambda^\star_{\max}/\lambda^\star_{\min}. \tag{11}$$

[3]This means that if $(i,j,k) \in \Omega$, then (j,i,k), (i,k,j), (j,k,i), (k,i,j), (k,j,i) are all in Ω.

[4]This means that $E_{i,j,k} = E_{j,i,k} = E_{i,k,j} = E_{j,k,i} = E_{k,i,j} = E_{k,j,i}$ for any $1 \leq i,j,k \leq d$.

To enable reliable tensor completion, we introduce the following incoherence definitions regarding the tensor $\boldsymbol{T}^\star$ and the tensor factors $\{\boldsymbol{u}_l^\star\}$.

Definition 1 (Incoherence): Define the incoherence parameters of $\boldsymbol{T}^\star$ and $\{\boldsymbol{u}_l^\star\}$ as follows

$$\mu_0 := \frac{d^3 \left\|\boldsymbol{T}^\star\right\|_\infty^2}{\left\|\boldsymbol{T}^\star\right\|_{\mathrm{F}}^2}, \tag{12a}$$

$$\mu_1 := \max_{1\le i\le r} \frac{d \left\|\boldsymbol{u}_i^\star\right\|_\infty^2}{\left\|\boldsymbol{u}_i^\star\right\|_2^2}, \tag{12b}$$

$$\mu_2 := \max_{1\le i\ne j\le r} \frac{d \left\langle \boldsymbol{u}_i^\star, \boldsymbol{u}_j^\star\right\rangle^2}{\left\|\boldsymbol{u}_i^\star\right\|_2^2 \left\|\boldsymbol{u}_j^\star\right\|_2^2}, \tag{12c}$$

Informally, when both μ_0 and μ_1 are small, the ℓ_2 energy of both $\boldsymbol{T}^\star$ and $\boldsymbol{u}_l^\star$ $(1 \le l \le r)$ is dispersed more or less evenly across their entries. In addition, a small μ_2 necessarily implies that every pair of the tensor factors of interest is nearly orthogonal to (and hence incoherent with) each other. Finally, the well-conditionedness assumption guarantees that no single tensor component has significantly higher energy compared to the rest of them. For the sake of notational simplicity, we shall combine them into a single incoherence parameter

$$\mu := \max\left\{\mu_0,\, \mu_1,\, \mu_2\right\}. \tag{13}$$

The focal point of this paper lies in obtaining distributional characterization of, and uncertainty assessment for, the nonconvex estimate (i.e. the solution $\boldsymbol{U}$ returned by Algorithm 1) in a strong entrywise sense. In particular, we set out the goal to

1) establish distributional representation of the estimate $\boldsymbol{U}$;
2) construct short yet valid confidence intervals for each entry of the tensor factor $\{\boldsymbol{u}_l^\star\}_{1\le l\le r}$ as well as each entry of the unknown tensor $\boldsymbol{T}^\star$.

To cast the latter task in more precise terms: given any target coverage level $0 < 1-\alpha < 1$, any $1 \le l \le r$ and any $1 \le i,j,k \le d$, the aim is to compute intervals $[c_{1,u}, c_{2,u}]$ and $[c_{1,T}, c_{2,T}]$ such that

$$\mathbb{P}\left\{u_{l,i}^\star \in [c_{1,u}, c_{2,u}]\right\} = 1-\alpha+o\left(1\right) \tag{14a}$$

$$\mathbb{P}\left\{T_{i,j,k}^\star \in [c_{1,T}, c_{2,T}]\right\} = 1-\alpha+o\left(1\right). \tag{14b}$$

Here, (14a) is phrased accounting for global permutation, since one cannot possibly distinguish $\{\boldsymbol{u}_l^\star\}_{1\le l\le r}$ and any permutation of them given only the observations (2). Ideally, the above tasks should be accomplished in a data-driven manner without requiring prior knowledge about the noise distributions.

III. Main Results

This section presents our distributional theory for nonconvex tensor completion, and demonstrates how to conduct data-driven and optimal uncertainty quantification. For notational convenience, in the sequel we denote by $\boldsymbol{U} = [\boldsymbol{u}_1, \cdots, \boldsymbol{u}_r] \in \mathbb{R}^{d\times r}$ the estimate returned by Algorithm 1, and let $\boldsymbol{T} \in \mathbb{R}^{d\times d\times d}$ indicate the resulting tensor estimate as follows

$$\boldsymbol{T} := \sum_{l=1}^r \boldsymbol{u}_l \otimes \boldsymbol{u}_l \otimes \boldsymbol{u}_l. \tag{15}$$

In addition, recognizing that one can only hope to recover $\boldsymbol{U}^\star$ up to global permutation, we introduce a permutation matrix as follows

$$\boldsymbol{\Pi} := \operatorname*{argmin}_{\boldsymbol{Q}\in \mathsf{perm}_r} \left\|\boldsymbol{U}\boldsymbol{Q} - \boldsymbol{U}^\star\right\|_{\mathrm{F}}, \tag{16}$$

where perm_r represents the set of permutation matrices in $\mathbb{R}^{r\times r}$. Additionally, in order to guarantee reliable convergence of Algorithm 1, there are several algorithmic parameters (e.g. the learning rates) that need to be properly chosen. We shall adopt the choices suggested by [2] throughout this paper. Given that our theory can be presented regardless of whether one understands these algorithmic choices, we defer the specification of these parameters to Appendix A-B to avoid distraction.

A. Distributional Guarantees for Nonconvex Estimates

We now establish distributional guarantees for the nonconvex estimate. For notational convenience, we introduce an auxiliary matrix $\widetilde{\boldsymbol{U}}^\star \in \mathbb{R}^{d^2\times r}$ as well as a collection of diagonal matrices $\boldsymbol{D}_k^\star \in \mathbb{R}^{d^2\times d^2}$ $(1\le k\le d)$ such that

$$\widetilde{\boldsymbol{U}}^\star := \left[\boldsymbol{u}_1^\star \otimes \boldsymbol{u}_1^\star, \cdots, \boldsymbol{u}_r^\star \otimes \boldsymbol{u}_r^\star\right] \in \mathbb{R}^{d^2\times r}; \tag{17}$$

$$\left(\boldsymbol{D}_k^\star\right)_{(i,j),(i,j)} := \sigma_{i,j,k}^2, \qquad 1\le i,j\le d; \tag{18}$$

here, we abuse the notation (i,j) to denote $(i-1)d+j$ whenever it is clear from the context. In words, $\widetilde{\boldsymbol{U}}^\star$ lifts the tensor factors to a higher order, and $\boldsymbol{D}_k^\star$ collects the noise variance in the k-th slice of $\boldsymbol{E}$. To simplify presentation, we begin with the case with independent Gaussian noise.

Theorem 1 (Distributional Guarantees for Tensor Factor Estimates (Gaussian Noise)): Suppose that the $E_{i,j,k}$'s are Gaussian, and that Assumptions 1-2 hold. Assume that $\mu, \kappa, r = O(1)$ and that $t_0 = c_0 \log d$,

$$p \ge c_1 \frac{\log^5 d}{d^{3/2}}, \quad \frac{c_2}{d^{100}} \le \frac{\sigma_{\min}}{\|\boldsymbol{T}^\star\|_\infty} \le \frac{\sigma_{\max}}{\|\boldsymbol{T}^\star\|_\infty} \le c_3 \sqrt{\frac{p d^{3/2}}{\log^4 d}} \tag{19}$$

for some sufficiently large (resp. small) constant $c_0, c_2, c_3 > 0$ (resp. $c_1 > 0$). Then with probability at least $1-o\left(1\right)$, one has the following decomposition:

$$\boldsymbol{U}\boldsymbol{\Pi} - \boldsymbol{U}^\star = \boldsymbol{Z} + \boldsymbol{W},$$

where $\boldsymbol{\Pi}$ is defined in (16), $\|\boldsymbol{W}\|_{2,\infty} = o\Big(\frac{\sigma_{\min}}{\lambda_{\max}^{\star 2/3}\sqrt{p}}\Big)$, and for any $1\le k\le d$ one has $\boldsymbol{Z}_{k,:} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_k^\star)$ with

$$\boldsymbol{\Sigma}_k^\star := \frac{2}{p}\big(\widetilde{\boldsymbol{U}}^{\star\top}\widetilde{\boldsymbol{U}}^\star\big)^{-1}\widetilde{\boldsymbol{U}}^{\star\top}\boldsymbol{D}_k^\star\widetilde{\boldsymbol{U}}^\star\big(\widetilde{\boldsymbol{U}}^{\star\top}\widetilde{\boldsymbol{U}}^\star\big)^{-1}. \tag{20}$$

Remark 1 As an interpretation of Condition (19): (i) the sample complexity is $pd^3 \gtrsim d^{3/2}\mathrm{poly}\log(d)$, which is widely conjectured to be computationally optimal (up to some log factor) [19]; (ii) the typical size of each noise component (as captured by $\{\sigma_{i,j,k}\}$) is allowed to be substantially larger than the maximum magnitude of the entries of $\boldsymbol{T}^\star$ under the sample size assumption stated here.

In words, Theorem 1 reveals that the estimation error of $\boldsymbol{U}$ can be decomposed into a Gaussian component $\boldsymbol{Z}$ and a residual term $\boldsymbol{W}$. Encouragingly, the residual term $\boldsymbol{W}$ is, in some sense, dominated by the Gaussian term and can be safely neglected. To see this, recall that $\sigma_{i,j,k} \geq \sigma_{\min}$, leading to a lower bound[5]

$$\begin{aligned}\boldsymbol{\Sigma}_k^\star &\succeq \frac{2\sigma_{\min}^2}{p}\big(\widetilde{\boldsymbol{U}}^{\star\top}\widetilde{\boldsymbol{U}}^\star\big)^{-1} \\ &\succeq \frac{(2-o(1))\sigma_{\min}^2}{p}\mathrm{diag}\Big(\big[\|\boldsymbol{u}_i^\star\|_2^{-4}\big]_{1\leq i\leq r}\Big) \\ &\succeq (1-o(1))\frac{2\sigma_{\min}^2}{p\lambda_{\max}^{\star 4/3}}\boldsymbol{I}.\end{aligned}$$

This tells us that the typical ℓ_2 norm of each row $\boldsymbol{Z}_{k,:}$ exceeds the order of $\frac{\sigma_{\min}\sqrt{r}}{\sqrt{p}\lambda_{\max}^{\star 2/3}}$, which is hence much larger than $\|\boldsymbol{W}\|_{2,\infty}$ (by virtue of Theorem 1). To conclude, the nonconvex estimate $\boldsymbol{U}$ is — up to global permutation — a nearly un-biased estimate of the true tensor factors $\boldsymbol{U}^\star$, with estimation errors being approximately Gaussian.

As it turns out, this distributional characterization can be extended to accommodate a much broader class of noise beyond Gaussian noise, as stated below.

Theorem 2 (Distributional Guarantees for Tensor Factor Estimates (General Noise)): Suppose that $\{E_{i,j,k}\}$ are not necessarily Gaussian but satisfy Assumption 2. Then the decomposition in Theorem 1 continues to hold, except that $\boldsymbol{Z}$ is not necessarily Gaussian but instead obeys

$$\Big|\mathbb{P}\big\{\boldsymbol{Z}_{k,:}\in\mathcal{A}\big\}-\mathbb{P}\left\{\boldsymbol{g}_k\in\mathcal{A}\right\}\Big| = o\left(1\right), \qquad 1\leq k\leq d$$

for any convex set $\mathcal{A}\subset\mathbb{R}^r$. Here, $\boldsymbol{g}_k \sim \mathcal{N}\left(\boldsymbol{0},\boldsymbol{\Sigma}_k^\star\right)$ with covariance matrix $\boldsymbol{\Sigma}_k^\star$ defined in (20).

Before continuing, there is another important point that is worth making (which is not included in Theorems 1-2 but will be made precise in the analysis): our theory is capable of characterizing not only the distribution for a single row of the tensor factor matrix, but also joint distributions of multiple rows of the tensor factor matrix; this will in turn help provide simultaneous coverage for multiple rows of the tensor factors (which we omit in this paper though for the sake of conciseness).

As it turns out, for any three different rows i,j,k, the corresponding errors $\boldsymbol{Z}_{i,:}$, $\boldsymbol{Z}_{j,:}$ and $\boldsymbol{Z}_{k,:}$ are nearly statistically independent. This is a crucial observation that immediately allows for entrywise distributional characterizations for the resulting tensor estimate $\boldsymbol{T}$, as summarized below.

Theorem 3 (Distributional Guarantees for Tensor Entry Estimates): Instate the assumptions of Theorem 2. Consider any $1\leq i\leq j\leq k\leq d$ obeying

$$\frac{\big\|\widetilde{\boldsymbol{U}}^\star_{(j,k),:}\big\|_2+\big\|\widetilde{\boldsymbol{U}}^\star_{(i,j),:}\big\|_2+\big\|\widetilde{\boldsymbol{U}}^\star_{(i,k),:}\big\|_2}{\|\widetilde{\boldsymbol{U}}^\star\|_{2,\infty}} \geq c_5\frac{\sigma_{\max}}{\|\boldsymbol{T}^\star\|_\infty}\sqrt{\frac{\log^3 d}{d^2 p}} \tag{21}$$

for some large constant $c_5>0$, with $\widetilde{\boldsymbol{U}}^\star$ defined in (17). Then the estimate $\boldsymbol{T}$ defined in (15) obeys

$$\sup_{\tau\in\mathbb{R}}\left|\mathbb{P}\Big\{T_{i,j,k}\leq T_{i,j,k}^\star+\tau\sqrt{v_{i,j,k}^\star}\Big\}-\Phi(\tau)\right| = o\left(1\right), \tag{22}$$

where $\Phi(\cdot)$ is the CDF of a standard Gaussian random variable. Here, the variance parameters $\{v_{i,j,k}^\star\}$ are defined such that for any three distinct numbers i,j,k,

$$\begin{aligned}v_{i,j,k}^\star &:= \widetilde{\boldsymbol{U}}^\star_{(j,k),:}\boldsymbol{\Sigma}_i^\star\big(\widetilde{\boldsymbol{U}}^\star_{(j,k),:}\big)^\top+\widetilde{\boldsymbol{U}}^\star_{(i,k),:}\boldsymbol{\Sigma}_j^\star\big(\widetilde{\boldsymbol{U}}^\star_{(i,k),:}\big)^\top \\ &\quad+\widetilde{\boldsymbol{U}}^\star_{(i,j),:}\boldsymbol{\Sigma}_k^\star\big(\widetilde{\boldsymbol{U}}^\star_{(i,j),:}\big)^\top, \end{aligned}\tag{23a}$$

$$v_{i,i,k}^\star := 4\,\widetilde{\boldsymbol{U}}^\star_{(i,k),:}\boldsymbol{\Sigma}_i^\star\big(\widetilde{\boldsymbol{U}}^\star_{(i,k),:}\big)^\top+\widetilde{\boldsymbol{U}}^\star_{(i,i),:}\boldsymbol{\Sigma}_k^\star\big(\widetilde{\boldsymbol{U}}^\star_{(i,i),:}\big)^\top, \tag{23b}$$

$$v_{i,i,i}^\star := 9\,\widetilde{\boldsymbol{U}}^\star_{(i,i),:}\boldsymbol{\Sigma}_i^\star\big(\widetilde{\boldsymbol{U}}^\star_{(i,i),:}\big)^\top, \tag{23c}$$

where $\boldsymbol{\Sigma}_k^\star$ is defined in (20).

In short, the above theorem indicates that: if the "strength" of a tensor entry $T_{i,j,k}^\star$ is not exceedingly small, then our nonconvex estimate of this entry is nearly unbiased, whose estimation error is approximately zero-mean Gaussian with variance $v_{i,j,k}^\star$. To see this, note that when (19) holds, the right-hand side of Condition (21) is at most $O\left(d^{-1/4}/\sqrt{\log d}\right)$, which is vanishingly small. In other words, the Gaussian approximation is nearly tight unless the energy $\big\|\widetilde{\boldsymbol{U}}^\star_{(j,k),:}\big\|_2+\big\|\widetilde{\boldsymbol{U}}^\star_{(i,k),:}\big\|_2+\big\|\widetilde{\boldsymbol{U}}^\star_{(i,j),:}\big\|_2$ is vanishingly small compared to the average size. This entrywise distributional theory allows one to accommodate a broad family of noise models.

B. Confidence Intervals

The preceding distributional guarantees pave the way for uncertainty quantification. However, an outstanding challenge remains in computing / estimating the covariance matrices $\{\boldsymbol{\Sigma}_k^\star\}$ and the variance parameters $\{v_{i,j,k}^\star\}$. In particular, these crucial parameters are functions of both the ground truth $\{\boldsymbol{u}_l^\star\}$ and the noise variance $\{\sigma_{i,j,k}^2\}$, which we do not have access to *a priori*. To further complicate matters, in the heteroscedastic case where $\{\sigma_{i,j,k}^2\}$ are location varying, it is in general infeasible to estimate all variance parameters reliably.

Variance and Covariance Estimation: Fortunately, despite the absence of prior knowledge about the truth and the noise parameters, we are still able to faithfully estimate these important parameters from the data in hand, using a simple plug-in procedure. Specifically:

1) Rather than estimating all $\{\sigma_{i,j,k}\}$ directly, we turn attention to estimating the noise components $\{E_{i,j.k}\}$ instead, with the assistance of our tensor estimate $\boldsymbol{T}$ as follows
$$\widehat{E}_{i,j,k} := T_{i,j,k}^{\mathsf{obs}} - T_{i,j,k}, \qquad (i,j,k)\in\Omega. \tag{24}$$
We then construct a diagonal matrix $\boldsymbol{D}_k\in\mathbb{R}^{d^2\times d^2}$ $(1\leq k\leq d)$ obeying
$$\big(\boldsymbol{D}_k\big)_{(i,j),(i,j)} = p^{-1}\widehat{E}_{i,j,k}^2\mathbb{1}_{\{(i,j,k)\in\Omega\}}. \tag{25}$$
Note that $\boldsymbol{D}_k$ is not really a faithful estimate of the $\boldsymbol{D}_k^\star$ defined in (18), but it suffices for our purpose.
2) Estimate $\widetilde{\boldsymbol{U}}^\star$ (cf. (17)) via the plug-in estimator $\widetilde{\boldsymbol{U}} := [\boldsymbol{u}_s\otimes\boldsymbol{u}_s]_{1\leq s\leq r}\in\mathbb{R}^{d^2\times r}$.

[5]To see why the penultimate inequality holds, note that under our assumptions,

$$\begin{aligned}\widetilde{\boldsymbol{U}}^{\star\top}\widetilde{\boldsymbol{U}}^\star &= \big[(\boldsymbol{u}_i^{\star\top}\boldsymbol{u}_j^\star)^2\big]_{1\leq i,j\leq r} \\ &\preceq \mathsf{diag}\big(\big[\|\boldsymbol{u}_i^\star\|_2^4\big]_{1\leq i\leq r}\big)+r\max\nolimits_{i\neq j}(\boldsymbol{u}_i^{\star\top}\boldsymbol{u}_j^\star)^2\boldsymbol{I}_r \\ &= (1+o\left(1\right))\,\mathsf{diag}\big(\big[\|\boldsymbol{u}_i^\star\|_2^4\big]_{1\leq i\leq r}\big).\end{aligned}$$

3) Substitute the above estimators into the expressions of the variance parameters to yield our estimate. Specifically, for any $1 \le k \le d$, we compute

$$\boldsymbol{\Sigma}_k = \frac{2}{p}(\widetilde{\boldsymbol{U}}^\top \widetilde{\boldsymbol{U}})^{-1} \widetilde{\boldsymbol{U}}^\top \boldsymbol{D}_k \widetilde{\boldsymbol{U}} (\widetilde{\boldsymbol{U}}^\top \widetilde{\boldsymbol{U}})^{-1} \tag{26}$$

as an estimate of $\boldsymbol{\Sigma}_k^\star$ (cf. (20)). We also compute the estimates for $\{v_{i,j,k}^\star\}$ such that: for any three distinct numbers $1 \le i, j, k \le d$,

$$\begin{aligned} v_{i,j,k} &:= \widetilde{\boldsymbol{U}}_{(j,k),:} \boldsymbol{\Sigma}_i \big(\widetilde{\boldsymbol{U}}_{(j,k),:}\big)^\top + \widetilde{\boldsymbol{U}}_{(i,k),:} \boldsymbol{\Sigma}_j \big(\widetilde{\boldsymbol{U}}_{(i,k),:}\big)^\top \\ &\quad + \widetilde{\boldsymbol{U}}_{(i,j),:} \boldsymbol{\Sigma}_k \big(\widetilde{\boldsymbol{U}}_{(i,j),:}\big)^\top; \end{aligned} \tag{27a}$$

$$v_{i,i,k} := 4\, \widetilde{\boldsymbol{U}}_{(i,k),:} \boldsymbol{\Sigma}_i \big(\widetilde{\boldsymbol{U}}_{(i,k),:}\big)^\top + \widetilde{\boldsymbol{U}}_{(i,i),:} \boldsymbol{\Sigma}_k \big(\widetilde{\boldsymbol{U}}_{(i,i),:}\big)^\top; \tag{27b}$$

$$v_{i,i,i} := 9\, \widetilde{\boldsymbol{U}}_{(i,i),:} \boldsymbol{\Sigma}_i \big(\widetilde{\boldsymbol{U}}_{(i,i),:}\big)^\top. \tag{27c}$$

Confidence Intervals: With the above variance / covariance estimates in place, we are positioned to introduce our uncertainty quantification procedure, which consists in constructing *entrywise* confidence intervals for both the tensor factors and the unknown tensor as follows.

- For each $1 \le k \le d$ and $1 \le l \le r$, we construct a $(1-\alpha)$-confidence interval for the k-th entry of the l-th tensor factor (up to global permutation) as follows

$$\mathsf{CI}_{u_{l,k}}^{1-\alpha} := \big[u_{l,k} \pm \sqrt{(\boldsymbol{\Sigma}_k)_{l,l}} \cdot \Phi^{-1}(1-\alpha/2)\big], \tag{28}$$

  where $\Phi^{-1}(\cdot)$ is the inverse CDF of a standard Gaussian, $[a \pm b] := [a-b, a+b]$, and $\boldsymbol{\Sigma}_k$ is constructed in (26).
- For each $1 \le i, j, k \le d$, we construct a $(1-\alpha)$-confidence interval for the (i,j,k)-th entry of $\boldsymbol{T}^\star$ as follows

$$\mathsf{CI}_{T_{i,j,k}}^{1-\alpha} := \big[T_{i,j,k} \pm \sqrt{v_{i,j,k}} \cdot \Phi^{-1}(1-\alpha/2)\big], \tag{29}$$

  where $v_{i,j,k}$ is constructed in (27).

As it turns out, the proposed (entrywise) confidence intervals are nearly accurate, as revealed by the following theorem. The proof is postponed to Appendix E.

Theorem 4 (Validity of Constructed Confidence Intervals): Instate the assumptions of Theorem 2. There is a permutation $\pi(\cdot) : [d] \mapsto [d]$ such that for any $0 < \alpha < 1$, the confidence interval constructed in (28) obeys

$$\mathbb{P}\left\{u_{\pi(l),k}^\star \in \mathsf{CI}_{u_{l,k}}^{1-\alpha}\right\} = 1-\alpha+o(1), \quad \forall 1 \le l \le r, 1 \le k \le d.$$

In addition, for any $1 \le i, j, k \le d$ obeying (21) and any $0 < \alpha < 1$, the confidence interval constructed in (29) obeys

$$\mathbb{P}\left\{T_{i,j,k}^\star \in \mathsf{CI}_{T_{i,j,k}}^{1-\alpha}\right\} = 1-\alpha+o(1).$$

This theorem justifies the validity of the uncertainty quantification procedure we propose. Several important features are worth emphasizing:

- *"Fine-grained" entrywise uncertainty quantification*. Our results enable trustworthy uncertainty quantification down to quite fine scale, namely, we are capable of assessing the uncertainty reliably at the entrywise level for both the tensor factors and the tensor of interest. To the best of our knowledge, accurate entrywise uncertainty characterization for tensor completion is previously unavailable.
- *Adaptivity to heterogeneous and unknown noise distributions*. The proposed confidence intervals do not require prior knowledge about the noise distributions, and automatically adapt to noise heteroscedasticity (i.e. the case when the noise variance varies across entries). Such model-free and adaptive features are of important practical value.
- *No need of sample splitting*. The whole procedure studied here — including both estimation and uncertainty quantification — does not rely on any sort of data splitting, thus effectively preventing unnecessary information loss due to sample splitting.

Remark 2: In practice, the rank r of the true tensor is often unknown *a priori*, and needs to be estimated first. Fortunately, rank estimation can often be accomplished in a data-driven manner. For instance, under the assumptions imposed in this paper, the largest r eigenvalues of the matrix $\mathcal{P}_{\textsf{off-diag}}(\boldsymbol{A}\boldsymbol{A}^\top)$ are provably much larger than its remaining eigenvalues (see [17], [18]). As a result, one can simply examine the eigenvalues of $\mathcal{P}_{\textsf{off-diag}}(\boldsymbol{A}\boldsymbol{A}^\top)$, and utilize the eigen-gap of this data matrix to obtain a faithful estimate of r.

Lower Bounds: One might naturally wonder whether the proposed confidence intervals can be further improved; concretely, is it possible to identify a shorter confidence interval that remains valid? As it turns out, our procedures are, in some sense, statistically optimal under Gaussian noise, as confirmed by the following fundamental lower bound.

Theorem 5 (Entrywise Lower Bounds): Consider any unbiased estimator $\widehat{\boldsymbol{u}}_l$ for $\boldsymbol{u}_l^\star$ $(1 \le l \le r)$ and any unbiased estimator $\widehat{\boldsymbol{T}}$ for $\boldsymbol{T}^\star$. Suppose that $\{E_{i,j,k}\}$ are i.i.d. Gaussians and that Assumptions 1-2 hold. If $\mu, \kappa, r = O(1)$ and

$$p \ge c_6 \frac{\log^2 d}{d^2}$$

for some sufficiently large constant $c_6 > 0$, then the following holds with probability at least $1 - O\left(d^{-10}\right)$:

$$\begin{aligned} \mathsf{Var}\big[\widehat{u}_{l,k}\big] &\ge (1-o(1))\left(\boldsymbol{\Sigma}_k^\star\right)_{l,l}, \qquad 1 \le k \le d; \\ \mathsf{Var}\big[\widehat{T}_{i,j,k}\big] &\ge (1-o(1))\, v_{i,j,k}^\star, \qquad 1 \le i, j, k \le d. \end{aligned}$$

Taken collectively with Theorems 2 and 3, the above result reveals that our nonconvex estimators $\{\boldsymbol{u}_l\}$ and $\boldsymbol{T}$ achieve minimal mean square estimation errors in a very sharp manner at the entrywise level. Recognizing that the proposed confidence intervals allow for accurate assessment of the uncertainty (by virtue of Theorem 4), we conclude that the proposed inferential procedures are, in some sense, un-improvable under i.i.d. Gaussian noise (including both the rates and the pre-constants).

C. Back to Estimation: ℓ_2 Optimality of Nonconvex Estimates

Thus far, we have established optimality of the estimators $u_{l,k}$ $(1 \le l \le r, 1 \le k \le d)$ and $T_{i,j,k}$ (for those i, j, k obeying (21)) in an entrywise sense. These results taken together allow one to uncover the ℓ_2 optimality of the nonconvex optimization approach as well. Our result is this:

Theorem 6 (Optimality w.r.t. ℓ_2 Estimation Accuracy): Instate the assumptions of Theorem 2. With probability exceeding $1-o(1)$, the estimates returned by Algorithm 1 obey

$$\left\|\boldsymbol{u}_{\pi(l)}-\boldsymbol{u}_l^\star\right\|_2^2=\frac{(2+o(1))\,\sigma_{\max}^2 d}{p\left\|\boldsymbol{u}_l^\star\right\|_2^4},\qquad \forall 1\le l\le r \tag{30a}$$

$$\left\|\boldsymbol{T}-\boldsymbol{T}^\star\right\|_{\mathrm{F}}^2=\frac{(6+o(1))\,\sigma_{\max}^2 dr}{p} \tag{30b}$$

for some permutation $\pi(\cdot):[d]\mapsto[d]$.

In addition, if $\{E_{i,j,k}\}$ are i.i.d. Gaussians, we have the following lower bound:

Theorem 7 (Lower Bound w.r.t ℓ_2 Estimation Accuracy): Instate the assumptions of Theorem 5. The following holds with probability at least $1-O\left(d^{-10}\right)$: any unbiased estimator $\widehat{\boldsymbol{u}}_l$ (resp. $\widehat{\boldsymbol{T}}$) for $\boldsymbol{u}_l^\star$ (resp. $\boldsymbol{T}^\star$) necessarily obeys

$$\mathbb{E}\left[\left\|\widehat{\boldsymbol{u}}_l-\boldsymbol{u}_l^\star\right\|_2^2\right]\ge\frac{(2-o(1))\,\sigma_{\min}^2 d}{p\left\|\boldsymbol{u}_l^\star\right\|_2^4}; \tag{31a}$$

$$\mathbb{E}\left[\left\|\widehat{\boldsymbol{T}}-\boldsymbol{T}^\star\right\|_{\mathrm{F}}^2\right]\ge\frac{(6-o(1))\,\sigma_{\min}^2 dr}{p}. \tag{31b}$$

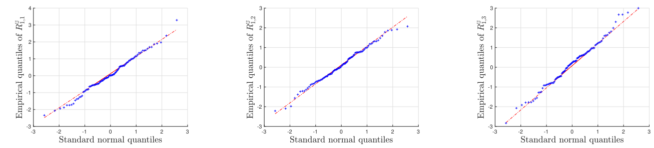
Here, the characterization of the ℓ_2 risk (37a) for $\boldsymbol{u}_l$ is a straightforward consequence of Theorems 1-2, and the lower bounds (31) follow immediately from Theorem 5. Establishing the ℓ_2 risk (30b) for $\boldsymbol{T}$ requires more work, as Theorem 3 is valid only for a set of entries obeying (21). Fortunately, a majority of the entries of $\boldsymbol{T}^\star$ satisfy (21), thus allowing for a nearly accurate approximation of the Euclidean risk of $\boldsymbol{T}$. All in all, Theorems 6-7 deliver an encouraging news: when the noise components are i.i.d. Gaussian, nonconvex optimization is information-theoretically optimal when estimating both the unknown tensor and its underlying tensor factors.

D. Numerical Experiments

To validate our theory and demonstrate the practical applicability of our inferential procedures, we perform a series of numerical experiments for a variety of settings. Specifically, we set $d=100$, $p=0.2$, and generate the ground-truth tensor $\boldsymbol{T}^\star=\sum_{l=1}^r\boldsymbol{u}_l^\star$ in a random fashion such that $\boldsymbol{u}_l^\star\overset{\text{i.i.d.}}{\sim}\mathcal{N}(\boldsymbol{0},\boldsymbol{I}_d)$. Regarding the algorithmic parameters for nonconvex optimization (i.e. Algorithm 1), we choose $L=r^2$, $\epsilon_{\mathsf{th}}=0.4$, $\eta_t\equiv 3\times10^{-5}/p$, and $t_0=100$. The noise components are independently drawn from Gaussian distributions, obeying $E_{i,j,k}\sim\mathcal{N}(0,\sigma_{i,j,k}^2), 1\le i\le j\le k\le d$ with variance $\sigma_{i,j,k}^2$ constructed as follows. We generate $w_{i,j,k}\overset{\text{i.i.d.}}{\sim}\mathsf{Unif}[0,1], 1\le i,j,k\le d$ and let

$$\sigma_{i,j,k}^2=\frac{\sigma^2 w_{i,j,k}^\beta}{\sum_{1\le i\le j\le k\le d}w_{i,j,k}^\beta}\frac{d^3}{6},$$

where β represents the degree of heteroscedasticity. The noise becomes more heteroscedastic as β increases, and setting $\beta=0$ reduces to the homoscedastic case where the noise variances are identical across all entries. In what follows, we set $\beta=5$.



(a) (b) (c)

Fig. 1. Q-Q (quantile-quantile) plots of $R_{1,1}^{\mathsf{U}}$, $R_{1,2}^{\mathsf{U}}$ and $R_{1,3}^{\mathsf{U}}$ vs. a standard Gaussian distribution (where $r=4$, $p=0.2$, $\sigma=0.1$ and $\beta=5$).

TABLE I

EMPIRICAL COVERAGE RATES OF TENSOR FACTOR ENTRIES FOR VARYING r AND σ

(r,σ)	$\mathsf{Mean(CR)}$	$\mathsf{Std(CR)}$
$(2,10^{-2})$	0.9481	0.0201
$(2,10^{-1})$	0.9477	0.0228
$(2,1)$	0.9478	0.0215
$(4,10^{-2})$	0.9450	0.0218
$(4,10^{-1})$	0.9472	0.0231
$(4,1)$	0.9462	0.0234

a) Tensor factor entries: We begin with inference for the entries of the tensor factors of interest. Consider the construction of 95% confidence intervals (i.e. $\alpha=0.05$). Define the normalized estimation error as follows

$$R_{l,k}^{\mathsf{U}}:=\frac{1}{\sqrt{(\boldsymbol{\Sigma}_k)_{l,l}}}\left(u_{l,k}-u_{l,k}^\star\right),\qquad 1\le l\le r,\ 1\le k\le d.$$

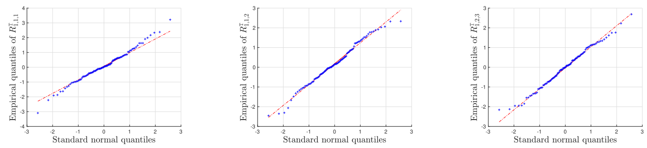
For each $1\le l\le r$ and $1\le k\le d$, we denote by $\mathsf{CR}_{l,k}$ the empirical coverage rate for the tensor factor entry $u_{l,k}^\star$ over 100 independent trials. Let $\mathsf{Mean(CR)}$ (resp. Std $\mathsf{(CR)}$) denote the average (resp. the standard deviation) of $\{\mathsf{CR}_{l,k}\}$ over all tensor factor entries. Figure 1 displays the Q-Q (quantile-quantile) plots of $R_{1,1}^{\mathsf{U}}$ vs. a standard Gaussian random variable, and Table I summarizes the numerical results for varying p,r and σ. Encouragingly, the empirical coverage rates are all very close to 95%, and the empirical distributions of the normalized estimation errors are all well approximated by a standard Gaussian distribution.

b) Tensor entries: Next, we turn to inference for tensor entries. Similar to the above case, we intend to construct 95% confidence intervals. Define

$$R_{i,j,k}^{\mathsf{T}}:=\frac{1}{\sqrt{v_{i,j,k}}}\left(T_{i,j,k}-T_{i,j,k}^\star\right),\qquad 1\le i\le j\le k\le d.$$

For each $1\le i\le j\le k\le d$, we record the empirical coverage rate $\mathsf{CR}_{i,j,k}$ for the tensor entry $T_{i,j,k}^\star$ over 100 Monte Carlo trials. Denote by $\mathsf{Mean(CR)}$ (resp. $\mathsf{Std(CR)}$) the average (resp. the standard deviation) of $\{\mathsf{CR}_{i,j,k}\}$ over entries $1\le i\le j\le k\le d$. Figure 2 depicts the Q-Q (quantile-quantile) plots of $R_{1,1,1}^{\mathsf{T}}$, $R_{1,1,2}^{\mathsf{T}}$ and $R_{1,2,3}^{\mathsf{T}}$ vs. a standard Gaussian random variable. Table II collects the numerical results $\mathsf{Mean(CR)}$ and $\mathsf{Std(CR)}$ for a variety of settings. Similar to previous experiments, the confidence intervals and the Q-Q plots match our theoretical prediction in a reasonably well manner.

c) ℓ_2 estimation accuracy: Finally, let us verify the Euclidean estimation guarantees we develop for Algorithm 1. Figure 3 plots the Euclidean estimation errors of the tensor

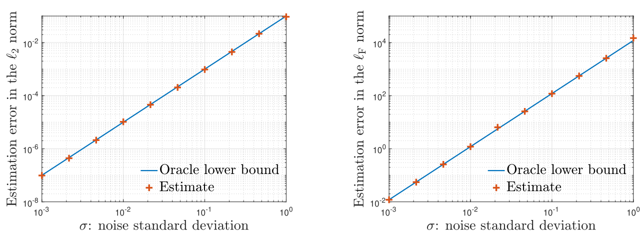


(a) (b) (c)

Fig. 2. Q-Q (quantile-quantile) plots of $R^{\mathsf{T}}_{1,1,1}$, $R^{\mathsf{T}}_{1,1,2}$ and $R^{\mathsf{T}}_{1,2,3}$ vs. a standard Gaussian distribution (where $r = 4$, $p = 0.2$, $\sigma = 0.1$ and $\beta = 5$).

TABLE II
EMPIRICAL COVERAGE RATES OF TENSOR ENTRIES FOR DIFFERENT r AND σ

(r, σ)	Mean(CR)	Std(CR)
$(2, 10^{-2})$	0.9494	0.0218
$(2, 10^{-1})$	0.9513	0.0218
$(2, 1)$	0.9475	0.0222
$(4, 10^{-2})$	0.9434	0.0225
$(4, 10^{-1})$	0.9494	0.0220
$(4, 1)$	0.9494	0.0219



(a) (b)

Fig. 3. (a) ℓ_2 estimation error of $\boldsymbol{u}_1$ vs. the Cramér–Rao lower bound; (b) Euclidean estimation errors of $\boldsymbol{T}$ vs. the Cramér–Rao lower bound (where $r = 4$, $p = 0.2$ and $\beta = 0$).

factor estimate $\boldsymbol{u}_1$ (resp. the tensor estimate $\boldsymbol{T}$). In this series of experiments, we focus on the homoskedastic case, i.e. $\beta = 0$. As one can see, the empirical ℓ_2 risks are exceedingly close to the Cramér–Rao lower bounds supplied in Theorem 6.

IV. PRIOR ART

Much progress has been made in the past few years towards understanding and solving low-rank tensor completion. Inspired by the success of convex relaxation for matrix completion [23], [24], [25], [26], [27], an estimate based on tensor nuclear norm minimization was proposed by [28] and [29], which enables information-theoretically optimal sample complexity. Unfortunately, the tensor nuclear norm is itself NP-hard to compute and hence computationally infeasible in practice. To allow for more economical algorithms, a widely adopted strategy is to unfold the tensor data into a matrix [11], [13], [30], [31], thus transforming it into a low-rank matrix completion problem [23], [32], [33]. However, unfolding a third-order tensor often leads to an extremely unbalanced matrix, thereby resulting in sub-optimal sample complexity when directly invoking matrix completion theory. To address this issue, a recent line of work [19], [20] suggested the use of sum-of-squares (SOS) hierarchy, which performs convex relaxation after lifting the data into higher dimension. The SOS-based algorithms achieve a sample complexity on the order of $rd^{3/2}$ for third-order tensors, which is widely conjectured to be optimal among all polynomial-time algorithms. However, despite their polynomial-time complexity, the SOS-based methods remain too expensive for solving large-scale practical problems, primarily due to the lifting operation.

Motivated by the above computational concerns, several nonconvex approaches have been developed, which often consist of two stages: (1) finding a rough initialization; (2) local refinement. Existing initialization schemes include unfolding-based spectral methods [2], [16], [17], [18], [22], [34], [35]. tensor power methods [21], tensor SVD [36], and so on. To improve the estimation accuracy, the local refinement stage invokes nonconvex optimization algorithms like alternating minimization [21], [22], gradient descent [2], [7], manifold-based optimization [35], block coordinate decent [37], etc. These were motivated in part by the effectiveness of nonconvex optimization in solving nonconvex low-complexity problems [32], [38], [39], [40], [41], [42], [43], [44], [45], [46], [47], [48], [49], [50], [51], [52], [53], [54], [55], [56], [57], [58], [59], [60]; see an overview of recent development in [33]. Various statistical and computational guarantees have been provided for these algorithms, all of which have been shown to run in polynomial time. In particular, (unfolding-based) spectral initialization followed by gradient descent converges linearly to an accuracy that is within a logarithmic factor from optimal [2]. It is also worth noting that the leave-one-out analysis framework adopted herein has been adopted in obtaining sharp analysis of these nonconvex optimization algorithms (e.g., [2], [18], [45], [47], [61], [62]) as well as other statistical and learning problems (e.g., [63], [64], [65], [66], [67], [68], [69]).

However, none of the above results suggested how to evaluate the uncertainty of the resulting estimates in a meaningful way. Despite a large body of work on statistical estimation for noisy tensor completion, it remains completely unclear how to exploit existing results to construct valid yet short confidence intervals for the unknown tensor. Perhaps the work closest to the current paper is inference and uncertainty quantification for noisy matrix completion and matrix denoising [70], [71], [72], which enables optimal construction of confidence intervals on the basis of nonconvex matrix completion algorithms. Inference for singular subspaces has also been investigated under both low-rank matrix regression and denoising settings [73], [74]. While these results might potentially be applicable here by first matricizing the data, the resulting sample complexity, as discussed above, could be pessimistic. Recently, inference for low-rank tensor models has been studied including Tucker low-rank tensor PCA [75], [76] and tensor regression [75]. However, the results therein did not consider the effect of missing data. Finally, we remark that construction of confidence intervals has been extensively studied in a variety of high-dimensional sparse estimation settings [77], [78], [79], [80], [81], [82], [83], [84], [85], [86], [87], [88]. Both the inferential approaches and the analysis techniques therein, however, are drastically different from the ones employed to perform inference for either tensor completion or matrix completion.

Finally, we would like to note that our results have been presented in part in ICML 2020 [1]. In comparison to [1] which was restricted to the case where $r, \mu = O(1)$, the current paper accommodates the more general case where both r and μ are allowed to grow with the problem dimension d. In addition, the current paper includes full analysis details that were not included in the short ICML version.

V. Analysis

This section outlines the proof for our main theorems.

A. A Set of More General Theorems

We begin by presenting a set of more general theorems that allow both r and μ to grow with the dimension d. As can be straightforwardly verified, the theorems stated below subsume as special cases the main theorems presented in Section III.

Theorem 8 (Distributional Guarantees for Tensor Factor Estimates (Gaussian Noise, General (r,μ))): Suppose that the $E_{i,j,k}$'s are Gaussian, and that Assumptions 1-2 hold. Assume that $\kappa \asymp 1$, and that $t_0 = c_0 \log d$,

$$p \geq c_1 \frac{\mu^4 r^4 \log^5 d}{d^{3/2}}, \quad \frac{c_2}{d^{100}} \leq \frac{\sigma_{\max}}{\lambda_{\min}^\star} \leq \frac{c_3\sqrt{p}}{\mu r^{3/2} d^{3/4}\log^2 d} \tag{32a}$$

$$r \leq c_4 \left(\frac{d}{\mu^6 \log^6 d}\right)^{1/6} \tag{32b}$$

hold for some sufficiently large (resp. small) constant $c_0, c_1, c_2 > 0$ (resp. $c_3, c_4 > 0$). Then with probability at least $1 - o(1)$, one has the following decomposition:

$$\boldsymbol{U}\boldsymbol{\Pi} - \boldsymbol{U}^\star = \boldsymbol{Z} + \boldsymbol{W},$$

where $\boldsymbol{\Pi}$ is defined in (16), $\|\boldsymbol{W}\|_{2,\infty} = o\Big(\frac{\sigma_{\min}}{\lambda_{\max}^{\star 2/3}\sqrt{p}}\Big)$, and for any $1 \leq k \leq d$, $\boldsymbol{Z}_{k,:} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_k^\star)$ with covariance matrix $\boldsymbol{\Sigma}_k^\star$ defined in (20).

Remark 3: Theorem 8 subsumes Theorem 1 as a special case.

Theorem 9 (Distributional Guarantees for Tensor Factor Estimates (General Noise, General (r,μ))): Suppose that Assumption 1 holds, and that $\{E_{i,j,k}\}$ are not necessarily Gaussian but satisfy Assumption 2. Then under the condition (32), the decomposition in Theorem 8 continues to hold, except that $\boldsymbol{Z}$ is not necessarily Gaussian but instead obeys

$$\left|\mathbb{P}\big\{\boldsymbol{Z}_{k,:} \in \mathcal{A}\big\} - \mathbb{P}\left\{\boldsymbol{g}_k \in \mathcal{A}\right\}\right| = o\left(1\right), \qquad 1 \leq k \leq d$$

for any convex set $\mathcal{A} \subset \mathbb{R}^r$. Here, $\boldsymbol{g}_k \sim \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{\Sigma}_k^\star\right)$ with covariance matrix $\boldsymbol{\Sigma}_k^\star$ defined in (20).

Remark 4: Theorem 9 subsumes Theorem 2 as a special case.

Theorem 10 (Distributional Guarantees for Tensor Entry Estimates (General (r,μ))): Instate the assumptions of Theorem 9. Consider any $1 \leq i \leq j \leq k \leq d$ obeying

$$\omega_{i,j,k} := \big\|\widetilde{\boldsymbol{U}}^\star_{(j,k),:}\big\|_2 + \big\|\widetilde{\boldsymbol{U}}^\star_{(i,j),:}\big\|_2 + \big\|\widetilde{\boldsymbol{U}}^\star_{(i,k),:}\big\|_2 \tag{33}$$

$$> c_5 \frac{\sigma_{\max}}{\lambda_{\min}^\star}\sqrt{\frac{\mu^5 r^3 \log^3 d}{dp}}\lambda_{\max}^{\star 2/3} \tag{34}$$

for some large constant $c_5 > 0$, with $\widetilde{\boldsymbol{U}}^\star$ defined in (17). Then the estimate $\boldsymbol{T}$ defined in (15) obeys

$$\sup_{\tau \in \mathbb{R}} \left|\mathbb{P}\Big\{T_{i,j,k} \leq T_{i,j,k}^\star + \tau\sqrt{v_{i,j,k}^\star}\Big\} - \Phi(\tau)\right| = o\left(1\right),$$

where $\Phi(\cdot)$ is the CDF of a standard Gaussian random variable. Here, the variance parameters $\{v_{i,j,k}^\star\}$ are defined in (23).

Remark 5: Theorem 10 subsumes Theorem 3 as a special case.

Theorem 11 (Validity of Confidence Intervals (General (r,μ))): Instate the assumptions of Theorem 9. There exists a permutation $\pi(\cdot) : [d] \mapsto [d]$ such that for any $0 < \alpha < 1$, the confidence interval constructed in (28) obeys

$$\mathbb{P}\left\{u_{\pi(l),k}^\star \in \mathsf{CI}_{u_{l,k}}^{1-\alpha}\right\} = 1-\alpha+o\left(1\right), \quad \forall 1 \leq l \leq r,\ 1 \leq k \leq d.$$

In addition, for any $1 \leq i,j,k \leq d$ obeying (34) and any $0 < \alpha < 1$, the confidence interval constructed in (29) obeys

$$\mathbb{P}\left\{T_{i,j,k}^\star \in \mathsf{CI}_{T_{i,j,k}}^{1-\alpha}\right\} = 1 - \alpha + o\left(1\right).$$

Remark 6: Theorem 11 subsumes Theorem 4 as a special case.

Theorem 12 (Entrywise Lower Bounds (General (r,μ))): Consider any unbiased estimator $\widehat{\boldsymbol{u}}_l$ for $\boldsymbol{u}_l^\star$ $(1 \leq l \leq r)$ and any unbiased estimator $\widehat{\boldsymbol{T}}$ for $\boldsymbol{T}^\star$. Suppose that $\{E_{i,j,k}\}$ are i.i.d. Gaussians and that Assumptions 1-2 hold. Assume that $\kappa \asymp 1$ and that

$$p \geq c_6 \frac{\mu^2 r \log^2 d}{d^2} \qquad r \leq c_7\sqrt{\frac{d}{\mu \log d}}$$

hold for some sufficiently large (resp. small) constant $c_6 > 0$ (resp. $c_7 > 0$). Then the following holds with probability at least $1 - O(d^{-10})$:

$$\mathsf{Var}\big[\widehat{u}_{l,k}\big] \geq \left(1 - o\left(1\right)\right)\left(\boldsymbol{\Sigma}_k^\star\right)_{l,l}, \qquad 1 \leq k \leq d;$$
$$\mathsf{Var}\big[\widehat{T}_{i,j,k}\big] \geq \left(1 - o\left(1\right)\right) v_{i,j,k}^\star, \qquad 1 \leq i,j,k \leq d.$$

Remark 7: Theorem 12 subsumes Theorem 5 as a special case.

Theorem 13 (Optimality w.r.t. ℓ_2 Estimation Accuracy (General (r,μ))): Instate the assumptions of Theorem 9. With probability exceeding $1 - o(1)$, the estimates returned by Algorithm 1 obey

$$\left\|\boldsymbol{u}_{\pi(l)} - \boldsymbol{u}_l^\star\right\|_2^2 = \frac{\left(2 + o\left(1\right)\right)\sigma_{\max}^2 d}{p\left\|\boldsymbol{u}_l^\star\right\|_2^4}, \qquad \forall 1 \leq l \leq r$$

$$\|\boldsymbol{T} - \boldsymbol{T}^\star\|_{\mathrm{F}}^2 = \frac{\left(6 + o\left(1\right)\right)\sigma_{\max}^2 dr}{p}$$

for some permutation $\pi(\cdot) : [d] \mapsto [d]$.

Remark 8: Theorem 13 subsumes Theorem 6 as a special case.

Theorem 14 (Lower Bound w.r.t ℓ_2 Estimation Accuracy (General (r,μ))): Instate the assumptions of Theorem 12. The following holds with probability at least $1 - O\left(d^{-10}\right)$: any unbiased estimator $\widehat{\boldsymbol{u}}_l$ (resp. $\widehat{\boldsymbol{T}}$) for $\boldsymbol{u}_l^\star$ (resp. $\boldsymbol{T}^\star$)

necessarily obeys

$$\mathbb{E}\left[\left\|\widehat{\boldsymbol{u}}_l - \boldsymbol{u}_l^\star\right\|_2^2\right] \geq \frac{(2-o(1))\,\sigma_{\min}^2 d}{p\left\|\boldsymbol{u}_l^\star\right\|_2^4};$$

$$\mathbb{E}\left[\left\|\widehat{\boldsymbol{T}} - \boldsymbol{T}^\star\right\|_{\mathrm{F}}^2\right] \geq \frac{(6-o(1))\,\sigma_{\min}^2 dr}{p}.$$

Remark 9: Theorem 14 subsumes Theorem 7 as a special case.

The rest of this section is dedicated to establishing Theorems 8-11. The proof of Theorems 12 and 14 (resp. Theorem 13) is deferred to Appendix G (resp. Appendix F). Before continuing, we introduce several notation for simplicity of presentation. First, we rescale the loss function as follows

$$g(\boldsymbol{U}) \coloneqq \frac{1}{6p} f(\boldsymbol{U}) = \frac{1}{6p}\left\|\mathcal{P}_\Omega\Big(\sum_{i=1}^r \boldsymbol{u}_i^{\otimes 3} - \boldsymbol{T}\Big)\right\|_{\mathrm{F}}^2$$

throughout the rest of the paper. By defining $\widetilde{\boldsymbol{U}} \coloneqq \left[\boldsymbol{u}_l^{\otimes 2}\right]_{1\leq l\leq r} \in \mathbb{R}^{d^2\times r}$ as before, we can express the gradient of $g(\boldsymbol{U})$ as follows

$$\nabla g(\boldsymbol{U}) = \Big[p^{-1}\mathcal{P}_\Omega\Big(\sum_{i=1}^r \boldsymbol{u}_i^{\otimes 3} - \boldsymbol{T}^\star - \boldsymbol{E}\Big) \times_1 \boldsymbol{u}_l \times_2 \boldsymbol{u}_l\Big]_{1\leq l\leq d}$$
$$= \mathsf{unfold}\Big(p^{-1}\mathcal{P}_\Omega\Big(\sum_{i=1}^r \boldsymbol{u}_i^{\otimes 3} - \boldsymbol{T}^\star - \boldsymbol{E}\Big)\Big)\widetilde{\boldsymbol{U}}, \qquad (35)$$

where we recall the tensor vector products $\times_1$ and $\times_2$ are both defined in Section I-E, and $\mathsf{unfold}(\cdot)$ denotes the mode-3 matricization of a three-order tensor. Here and throughout, for any matrix $\boldsymbol{A} = [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_r] \in \mathbb{R}^{d\times r}$, we denote

$$\widetilde{\boldsymbol{A}} \coloneqq [\boldsymbol{a}_1 \otimes \boldsymbol{a}_1, \ldots, \boldsymbol{a}_r \otimes \boldsymbol{a}_r] \in \mathbb{R}^{d^2\times r}, \qquad (36)$$

where for any $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^d$, we let $\boldsymbol{a}\otimes\boldsymbol{b} \coloneqq \begin{bmatrix} a_1\boldsymbol{b} \\ \vdots \\ a_d\boldsymbol{b}\end{bmatrix} \in \mathbb{R}^{d^2}$.

In addition, we define an event $\mathcal{E}$ on which several important properties (59)-(63) (which we defer to Appendix B to streamline presentation) hold. In what follows, we shall primarily work with this event $\mathcal{E}$, which happens with probability exceeding $1-o(1)$ as guaranteed by Lemmas 11-14 in Appendix B.

Remark 10 (Sub-Optimality in r and κ): Before moving on to the analysis details, we briefly remark on the dependency of our theory on the rank r and the condition number κ of the ground truth. To begin with, our theory is only optimal when the rank r is a constant independent of the ambient dimension d, and becomes loose if r is allowed to grow with d. This suboptimality arises since the state-of-the-art analysis framework for nonconvex algorithms remains suboptimal in this aspect [2], [22], [35]. For instance, our local convergence analysis relies heavily on (restricted) strong convexity and smoothness of the objective function, but the present theory is only able to establish these desirable geometric properties within a neighborhood around the truth of radius $o(1/r)$. Improving the rank dependency requires more refined analysis ideas, which forms an important future direction.

Remark 11 (Extension to Asymmetric Tensors): In this paper, we focus on symmetric tensors for simplicity of presentation. The statistical inferential procedure can be generalized to accommodate the asymmetric case. Briefly speaking, one can run Algorithms 7-8 in [2] to obtain nonconvex estimates for all tensor factors of an asymmetric tensor (which enjoy similar theoretical guarantees as those in the symmetric case). Then one can apply the same analysis framework in Section V to characterize their distributions and construct the corresponding confidence intervals in the same manner. The interested readers are referred to [2, Appendix E] for more details about how to extend the estimation algorithms to handle asymmetric tensors.

Remark 12 (Extension to High-Order Tensors): Finally, let us briefly discuss how to generalize the algorithmic idea proposed herein to accommodate higher-order tensors. In fact, one can see from the analysis in Section V that the key for accomplishing the task of uncertainty quantification lies in deriving entrywise distributional guarantees for the nonconvex estimates of tensor factors. Therefore, the primary step reduces to how to adapt the nonconvex algorithm here to recover an order-k symmetric tensor $\boldsymbol{T}^\star$ with low rank ($r \leq d$). Towards this end, one can first unfold the observed tensor into a $d\times d^{k-1}$ matrix and invoke the spectral initialization algorithm as Algorithm 2 to obtain the estimates of the subspace spanned by the tensor factors. Then we can generate random vectors from the estimated subspace and use them to project the observed tensor onto the matrix space, and then apply the spectral method to retrieve all tensor factors as in Algorithm 3. After obtaining the initial estimates, we can run gradient descent to further refine them, and show that the estimation error of the nonconvex estimates are statistically optimal in both ℓ_2 and ℓ_∞ senses. Once these theoretical guarantees are in place, the inferential procedure is exactly the same as that in the order-3 case. In addition, one can repeat the analysis framework as that in Theorem 6 to show that with high probability, the noconvex estimates satisfy

$$\min_{s=\pm 1}\left\|\boldsymbol{u}_{\pi(l)} - s\boldsymbol{u}_l^\star\right\|_2^2 = \frac{(2+o(1))\,\sigma_{\max}^2 d}{p\left\|\boldsymbol{u}_l^\star\right\|_2^4}, \qquad \forall 1\leq l\leq r \qquad (37a)$$

$$\left\|\boldsymbol{T} - \boldsymbol{T}^\star\right\|_{\mathrm{F}}^2 = \frac{(2k+o(1))\,\sigma_{\max}^2 dr}{p} \qquad (37b)$$

for some permutation $\pi(\cdot) : [d] \mapsto [d]$. Here, it is worth noting that when the tensor order is even, one can only hope to retrieve tensor factors up to global signs. It is also straightforward to apply the same argument as in Theorem 7 to demonstrate that the estimation error above matches the information-theoretic lower bound.

B. Proof Outline for the Distributional Theory

We now outline the proof strategy for our distributional theory, namely, Theorems 8-10.

1) Distributional Theory for Tensor Factors: Recall the definition $\widetilde{\boldsymbol{U}} \coloneqq [\boldsymbol{u}_1^{\otimes 2}, \cdots, \boldsymbol{u}_r^{\otimes 2}] \in \mathbb{R}^{d^2\times r}$. We start by making note of the following crucial decomposition of $\boldsymbol{U}\boldsymbol{\Pi}$:

$$\begin{aligned}\boldsymbol{U}\boldsymbol{\Pi} = \; & \mathsf{unfold}\left(p^{-1}\mathcal{P}_{\Omega}\left(\boldsymbol{E}\right)\right)\widetilde{\boldsymbol{U}}\boldsymbol{\Pi}\big((\widetilde{\boldsymbol{U}}\boldsymbol{\Pi})^{\top}\widetilde{\boldsymbol{U}}\boldsymbol{\Pi}\big)^{-1} \\ & + \boldsymbol{U}^{\star}\widetilde{\boldsymbol{U}}^{\star\top}\widetilde{\boldsymbol{U}}\boldsymbol{\Pi}\big((\widetilde{\boldsymbol{U}}\boldsymbol{\Pi})^{\top}\widetilde{\boldsymbol{U}}\boldsymbol{\Pi}\big)^{-1} \\ & + \mathsf{unfold}\left(\left(\mathcal{I}-p^{-1}\mathcal{P}_{\Omega}\right)\left(\boldsymbol{T}-\boldsymbol{T}^{\star}\right)\right)\widetilde{\boldsymbol{U}}\boldsymbol{\Pi}\big((\widetilde{\boldsymbol{U}}\boldsymbol{\Pi})^{\top}\widetilde{\boldsymbol{U}}\boldsymbol{\Pi}\big)^{-1} \\ & + \nabla g(\boldsymbol{U})\big(\widetilde{\boldsymbol{U}}^{\top}\widetilde{\boldsymbol{U}}\big)^{-1}\boldsymbol{\Pi}, \end{aligned} \tag{38}$$

where $\boldsymbol{\Pi}$ is defined in (16), $\mathcal{I}$ stands for the identity operator, and $\nabla g(\boldsymbol{U})$ is given in (35). As a result, we arrive at the following key decomposition

$$\boldsymbol{U}\boldsymbol{\Pi}-\boldsymbol{U}^{\star} = \underbrace{\mathsf{unfold}\big(p^{-1}\mathcal{P}_{\Omega}(\boldsymbol{E})\big)\widetilde{\boldsymbol{U}}^{\star}\big(\widetilde{\boldsymbol{U}}^{\star\top}\widetilde{\boldsymbol{U}}^{\star}\big)^{-1}}_{=:\boldsymbol{X}} + \sum_{1\leq i\leq 4}\boldsymbol{W}_i, \tag{39}$$

where the $\boldsymbol{W}_i$'s are given by

$$\boldsymbol{W}_1 \coloneqq \boldsymbol{U}^{\star}\big(\widetilde{\boldsymbol{U}}^{\star\top}\widetilde{\boldsymbol{U}}\boldsymbol{\Pi}\big((\widetilde{\boldsymbol{U}}\boldsymbol{\Pi})^{\top}\widetilde{\boldsymbol{U}}\boldsymbol{\Pi}\big)^{-1} - \boldsymbol{I}_r\big); \tag{40a}$$

$$\begin{aligned}\boldsymbol{W}_2 \coloneqq \; & \mathsf{unfold}\big(p^{-1}\mathcal{P}_{\Omega}(\boldsymbol{E})\big) \\ & \cdot\Big(\widetilde{\boldsymbol{U}}\boldsymbol{\Pi}\big((\widetilde{\boldsymbol{U}}\boldsymbol{\Pi})^{\top}\widetilde{\boldsymbol{U}}\boldsymbol{\Pi}\big)^{-1} - \widetilde{\boldsymbol{U}}^{\star}(\widetilde{\boldsymbol{U}}^{\star\top}\widetilde{\boldsymbol{U}}^{\star})^{-1}\Big);\end{aligned} \tag{40b}$$

$$\boldsymbol{W}_3 \coloneqq \mathsf{unfold}\big((\mathcal{I}-p^{-1}\mathcal{P}_{\Omega})(\boldsymbol{T}-\boldsymbol{T}^{\star})\big)\widetilde{\boldsymbol{U}}\boldsymbol{\Pi}\big((\widetilde{\boldsymbol{U}}\boldsymbol{\Pi})^{\top}\widetilde{\boldsymbol{U}}\boldsymbol{\Pi}\big)^{-1}; \tag{40c}$$

$$\boldsymbol{W}_4 \coloneqq \nabla g(\boldsymbol{U})\big(\widetilde{\boldsymbol{U}}^{\top}\widetilde{\boldsymbol{U}}\big)^{-1}\boldsymbol{\Pi}. \tag{40d}$$

Remark 13: Let us make a few remarks about the quantities in the key decomposition of $\boldsymbol{U}\boldsymbol{\Pi}$ in (38) and (39). The first term $\boldsymbol{X}$ in (39) consists of a collection of sub-Gaussian random variables, which leads to the main uncertainty term. In comparison, $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ are both expected to be small when the estimate $\boldsymbol{U}$ is close to the ground truth $\boldsymbol{U}^{\star}$. The third term $\boldsymbol{W}_3$ is concerned with the missing data, which disappears when we have full observations (i.e. the case with $p = 1$). Finally, the last term $\boldsymbol{W}_4$ is related to the gradient of our nonconvex estimate $\boldsymbol{U}$, which is also vanishingly small due to the contraction property of the Eucilidean norm of the gradient of a locally strongly convex and smooth loss function.

In what follows, we shall demonstrate through a set of auxiliary lemmas that $\boldsymbol{U}\boldsymbol{\Pi} - \boldsymbol{U}^{\star}$ is approximately characterized by the term $\boldsymbol{X}$ defined in (39). More specifically,

- Lemma 1 reveals that, under Gaussian noise, each row of $\boldsymbol{X}$ is approximately a Gaussian random vector.
- Lemma 2 extends the above (approximate) normality result to the case with non-Gaussian noise.
- Lemmas 3-6 deliver upper bounds on the $\ell_{2,\infty}$ norms of the remaining quantities $\boldsymbol{W}_1$, $\boldsymbol{W}_2$, $\boldsymbol{W}_3$ and $\boldsymbol{W}_4$, respectively (in particular, they are provably negligible compared to the typical size of each row of $\boldsymbol{X}$).

Theorems 8-9 then follow immediately by combining Lemmas 1-6.

Lemma 1: Instate the assumptions of Theorem 8. Conditional on the event $\mathcal{E}$ where (59)-(63) hold, with probability at least $1 - O\left(d^{-10}\right)$ we can decompose $\boldsymbol{X} = \boldsymbol{Z} + \boldsymbol{W}_0$ such that (i) for any $1 \leq k \leq d$, $\boldsymbol{Z}_{k,:} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_k^{\star})$ with covariance matrix $\boldsymbol{\Sigma}_k^{\star}$ defined in (20), and (ii)

$$\|\boldsymbol{W}_0\|_{2,\infty} \lesssim \frac{\sigma_{\max}}{\lambda_{\min}^{\star 2/3}\sqrt{p}}\left\{\frac{\mu r\log^2 d}{d\sqrt{p}} + \sqrt{\frac{\mu r \log d}{d}}\right\}. \tag{41}$$

Proof: See Appendix C-A. □

Lemma 2: Instate the assumptions of Theorem 9. Conditional on the event $\mathcal{E}$ where (59)-(63) hold, with probability at least $1 - O\left(d^{-10}\right)$, the decomposition $\boldsymbol{X} = \boldsymbol{Z} + \boldsymbol{W}_0$ in Lemma 1 and (41) continues to hold, except that $\boldsymbol{Z}$ is not necessarily Gaussian but instead obeys

$$\left|\mathbb{P}\left\{\boldsymbol{Z}_{k,:} \in \mathcal{A}\right\} - \mathbb{P}\{\boldsymbol{g}_k \in \mathcal{A}\}\right| \lesssim \frac{\mu r^{3/2}}{\sqrt{d^{3/2}p}}$$

for any convex set $A \subset \mathbb{R}^d$. Here, $\boldsymbol{g}_k \sim \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{\Sigma}_k^{\star}\right)$ with covariance matrix $\boldsymbol{\Sigma}_k^{\star}$ defined in (20).

Proof: See Appendix C-B. □

Remark 14: We pause to make an observation regarding the statistical dependency across the rows of $\boldsymbol{Z}$. As will be made clear in the proof, the randomness of $\boldsymbol{Z}_{k,:}$ arises from the random components $\{E_{i,j,k}\}_{1\leq i<j\leq d}$. Given that $\boldsymbol{E}$ is assumed to be a symmetric tensor (in the sense that $E_{i,j,k} = E_{j,i,k} = E_{i,k,j} = E_{j,k,i} = E_{k,i,j} = E_{k,j,i}$ for any $1 \leq i,j,k \leq d$), one can easily see that: for any $k \neq l$, $\boldsymbol{Z}_{k,:}$ and $\boldsymbol{Z}_{l,:}$ are not statistically independent and instead share some source of randomness. Fortunately, each $\boldsymbol{Z}_{k,:}$ is a sum of around d^2 independent random vectors, whereas the number of overlapping terms between $\boldsymbol{Z}_{k,:}$ and $\boldsymbol{Z}_{l,:}$ is only on the order of d and hence accounts for a vanishingly small fraction of the total amount of randomness. Therefore, this allows one to justify that $\boldsymbol{Z}_{k,:}$ and $\boldsymbol{Z}_{l,:}$ are "nearly" statistically independent for any $k \neq l$.

Lemma 3: Instate the assumptions of Theorem 9. Conditional on the event $\mathcal{E}$ where (59)-(63) hold, the matrix $\boldsymbol{W}_1$ defined in (40a) obeys

$$\|\boldsymbol{W}_1\|_{2,\infty} \lesssim \frac{\sigma_{\max}}{\lambda_{\min}^{\star 2/3}\sqrt{p}}\cdot\zeta \tag{42}$$

with probability at least $1 - O\left(d^{-10}\right)$, where ζ is defined as

$$\begin{aligned}\zeta \coloneqq \; & \frac{\mu^2 r^2\log^{7/2} d}{d^{3/2}p} + \frac{\mu^2 r^2 \log^3 d}{d\sqrt{p}} + \sqrt{\frac{\mu^2 r^2 \log d}{d}} \\ & + \frac{\sigma_{\max}}{\lambda_{\min}^{\star}}\sqrt{\frac{\mu r^3 d\log^2 d}{p}}\end{aligned} \tag{43}$$

Proof: See Appendix C-C. □

Lemma 4: Instate the assumptions of Theorem 9. Conditional on the event $\mathcal{E}$ where (59)-(63) hold, the matrix $\boldsymbol{W}_2$ defined in (40b) obeys

$$\|\boldsymbol{W}_2\|_{2,\infty} \lesssim \frac{\sigma_{\max}}{\lambda_{\min}^{\star 2/3}\sqrt{p}}\frac{\sigma_{\max}}{\lambda_{\min}^{\star}}\sqrt{\frac{\mu r^2 d\log d}{p}}$$

with probability at least $1 - O\left(d^{-10}\right)$.

Proof: See Appendix C-D. □

Lemma 5: Instate the assumptions of Theorem 9. Conditional on the event $\mathcal{E}$ where (59)-(63) hold, the matrix $\boldsymbol{W}_3$

defined in (40c) obeys

$$\|\boldsymbol{W}_3\|_{2,\infty} \lesssim \frac{\sigma_{\max}}{\lambda_{\min}^{\star 2/3}\sqrt{p}}\sqrt{\frac{\mu^3 r^2 \log^2 d}{d^2 p}}$$

with probability at least $1 - O\left(d^{-10}\right)$.

Proof: See Appendix C-E. □

Lemma 6: Instate the assumptions of Theorem 9. Conditional on the event $\mathcal{E}$ where (59)-(63) hold, the matrix $\boldsymbol{W}_4$ defined in (40d) obeys

$$\|\boldsymbol{W}_4\|_{2,\infty} \lesssim \frac{\sigma_{\max}}{\lambda_{\min}^{\star 2/3}\sqrt{p}}\frac{1}{\sqrt{d}}$$

with probability at least $1 - O\left(d^{-10}\right)$.

Proof: See Appendix C-F. □

Before we move on to the distributions of the tensor entries, we make note of the following observation that will play a useful role later. Define

$$\boldsymbol{W} := \boldsymbol{W}_0 + \boldsymbol{W}_1 + \boldsymbol{W}_2 + \boldsymbol{W}_3 + \boldsymbol{W}_4. \tag{44}$$

Taking Lemmas 1-6 collectively, we obtain

$$\|\boldsymbol{W}\|_{2,\infty} \lesssim \zeta \frac{\sigma_{\max}}{\lambda_{\min}^{\star 2/3}\sqrt{p}} = o\left(\frac{\sigma_{\max}}{\lambda_{\min}^{\star 2/3}\sqrt{p}}\right), \tag{45}$$

where ζ is defined in (43). The last relation holds true due to our assumptions (32) on the sample size, the noise level, and the rank.

2) Distributional Theory for Tensor Entries: As it turns out, the theoretical guarantees for the tensor factors enable us to characterize the distribution of tensor entries. Towards this, let us first define

$$\boldsymbol{\Delta} := \boldsymbol{U}\boldsymbol{\Pi} - \boldsymbol{U}^\star, \qquad \widetilde{\boldsymbol{\Delta}} := \left[\boldsymbol{\Delta}_l^{\otimes 2}\right]_{1 \le l \le r} \in \mathbb{R}^{d^2 \times r},$$

and recall the decomposition in (39), (40), Lemma 1, and (44), i.e. $\boldsymbol{\Delta} = \boldsymbol{Z} + \boldsymbol{W}$. With these in mind, we can expand

$$\begin{aligned} T_{i,j,k} - T^\star_{i,j,k} &= \left\langle \boldsymbol{U}_{i,:}, \widetilde{\boldsymbol{U}}_{(j,k),:}\right\rangle - \left\langle \boldsymbol{U}^\star_{i,:}, \widetilde{\boldsymbol{U}}^\star_{(j,k),:}\right\rangle \\ &= \left\langle \boldsymbol{\Delta}_{i,:}, \widetilde{\boldsymbol{U}}^\star_{(j,k),:}\right\rangle + \left\langle \boldsymbol{\Delta}_{j,:}, \widetilde{\boldsymbol{U}}^\star_{(i,k),:}\right\rangle + \left\langle \boldsymbol{\Delta}_{k,:}, \widetilde{\boldsymbol{U}}^\star_{(i,j),:}\right\rangle \\ &\quad + \left\langle \boldsymbol{U}^\star_{i,:}, \widetilde{\boldsymbol{\Delta}}_{(j,k),:}\right\rangle + \left\langle \boldsymbol{U}^\star_{j,:}, \widetilde{\boldsymbol{\Delta}}_{(i,k),:}\right\rangle + \left\langle \boldsymbol{U}^\star_{k,:}, \widetilde{\boldsymbol{\Delta}}_{(i,k),:}\right\rangle \\ &\quad + \left\langle \boldsymbol{\Delta}_{i,:}, \widetilde{\boldsymbol{\Delta}}_{(j,k),:}\right\rangle \\ &= \underbrace{\left\langle \boldsymbol{Z}_{i,:}, \widetilde{\boldsymbol{U}}^\star_{(j,k),:}\right\rangle + \left\langle \boldsymbol{Z}_{j,:}, \widetilde{\boldsymbol{U}}^\star_{(i,k),:}\right\rangle + \left\langle \boldsymbol{Z}_{k,:}, \widetilde{\boldsymbol{U}}^\star_{(i,j),:}\right\rangle}_{=: Y_{i,j,k}} \\ &\quad + R_{i,j,k} \end{aligned} \tag{46}$$

for any $1 \le i, j, k \le d$, with the residual term $R_{i,j,k}$ given by

$$\begin{aligned} R_{i,j,k} := &\left\langle \boldsymbol{W}_{i,:}, \widetilde{\boldsymbol{U}}^\star_{(j,k),:}\right\rangle + \left\langle \boldsymbol{W}_{j,:}, \widetilde{\boldsymbol{U}}^\star_{(i,k),:}\right\rangle + \left\langle \boldsymbol{W}_{k,:}, \widetilde{\boldsymbol{U}}^\star_{(i,j),:}\right\rangle \\ &+ \left\langle \boldsymbol{U}^\star_{i,:}, \widetilde{\boldsymbol{\Delta}}_{(j,k),:}\right\rangle + \left\langle \boldsymbol{U}^\star_{j,:}, \widetilde{\boldsymbol{\Delta}}_{(i,k),:}\right\rangle + \left\langle \boldsymbol{U}^\star_{k,:}, \widetilde{\boldsymbol{\Delta}}_{(i,j),:}\right\rangle \\ &+ \left\langle \boldsymbol{\Delta}_{i,:}, \widetilde{\boldsymbol{\Delta}}_{(j,k),:}\right\rangle. \end{aligned} \tag{47}$$

Armed with the distributional characterization for $\boldsymbol{Z}$ (cf. Lemma 2), we can show that $Y_{i,j,k}$ is approximately Gaussian, as formalized by the lemma below.

Lemma 7: Instate the assumptions of Theorem 9. On the event $\mathcal{E}$ where (59)-(63) hold, one can decompose $Y_{i,j,k} = G_{i,j,k} + H_{i,j,k}$ for each $1 \le i, j, k \le d$ such that

$$\sup_{\tau \in \mathbb{R}} \left| \mathbb{P}\left\{ G_{i,j,k} \le \tau \sqrt{v^\star_{i,j,k}} \right\} - \Phi(\tau) \right| \lesssim \frac{\mu\sqrt{r}}{d\sqrt{p}}, \tag{48}$$

where $\Phi(\cdot)$ is the CDF of a standard Gaussian random variable; further, with probability at least $1 - O\left(d^{-10}\right)$ one has

$$\frac{|H_{i,j,k}|}{\sqrt{v^\star_{i,j,k}}} \lesssim \frac{\mu\sqrt{r}\log^2 d}{d\sqrt{p}} + \sqrt{\frac{\mu r \log d}{d}} + \frac{\mu r^{3/2}\sqrt{\log d}}{d}. \tag{49}$$

Proof: The key step boils down to proving that $\boldsymbol{Z}_{i,:}$, $\boldsymbol{Z}_{j,:}$ and $\boldsymbol{Z}_{k,:}$ are nearly statistically independent (as alluded to previously). See Appendix D-A. □

In addition, given the $\ell_{2,\infty}$ bounds of the residual term $\boldsymbol{W}$ (cf. (45)) and the estimation error $\boldsymbol{\Delta}$ (cf. (59b)), we can demonstrate that $R_{i,j,k}$ (cf. (47)) is negligible in magnitude, as stated in the following lemma.

Lemma 8: Instate the assumptions of Theorem 9. Conditional on the event $\mathcal{E}$ where (59)-(63) hold, one has

$$\frac{|R_{i,j,k}|}{\sqrt{v^\star_{i,j,k}}} \lesssim \zeta + \frac{\sigma_{\max}}{\lambda_{\min}^{\star 1/3}} \frac{\mu^{3/2} r \log d}{\sqrt{dp}} \frac{1}{\omega_{i,j,k}} \tag{50}$$

for any $1 \le i, j, k \le d$ with probability at least $1 - O\left(d^{-10}\right)$, where ζ and $\omega_{i,j,k}$ is defined in (43) and (33), respectively.

Proof: See Appendix D-B. □

Proof of Theorem 10: With Lemmas 7 and 8 in place, one can readily prove Theorem 10. Applying the union bound yields that: for any $\tau \in \mathbb{R}$ and any $\varepsilon_1, \varepsilon_2 > 0$,

$$\begin{aligned} &\mathbb{P}\left\{ T_{i,j,k} - T^\star_{i,j,k} \le \tau \sqrt{v^\star_{i,j,k}} \right\} \\ &\quad \le \mathbb{P}\left\{ G_{i,j,k} \le (\tau + \varepsilon_1 + \varepsilon_2) \sqrt{v^\star_{i,j,k}} \right\} \\ &\qquad + \mathbb{P}\left\{ \left|H_{i,j,k}\right| > \varepsilon_1 \sqrt{v^\star_{i,j,k}} \right\} + \mathbb{P}\left\{ \left|R_{i,j,k}\right| > \varepsilon_2 \sqrt{v^\star_{i,j,k}} \right\} \\ &\quad \le \Phi(\tau + \varepsilon_1 + \varepsilon_2) + o(1) \\ &\qquad + \mathbb{P}\left\{ \left|H_{i,j,k}\right| > \varepsilon_1 \sqrt{v^\star_{i,j,k}} \right\} + \mathbb{P}\left\{ \left|R_{i,j,k}\right| > \varepsilon_2 \sqrt{v^\star_{i,j,k}} \right\}, \end{aligned}$$

where the last line results from (48) and the sample size condition that $p \gg \mu^2 r d^{-3/2}$. By setting

$$\begin{aligned} \varepsilon_1 &\asymp \frac{\mu\sqrt{r}\log^2 d}{d\sqrt{p}} + \sqrt{\frac{\mu r \log d}{d}} + \frac{\mu r^{3/2}\sqrt{\log d}}{d}, \\ \varepsilon_2 &\asymp \zeta + \frac{\sigma_{\max}}{\lambda_{\min}^{\star 1/3}} \frac{\mu^{3/2} r \log d}{\sqrt{dp}} \frac{1}{\omega_{i,j,k}}, \end{aligned}$$

one sees from (49) and (50) that

$$\begin{aligned} \mathbb{P}\left\{ |H_{i,j,k}| > \varepsilon_1 \sqrt{v^\star_{i,j,k}} \right\} &\lesssim d^{-10} \\ \mathbb{P}\left\{ |R_{i,j,k}| > \varepsilon_2 \sqrt{v^\star_{i,j,k}} \right\} &\lesssim d^{-10}. \end{aligned}$$

In particular, in view of the assumptions (32) and (34), one has $\max\{\varepsilon_1, \varepsilon_2\} = o(1)$. Consequently, we can obtain

$$\begin{aligned} &\mathbb{P}\left\{ T_{i,j,k} - T^\star_{i,j,k} \le \tau \sqrt{v^\star_{i,j,k}} \right\} - \Phi(\tau) \\ &\quad \le \Phi(\tau + \varepsilon_1 + \varepsilon_2) - \Phi(\tau) + o(1) \\ &\quad \le \varepsilon_1 + \varepsilon_2 + o(1) = o(1) \end{aligned}$$

for any $\tau \in \mathbb{R}$, where the last step arises from the property of the CDF of a standard Gaussian. The lower bound on $\mathbb{P}\{T_{i,j,k} - T^\star_{i,j,k} \le \tau\sqrt{v^\star_{i,j,k}}\} - \Phi(\tau)$ can be obtained analogously. These taken together lead to the advertised claim

$$\sup_{\tau\in\mathbb{R}} \left|\mathbb{P}\left\{T_{i,j,k} - T^\star_{i,j,k} \le \tau\sqrt{v^\star_{i,j,k}}\right\} - \Phi(\tau)\right| = o(1).$$

C. Proof Outline for the Validity of Confidence Intervals

With the above distributional guarantees in place, the validity of our confidence intervals can be established as long as the proposed variance/covariance estimates are sufficiently accurate. Before proceeding, we make note of the following crucial observation:

$$\begin{aligned}\max_{(i,j,k)\in\Omega} \left|\widehat{E}_{i,j,k} - E_{i,j,k}\right| &= \max_{(i,j,k)\in\Omega} \left|T^{\mathsf{obs}}_{i,j,k} - T_{i,j,k} - E_{i,j,k}\right| \\ &\le \|\boldsymbol{T} - \boldsymbol{T}^\star\|_\infty \\ &\lesssim \frac{\sigma_{\max}}{\lambda^\star_{\min}}\sqrt{\frac{\mu^3 r^2 \log d}{d^2 p}}\,\lambda^\star_{\max}, \end{aligned} \quad (51)$$

where $\widehat{E}_{i,j,k}$ is defined in (24). Here, we have used the relation (59c) provided in Appendix B. As we shall see momentarily, this simple fact plays a crucial role in ensuring that our procedure returns faithful variance estimates.

1) Confidence Intervals for Tensor Factors: We start with the tensor factors. For each $1 \le l \le r$ and $1 \le k \le d$, we can decompose

$$\frac{u_{l,k} - u^\star_{l,k}}{\sqrt{(\boldsymbol{\Sigma}_k)_{l,l}}} = \frac{u_{l,k} - u^\star_{l,k}}{\sqrt{(\boldsymbol{\Sigma}^\star_k)_{l,l}}} + \underbrace{\frac{u_{l,k} - u^\star_{l,k}}{\sqrt{(\boldsymbol{\Sigma}_k)_{l,l}}} - \frac{u_{l,k} - u^\star_{l,k}}{\sqrt{(\boldsymbol{\Sigma}^\star_k)_{l,l}}}}_{=:\,J_{l,k}}. \quad (52)$$

As it turns out, the approximation error term $J_{l,k}$ is quite small, as formalized in Lemma 9 below. The proof is postponed to Appendix E-A.

Lemma 9: Instate the assumptions of Theorem 11. Conditional on the event $\mathcal{E}$ where (59)-(63) hold, the following holds simultaneously for all $1 \le l \le r$ and $1 \le k \le d$ with probability at least $1 - O(d^{-10})$:

$$|J_{l,k}| \lesssim \sqrt{\frac{\mu^4 r^3 \log^3 d}{d^2 p}} + \frac{\sigma_{\max}}{\lambda^\star_{\min}}\sqrt{\frac{\mu^3 r^2 d \log^2 d}{p}}, \quad (53)$$

Proof of Theorem 11 (the part w.r.t. $u^\star_{l,k}$): Fix arbitrary $1 \le l \le r$ and $1 \le k \le d$. By virtue of Theorem 9 and the continuous mapping theorem, we know that

$$\sup_{\tau\in\mathbb{R}} \left|\mathbb{P}\left\{u_{l,k} - u^\star_{l,k} \le \tau\sqrt{(\boldsymbol{\Sigma}^\star_k)_{l,l}}\right\} - \Phi(\tau)\right| = o(1). \quad (54)$$

Given the decomposition in (52), one can use the union bound to find that for any $\tau \in \mathbb{R}$ and any $\varepsilon > 0$,

$$\begin{aligned}&\mathbb{P}\left\{u_{l,k} - u^\star_{l,k} \le \tau\sqrt{(\boldsymbol{\Sigma}_k)_{l,l}}\right\} - \Phi(\tau) \\ &\le \mathbb{P}\left\{u_{l,k} - u^\star_{l,k} \le (\tau+\varepsilon)\sqrt{(\boldsymbol{\Sigma}^\star_k)_{l,l}}\right\} \\ &\quad + \mathbb{P}\{|J_{l,k}| > \varepsilon\} - \Phi(\tau) \\ &\overset{\text{(i)}}{\le} \Phi(\tau+\varepsilon) - \Phi(\tau) + o(1) + \mathbb{P}\{|J_{l,k}| > \varepsilon\} \\ &\overset{\text{(ii)}}{\le} \varepsilon + o(1) + \mathbb{P}\{|J_{l,k}| > \varepsilon\},\end{aligned}$$

where (i) follows from (54), and (ii) arises from the property of the CDF of $\mathcal{N}(0,1)$. Set

$$\varepsilon \asymp \sqrt{\frac{\mu^4 r^3 \log^3 d}{d^2 p}} + \frac{\sigma_{\max}}{\lambda^\star_{\min}}\sqrt{\frac{\mu^3 r^2 d \log^2 d}{p}} = o(1),$$

where the last identity is valid as long as $p \gg \mu^4 r^3 d^{-2}\log^4 d$ and $\sigma_{\max}/\lambda^\star_{\min} \ll \sqrt{p/(\mu^3 r^2 d \log^3 d)}$. By Lemma 9, we have $\mathbb{P}\{|J_{l,k}| > \varepsilon\} \lesssim d^{-10}$. Applying a similar argument for the lower bound, one arrives at

$$\begin{aligned}&\sup_{\tau\in\mathbb{R}} \left|\mathbb{P}\left\{u_{l,k} - u^\star_{l,k} \le \tau\sqrt{(\boldsymbol{\Sigma}_k)_{l,l}}\right\} - \Phi(\tau)\right| \\ &\quad \le \varepsilon + o(1) + \mathbb{P}\{|J_{l,k}| > \varepsilon\} = o(1)\end{aligned}$$

as claimed.

2) Confidence Intervals for Tensor Entries: Next, we turn to the constructed confidence intervals for tensor entries. As before, let us decompose

$$\begin{aligned}\frac{T_{i,j,k} - T^\star_{i,j,k}}{\sqrt{v_{i,j,k}}} &= \frac{T_{i,j,k} - T^\star_{i,j,k}}{\sqrt{v^\star_{i,j,k}}} \\ &\quad + \underbrace{\frac{T_{i,j,k} - T^\star_{i,j,k}}{\sqrt{v_{i,j,k}}} - \frac{T_{i,j,k} - T^\star_{i,j,k}}{\sqrt{v^\star_{i,j,k}}}}_{=:\,K_{i,j,k}}\end{aligned} \quad (55)$$

for each $1 \le i,j,k \le d$. The following lemma reveals that the residual term $K_{i,j,k}$ is considerably small; the proof is deferred to Appendix 10.

Lemma 10: Instate the assumptions and notation of Theorem 11. Conditional on the event $\mathcal{E}$ where (59)-(63) hold, the following holds simultaneously for all $1 \le i,j,k \le d$ with probability at least $1 - O(d^{-10})$:

$$|K_{i,j,k}| \lesssim \frac{1}{\omega_{i,j,k}}\frac{\sigma_{\max}}{\lambda^{\star 1/3}_{\min}}\sqrt{\frac{\mu^5 r^3 \log^2 d}{dp}} + \sqrt{\frac{\mu^4 r^3 \log^3 d}{d^2 p}},$$

Proof of Theorem 11 (the part w.r.t. $T^\star_{i,j,k}$): Fix arbitrary $1 \le i \le j \le k \le d$. Recalling the decomposition in (55), we can apply the union bound to show that: for any $\tau \in \mathbb{R}$ and any $\varepsilon > 0$,

$$\begin{aligned}&\mathbb{P}\left\{T_{i,j,k} - T^\star_{i,j,k} \le \tau\sqrt{v_{i,j,k}}\right\} - \Phi(\tau) \\ &\le \mathbb{P}\left\{T_{i,j,k} - T^\star_{i,j,k} \le (\tau+\varepsilon)\sqrt{v^\star_{i,j,k}}\right\} \\ &\quad + \mathbb{P}\left\{|K_{i,j,k}| > \varepsilon\right\} - \Phi(\tau) \\ &\overset{\text{(i)}}{\le} \Phi(\tau+\varepsilon) - \Phi(\tau) + o(1) + \mathbb{P}\left\{|K_{i,j,k}| > \varepsilon\right\} \\ &\overset{\text{(ii)}}{\le} \varepsilon + o(1) + \mathbb{P}\left\{|K_{i,j,k}| > \varepsilon\right\},\end{aligned}$$

where (i) follows from Theorem 10, and (ii) arises from the property of the CDF of a standard Gaussian. Set

$$\varepsilon \asymp \frac{1}{\omega_{i,jk}}\frac{\sigma_{\max}}{\lambda^{\star 1/3}_{\min}}\sqrt{\frac{\mu^5 r^3 \log^2 d}{dp}} + \sqrt{\frac{\mu^4 r^3 \log^3 d}{d^2 p}} = o(1),$$

where the last equality holds due to our conditions $p \gg \mu^4 r^3 d^{-2} \log^4 d$ and (34). Then Lemma 10 guarantees that $\mathbb{P}\big\{|K_{i,j,k}| > \varepsilon\big\} \lesssim d^{-10}$, allowing us to reach

$$\begin{aligned}&\mathbb{P}\left\{T_{i,j,k} - T^\star_{i,j,k} \le \tau\sqrt{v_{i,j,k}}\right\} - \Phi(\tau)\\ &\quad\le \varepsilon + o\left(1\right) + \mathbb{P}\big\{|K_{i,j,k}| > \varepsilon\big\} = o\left(1\right).\end{aligned}$$

The lower bound can be obtained analogously. The proof is thus complete.

VI. Discussions

This paper has explored the problem of uncertainty quantification for nonconvex tensor completion. The main contributions lie in establishing (nearly) precise distributional guarantees for the nonconvex estimates down to an entrywise level. Our distributional representation enables data-driven construction of confidence intervals for both the unknown tensor and its underlying tensor factors. Our inferential procedure and the accompanying theory are model-agnostic, which do not require prior knowledge about the noise distributions and are automatically adaptive to location-varying noise levels. Our results uncover the effectiveness of nonconvex optimization, which is statistically optimal for both estimation and confidence interval construction.

The findings of the current paper further suggest numerous possible extensions that are worth pursuing. To begin with, our current results are only optimal when both the rank r and the condition number κ are constants independent of the ambient dimension d. Can we further refine the analysis to enable optimal inference for more general settings? In addition, our theory falls short of providing valid confidence intervals for tensor entries with very small "strength". This calls for further investigation in order to complete the picture.

Further, the current algorithmic and analysis frameworks rely heavily upon the assumption of uniform random sampling. In practice, there is no shortage of applications where the sampling patterns are highly non-uniform. It would of great interest to accommodate more general types of sampling patterns with heterogeneous missingness mechanisms.

Appendix A
More Details About Algorithm 1

A. The Initialization Scheme

For self-completeness, we record in this section the detailed initialization procedure employed in the two-stage nonconvex algorithm proposed in [2] (namely, Algorithm 1). This is summarized in Algorithm 2, with auxiliary procedures detailed in Algorithm 3. As a high-level interpretation, Algorithm 2 estimates the subspace spanned by the tensor factor $\{\boldsymbol{u}_l^\star\}_{1\le l\le r}$ via a spectral method (similar to PCA-type methods [16], [17], [89]), whereas Algorithm 3 attempts to retrieve estimates for individual tensor factors from this subspace estimate $\boldsymbol{U}_{\mathsf{space}}$. Here and throughout, we denote $\boldsymbol{T}^{\mathsf{obs}} := [T^{\mathsf{obs}}_{i,j,k}]_{1\le i,j,k\le d}$, where we set $T^{\mathsf{obs}}_{i,j,k} = 0$ for any $(i,j,k) \notin \Omega$.

Remark 15: In practice, the rank r of the true tensor is often unknown *a priori*, and needs to be estimated first. Fortunately, rank estimation can often be accomplished in a data-driven manner. For instance, under the assumptions imposed in this paper, the largest r eigenvalues of the matrix $\mathcal{P}_{\mathsf{off\text{-}diag}}(\boldsymbol{A}\boldsymbol{A}^\top)$ (cf. Algorithm 2) are provably much larger than its remaining eigenvalues (see [17], [18]). As a result, one can simply examine the eigenvalues of $\mathcal{P}_{\mathsf{off\text{-}diag}}(\boldsymbol{A}\boldsymbol{A}^\top)$, and utilize the eigen-gap of this data matrix to obtain a faithful estimate of r.

Algorithm 2 Spectral Initialization for Nonconvex Tensor Completion

Let $\boldsymbol{U}_{\mathsf{space}}\boldsymbol{\Lambda}\boldsymbol{U}_{\mathsf{space}}^\top$ be the rank-r eigen-decomposition of $\mathcal{P}_{\mathsf{off\text{-}diag}}(\boldsymbol{A}\boldsymbol{A}^\top)$, where $\boldsymbol{A} = \mathsf{unfold}\big(p^{-1}\boldsymbol{T}^{\mathsf{obs}}\big)$ is the mode-1 matricization of $p^{-1}\boldsymbol{T}^{\mathsf{obs}}$, and $\mathcal{P}_{\mathsf{off\text{-}diag}}(\boldsymbol{Z})$ extracts out the off-diagonal entries of $\boldsymbol{Z}$.
Output: an initial estimate $\boldsymbol{U}^0 \in \mathbb{R}^{d\times r}$ on the basis of $\boldsymbol{U}_{\mathsf{space}} \in \mathbb{R}^{d\times r}$ using Algorithm 3.

Algorithm 3 Retrieval of Low-Rank Tensor Factors From a Given Subspace Estimate

Input: number of restarts L, pruning threshold ϵ_{th}, subspace estimate $\boldsymbol{U}_{\mathsf{space}} \in \mathbb{R}^{d\times r}$ given by Algorithm 2.
for $\tau = 1, \dots, L$ **do**
  Generate an independent Gaussian vector $\boldsymbol{g}^\tau \sim \mathcal{N}(0, \boldsymbol{I}_d)$.
  $\big(\boldsymbol{\nu}^\tau, \lambda_\tau, \mathsf{gap}_\tau\big)$
$$\leftarrow \textsc{Retrieve-one-factor}(\boldsymbol{T}^{\mathsf{obs}}, p, \boldsymbol{U}_{\mathsf{space}}, \boldsymbol{g}^\tau).$$
Generate tensor factor estimates
$$\begin{aligned}&\big\{(\boldsymbol{w}^1,\lambda_1),\dots,(\boldsymbol{w}^r,\lambda_r)\big\}\\ &\qquad\leftarrow \textsc{Prune}\big(\big\{\big(\boldsymbol{\nu}^\tau,\lambda_\tau,\mathsf{gap}_\tau\big)\big\}_{\tau=1}^{L}, \epsilon_{\mathsf{th}}\big).\end{aligned}$$
Output: initial estimate $\boldsymbol{U}^0 = \big[\lambda_1^{1/3}\boldsymbol{w}^1, \dots, \lambda_r^{1/3}\boldsymbol{w}^r\big]$.

Remark 16: Let us take a moment to elucidate the role of the initialization stage. To begin with, it is widely observed that careful initialization plays a crucial role in many nonconvex tensor estimation algorithms [90]. As has been discussed in [2], delicate initialization is critical in enabling fast convergence and optimal statistical accuracy of the nonconvex estimator. In addition, it is natural to wonder whether one can perform statistical inference based solely on the initial estimate returned by Algorithm 2. As one can see from [2, Section 5.3], however, the statistical error of the initial estimates might be substantially larger than that of the final nonconvex estimates, thus resulting in a much higher degree of uncertainty. Consequently, the nonconvex refinement stage is essential in achieving statistically efficient inference.

B. Choices of Algorithmic Parameters

To guarantee fast convergence of Algorithm 1, there are a couple of algorithmic parameters — namely, the number of restart attempts L, the pruning threshold ϵ_{th} in Algorithm 3, as well as the learning rates η_t — that need to be properly chosen. Unless otherwise noted, this paper adopts the following

1: **function** RETRIEVE-ONE-FACTOR($\boldsymbol{T}, p, \boldsymbol{U}_{\mathsf{space}}, \boldsymbol{g}$)

2: Compute

$$\boldsymbol{\theta} = \boldsymbol{U}_{\mathsf{space}} \boldsymbol{U}_{\mathsf{space}}^{\top} \boldsymbol{g}, \tag{56a}$$

$$\boldsymbol{M} = p^{-1} \boldsymbol{T}^{\mathsf{obs}} \times_3 \boldsymbol{\theta}, \tag{56b}$$

where $\times_3$ is defined in Section I-E.

3: Let $\boldsymbol{\nu}$ be the leading singular vector of $\boldsymbol{M}$ obeying $\langle \boldsymbol{T}^{\mathsf{obs}}, \boldsymbol{\nu}^{\otimes 3} \rangle \geq 0$, and set $\lambda = \langle p^{-1} \boldsymbol{T}^{\mathsf{obs}}, \boldsymbol{\nu}^{\otimes 3} \rangle$.

4: **return** $\big(\boldsymbol{\nu}, \lambda, \sigma_1(\boldsymbol{M}) - \sigma_2(\boldsymbol{M})\big)$.

1: **function** PRUNE($\big\{\big(\boldsymbol{\nu}^{\tau}, \lambda_{\tau}, \mathsf{gap}_{\tau}\big)\big\}_{\tau=1}^{L}, \epsilon_{\mathsf{th}}$)

2: Set $\Theta = \big\{\big(\boldsymbol{\nu}^{\tau}, \lambda_{\tau}, \mathsf{gap}_{\tau}\big)\big\}_{\tau=1}^{L}$.

3: **for** $i = 1, \ldots, r$ **do**

4: Choose $(\boldsymbol{\nu}^{\tau}, \lambda_{\tau}, \mathsf{gap}_{\tau})$ from Θ with the largest gap_{τ}; set $\boldsymbol{w}^{i} = \boldsymbol{\nu}^{\tau}$ and $\lambda_i = \lambda_{\tau}$.

5: Update

$$\Theta \leftarrow \Theta \setminus \big\{\big(\boldsymbol{\nu}^{\tau}, \lambda_{\tau}, \mathsf{gap}_{\tau}\big) \in \Theta : |\langle \boldsymbol{\nu}^{\tau}, \boldsymbol{w}^{i} \rangle| > 1 - \epsilon_{\mathsf{th}}\big\}.$$

6: **return** $\big\{(\boldsymbol{w}^{1}, \lambda_1), \ldots, (\boldsymbol{w}^{r}, \lambda_r)\big\}$.

---

choices suggested by [2]:

$$L = c_4 r^{2\kappa^2} \log^{3/2} d, \quad \eta_t \equiv \frac{c_5 \lambda_{\min}^{\star 4/3}}{p \lambda_{\max}^{\star 8/3}}$$

$$\epsilon_{\mathsf{th}} = c_6 \bigg\{ \frac{\mu r \log d}{d\sqrt{p}} + \frac{\sigma_{\min}}{\lambda_{\min}^{\star}} \sqrt{\frac{rd \log^2 d}{p}} + \sqrt{\frac{\mu r \log d}{d}} \bigg\}, \tag{57}$$

where $c_4 > 0$ is some sufficiently large constant, and $c_5, c_6 > 0$ are some sufficiently small constants. The interested reader is referred to [2] for justification.

APPENDIX B
PRELIMINARY FACTS

In this section, we gather a few preliminary facts that prove useful throughout the analysis.

A. Leave-One-Out Sequences

To facilitate the analysis and decouple statistical dependency, we introduce the following set of auxiliary tensors and loss functions for all $1 \leq m \leq d$:

$$\boldsymbol{T}^{\mathsf{obs},(m)} \coloneqq \mathcal{P}_{\Omega_{-m}}(\boldsymbol{T}^{\mathsf{obs}}) + p\, \mathcal{P}_m(\boldsymbol{T}^{\star}),$$

$$f^{(m)}(\boldsymbol{U}) \coloneqq \Big\| \mathcal{P}_{\Omega_{-m}} \Big(\sum_{i=1}^{r} \boldsymbol{u}_i^{\otimes 3} - \boldsymbol{T}^{\star} - \boldsymbol{E} \Big) \Big\|_{\mathrm{F}}^2 + p \Big\| \mathcal{P}_m \Big(\sum_{i=1}^{r} \boldsymbol{u}_i^{\otimes 3} - \boldsymbol{T}^{\star} \Big) \Big\|_{\mathrm{F}}^2,$$

where $\mathcal{P}_{\Omega_{-m}}$ (resp. $\mathcal{P}_m$) is the Euclidean projection onto the subspace of tensors supported on $\{(i,j,k) \in \Omega\colon i \neq m \text{ and } j \neq m \text{ and } k \neq m\}$ (resp. $\{(i,j,k) \in [d]^3\colon i = m \text{ or } j = m \text{ or } k = m\}$). We shall denote by $\boldsymbol{U}^{(m)}$ the leave-one-out estimate returned by Algorithm 4.

Algorithm 4 The m-th Leave-One-Out Estimate

Initialize $\boldsymbol{U}^{0,(m)} = \big[\boldsymbol{u}_1^{0,(m)}, \cdots, \boldsymbol{u}_r^{0,(m)}\big]$ via Algorithm 5.

Gradient updates: **for** $t = 0, 1, \ldots, t_0 - 1$ **do**

$$\boldsymbol{U}^{t+1,(m)} = \boldsymbol{U}^{t,(m)} - \eta_t \nabla f^{(m)}(\boldsymbol{U}^{t,(m)}). \tag{58}$$

Output $\boldsymbol{U}^{(m)} = \big[\boldsymbol{u}_1^{(m)}, \cdots, \boldsymbol{u}_r^{(m)}\big] \coloneqq \boldsymbol{U}^{t_0,(m)}$.

Algorithm 5 The m-th Leave-One-Out Sequence for Spectral Initialization

1: Let $\boldsymbol{U}_{\mathsf{space}}^{(m)} \boldsymbol{\Lambda}^{(m)} \boldsymbol{U}_{\mathsf{space}}^{(m)\top}$ be the rank-r eigen-decomposition of $\mathcal{P}_{\mathsf{off\text{-}diag}}(\boldsymbol{A}^{(m)} \boldsymbol{A}^{(m)\top})$, where $\boldsymbol{A}^{(m)} = \mathsf{unfold}\big(p^{-1} \boldsymbol{T}^{\mathsf{obs},(m)}\big)$ is the mode-1 matricization of $p^{-1} \boldsymbol{T}^{\mathsf{obs},(m)}$, and $\mathcal{P}_{\mathsf{off\text{-}diag}}(\boldsymbol{Z})$ extracts out the off-diagonal entries of $\boldsymbol{Z}$.

2: **Output:** an estimate $\boldsymbol{U}^{0,(m)} \in \mathbb{R}^{d \times r}$ on the basis of $\boldsymbol{U}_{\mathsf{space}}^{(m)} \in \mathbb{R}^{d \times r}$ using Algorithm 6.

Algorithm 6 The m-th Leave-One-Out Sequence for Retrieving Individual Tensor Components

1: **Input:** number of restarts L, pruning threshold ϵ_{th}, subspace estimate $\boldsymbol{U}_{\mathsf{space}}^{(m)} \in \mathbb{R}^{d \times r}$ given by Algorithm 5.

2: **for** $\tau = 1, \ldots, L$ **do**

3: Recall the Gaussian vector $\boldsymbol{g}^{\tau} \sim \mathcal{N}(0, \boldsymbol{I}_d)$ generated in Algorithm 3.

4: $\big(\boldsymbol{\nu}^{\tau,(m)}, \lambda_{\tau}^{(m)}, \mathsf{gap}_{\tau}^{(m)}\big) \leftarrow$ RETRIEVE-ONE-FACTOR$(\boldsymbol{T}^{(m)}, p, \boldsymbol{U}_{\mathsf{space}}^{(m)}, \boldsymbol{g}^{\tau})$.

5: Generate tensor factor estimates

$$\big\{\big(\boldsymbol{w}^{1,(m)}, \lambda_1^{(m)}\big), \ldots, \big(\boldsymbol{w}^{r,(m)}, \lambda_r^{(m)}\big)\big\} \leftarrow \mathrm{PRUNE}\big(\big\{\big(\boldsymbol{\nu}^{\tau,(m)}, \lambda_{\tau}^{(m)}, \mathsf{gap}_{\tau}^{(m)}\big)\big\}_{\tau=1}^{L}, \epsilon_{\mathsf{th}}\big).$$

6: **Output:** an estimate

$$\boldsymbol{U}^{0,(m)} = \big[\big(\lambda_1^{(m)}\big)^{1/3} \boldsymbol{w}^{1,(m)}, \ldots, \big(\lambda_r^{(m)}\big)^{1/3} \boldsymbol{w}^{r,(m)}\big].$$

While the algorithms look somewhat complex, the idea is very simple. In words, the new estimate $\boldsymbol{U}^{(m)}$ is obtained by dropping all randomness from the m-th slice (namely, those data from the index set $\{(i,j,k) \in [d]^3\colon i = m \text{ or } j = m \text{ or } k = m\}$). This means that $\boldsymbol{U}^{(m)}$ is statistically independent from the data coming from the m-th slice. These leave-one-out sequences enjoy several useful properties that have been established in [2], which we shall present in the next subsection.

B. Properties of the Nonconvex Estimates

We now collect several important properties of our tensor estimates as well as the associated leave-one-out estimates, most of which have been established in [2]. To begin with, Lemma 11 quantifies the estimation error of $\boldsymbol{U}$ and $\boldsymbol{T}$.

Lemma 11: Instate the assumptions and notations of Theorem 9. With probability at least $1 - o(1)$,

$$\left\|\boldsymbol{U\Pi} - \boldsymbol{U}^{\star}\right\|_{\mathrm{F}} \lesssim \frac{\sigma_{\max}}{\lambda_{\min}^{\star}}\sqrt{\frac{rd\log d}{p}}\,\lambda_{\max}^{\star 1/3}; \tag{59a}$$

$$\left\|\boldsymbol{U\Pi} - \boldsymbol{U}^{\star}\right\|_{2,\infty} \lesssim \frac{\sigma_{\max}}{\lambda_{\min}^{\star}}\sqrt{\frac{\mu r\log d}{p}}\,\lambda_{\max}^{\star 1/3}; \tag{59b}$$

$$\left\|\boldsymbol{T} - \boldsymbol{T}^{\star}\right\|_{\infty} \lesssim \frac{\sigma_{\max}}{\lambda_{\min}^{\star}}\sqrt{\frac{\mu^3 r^2\log d}{d^2 p}}\,\lambda_{\max}^{\star}. \tag{59c}$$

The next lemma demonstrates that the leave-one-out sequences $\{\boldsymbol{U}^{(m)}\}_{1\le m\le d}$ constructed in Algorithm 4 are sufficiently close to the true estimate $\boldsymbol{U}$. As a result, $\boldsymbol{U}^{(m)}$ (resp. $\boldsymbol{T}^{(m)}$) also serves as a faithful estimate of the ground truth $\boldsymbol{U}^{\star}$ (resp. $\boldsymbol{T}^{\star}$), where

$$\boldsymbol{T}^{(m)} := \sum_{1\le l\le r}\left(\boldsymbol{u}_l^{(m)}\right)^{\otimes 3}. \tag{60}$$

The results are summarized as follows.

Lemma 12: Instate the assumptions and notation of Theorem 9. With probability at least $1 - o(1)$, for all $1 \le m \le d$ one has

$$\left\|\boldsymbol{U} - \boldsymbol{U}^{(m)}\right\|_{\mathrm{F}} \lesssim \frac{\sigma_{\max}}{\lambda_{\min}^{\star}}\sqrt{\frac{\mu r\log d}{p}}\,\lambda_{\max}^{\star 1/3}. \tag{61a}$$

$$\left\|\boldsymbol{U}^{(m)}\boldsymbol{\Pi} - \boldsymbol{U}^{\star}\right\|_{\mathrm{F}} \lesssim \frac{\sigma_{\max}}{\lambda_{\min}^{\star}}\sqrt{\frac{rd\log d}{p}}\,\lambda_{\max}^{\star 1/3}; \tag{61b}$$

$$\left\|\boldsymbol{U}^{(m)}\boldsymbol{\Pi} - \boldsymbol{U}^{\star}\right\|_{2,\infty} \lesssim \frac{\sigma_{\max}}{\lambda_{\min}^{\star}}\sqrt{\frac{\mu r\log d}{p}}\,\lambda_{\max}^{\star 1/3}; \tag{61c}$$

$$\left\|\boldsymbol{T}^{(m)} - \boldsymbol{T}^{\star}\right\|_{\infty} \lesssim \frac{\sigma_{\max}}{\lambda_{\min}^{\star}}\sqrt{\frac{\mu^3 r^2\log d}{d^2 p}}\,\lambda_{\max}^{\star}. \tag{61d}$$

In addition, Lemma 13 collects several simple properties about the true tensor factors and their corresponding estimates. The proof can be found in Appendix H-A.

Lemma 13: Instate the assumptions and notation of Theorem 9. With probability at least $1 - o(1)$, there is a permutation $\pi(\cdot) : [d] \mapsto [d]$ such that

$$\|\boldsymbol{U}^{\star}\|_{\mathrm{F}} \le \sqrt{r}\quad \lambda_{\max}^{\star 1/3},\ \ \|\boldsymbol{U}^{\star}\|_{2,\infty} \le \sqrt{\frac{\mu r}{d}}\,\lambda_{\max}^{\star 1/3}; \tag{62a}$$

$$\sigma_1(\boldsymbol{U}^{\star}) = \lambda_{\max}^{\star 1/3}\left(1 + o(1)\right),\ \ \sigma_r(\boldsymbol{U}^{\star}) = \lambda_{\min}^{\star}\left(1 + o(1)\right); \tag{62b}$$

$$\left\|\boldsymbol{u}_{\pi(i)} - \boldsymbol{u}_i^{\star}\right\|_2 = o(1)\left\|\boldsymbol{u}_i^{\star}\right\|_2,\ 1 \le i \le r; \tag{62c}$$

$$\left\|\boldsymbol{u}_{\pi(i)} - \boldsymbol{u}_i^{\star}\right\|_{\infty} = o(1)\left\|\boldsymbol{u}_i^{\star}\right\|_{\infty},\ 1 \le i \le r; \tag{62d}$$

$$\lambda_{\min}^{\star 1/3} \lesssim \|\boldsymbol{u}_i\|_2 \lesssim \lambda_{\max}^{\star 1/3},\ \ 1 \le i \le r; \tag{62e}$$

$$\sqrt{\frac{1}{d}}\,\lambda_{\min}^{\star 1/3} \lesssim \|\boldsymbol{u}_i\|_{\infty} \lesssim \sqrt{\frac{\mu}{d}}\,\lambda_{\max}^{\star 1/3},\ \ 1 \le i \le r; \tag{62f}$$

$$\max_{1\le i\ne j\le r}\left|\langle \boldsymbol{u}_i, \boldsymbol{u}_j\rangle\right| \lesssim \left\{\sqrt{\frac{\mu}{d}} + \frac{\sigma_{\max}}{\lambda_{\min}^{\star}}\sqrt{\frac{rd\log d}{p}}\right\}\lambda_{\max}^{\star 2/3}, \tag{62g}$$

$$\sigma_1(\boldsymbol{U}) = \lambda_{\max}^{\star 1/3}\left(1 + o(1)\right),\ \ \sigma_r(\boldsymbol{U}) = \lambda_{\min}^{\star}\left(1 + o(1)\right). \tag{62h}$$

In addition, these results hold unchanged if we replace $\boldsymbol{u}_i$ with $\boldsymbol{u}_i^{(m)}$ for all $1 \le m \le d$.

Finally, Lemma 14 summarizes several useful bounds regarding $\widetilde{\boldsymbol{U}}^{\star} := [\boldsymbol{u}_l^{\star\otimes 2}]_{1\le l\le r} \in \mathbb{R}^{d^2\times r}$ and $\widetilde{\boldsymbol{U}} := [\boldsymbol{u}_l^{\otimes 2}]_{1\le l\le r} \in \mathbb{R}^{d^2\times r}$. The proof is deferred to Appendix H-B.

Lemma 14: Instate the assumptions and notation of Theorem 9. With probability at least $1 - o(1)$,

$$\left\|\widetilde{\boldsymbol{U}}^{\star}\right\|_{2,\infty} \le \frac{\mu\sqrt{r}}{d}\lambda_{\max}^{\star 2/3},\quad \left\|\widetilde{\boldsymbol{U}}^{\star}\right\|_{\mathrm{F}} \le \sqrt{r}\,\lambda_{\max}^{\star 2/3}; \tag{63a}$$

$$\sigma_1\left(\widetilde{\boldsymbol{U}}^{\star}\right) = \lambda_{\max}^{\star 2/3}\left(1 + o(1)\right),\ \sigma_r\left(\widetilde{\boldsymbol{U}}^{\star}\right) = \lambda_{\min}^{\star 2/3}\left(1 + o(1)\right); \tag{63b}$$

$$\sigma_1\left(\widetilde{\boldsymbol{U}}\right) = \lambda_{\max}^{\star 2/3}\left(1 + o(1)\right),\ \sigma_r\left(\widetilde{\boldsymbol{U}}\right) = \lambda_{\min}^{\star 2/3}\left(1 + o(1)\right); \tag{63c}$$

$$\left\|\widetilde{\boldsymbol{U}}\boldsymbol{\Pi} - \widetilde{\boldsymbol{U}}^{\star}\right\|_{\mathrm{F}} \lesssim \frac{\sigma_{\max}}{\lambda_{\min}^{\star}}\sqrt{\frac{rd\log d}{p}}\,\lambda_{\max}^{\star 2/3}; \tag{63d}$$

$$\left\|\widetilde{\boldsymbol{U}}\boldsymbol{\Pi} - \widetilde{\boldsymbol{U}}^{\star}\right\|_{2,\infty} \lesssim \frac{\sigma_{\max}}{\lambda_{\min}^{\star}}\sqrt{\frac{\mu^2 r\log d}{dp}}\,\lambda_{\max}^{\star 2/3}. \tag{63e}$$

In addition, the above results continue to hold if $\widetilde{\boldsymbol{U}}$ is replaced by $\widetilde{\boldsymbol{U}}^{(m)} = \left[\boldsymbol{u}_l^{(m)} \otimes \boldsymbol{u}_l^{(m)}\right]_{1\le l\le r}$ for all $1 \le m \le d$.

C. A Berry-Esseen-Type Theorem

The distributional guarantees are built upon the Berry-Esseen-type inequality [91, Theorem 1.1], which will be used multiple times in the analysis.

Theorem 15: Let $\{\boldsymbol{x}_i\}_{1\le i\le n}$ be a sequence of independent zero-mean random vectors in $\mathbb{R}^d$. Denote by $\boldsymbol{\Sigma}$ the covariance matrix of $\sum_{1\le i\le n}\boldsymbol{x}_i$, and let $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$ be a Gaussian vector in $\mathbb{R}^d$. Then one has

$$\sup_{\mathcal{A}\in\mathcal{C}}\left|\mathbb{P}\Big\{\sum_{1\le i\le n}\boldsymbol{x}_i \in \mathcal{A}\Big\} - \mathbb{P}\left\{\boldsymbol{z}\in\mathcal{A}\right\}\right| \lesssim d^{1/4}\rho, \tag{64}$$

where $\mathcal{C}$ is the set of all convex subsets of $\mathbb{R}^d$, and ρ is defined as follows

$$\rho := \sum_{1\le i\le n}\mathbb{E}\left[\left\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{x}_i\right\|_2^3\right]. \tag{65}$$

APPENDIX C
PROOF OF AUXILIARY LEMMAS: DISTRIBUTIONAL THEORY FOR TENSOR FACTORS

Given a symmetric random tensor, we can always partition it into six sub-tensors such that the entries within each sub-tensor are independent. Therefore, whenever orderwise bounds are sufficient, we shall treat $\{E_{i,j,k}\}_{1\le i,j,k\le d}$ and $\{\chi_{i,j,k}\}_{1\le i,j,k\le d}$ (see the definition in (9)) as independent random variables in order to simplify presentation.

A. Proof of Lemma 1

Fix an arbitrary $m \in [d]$. Let us define a sequence of random vectors $\{\boldsymbol{z}_{i,j,k}\}_{1\le i,j,k\le d}$ in $\mathbb{R}^r$ as follows:

$$\boldsymbol{z}_{i,j,k} := p^{-1}E_{i,j,k}\chi_{i,j,k}\widetilde{\boldsymbol{U}}^{\star}_{(i,j),:}(\widetilde{\boldsymbol{U}}^{\star\top}\widetilde{\boldsymbol{U}}^{\star})^{-1}, \tag{66}$$

where we recall the notation $(i,j) \coloneqq (i-1)d+j$ defined in Section I-E. Then we can express

$$\boldsymbol{X}_{m,:} = \sum_{1\le i\le d} \boldsymbol{z}_{i,i,m} + 2\sum_{1\le i<j\le d} \boldsymbol{z}_{i,j,m}$$

as a sum of independent zero-mean random vectors in $\mathbb{R}^r$. Let us further define a matrix $\boldsymbol{Z}\in\mathbb{R}^{d\times r}$ whose k-th row is given by

$$\boldsymbol{Z}_{k,:} = \sqrt{2}\sum_{1\le i\le d} \boldsymbol{z}_{i,i,k} + 2\sum_{1\le i<j\le d} \boldsymbol{z}_{i,j,k}, \tag{67}$$

and let $\boldsymbol{W}_0 \coloneqq \boldsymbol{X}-\boldsymbol{Z}$. Straightforward calculation gives

$$\begin{aligned}
&\mathbb{E}\big[(\boldsymbol{X}_{m,:})^\top \boldsymbol{X}_{m,:}\big] \\
&\quad= \frac{1}{p}(\widetilde{\boldsymbol{U}}^{\star\top}\widetilde{\boldsymbol{U}}^\star)^{-1}\widetilde{\boldsymbol{U}}^{\star\top}(2\boldsymbol{D}_m^\star - \boldsymbol{C}_m^\star)\widetilde{\boldsymbol{U}}^\star(\widetilde{\boldsymbol{U}}^{\star\top}\widetilde{\boldsymbol{U}}^\star)^{-1}, \\
&\mathbb{E}\big[(\boldsymbol{Z}_{m,:})^\top \boldsymbol{Z}_{m,:}\big] \\
&\quad= \frac{2}{p}(\widetilde{\boldsymbol{U}}^{\star\top}\widetilde{\boldsymbol{U}}^\star)^{-1}\widetilde{\boldsymbol{U}}^{\star\top}\boldsymbol{D}_m^\star\widetilde{\boldsymbol{U}}^\star(\widetilde{\boldsymbol{U}}^{\star\top}\widetilde{\boldsymbol{U}}^\star)^{-1} = \boldsymbol{\Sigma}_m^\star,
\end{aligned}$$

where we recall that $\boldsymbol{D}_m^\star$ (resp. $\boldsymbol{\Sigma}_m^\star$) is defined in (18) (resp. (20)), and $\boldsymbol{C}_m^\star$ is a diagonal matrix in $\mathbb{R}^{d^2\times d^2}$ with entries

$$(\boldsymbol{C}_m^\star)_{(i,j),(i,j)} = \begin{cases}\sigma_{i,j,m}^2, & \text{if} \quad i=j, \\ 0, & \text{if} \quad i\ne j.\end{cases}$$

In what follows, we will prove (i) $\boldsymbol{X}_{m,:}$ and $\boldsymbol{Z}_{m,:}$ are sufficiently close, i.e. the $\ell_{2,\infty}$ norm of $\boldsymbol{W}_0$ is considerably small; (ii) $\boldsymbol{Z}_{m,:}$ is a Gaussian vector with mean zero and covariance matrix $\boldsymbol{\Sigma}_m^\star$ with high probability.

We begin with the first claim. Observe that

$$(\boldsymbol{W}_0)_{m,:} = (\sqrt{2}-1)\sum_{1\le i\le d} p^{-1}E_{i,i,m}\chi_{i,i,m}\widetilde{\boldsymbol{U}}^\star_{(i,i),:}(\widetilde{\boldsymbol{U}}^{\star\top}\widetilde{\boldsymbol{U}}^\star)^{-1}$$

is a sum of independent zero-mean random vectors in $\mathbb{R}^r$. By (63a) and (63b), it is straightforward to compute

$$\begin{aligned}
B_1 &\coloneqq \max_{1\le i\le d}\big\|p^{-1}E_{i,i,m}\chi_{i,i,m}\widetilde{\boldsymbol{U}}^\star_{(i,i),:}(\widetilde{\boldsymbol{U}}^{\star\top}\widetilde{\boldsymbol{U}}^\star)^{-1}\big\|_{\psi_1} \\
&\lesssim \frac{\sigma_{\max}}{p}\big\|\widetilde{\boldsymbol{U}}^\star\big\|_{2,\infty}\big\|(\widetilde{\boldsymbol{U}}^{\star\top}\widetilde{\boldsymbol{U}}^\star)^{-1}\big\| \lesssim \frac{\sigma_{\max}}{p}\cdot\frac{\mu\sqrt{r}\lambda_{\max}^{\star 2/3}}{d\lambda_{\min}^{\star 4/3}}, \\
V_1 &\coloneqq \sum_{1\le i\le d}\frac{1}{p^2}\mathbb{E}\big[E_{i,i,m}^2\chi_{i,i,m}\big]\big\|\widetilde{\boldsymbol{U}}^\star_{(i,i),:}(\widetilde{\boldsymbol{U}}^{\star\top}\widetilde{\boldsymbol{U}}^\star)^{-1}\big\|_2^2 \\
&\lesssim \frac{\sigma_{\max}^2}{p}\sum_{1\le i\le d}\big\|\widetilde{\boldsymbol{U}}^\star_{(i,i),:}\big\|_2^2\big\|(\widetilde{\boldsymbol{U}}^{\star\top}\widetilde{\boldsymbol{U}}^\star)^{-1}\big\|^2 \\
&\lesssim \frac{\sigma_{\max}^2}{p}\cdot\frac{\mu r\lambda_{\max}^{\star 4/3}}{d\lambda_{\min}^{\star 8/3}},
\end{aligned}$$

where $\|\cdot\|_{\psi_1}$ denotes the sub-exponential norm, and we use the following bound in the second line:

$$\begin{aligned}
\sum_{1\le i\le d}\big\|\widetilde{\boldsymbol{U}}^\star_{(i,i),:}\big\|_2^2 &= \sum_{1\le i\le d}\sum_{1\le l\le r} u_{l,i}^{\star 4} \le \max_{1\le l\le r}\|\boldsymbol{u}_l^\star\|_\infty^2\|\boldsymbol{U}^\star\|_{\mathrm{F}}^2 \\
&\lesssim \frac{\mu r}{d}\lambda_{\max}^{\star 4/3}.
\end{aligned}$$

We then invoke the matrix Bernstein inequality [92, Corollary 2.1] to find that with probability exceeding $1-O\left(d^{-11}\right)$,

$$\begin{aligned}
\|(\boldsymbol{W}_0)_{m,:}\|_2 &\lesssim B_1\log^2 d + \sqrt{V_1\log d} \\
&\lesssim \frac{\sigma_{\max}}{\lambda_{\min}^{\star 2/3}\sqrt{p}}\left\{\frac{\mu\sqrt{r}\log^2 d}{d\sqrt{p}} + \sqrt{\frac{\mu r\log d}{d}}\right\},
\end{aligned} \tag{68}$$

where we have used the assumption $\kappa\asymp 1$.

We move on to consider the distribution of $\boldsymbol{Z}_{m,:}$. Conditional on $\{\chi_{i,j,m}\}_{1\le i,j\le d}$, the vector $\boldsymbol{Z}_{m,:}$ is zero-mean Gaussian with covariance matrix

$$\begin{aligned}
\boldsymbol{S}_m^\star \coloneqq \frac{2}{p^2}\sum_{1\le i,j\le d}\Big(&\sigma_{i,j,m}^2\chi_{i,j,m} \\
&\cdot(\widetilde{\boldsymbol{U}}^{\star\top}\widetilde{\boldsymbol{U}}^\star)^{-1}(\widetilde{\boldsymbol{U}}^\star_{(i,j),:})^\top\widetilde{\boldsymbol{U}}^\star_{(i,j),:}(\widetilde{\boldsymbol{U}}^{\star\top}\widetilde{\boldsymbol{U}}^\star)^{-1}\Big),
\end{aligned}$$

which satisfies

$$\mathbb{E}[\boldsymbol{S}_m^\star] = \boldsymbol{\Sigma}_m^\star.$$

Lemma 15 below demonstrates that $\boldsymbol{S}_m^\star$ and $\boldsymbol{\Sigma}_m^\star$ are, with high probability, sufficiently close in the spectral norm; the proof is deferred to the end of the section.

Lemma 15: Instate the assumptions of Lemma 1. With probability exceeding $1-O\left(d^{-10}\right)$,

$$\max_{1\le m\le d}\|\boldsymbol{S}_m^\star - \boldsymbol{\Sigma}_m^\star\| \lesssim \frac{\sigma_{\max}^2}{\lambda_{\min}^{\star 4/3}p}\sqrt{\frac{\mu^2 r\log d}{d^2 p}} = o\left(\frac{\sigma_{\max}^2}{\lambda_{\min}^{\star 4/3}p}\right). \tag{69}$$

In what follows, we shall work on the high-probability event where (69) holds. From the definition of the covariance matrix $\boldsymbol{\Sigma}_m^\star$, it is easily seen that

$$\frac{2\sigma_{\min}^2}{p}(\widetilde{\boldsymbol{U}}^{\star\top}\widetilde{\boldsymbol{U}}^\star)^{-1} \preceq \boldsymbol{\Sigma}_m^\star \preceq \frac{2\sigma_{\max}^2}{p}(\widetilde{\boldsymbol{U}}^{\star\top}\widetilde{\boldsymbol{U}}^\star)^{-1}.$$

Additionally, it follows from (63b) that

$$\lambda_{\min}(\boldsymbol{\Sigma}_m^\star) \gtrsim \frac{\sigma_{\min}^2}{\lambda_{\max}^{\star 4/3}p} \quad\text{and}\quad \lambda_{\max}(\boldsymbol{\Sigma}_m^\star) \lesssim \frac{\sigma_{\max}^2}{\lambda_{\max}^{\star 4/3}p}. \tag{70}$$

We then know from Weyl's inequality and the conditions $\sigma_{\max}/\sigma_{\min}\asymp 1$ and $\kappa\asymp 1$ that

$$\lambda_{\min}(\boldsymbol{S}_m^\star) \ge \lambda_{\min}(\boldsymbol{\Sigma}_m^\star) - \|\boldsymbol{S}_m^\star - \boldsymbol{\Sigma}_m^\star\| \gtrsim \frac{\sigma_{\min}^2}{\lambda_{\max}^{\star 4/3}p}. \tag{71}$$

This implies that $\boldsymbol{S}_m^\star$ is positive semidefinite and, therefore, $\boldsymbol{S}_m^{\star -1/2}$ is well-defined. As a result, $\boldsymbol{Z}_{m,:}\boldsymbol{S}_m^{\star -1/2}\boldsymbol{\Sigma}_m^{\star 1/2}$ is a zero-mean Gaussian vector with covariance matrix $\boldsymbol{\Sigma}_m^\star$.

With slight abuse of notation, we will treat $\boldsymbol{Z}_{m,:} - \boldsymbol{Z}_{m,:}\boldsymbol{S}_m^{\star -1/2}\boldsymbol{\Sigma}_m^{\star 1/2} + \boldsymbol{W}_0$ as the residual term for the Gaussian approximation. Hence, it remains to show that $\boldsymbol{Z}_{m,:}\boldsymbol{S}_m^{\star -1/2}\boldsymbol{\Sigma}_m^{\star 1/2}$ and $\boldsymbol{Z}_{m,:}$ are exceedingly close in the ℓ_2 norm. To this end, we observe an upper bound

$$\begin{aligned}
&\big\|\boldsymbol{Z}_{m,:} - \boldsymbol{Z}_{m,:}\boldsymbol{S}_m^{\star -1/2}\boldsymbol{\Sigma}_m^{\star 1/2}\big\|_2 \\
&\quad= \big\|\boldsymbol{Z}_{m,:}\boldsymbol{S}_m^{\star -1/2}\big(\boldsymbol{S}_m^{\star 1/2} - \boldsymbol{\Sigma}_m^{\star 1/2}\big)\big\|_2
\end{aligned}$$

$$\lesssim \|\boldsymbol{Z}_{m,:}\|_2 \left\|\boldsymbol{S}_m^{\star-1/2}\right\| \left\|\boldsymbol{S}_m^{\star1/2} - \boldsymbol{\Sigma}_m^{\star1/2}\right\|.$$

By the perturbation bounds for matrix square roots [93, Lemma 2.1], one knows from (69), (70) and (63b) that

$$\left\|\boldsymbol{S}_m^{\star1/2} - \boldsymbol{\Sigma}_m^{\star1/2}\right\| \leq \frac{\|\boldsymbol{S}_m^\star - \boldsymbol{\Sigma}_m^\star\|}{\lambda_{\min}\big(\boldsymbol{S}_m^{\star1/2}\big) + \lambda_{\min}\big(\boldsymbol{\Sigma}_m^{\star1/2}\big)} \lesssim \frac{\sigma_{\max}}{\lambda_{\min}^{\star2/3}\sqrt{p}}\sqrt{\frac{\mu^2 r\log d}{d^2 p}},$$

where we use the conditions $\sigma_{\max}/\sigma_{\min} \asymp 1$ and $\kappa \asymp 1$. In addition, $\boldsymbol{Z}_{m,:}$ is a sum of independent zero-mean random vectors with bounds

$$B_2 := \max_{1\leq i,j\leq d}\left\|p^{-1}E_{i,j,m}\chi_{i,j,m}\widetilde{\boldsymbol{U}}^\star_{(i,j),:}(\widetilde{\boldsymbol{U}}^{\star\top}\widetilde{\boldsymbol{U}}^\star)^{-1}\right\|_{\psi_1} \lesssim \frac{\sigma_{\max}}{p}\left\|\widetilde{\boldsymbol{U}}^\star\right\|_{2,\infty}\left\|(\widetilde{\boldsymbol{U}}^{\star\top}\widetilde{\boldsymbol{U}}^\star)^{-1}\right\| \lesssim \frac{\sigma_{\max}}{p}\cdot\frac{\mu\sqrt{r}\lambda_{\max}^{\star2/3}}{d\lambda_{\min}^{\star4/3}},$$

$$V_2 := \sum_{1\leq i,j\leq d}\frac{1}{p^2}\mathbb{E}\big[E_{i,j,m}^2\chi_{i,j,m}\big]\left\|\widetilde{\boldsymbol{U}}^\star_{(i,j),:}(\widetilde{\boldsymbol{U}}^{\star\top}\widetilde{\boldsymbol{U}}^\star)^{-1}\right\|_2^2 \lesssim \frac{\sigma_{\max}^2}{p}\left\|\widetilde{\boldsymbol{U}}^\star\right\|_{\mathrm{F}}^2\left\|(\widetilde{\boldsymbol{U}}^{\star\top}\widetilde{\boldsymbol{U}}^\star)^{-1}\right\|^2 \lesssim \frac{\sigma_{\max}^2}{p}\cdot\frac{r\lambda_{\max}^{\star4/3}}{\lambda_{\min}^{\star8/3}},$$

which rely on (63a) and (63b). It then follows from the matrix Bernstein inequality that

$$\|\boldsymbol{Z}_{m,:}\|_2 \lesssim B_2\log^2 d + \sqrt{V_2\log d} \lesssim \frac{\sigma_{\max}\lambda_{\max}^{\star2/3}}{\lambda_{\min}^{\star4/3}\sqrt{p}}\left\{\frac{\mu\sqrt{r}\,\log^2 d}{d\sqrt{p}} + \sqrt{r\log d}\right\} \asymp \frac{\sigma_{\max}\sqrt{r\log d}}{\lambda_{\min}^{\star2/3}\sqrt{p}} \tag{72}$$

with probability at least $1 - O\big(d^{-11}\big)$, as long as $p \gtrsim \mu^2 d^{-2}\log^3$ and $\kappa \asymp 1$. Therefore, we arrive at

$$\left\|\boldsymbol{Z}_{m,:} - \boldsymbol{Z}_{m,:}\boldsymbol{S}_m^{\star-1/2}\boldsymbol{\Sigma}_m^{\star1/2}\right\|_2 \lesssim \frac{\sigma_{\max}\sqrt{r\log d}}{\lambda_{\min}^{\star2/3}\sqrt{p}}\cdot\sqrt{\frac{\lambda_{\max}^{\star4/3}p}{\sigma_{\min}^2}}\cdot\frac{\sigma_{\max}}{\lambda_{\min}^{\star2/3}\sqrt{p}}\sqrt{\frac{\mu^2 r\log d}{d^2 p}} \lesssim \frac{\sigma_{\max}}{\lambda_{\min}^{\star2/3}\sqrt{p}}\frac{\mu r\log d}{d\sqrt{p}}. \tag{73}$$

Combining (68) and (73) and then taking a union bound over $1 \leq m \leq d$, we reach the advertised bound on the $\ell_{2,\infty}$ norm of the residual term $\boldsymbol{W}_0$.

1) Proof of Lemma 15: Recalling the definitions of $\boldsymbol{S}_m^\star$ and $\boldsymbol{\Sigma}_m^\star$, we can express

$$\boldsymbol{S}_m^\star - \boldsymbol{\Sigma}_m^\star = \frac{2}{p}(\widetilde{\boldsymbol{U}}^{\star\top}\widetilde{\boldsymbol{U}}^\star)^{-1}\sum_{1\leq i,j\leq d}\Big(\sigma_{i,j,m}^2(p^{-1}\chi_{i,j,m}-1)\cdot(\widetilde{\boldsymbol{U}}^\star_{(i,j),:})^\top\widetilde{\boldsymbol{U}}^\star_{(i,j),:}(\widetilde{\boldsymbol{U}}^{\star\top}\widetilde{\boldsymbol{U}}^\star)^{-1}\Big)$$

as a sum of independent zero-mean random matrices in $\mathbb{R}^{d\times d}$. By (63), it is straightforward to bound

$$B := \max_{1\leq i,j\leq d}\left\|\sigma_{i,j,m}^2(p^{-1}\chi_{i,j,m}-1)(\widetilde{\boldsymbol{U}}^\star_{(i,j),:})^\top\widetilde{\boldsymbol{U}}^\star_{(i,j),:}\right\| \leq \frac{\sigma_{\max}^2}{p}\left\|\widetilde{\boldsymbol{U}}^\star\right\|_{2,\infty}^2 \lesssim \frac{\sigma_{\max}^2}{p}\cdot\frac{\mu^2 r}{d^2}\lambda_{\max}^{\star4/3};$$

and

$$V := \Big\|\sum_{1\leq i,j\leq d}\sigma_{i,j,m}^2\mathbb{E}\big[(p^{-1}\chi_{i,j,m}-1)^2\big]\cdot(\widetilde{\boldsymbol{U}}^\star_{(i,j),:})^\top\widetilde{\boldsymbol{U}}^\star_{(i,j),:}(\widetilde{\boldsymbol{U}}^\star_{(i,j),:})^\top\widetilde{\boldsymbol{U}}^\star_{(i,j),:}\Big\| \leq \frac{\sigma_{\max}^2}{p}\left\|\widetilde{\boldsymbol{U}}^\star\right\|_{2,\infty}^2\left\|\widetilde{\boldsymbol{U}}^{\star\top}\widetilde{\boldsymbol{U}}^\star\right\| \lesssim \frac{\sigma_{\max}^2}{p}\cdot\frac{\mu^2 r}{d^2}\lambda_{\max}^{\star4/3}\cdot\lambda_{\max}^{\star4/3}.$$

Invoke the matrix Bernstein inequality to reveal: with probability at least $1 - O\big(d^{-11}\big)$,

$$\Big\|\sum_{1\leq i,j\leq d}\sigma_{i,j,m}^2(p^{-1}\chi_{i,j,m}-1)(\widetilde{\boldsymbol{U}}^\star_{(i,j),:})^\top\widetilde{\boldsymbol{U}}^\star_{(i,j),:}\Big\| \lesssim B\log d + \sqrt{V\log d} \lesssim \sigma_{\max}^2\lambda_{\max}^{\star4/3}\left\{\frac{\mu^2 r\log d}{d^2 p} + \sqrt{\frac{\mu^2 r\log d}{d^2 p}}\right\} \asymp \sigma_{\max}^2\lambda_{\max}^{\star4/3}\sqrt{\frac{\mu^2 r\log d}{d^2 p}},$$

where the last line holds as long as $p \gtrsim \mu^2 r d^{-2}\log d$. Combined with (63b) and the condition $\kappa \asymp 1$, we conclude that

$$\|\boldsymbol{S}_m^\star - \boldsymbol{\Sigma}_m^\star\| \lesssim \frac{\sigma_{\max}^2\lambda_{\max}^{\star4/3}}{p}\sqrt{\frac{\mu^2 r\log d}{d^2 p}}\left\|(\widetilde{\boldsymbol{U}}^{\star\top}\widetilde{\boldsymbol{U}}^\star)^{-1}\right\|^2 \lesssim \frac{\sigma_{\max}^2}{\lambda_{\min}^{\star4/3}p}\sqrt{\frac{\mu^2 r\log d}{d^2 p}} = o\left(\frac{\sigma_{\max}^2}{\lambda_{\min}^{\star4/3}p}\right),$$

where the last step arises from the assumption that $p \gg \mu^2 r d^{-2}\log^2 d$.

B. Proof of Lemma 2

As before, let us use the notation $\boldsymbol{Z}$ and $\boldsymbol{W}_0 := \boldsymbol{X} - \boldsymbol{Z}$ as defined in (67) in Lemma 1. It is easily seen that (68) continues to hold in the non-Gaussian noise case (using the same proof). Therefore, it suffices to show that $\boldsymbol{Z}_{m,:}$ converges in distribution to a Gaussian random vector $\boldsymbol{g}_m \sim \mathcal{N}(\boldsymbol{0},\boldsymbol{\Sigma}_m^\star)$ in $\mathbb{R}^r$, towards which we resort to the Berry–Esseen-type theorem in Appendix B-C. In order to do so, we need to upper bound the quantity ρ defined in (65), which we proceed as follows

$$\begin{aligned}
\rho &\lesssim \sum_{1\leq i,j\leq d}\mathbb{E}\left[\left\|\frac{1}{p}E_{i,j,m}\chi_{i,j,m}\boldsymbol{\Sigma}_m^{\star-1/2}\widetilde{\boldsymbol{U}}^\star_{(i,j),:}(\widetilde{\boldsymbol{U}}^{\star\top}\widetilde{\boldsymbol{U}}^\star)^{-1}\right\|_2^3\right] \\
&= \frac{1}{p^2}\sum_{1\leq i,j\leq d}\mathbb{E}\big[|E_{i,j,m}|^3\big]\left\|\widetilde{\boldsymbol{U}}^\star_{(i,j),:}(\widetilde{\boldsymbol{U}}^{\star\top}\widetilde{\boldsymbol{U}}^\star)^{-1}\boldsymbol{\Sigma}_m^{\star-1/2}\right\|_2^3 \\
&\overset{\text{(i)}}{\lesssim} \frac{\sigma_{\max}^3}{p^2}\sum_{1\leq i,j\leq d}\left\|\widetilde{\boldsymbol{U}}^\star_{(i,j),:}\right\|_2^3\left\|(\widetilde{\boldsymbol{U}}^{\star\top}\widetilde{\boldsymbol{U}}^\star)^{-1}\right\|^3\left\|\boldsymbol{\Sigma}_m^{\star-1/2}\right\|^3 \\
&\lesssim \frac{\sigma_{\max}^3}{p^2}\left\|\widetilde{\boldsymbol{U}}^\star\right\|_{2,\infty}\left\|\widetilde{\boldsymbol{U}}^\star\right\|_{\mathrm{F}}^2\left\|(\widetilde{\boldsymbol{U}}^{\star\top}\widetilde{\boldsymbol{U}}^\star)^{-1}\right\|^3\left\|\boldsymbol{\Sigma}_m^{\star-1/2}\right\|^3 \\
&\overset{\text{(ii)}}{\lesssim} \frac{\sigma_{\max}^3}{p^2}\cdot\frac{\mu\sqrt{r}}{d}\lambda_{\max}^{\star2/3}\cdot r\lambda_{\max}^{\star4/3}\cdot\frac{1}{\lambda_{\min}^{\star4}}\cdot\frac{\lambda_{\max}^{\star2}p^{3/2}}{\sigma_{\min}^3} \overset{\text{(iii)}}{\lesssim} \frac{\mu r^{3/2}}{d\sqrt{p}}.
\end{aligned}$$

Here, (i) follows from the property of sub-Gaussian random variables, (ii) arises from (63a), (63b) and (70), whereas

(iii) results from the assumptions $\sigma_{\max}/\sigma_{\min} \asymp 1$ and $\kappa \asymp 1$. Therefore, invoke the Berry-Esseen theorem in Appendix B-C to conclude that: for any convex set $\mathcal{A} \subset \mathbb{R}^d$,

$$\left|\mathbb{P}\left\{\boldsymbol{Z}_{m,:} \in \mathcal{A}\right\} - \mathbb{P}\{\boldsymbol{g}_m \in \mathcal{A}\}\right| \lesssim \frac{\mu r^{3/2}}{\sqrt{d^{3/2} p}},$$

where $\boldsymbol{g}_m \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_m^\star)$ is a Gaussian random vector in $\mathbb{R}^r$.

C. Proof of Lemma 3

Without loss of generality, assume that $\boldsymbol{\Pi} = \boldsymbol{I}_r$ to simplify presentation. Fix an arbitrary $m \in [d]$. One can use (63a) and (63b) to upper bound

$$\begin{aligned}
&\left\|\boldsymbol{U}^\star_{m,:}\big(\widetilde{\boldsymbol{U}}^{\star\top}\widetilde{\boldsymbol{U}}(\widetilde{\boldsymbol{U}}^\top\widetilde{\boldsymbol{U}})^{-1} - \boldsymbol{I}_r\big)\right\|_2 \\
&\quad = \left\|\boldsymbol{U}^\star_{m,:}(\widetilde{\boldsymbol{U}} - \widetilde{\boldsymbol{U}}^\star)^\top\widetilde{\boldsymbol{U}}(\widetilde{\boldsymbol{U}}^\top\widetilde{\boldsymbol{U}})^{-1}\right\|_2 \\
&\quad \le \|\boldsymbol{U}^\star\|_{2,\infty}\left\|(\widetilde{\boldsymbol{U}} - \widetilde{\boldsymbol{U}}^\star)^\top\widetilde{\boldsymbol{U}}\right\|_2\left\|(\widetilde{\boldsymbol{U}}^\top\widetilde{\boldsymbol{U}})^{-1}\right\| \\
&\quad \lesssim \frac{1}{\lambda_{\min}^{\star 4/3}}\sqrt{\frac{\mu r}{d}}\,\lambda_{\max}^{\star 1/3}\left\|(\widetilde{\boldsymbol{U}} - \widetilde{\boldsymbol{U}}^\star)^\top\widetilde{\boldsymbol{U}}\right\|.
\end{aligned}$$

It then suffices to bound the spectral norm of $(\widetilde{\boldsymbol{U}} - \widetilde{\boldsymbol{U}}^\star)^\top\widetilde{\boldsymbol{U}}$. For notational convenience, we define

$$\boldsymbol{\Delta}_s := \boldsymbol{u}_s - \boldsymbol{u}_s^\star, \qquad 1 \le s \le r; \tag{74a}$$

$$\boldsymbol{\Delta} := \boldsymbol{U} - \boldsymbol{U}^\star. \tag{74b}$$

Let us decompose

$$\begin{aligned}
(\widetilde{\boldsymbol{U}} - \widetilde{\boldsymbol{U}}^\star)^\top\widetilde{\boldsymbol{U}} &= (\widetilde{\boldsymbol{U}} - \widetilde{\boldsymbol{U}}^\star)^\top\widetilde{\boldsymbol{U}}^\star \\
&\quad + (\widetilde{\boldsymbol{U}} - \widetilde{\boldsymbol{U}}^\star)^\top(\widetilde{\boldsymbol{U}} - \widetilde{\boldsymbol{U}}^\star),
\end{aligned} \tag{75}$$

and look at these two matrices separately.

1. We begin with the first term $(\widetilde{\boldsymbol{U}} - \widetilde{\boldsymbol{U}}^\star)^\top\widetilde{\boldsymbol{U}}^\star$ in (75), whose entries are given by

$$\begin{aligned}
\big((\widetilde{\boldsymbol{U}} - \widetilde{\boldsymbol{U}}^\star)^\top\widetilde{\boldsymbol{U}}^\star\big)_{i,j} &= \big\langle\boldsymbol{u}_i, \boldsymbol{u}_j^\star\big\rangle^2 - \big\langle\boldsymbol{u}_i^\star, \boldsymbol{u}_j^\star\big\rangle^2 \\
&= \big\langle\boldsymbol{u}_i^\star + \boldsymbol{\Delta}_i, \boldsymbol{u}_j^\star\big\rangle^2 - \big\langle\boldsymbol{u}_i^\star, \boldsymbol{u}_j^\star\big\rangle^2 \\
&= 2\big\langle\boldsymbol{u}_i^\star, \boldsymbol{u}_j^\star\big\rangle\big\langle\boldsymbol{\Delta}_i, \boldsymbol{u}_j^\star\big\rangle + \big\langle\boldsymbol{\Delta}_i, \boldsymbol{u}_j^\star\big\rangle^2
\end{aligned} \tag{76}$$

for all $1 \le i, j \le r$. Here, we have used the fact that $\big\langle\boldsymbol{a}^{\otimes 2}, \boldsymbol{b}^{\otimes 2}\big\rangle = \langle\boldsymbol{a}, \boldsymbol{b}\rangle^2$ for any $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^d$. Therefore, one can express

$$\begin{aligned}
(\widetilde{\boldsymbol{U}} - \widetilde{\boldsymbol{U}}^\star)^\top\widetilde{\boldsymbol{U}}^\star &= 2\,(\boldsymbol{U}^{\star\top}\boldsymbol{U}^\star) \odot (\boldsymbol{\Delta}^\top\boldsymbol{U}^\star) \\
&\quad + (\boldsymbol{\Delta}^\top\boldsymbol{U}^\star) \odot (\boldsymbol{\Delta}^\top\boldsymbol{U}^\star),
\end{aligned} \tag{77}$$

where we recall that $\odot$ is the Hadamard (entrywise) product. In the sequel, we shall treat these two terms individually.

- With regards to $(\boldsymbol{\Delta}^\top\boldsymbol{U}^\star) \odot (\boldsymbol{\Delta}^\top\boldsymbol{U}^\star)$, we can simply bound

$$\begin{aligned}
&\left\|(\boldsymbol{\Delta}^\top\boldsymbol{U}^\star) \odot (\boldsymbol{\Delta}^\top\boldsymbol{U}^\star)\right\| \\
&\quad \le \left\|(\boldsymbol{\Delta}^\top\boldsymbol{U}^\star) \odot (\boldsymbol{\Delta}^\top\boldsymbol{U}^\star)\right\|_{\mathrm{F}} \\
&\quad \overset{\text{(i)}}{\le} \left\|\boldsymbol{\Delta}^\top\boldsymbol{U}^\star\right\|_\infty\left\|\boldsymbol{\Delta}^\top\boldsymbol{U}^\star\right\|_{\mathrm{F}} \\
&\quad \overset{\text{(ii)}}{\le} \max_{1\le i\le r}\|\boldsymbol{\Delta}_i\|_2\max_{1\le i\le r}\|\boldsymbol{u}_i^\star\|_2\,\|\boldsymbol{\Delta}\|_{\mathrm{F}}\,\|\boldsymbol{U}^\star\| \\
&\quad \overset{\text{(iii)}}{\lesssim} \frac{\sigma_{\max}}{\lambda_{\min}^\star}\sqrt{\frac{rd\log d}{p}}\frac{\sigma_{\max}}{\lambda_{\min}^\star}\sqrt{\frac{rd\log d}{p}}\,\lambda_{\max}^{\star 4/3} \\
&\quad \overset{\text{(iv)}}{\lesssim} \frac{\sigma_{\max}^2}{\lambda_{\min}^{\star 2/3}}\frac{rd\log d}{p}.
\end{aligned} \tag{78}$$

Here, (i) is due to $\|\boldsymbol{A} \odot \boldsymbol{A}\|_{\mathrm{F}}^2 = \sum_{i,j}A_{i.j}^4 \le \max_{i,j}A_{i,j}^2\sum_{i,j}A_{i,j}^2 \le \|\boldsymbol{A}\|_\infty^2\|\boldsymbol{A}\|_{\mathrm{F}}^2$ for any matrix $\boldsymbol{A}$; (ii) arises from the inequality that $\|\boldsymbol{A}\boldsymbol{B}\|_{\mathrm{F}} \le \|\boldsymbol{A}\|\|\boldsymbol{B}\|_{\mathrm{F}}$ for any matrices $\boldsymbol{A}, \boldsymbol{B}$; (iii) uses (59); and (iv) arises from the condition that $\kappa \asymp 1$.

- Bounding the term $(\boldsymbol{U}^{\star\top}\boldsymbol{U}^\star) \odot (\boldsymbol{\Delta}^\top\boldsymbol{U}^\star)$ turns out to be more challenging. Towards this, we shall look at its diagonal and off-diagonal parts separately. For the off-diagonal part, by the incoherence condition (12c), one can bound

$$\begin{aligned}
&\left\|\mathcal{P}_{\textsf{off-diag}}\big((\boldsymbol{U}^{\star\top}\boldsymbol{U}^\star) \odot (\boldsymbol{\Delta}^\top\boldsymbol{U}^\star)\big)\right\| \\
&\quad \le \left\|\mathcal{P}_{\textsf{off-diag}}\big((\boldsymbol{U}^{\star\top}\boldsymbol{U}^\star) \odot (\boldsymbol{\Delta}^\top\boldsymbol{U}^\star)\big)\right\|_{\mathrm{F}} \\
&\quad \le \left\|\mathcal{P}_{\textsf{off-diag}}\Big(\boldsymbol{U}^{\star\top}\boldsymbol{U}^\star\Big)\right\|_\infty\left\|\boldsymbol{\Delta}^\top\boldsymbol{U}^\star\right\|_{\mathrm{F}} \\
&\quad \le \max_{1\le i\ne j\le r}\left|\big\langle\boldsymbol{u}_i^\star, \boldsymbol{u}_j^\star\big\rangle\right|\,\|\boldsymbol{\Delta}\|_{\mathrm{F}}\,\|\boldsymbol{U}^\star\| \\
&\quad \lesssim \sqrt{\frac{\mu}{d}}\,\lambda_{\max}^{\star 2/3}\cdot\frac{\sigma_{\max}}{\lambda_{\min}^\star}\sqrt{\frac{rd\log d}{p}}\,\lambda_{\max}^{\star 1/3}\cdot\lambda_{\max}^{\star 1/3} \\
&\quad \lesssim \sigma_{\max}\lambda_{\max}^{\star 1/3}\sqrt{\frac{\mu r\log d}{p}},
\end{aligned}$$

where we also use (59) and $\kappa \asymp 1$. Turning to the diagonal part, one observes that

$$\begin{aligned}
&\left\|\mathcal{P}_{\textsf{diag}}\big((\boldsymbol{U}^{\star\top}\boldsymbol{U}^\star) \odot (\boldsymbol{\Delta}^\top\boldsymbol{U}^\star)\big)\right\| \\
&\quad = \max_{1\le i\le r}\|\boldsymbol{u}_i\|_2^2\,|\langle\boldsymbol{\Delta}_i, \boldsymbol{u}_i^\star\rangle| \\
&\quad \le \max_{1\le i\le r}|\langle\boldsymbol{\Delta}_i, \boldsymbol{u}_i^\star\rangle|\,\lambda_{\max}^{\star 2/3}.
\end{aligned}$$

As result, the key step lies in upper bounding $\max_{1\le i\le r}|\langle\boldsymbol{\Delta}_i, \boldsymbol{u}_i^\star\rangle|$, which will be accomplished in the lemma below. The proof is deferred to the end of this section.

Lemma 16: Instate the assumptions of Lemma 3. With probability at least $1 - O\left(d^{-10}\right)$, one has

$$\max_{1\le i\le d}|\langle\boldsymbol{\Delta}_i, \boldsymbol{u}_i^\star\rangle| \lesssim \frac{\sigma_{\max}}{\lambda_{\min}^{\star 1/3}}\sqrt{\frac{d}{p}}\frac{\zeta}{\sqrt{\mu r}}, \tag{79}$$

where we recall the definition of ζ in (43) and the definition of $\boldsymbol{\Delta}_i$ in (74a).

With the above results in place, we conclude that

$$\begin{aligned}
&\left\|(\boldsymbol{U}^{\star\top}\boldsymbol{U}^\star) \odot (\boldsymbol{\Delta}^\top\boldsymbol{U}^\star)\right\| \\
&\quad \le \left\|\mathcal{P}_{\textsf{off-diag}}\big((\boldsymbol{U}^{\star\top}\boldsymbol{U}^\star) \odot (\boldsymbol{\Delta}^\top\boldsymbol{U}^\star)\big)\right\| \\
&\qquad + \left\|\mathcal{P}_{\textsf{diag}}\big((\boldsymbol{U}^{\star\top}\boldsymbol{U}^\star) \odot (\boldsymbol{\Delta}^\top\boldsymbol{U}^\star)\big)\right\| \\
&\quad \lesssim \sigma_{\max}\lambda_{\max}^{\star 1/3}\sqrt{\frac{\mu r\log d}{p}} + \frac{\sigma_{\max}}{\lambda_{\min}^{\star 1/3}}\sqrt{\frac{d}{p}}\frac{\zeta}{\sqrt{\mu r}}\lambda_{\max}^{\star 2/3} \\
&\quad \asymp \sigma_{\max}\lambda_{\max}^{\star 1/3}\sqrt{\frac{d}{p}}\frac{\zeta}{\sqrt{\mu r}},
\end{aligned}$$

where we use the condition $\kappa \asymp 1$, as well as the definition of ζ in (43) (which indicates that $\zeta \gtrsim \sqrt{\frac{\mu^2 r^2 \log d}{d}}$) in the last step.

- Combining the bounds above demonstrates that

$$\begin{aligned}
&\big\|(\widetilde{\boldsymbol{U}}-\widetilde{\boldsymbol{U}}^\star)^\top \widetilde{\boldsymbol{U}}^\star\big\| \\
&\lesssim \Big\|(\boldsymbol{U}^{\star\top}\boldsymbol{U}^\star)\odot(\boldsymbol{\Delta}^\top\boldsymbol{U}^\star)\Big\| + \Big\|(\boldsymbol{\Delta}^\top\boldsymbol{U}^\star)\odot(\boldsymbol{\Delta}^\top\boldsymbol{U}^\star)\Big\| \\
&\lesssim \sigma_{\max}\lambda_{\max}^{\star 1/3}\sqrt{\frac{d}{p}}\frac{\zeta}{\sqrt{\mu r}} + \frac{\sigma_{\max}^2}{\lambda_{\min}^{\star 2/3}}\frac{rd\log d}{p} \\
&\asymp \sigma_{\max}\lambda_{\max}^{\star 1/3}\sqrt{\frac{d}{p}}\frac{\zeta}{\sqrt{\mu r}},
\end{aligned} \tag{80}$$

where we have used the fact that $\zeta \gtrsim \frac{\sigma_{\max}}{\lambda_{\min}^\star}\sqrt{\frac{\mu r^3 d \log^2 d}{p}}$. In particular, we obtain the following upper bound for the spectral norm of $\widetilde{\boldsymbol{U}}-\widetilde{\boldsymbol{U}}^\star$:

$$\begin{aligned}
\big\|\widetilde{\boldsymbol{U}}-\widetilde{\boldsymbol{U}}^\star\big\| &= \big\|(\widetilde{\boldsymbol{U}}-\widetilde{\boldsymbol{U}}^\star)^\top \widetilde{\boldsymbol{U}}^\star \widetilde{\boldsymbol{U}}^{\star -1}\big\| \\
&\le \big\|\widetilde{\boldsymbol{U}}^{\star -1}\big\|\big\|(\widetilde{\boldsymbol{U}}-\widetilde{\boldsymbol{U}}^\star)^\top \widetilde{\boldsymbol{U}}^\star\big\| \\
&\lesssim \frac{\sigma_{\max}}{\lambda_{\min}^\star}\sqrt{\frac{d}{p}}\,\lambda_{\max}^{\star 2/3},
\end{aligned} \tag{81}$$

where we use the conditions that $p \gg \mu^3 r^3 d^{-3/2}\log^3 d$, $\sigma_{\max}/\lambda_{\min}^\star \ll \sqrt{p/(r^2 d^{3/2})}$ and $r \ll d/(\mu \log d)$.

2. Turning to the second term $(\widetilde{\boldsymbol{U}}-\widetilde{\boldsymbol{U}}^\star)^\top(\widetilde{\boldsymbol{U}}-\widetilde{\boldsymbol{U}}^\star)$ in (75), one can use (81) and $\kappa \asymp 1$ to upper bound

$$\big\|(\widetilde{\boldsymbol{U}}-\widetilde{\boldsymbol{U}}^\star)^\top(\widetilde{\boldsymbol{U}}-\widetilde{\boldsymbol{U}}^\star)\big\| \le \left(\frac{\sigma_{\max}}{\lambda_{\min}^\star}\sqrt{\frac{d}{p}}\,\lambda_{\max}^{\star 2/3}\right)^2. \tag{82}$$

3. Taking (78) and (80) together leads to

$$\begin{aligned}
&\big\|(\widetilde{\boldsymbol{U}}-\widetilde{\boldsymbol{U}}^\star)^\top \widetilde{\boldsymbol{U}}\big\| \\
&\le 2\big\|(\widetilde{\boldsymbol{U}}-\widetilde{\boldsymbol{U}}^\star)^\top \widetilde{\boldsymbol{U}}^\star\big\| + \big\|(\widetilde{\boldsymbol{U}}-\widetilde{\boldsymbol{U}}^\star)^\top(\widetilde{\boldsymbol{U}}-\widetilde{\boldsymbol{U}}^\star)\big\| \\
&\lesssim \sigma_{\max}\lambda_{\max}^{\star 1/3}\sqrt{\frac{d}{p}}\frac{\zeta}{\sqrt{\mu r}} + \left(\frac{\sigma_{\max}}{\lambda_{\min}^\star}\sqrt{\frac{d}{p}}\,\lambda_{\max}^{\star 2/3}\right)^2 \\
&\asymp \sigma_{\max}\lambda_{\max}^{\star 1/3}\sqrt{\frac{d}{p}}\frac{\zeta}{\sqrt{\mu r}},
\end{aligned}$$

where the last step arises from the definition of ζ (cf. (43)) that $\zeta \gtrsim \frac{\sigma_{\max}}{\lambda_{\min}^\star}\sqrt{\frac{\mu r d}{p}}$. Therefore, one can use the condition $\kappa \asymp 1$ to establish that

$$\begin{aligned}
&\big\|\boldsymbol{U}_{m,:}^\star\big(\widetilde{\boldsymbol{U}}^{\star\top}\widetilde{\boldsymbol{U}}(\widetilde{\boldsymbol{U}}^\top\widetilde{\boldsymbol{U}})^{-1} - \boldsymbol{I}_r\big)\big\|_2 \\
&\lesssim \frac{1}{\lambda_{\min}^{\star 4/3}}\sqrt{\frac{\mu r}{d}}\,\lambda_{\max}^{\star 1/3}\cdot\sigma_{\max}\lambda_{\max}^{\star 1/3}\sqrt{\frac{d}{p}}\frac{\zeta}{\sqrt{\mu r}} \lesssim \frac{\sigma_{\max}}{\lambda_{\min}^{\star 2/3}\sqrt{p}}\zeta
\end{aligned}$$

as claimed.

1) Proof of Lemma 16: Fix any $1 \le i \le r$. Recall the decomposition in (39) and (40) as well as the assumption $\boldsymbol{\Pi} = \boldsymbol{I}_r$ (without loss of generality). Left multiplying it by $\boldsymbol{U}^\star$ and right multiplying it by $\widetilde{\boldsymbol{U}}^\top\widetilde{\boldsymbol{U}}$, we arrive at

$$\begin{aligned}
&\boldsymbol{U}^{\star\top}(\boldsymbol{U}-\boldsymbol{U}^\star)\widetilde{\boldsymbol{U}}^\top\widetilde{\boldsymbol{U}} \\
&= -\boldsymbol{U}^{\star\top}\boldsymbol{U}^\star(\widetilde{\boldsymbol{U}}-\widetilde{\boldsymbol{U}}^\star)^\top\widetilde{\boldsymbol{U}}^\star + \boldsymbol{B},
\end{aligned} \tag{83}$$

where we use the following fact:

$$\begin{aligned}
&\boldsymbol{U}^{\star\top}\boldsymbol{U}^\star\big(\widetilde{\boldsymbol{U}}^{\star\top}\widetilde{\boldsymbol{U}}(\widetilde{\boldsymbol{U}}^\top\widetilde{\boldsymbol{U}})^{-1} - \boldsymbol{I}_r\big) \\
&\quad= \boldsymbol{U}^{\star\top}\boldsymbol{U}^\star\big(\widetilde{\boldsymbol{U}}^{\star\top}\widetilde{\boldsymbol{U}}(\widetilde{\boldsymbol{U}}^\top\widetilde{\boldsymbol{U}})^{-1} - \widetilde{\boldsymbol{U}}^\top\widetilde{\boldsymbol{U}}(\widetilde{\boldsymbol{U}}^\top\widetilde{\boldsymbol{U}})^{-1}\big) \\
&\quad= -\boldsymbol{U}^{\star\top}\boldsymbol{U}^\star(\widetilde{\boldsymbol{U}}-\widetilde{\boldsymbol{U}}^\star)^\top\widetilde{\boldsymbol{U}}(\widetilde{\boldsymbol{U}}^\top\widetilde{\boldsymbol{U}})^{-1},
\end{aligned}$$

and $\boldsymbol{B}$ is given by

$$\begin{aligned}
\boldsymbol{B} := &\underbrace{-\boldsymbol{U}^{\star\top}\boldsymbol{U}^\star(\widetilde{\boldsymbol{U}}-\widetilde{\boldsymbol{U}}^\star)^\top(\widetilde{\boldsymbol{U}}-\widetilde{\boldsymbol{U}}^\star)}_{=:\boldsymbol{B}_1} \\
&+ \underbrace{\boldsymbol{U}^{\star\top}\mathsf{unfold}\left(p^{-1}\mathcal{P}_\Omega(\boldsymbol{E})\right)\widetilde{\boldsymbol{U}}}_{=:\boldsymbol{B}_2} \\
&+ \underbrace{\boldsymbol{U}^{\star\top}\mathsf{unfold}\left(\left(\mathcal{I}-p^{-1}\mathcal{P}_\Omega\right)(\boldsymbol{T}-\boldsymbol{T}^\star)\right)\widetilde{\boldsymbol{U}}}_{=:\boldsymbol{B}_3} \\
&+ \underbrace{\boldsymbol{U}^{\star\top}\nabla g(\boldsymbol{U})}_{=:\boldsymbol{B}_4}.
\end{aligned} \tag{84}$$

One can compute the (i,i)-th entry of $\boldsymbol{U}^{\star\top}(\boldsymbol{U}-\boldsymbol{U}^\star)\widetilde{\boldsymbol{U}}^\top\widetilde{\boldsymbol{U}}$ on the left-hand side of (83) as follows

$$\begin{aligned}
\big(\boldsymbol{U}^{\star\top}(\boldsymbol{U}-\boldsymbol{U}^\star)\widetilde{\boldsymbol{U}}^\top\widetilde{\boldsymbol{U}}\big)_{i,i} &= \boldsymbol{u}_i^{\star\top}(\boldsymbol{U}-\boldsymbol{U}^\star)\widetilde{\boldsymbol{U}}^\top\widetilde{\boldsymbol{U}}_{:,i} \\
&= \sum_{1\le s\le r}\langle \boldsymbol{u}_i^\star, \boldsymbol{\Delta}_s\rangle\langle \boldsymbol{u}_s, \boldsymbol{u}_i\rangle^2,
\end{aligned}$$

where we recall that $\boldsymbol{\Delta}_s := \boldsymbol{u}_s - \boldsymbol{u}_s^\star$. In view of (76), the (i,i)-th entry of $\boldsymbol{U}^{\star\top}\boldsymbol{U}^\star(\widetilde{\boldsymbol{U}}-\widetilde{\boldsymbol{U}}^\star)^\top\widetilde{\boldsymbol{U}}^\star$ on the right-hand side of (83) is given by

$$\begin{aligned}
&\big(\boldsymbol{U}^{\star\top}\boldsymbol{U}^\star(\widetilde{\boldsymbol{U}}-\widetilde{\boldsymbol{U}}^\star)^\top\widetilde{\boldsymbol{U}}^\star\big)_{i,i} \\
&= (\boldsymbol{U}^{\star\top}\boldsymbol{U}^\star)_{i,:}(\widetilde{\boldsymbol{U}}-\widetilde{\boldsymbol{U}}^\star)^\top\widetilde{\boldsymbol{U}}_{:,i}^\star \\
&= \sum_{1\le s\le r}\langle \boldsymbol{u}_s^\star, \boldsymbol{u}_i^\star\rangle\big(2\langle \boldsymbol{u}_s^\star, \boldsymbol{u}_i^\star\rangle\langle \boldsymbol{\Delta}_s, \boldsymbol{u}_i^\star\rangle + \langle \boldsymbol{\Delta}_s, \boldsymbol{u}_i^\star\rangle^2\big),
\end{aligned}$$

Therefore, substituting these into (83) and rearranging terms lead to

$$\begin{aligned}
&\big(\|\boldsymbol{u}_i\|_2^4 + 2\|\boldsymbol{u}_i^\star\|_2^4\big)\langle \boldsymbol{\Delta}_i, \boldsymbol{u}_i^\star\rangle \\
&\quad= -\sum_{s:s\ne i}\big(\langle \boldsymbol{u}_s, \boldsymbol{u}_i\rangle^2 + 2\langle \boldsymbol{u}_s^\star, \boldsymbol{u}_i^\star\rangle^2\big)\langle \boldsymbol{\Delta}_s, \boldsymbol{u}_i^\star\rangle \\
&\qquad - \sum_{1\le s\le r}\langle \boldsymbol{u}_s^\star, \boldsymbol{u}_i^\star\rangle\langle \boldsymbol{\Delta}_s, \boldsymbol{u}_i^\star\rangle^2 + B_{i,i}.
\end{aligned} \tag{85}$$

It then suffices to control the quantities on the right-hand side of (85).

For the first term of (85), apply the Cauchy-Schwartz inequality to yield

$$\begin{aligned}
&\Big|\sum_{s:s\ne i}\big(\langle \boldsymbol{u}_s, \boldsymbol{u}_i\rangle^2 + 2\langle \boldsymbol{u}_s^\star, \boldsymbol{u}_i^\star\rangle^2\big)\langle \boldsymbol{\Delta}_s, \boldsymbol{u}_i^\star\rangle\Big| \\
&\le \max_{s:s\ne i}\big\{\langle \boldsymbol{u}_s, \boldsymbol{u}_i\rangle^2 + \langle \boldsymbol{u}_s^\star, \boldsymbol{u}_i^\star\rangle^2\big\}\|\boldsymbol{u}_i^\star\|_2\sum_{s\ne i}\|\boldsymbol{\Delta}_s\|_2
\end{aligned}$$

$$\leq \max_{s:s\neq i}\left\{\langle \boldsymbol{u}_s,\boldsymbol{u}_i\rangle^2+\langle \boldsymbol{u}_s^\star,\boldsymbol{u}_i^\star\rangle^2\right\}\|\boldsymbol{u}_i^\star\|_2\sqrt{r}\,\|\boldsymbol{U}-\boldsymbol{U}^\star\|_{\mathrm{F}}$$
$$\overset{\text{(i)}}{\lesssim}\left\{\frac{\mu}{d}+\frac{\sigma_{\max}^2}{\lambda_{\min}^{\star 2}}\frac{rd\log d}{p}\right\}\lambda_{\max}^{\star 4/3}\cdot\sqrt{r}\cdot\frac{\sigma_{\max}}{\lambda_{\min}^\star}\sqrt{\frac{rd\log d}{p}}\lambda_{\max}^{\star 2/3}$$
$$\overset{\text{(ii)}}{\lesssim}\sigma_{\max}\lambda_{\max}^\star\sqrt{\frac{\mu^2r^2\log d}{dp}}+\frac{\sigma_{\max}^2 rd\log d}{p}.$$

Here, (i) is due to the incoherence condition (12b) as well as (59a) and (62g), whereas (ii) arises from the noise condition $\sigma_{\max}/\lambda_{\min}^\star \ll \sqrt{p/(r^2 d\log d)}$. Turning to the second term of (85), the Cauchy-Schwartz inequality tells us that

$$\Big|\sum_{1\leq s\leq r}\langle \boldsymbol{u}_s^\star,\boldsymbol{u}_i^\star\rangle\langle \boldsymbol{\Delta}_s,\boldsymbol{u}_i^\star\rangle^2\Big|$$
$$\leq\|\boldsymbol{u}_i^\star\|_2^2\langle \boldsymbol{\Delta}_i,\boldsymbol{u}_i^\star\rangle^2+\max_{s:s\neq i}|\langle \boldsymbol{u}_s^\star,\boldsymbol{u}_i^\star\rangle|\,\|\boldsymbol{u}_i^\star\|_2^2\sum_{s:s\neq i}\|\boldsymbol{\Delta}_i\|_2^2$$
$$\leq\|\boldsymbol{\Delta}_i\|_2\|\boldsymbol{u}_i^\star\|_2^3|\langle \boldsymbol{\Delta}_i,\boldsymbol{u}_i^\star\rangle|+\max_{1\leq i\leq r}\|\boldsymbol{u}_i^\star\|_2^4\|\boldsymbol{U}-\boldsymbol{U}^\star\|_{\mathrm{F}}^2$$
$$\lesssim o(1)\|\boldsymbol{u}_i^\star\|_2^4|\langle \boldsymbol{\Delta}_i,\boldsymbol{u}_i^\star\rangle|+\lambda_{\max}^{\star 4/3}\cdot\frac{\sigma_{\max}^2}{\lambda_{\min}^{\star 2}}\frac{rd\log d}{p}\lambda_{\max}^{\star 2/3},$$

where we use (59a) and (62c). Substituting these into (85) and using (62e) and $\kappa\asymp 1$, we arrive at

$$\lambda_{\min}^{\star 4/3}|\langle \boldsymbol{\Delta}_i,\boldsymbol{u}_i^\star\rangle|\lesssim\sigma_{\max}\lambda_{\max}^\star\sqrt{\frac{\mu^2r^2\log d}{dp}}+\frac{\sigma_{\max}^2rd\log d}{p}+|B_{i,i}|. \tag{86}$$

It remains to bound $|B_{i,i}|$. Towards this end, we claim for the moment that

$$|B_{i,i}|\lesssim\sigma_{\max}\lambda_{\max}^\star\sqrt{\frac{d}{p}}\frac{\zeta}{\sqrt{\mu r}}, \tag{87}$$

where ζ is defined in (43). If this were true, then one could use (86) and the condition $\kappa\asymp 1$ to obtain the advertised bound

$$|\langle \boldsymbol{\Delta}_i,\boldsymbol{u}_i^\star\rangle|\lesssim\frac{\sigma_{\max}}{\lambda_{\min}^{\star 1/3}}\sqrt{\frac{d}{p}}\cdot\left\{\frac{\zeta}{\sqrt{\mu r}}+\frac{\mu r\sqrt{\log d}}{d}+\frac{\sigma_{\max}}{\lambda_{\min}^\star}\sqrt{\frac{r^2d\log^2 d}{p}}\right\}$$
$$\asymp\frac{\sigma_{\max}}{\lambda_{\min}^{\star 1/3}}\sqrt{\frac{d}{p}}\frac{\zeta}{\sqrt{\mu r}},$$

where the last step holds due to the fact that $\zeta\gtrsim\frac{\mu^{3/2}r^{3/2}\sqrt{\log d}}{d}+\frac{\sigma_{\max}}{\lambda_{\min}^\star}\sqrt{\frac{\mu r^3d\log^2 d}{p}}$.

The remainder of the proof is thus devoted to proving the claim (87). Recalling the decomposition in (84), we shall control $(\boldsymbol{B}_j)_{i,i}$, $1\leq j\leq 4$ separately.

- For $\boldsymbol{B}_1$, using (62b), (63d) and $\kappa\asymp 1$, we can simply upper bound

$$\big|(\boldsymbol{B}_1)_{i,i}\big|\leq\big\|\boldsymbol{U}^{\star\top}\boldsymbol{U}^\star(\widetilde{\boldsymbol{U}}-\widetilde{\boldsymbol{U}}^\star)^\top(\widetilde{\boldsymbol{U}}-\widetilde{\boldsymbol{U}}^\star)\big\|$$
$$\leq\big\|\boldsymbol{U}^\star\big\|^2\big\|\widetilde{\boldsymbol{U}}-\widetilde{\boldsymbol{U}}^\star\big\|^2\leq\big\|\boldsymbol{U}^\star\big\|^2\big\|\widetilde{\boldsymbol{U}}-\widetilde{\boldsymbol{U}}^\star\big\|_{\mathrm{F}}^2$$
$$\lesssim\left(\lambda_{\max}^{\star 1/3}\cdot\frac{\sigma_{\max}}{\lambda_{\min}^{\star 1/3}}\sqrt{\frac{rd\log d}{p}}\right)^2$$
$$\lesssim\frac{\sigma_{\max}^2rd\log d}{p}. \tag{88}$$

- Regarding $\boldsymbol{B}_2$, we can decompose

$$(\boldsymbol{B}_2)_{i,i}=\big\langle p^{-1}\mathcal{P}_\Omega(\boldsymbol{E}),\boldsymbol{u}_i^\star\otimes\boldsymbol{u}_i\otimes\boldsymbol{u}_i\big\rangle$$
$$=\underbrace{\big\langle p^{-1}\mathcal{P}_\Omega(\boldsymbol{E}),\boldsymbol{u}_i^{\star\otimes 3}\big\rangle}_{=:\gamma_1}+\underbrace{\big\langle p^{-1}\mathcal{P}_\Omega(\boldsymbol{E}),\boldsymbol{u}_i^\star\otimes\boldsymbol{\Delta}_i\otimes\boldsymbol{u}_i\big\rangle}_{=:\gamma_2}+\underbrace{\big\langle p^{-1}\mathcal{P}_\Omega(\boldsymbol{E}),\boldsymbol{u}_i^\star\otimes\boldsymbol{u}_i^\star\otimes\boldsymbol{\Delta}_i\big\rangle}_{=:\gamma_3},$$

leaving us with three terms to control.

  - For γ_1, observe that $\gamma_1=\sum_{1\leq j,k,l\leq d}p^{-1}E_{j,k,l}\chi_{jkl}u_{i,j}^\star u_{i,k}^\star u_{i,l}^\star$ is a sum of independent zero-mean random variables. Applying the Bernstein inequality shows that with probability at least $1-O\left(d^{-20}\right)$,

$$|\gamma_1|\lesssim\sigma_{\max}\left\{\frac{\log^2 d}{p}\|\boldsymbol{u}_i^\star\|_\infty^3+\sqrt{\frac{\log d}{p}}\|\boldsymbol{u}_i^\star\|_2^3\right\}$$
$$\lesssim\sigma_{\max}\lambda_{\max}^\star\left\{\frac{\mu^{3/2}\log^2 d}{d^{3/2}p}+\sqrt{\frac{\log d}{p}}\right\}$$
$$\asymp\sigma_{\max}\lambda_{\max}^\star\sqrt{\frac{\log d}{p}},$$

where we use the incoherence condition (12b), and the last step holds true as long as $p\gtrsim\mu^3d^{-3}\log^3 d$.

  - Regarding γ_2 and γ_3, we know from [2, Lemma D.4] that with probability at least $1-O\left(d^{-20}\right)$,

$$\big\|p^{-1}\mathcal{P}_\Omega(\boldsymbol{E})\times_1\boldsymbol{u}_i^\star\big\|$$
$$\lesssim\sigma_{\max}\left\{\frac{\log^{5/2}d}{p}\|\boldsymbol{u}_i^\star\|_\infty+\sqrt{\frac{d\log d}{p}}\|\boldsymbol{u}_i^\star\|_2\right\}$$
$$\lesssim\sigma_{\max}\lambda_{\max}^{\star 1/3}\left\{\frac{\log^{5/2}d}{p}\sqrt{\frac{\mu}{d}}+\sqrt{\frac{d\log d}{p}}\right\}$$
$$\asymp\sigma_{\max}\lambda_{\max}^{\star 1/3}\sqrt{\frac{d\log d}{p}},$$

where we use the incoherence condition (12b) and the assumption that $p\gtrsim\mu d^{-2}\log^4 d$. Consequently, we can use (62e) and (59a) to obtain

$$|\gamma_2|+|\gamma_3|$$
$$\leq\big\|p^{-1}\mathcal{P}_\Omega(\boldsymbol{E})\times_1\boldsymbol{u}_i^\star\big\|\left(\|\boldsymbol{u}_i\|_2+\|\boldsymbol{u}_i^\star\|_2\right)\|\boldsymbol{\Delta}_i\|_2$$
$$\lesssim\sigma_{\max}\lambda_{\max}^\star\sqrt{\frac{d\log d}{p}}\frac{\sigma_{\max}}{\lambda_{\min}^\star}\sqrt{\frac{rd\log d}{p}}.$$

– Combining the bounds above and using the condition $\kappa \asymp 1$, we find that

$$\left| (\boldsymbol{B}_2)_{i,i} \right| \lesssim \sigma_{\max}\lambda_{\max}^{\star}\sqrt{\frac{\log d}{p}} + \frac{\sigma_{\max}^2\sqrt{r}\, d\log d}{p}. \tag{89}$$

- Turning to $\boldsymbol{B}_3$, we can upper bound

$$\begin{aligned}\left|(\boldsymbol{B}_3)_{i,i}\right| &= \left|\left\langle\left(\mathcal{I} - p^{-1}\mathcal{P}_{\Omega}\right)\left(\boldsymbol{T}-\boldsymbol{T}^{\star}\right), \boldsymbol{u}_i^{\star}\otimes \boldsymbol{u}_i^{\otimes 2}\right\rangle\right| \\ &\le \left\|\left(p^{-1}\mathcal{P}_{\Omega} - \mathcal{I}\right)\left(\boldsymbol{T}-\boldsymbol{T}^{\star}\right)\right\| \left\|\boldsymbol{u}_i^{\star}\right\|_2 \left\|\boldsymbol{u}_i\right\|_2^2.\end{aligned}$$

Hence, it boils down to upper bounding the spectral norm of $\left(p^{-1}\mathcal{P}_{\Omega} - \mathcal{I}\right)\left(\boldsymbol{T}-\boldsymbol{T}^{\star}\right)$. Towards this, we decompose

$$\boldsymbol{T}-\boldsymbol{T}^{\star} = \sum_{1\le s\le r} \boldsymbol{\Delta}_s\otimes\boldsymbol{u}_s^{\otimes 2} + \boldsymbol{u}_s^{\star}\otimes\boldsymbol{\Delta}_s\otimes\boldsymbol{u}_s + \boldsymbol{u}_s^{\star\otimes 2}\otimes\boldsymbol{\Delta}_s.$$

Applying Lemma 21 in Appendix I, one obtains

$$\begin{aligned}&\left\|\left(p^{-1}\mathcal{P}_{\Omega} - \mathcal{I}\right)\left(\boldsymbol{T}-\boldsymbol{T}^{\star}\right)\right\| \\ &\quad\le \left\|\left(p^{-1}\mathcal{P}_{\Omega} - \mathcal{I}\right)\mathbf{1}^{\otimes 3}\right\| \\ &\qquad\cdot\sum_{1\le s\le r}\left\|\boldsymbol{\Delta}_s\right\|_{\infty}\left(\left\|\boldsymbol{u}_s\right\|_{\infty}^2 + \left\|\boldsymbol{u}_s^{\star}\right\|_{\infty}\left\|\boldsymbol{u}_s\right\|_{\infty} + \left\|\boldsymbol{u}_s\right\|_{\infty}^2\right) \\ &\quad\lesssim \left\|\left(p^{-1}\mathcal{P}_{\Omega} - \mathcal{I}\right)\mathbf{1}^{\otimes 3}\right\| r\max_{1\le s\le r}\left\|\boldsymbol{\Delta}_s\right\|_{\infty}\max_{1\le s\le r}\left\|\boldsymbol{u}_s^{\star}\right\|_{\infty}^2,\end{aligned}$$

where we use (62d) that $\left\|\boldsymbol{\Delta}_i\right\|_{\infty} \ll \left\|\boldsymbol{u}_i^{\star}\right\|_{\infty}$. In addition, from [2, Lemma D.2], we know that

$$\left\|\mathcal{P}_{\Omega}\left(\mathbf{1}^{\otimes 3}\right) - p\mathbf{1}^{\otimes 3}\right\| \lesssim \log^3 d + \sqrt{dp}\,\log^{5/2} d$$

with probability at least $1 - O\left(d^{-20}\right)$. Consequently, we obtain

$$\begin{aligned}&\left\|\left(p^{-1}\mathcal{P}_{\Omega} - \mathcal{I}\right)\left(\boldsymbol{T}-\boldsymbol{T}^{\star}\right)\right\| \\ &\quad\lesssim \frac{r}{p}\left(\log^3 d + \sqrt{dp}\,\log^{5/2} d\right)\frac{\sigma_{\max}}{\lambda_{\min}^{\star}}\sqrt{\frac{\mu r\log d}{p}}\cdot\frac{\mu}{d}\lambda_{\max}^{\star} \\ &\quad\lesssim \sigma_{\max}\sqrt{\frac{d}{p}}\left\{\frac{\mu^{3/2}r^{3/2}\log^{7/2} d}{d^{3/2}p} + \frac{\mu^{3/2}r^{3/2}\log^3 d}{d\sqrt{p}}\right\}.\end{aligned}$$

Together with (62e), this enables us to conclude that

$$\begin{aligned}\left|(\boldsymbol{B}_3)_{i,i}\right| &\lesssim \lambda_{\max}^{\star}\left\|\left(p^{-1}\mathcal{P}_{\Omega} - \mathcal{I}\right)\left(\boldsymbol{T}-\boldsymbol{T}^{\star}\right)\right\| \\ &\lesssim \sigma_{\max}\lambda_{\max}^{\star}\sqrt{\frac{d}{p}} \\ &\quad\cdot\left\{\frac{\mu^{3/2}r^{3/2}\log^{7/2} d}{d^{3/2}p} + \frac{\mu^{3/2}r^{3/2}\log^3 d}{d\sqrt{p}}\right\}.\end{aligned} \tag{90}$$

- It remains to look at $\boldsymbol{B}_4$. As shown in [2], the loss function $g(\boldsymbol{U})$ is locally strong convex and smooth with respect to the initial estimate $\boldsymbol{U}^0$. By the standard result of convex optimization, we know that the Euclidean norm of the gradient undergoes contraction at each iteration, in the sense that

$$\left\|\nabla g\left(\boldsymbol{U}^{t+1}\right)\right\|_{\mathrm{F}} \le \rho\left\|\nabla g\left(\boldsymbol{U}^{t}\right)\right\|_{\mathrm{F}}$$

for some constant $0<\rho<1$. By the construction of our estimate $\boldsymbol{U} := \boldsymbol{U}^{t_0}$ and the assumptions $\sigma_{\max}/\lambda_{\min}^{\star} \gtrsim d^{-100}$ and $t_0 \lesssim \log d$, one has

$$\left\|\nabla g(\boldsymbol{U})\right\|_{\mathrm{F}} \lesssim \sigma_{\max}\lambda_{\max}^{\star 2/3}\sqrt{\frac{d}{p}}\,\frac{1}{d}. \tag{91}$$

Consequently, we can upper bound

$$\begin{aligned}\left| (\boldsymbol{B}_4)_{i,i} \right| &= \left|\boldsymbol{u}_i^{\star\top}\nabla_{\boldsymbol{u}_i} g(\boldsymbol{U})\right| \le \left\|\boldsymbol{u}_i^{\star}\right\|_2\left\|\nabla g(\boldsymbol{U})\right\|_{\mathrm{F}} \\ &\lesssim \sigma_{\max}\lambda_{\max}^{\star}\sqrt{\frac{d}{p}}\frac{1}{d}.\end{aligned} \tag{92}$$

- Taking collectively (88), (89), (90) and (92) yields that

$$\begin{aligned}\left|B_{i,i}\right| &\le \left|(\boldsymbol{B}_1)_{i,i}\right| + \left|(\boldsymbol{B}_2)_{i,i}\right| + \left|(\boldsymbol{B}_3)_{i,i}\right| + \left|(\boldsymbol{B}_4)_{i,i}\right| \\ &\lesssim \sigma_{\max}\lambda_{\max}^{\star}\sqrt{\frac{d}{p}} \\ &\quad\cdot\left\{\frac{\mu^{3/2}r^{3/2}\log^{7/2} d}{d^{3/2}p} + \frac{\mu^{3/2}r^{3/2}\log^3 d}{d\sqrt{p}}\right. \\ &\quad\left. + \sqrt{\frac{\log d}{d}} + \frac{\sigma_{\max}}{\lambda_{\max}^{\star}}\sqrt{\frac{r^2 d\log^2 d}{p}} + \frac{1}{d}\right\} \\ &\lesssim \sigma_{\max}\lambda_{\max}^{\star}\sqrt{\frac{d}{p}}\frac{\zeta}{\sqrt{\mu r}},\end{aligned}$$

where we recall the definition of ζ in (43).

D. Proof of Lemma 4

Without loss of generality, we assume that $\boldsymbol{\Pi} = \boldsymbol{I}_r$ for simplicity of presentation. For any fixed $m\in[d]$, it is straightforward to decompose

$$\begin{aligned}&\boldsymbol{e}_m^{\top}\mathsf{unfold}\left(\mathcal{P}_{\Omega}\left(\boldsymbol{E}\right)\right)\left(\widetilde{\boldsymbol{U}}(\widetilde{\boldsymbol{U}}^{\top}\widetilde{\boldsymbol{U}})^{-1} - \widetilde{\boldsymbol{U}}^{\star}(\widetilde{\boldsymbol{U}}^{\star\top}\widetilde{\boldsymbol{U}}^{\star})^{-1}\right) \\ &\quad= \underbrace{\boldsymbol{e}_m^{\top}\mathsf{unfold}\left(\mathcal{P}_{\Omega}\left(\boldsymbol{E}\right)\right)\left(\widetilde{\boldsymbol{U}} - \widetilde{\boldsymbol{U}}^{\star}\right)(\widetilde{\boldsymbol{U}}^{\top}\widetilde{\boldsymbol{U}})^{-1}}_{=:\,\beta_1} \\ &\qquad+ \underbrace{\boldsymbol{e}_m^{\top}\mathsf{unfold}\left(\mathcal{P}_{\Omega}\left(\boldsymbol{E}\right)\right)\widetilde{\boldsymbol{U}}^{\star}\left((\widetilde{\boldsymbol{U}}^{\top}\widetilde{\boldsymbol{U}})^{-1} - (\widetilde{\boldsymbol{U}}^{\star\top}\widetilde{\boldsymbol{U}}^{\star})^{-1}\right)}_{=:\,\beta_2}.\end{aligned}$$

In the sequel, we shall upper bound β_1 and β_2 separately.

1) Controlling β_1: From the fact (63c) that $\lambda_{\min}(\widetilde{\boldsymbol{U}}^{\top}\widetilde{\boldsymbol{U}}) \asymp \lambda_{\min}^{\star 4/3}$, one has

$$\begin{aligned}&\left\|\boldsymbol{e}_m^{\top}\mathsf{unfold}\left(\mathcal{P}_{\Omega}(\boldsymbol{E})\right)(\widetilde{\boldsymbol{U}} - \widetilde{\boldsymbol{U}}^{\star})(\widetilde{\boldsymbol{U}}^{\top}\widetilde{\boldsymbol{U}})^{-1}\right\|_2 \\ &\quad\le \left\|\boldsymbol{e}_m^{\top}\mathsf{unfold}\left(\mathcal{P}_{\Omega}(\boldsymbol{E})\right)(\widetilde{\boldsymbol{U}} - \widetilde{\boldsymbol{U}}^{\star})\right\|_2\left\|(\widetilde{\boldsymbol{U}}^{\top}\widetilde{\boldsymbol{U}})^{-1}\right\| \\ &\quad\asymp \frac{1}{\lambda_{\min}^{\star 4/3}}\left\|\boldsymbol{e}_m^{\top}\mathsf{unfold}\left(\mathcal{P}_{\Omega}(\boldsymbol{E})\right)(\widetilde{\boldsymbol{U}} - \widetilde{\boldsymbol{U}}^{\star})\right\|_2.\end{aligned} \tag{93}$$

Hence, it suffices to control the ℓ_2 norm of the m-th row of

$$\begin{aligned}&\mathsf{unfold}\left(\mathcal{P}_{\Omega}(\boldsymbol{E})\right)(\widetilde{\boldsymbol{U}} - \widetilde{\boldsymbol{U}}^{\star}) \\ &\quad= \left[\mathcal{P}_{\Omega}(\boldsymbol{E})\times_1\boldsymbol{u}_s\times_2\boldsymbol{u}_s - \mathcal{P}_{\Omega}(\boldsymbol{E})\times_1\boldsymbol{u}_s^{\star}\times_2\boldsymbol{u}_s^{\star}\right]_{1\le s\le r},\end{aligned}$$

which admits the following decomposition

$$\left[\mathcal{P}_{\Omega}(\boldsymbol{E})\times_1\boldsymbol{u}_s\times_2\boldsymbol{u}_s - \mathcal{P}_{\Omega}(\boldsymbol{E})\times_1\boldsymbol{u}_s^{\star}\times_2\boldsymbol{u}_s^{\star}\right]_{1\le s\le r}$$

$$= \underbrace{\left[\mathcal{P}_\Omega(\boldsymbol{E})\times_1 \boldsymbol{u}_s^{(m)}\times_2 \boldsymbol{u}_s^{(m)} - \mathcal{P}_\Omega(\boldsymbol{E})\times_1 \boldsymbol{u}_s^{\star}\times_2 \boldsymbol{u}_s^{\star}\right]_{1\le s\le r}}_{=:\gamma_1} + \underbrace{\left[\mathcal{P}_\Omega(\boldsymbol{E})\times_1 \boldsymbol{u}_s\times_2 \boldsymbol{u}_s - \mathcal{P}_\Omega(\boldsymbol{E})\times_1 \boldsymbol{u}_s^{(m)}\times_2 \boldsymbol{u}_s^{(m)}\right]_{1\le s\le r}}_{=:\gamma_2}.$$

Here, we recall the leave-one-out matrix $\boldsymbol{U}^{(m)} = \left[\boldsymbol{u}_s^{(m)}\right]_{1\le s\le r} \in \mathbb{R}^{d\times r}$ returned by Algorithm 4. In what follow, we shall control γ_1 and γ_2 separately.

- Let us start with γ_1. For notational convenience, we denote $\boldsymbol{\Delta}_s := \boldsymbol{u}_s - \boldsymbol{u}_s^\star$, $\boldsymbol{\Delta}_s^{(m)} := \boldsymbol{u}_s^{(m)} - \boldsymbol{u}_s^\star$, $\Delta_{s,i} = (\boldsymbol{\Delta}_s)_i$ and $\Delta_{s,i}^{(m)} = \big(\boldsymbol{\Delta}_s^{(m)}\big)_i$ for each $1\le s\le r$ and $1\le i\le d$. With this notation in place, for each $1\le s\le r$ we can expand

$$\begin{aligned}&\left(\mathcal{P}_\Omega(\boldsymbol{E})\times_1 \boldsymbol{u}_s^{(m)}\times_2 \boldsymbol{u}_s^{(m)} - \mathcal{P}_\Omega(\boldsymbol{E})\times_1 \boldsymbol{u}_s^{\star}\times_2 \boldsymbol{u}_s^{\star}\right)_m\\ &= \boldsymbol{u}_s^{(m)\top}(\mathcal{P}_\Omega(\boldsymbol{E}))_{:,:,m}\boldsymbol{u}_s^{(m)} - \boldsymbol{u}_s^{\star\top}(\mathcal{P}_\Omega(\boldsymbol{E}))_{:,:,m}\boldsymbol{u}_s^{\star}\\ &= 2\boldsymbol{\Delta}_s^{(m)\top}(\mathcal{P}_\Omega(\boldsymbol{E}))_{:,:,m}\boldsymbol{u}_s^{\star} + \boldsymbol{\Delta}_s^{(m)\top}(\mathcal{P}_\Omega(\boldsymbol{E}))_{:,:,m}\boldsymbol{\Delta}_s^{(m)},\end{aligned}$$

  where $\boldsymbol{R}_{:,:,m}\in\mathbb{R}^{d\times d}$ denotes the m-th mode-3 slice of a tensor $\boldsymbol{R}\in\mathbb{R}^{d\times d\times d}$ as defined in Section I-E.

  We first look at $\boldsymbol{\Delta}_s^{(m)\top}(\mathcal{P}_\Omega(\boldsymbol{E}))_{:,:,m}\boldsymbol{u}_s^{\star}$. By construction, $\boldsymbol{\Delta}_s^{(m)}$ is independent of the m-th mode-3 slice of $\mathcal{P}_\Omega(\boldsymbol{E})$. Consequently, $\boldsymbol{\Delta}_s^{(m)\top}(\mathcal{P}_\Omega(\boldsymbol{E}))_{:,:,m}\boldsymbol{u}_s^{\star} = \sum_{1\le i,j\le d} E_{i,j,m}\chi_{i,j,m}\Delta_{s,i}^{(m)}u_{s,j}^\star$ is a sum of independent zero-mean random variables (conditional on $\mathcal{P}_{\Omega_{-m}}(\boldsymbol{E})$). Using the incoherence condition (12b), straightforward calculation gives

$$\begin{aligned}B_1 &:= \max_{1\le i,j\le d}\left\|E_{i,j,m}\chi_{i,j,m}\Delta_{s,i}^{(m)}u_{s,j}^\star\right\|_{\psi_1}\\ &\lesssim \sigma_{\max}\left\|\boldsymbol{\Delta}_s^{(m)}\right\|_\infty\left\|\boldsymbol{u}_s^\star\right\|_\infty\\ &\le \sigma_{\max}\lambda_{\max}^{\star 1/3}\sqrt{\frac{\mu}{d}}\left\|\boldsymbol{\Delta}_s^{(m)}\right\|_\infty;\\ V_1 &:= \sum_{1\le i,j\le d}\mathbb{E}\left[E_{i,j,m}^2\chi_{i,j,m}^2\right]\left(\Delta_{s,i}^{(m)}\right)_i^2 u_{s,j}^{\star 2}\\ &\lesssim \sigma_{\max}^2 p\left\|\boldsymbol{\Delta}_s^{(m)}\right\|_2^2\left\|\boldsymbol{u}_s^\star\right\|_2^2\\ &\le \sigma_{\max}^2\lambda_{\max}^{\star 2/3}p\left\|\boldsymbol{\Delta}_s^{(m)}\right\|_2^2.\end{aligned}$$

  We then apply the Bernstein inequality to find that with probability at least $1-O\left(d^{-20}\right)$,

$$\begin{aligned}&\left|\boldsymbol{\Delta}_s^{(m)\top}(\mathcal{P}_\Omega(\boldsymbol{E}))_{:,:,m}\boldsymbol{u}_s^{\star}\right|\\ &\lesssim B_1\log^2 d + \sqrt{V_1\log d}\\ &\lesssim \sigma_{\max}\lambda_{\max}^{\star 1/3}\left\{\sqrt{\frac{\mu}{d}}\log^2 d\left\|\boldsymbol{\Delta}_s^{(m)}\right\|_\infty + \sqrt{p\log d}\left\|\boldsymbol{\Delta}_s^{(m)}\right\|_2\right\}.\end{aligned} \tag{94}$$

  Applying a similar argument, we can also upper bound

$$\begin{aligned}&\left|\boldsymbol{\Delta}_s^{(m)\top}(\mathcal{P}_\Omega(\boldsymbol{E}))_{:,:,m}\boldsymbol{\Delta}_s^{(m)}\right|\\ &\lesssim \sigma_{\max}\left\{\log^2 d\left\|\boldsymbol{\Delta}_s^{(m)}\right\|_\infty^2 + \sqrt{p\log d}\left\|\boldsymbol{\Delta}_s^{(m)}\right\|_2^2\right\}\\ &\ll \sigma_{\max}\lambda_{\max}^{\star 1/3}\cdot\left\{\sqrt{\frac{\mu}{d}}\log^2 d\left\|\boldsymbol{\Delta}_s^{(m)}\right\|_\infty + \sqrt{p\log d}\left\|\boldsymbol{\Delta}_s^{(m)}\right\|_2\right\},\end{aligned} \tag{95}$$

  where we utilize (62) in Lemma 13 that $\left\|\boldsymbol{\Delta}_s^{(m)}\right\|_\infty \ll \sqrt{\mu/d}\,\lambda_{\max}^{\star 1/3}$ and $\left\|\boldsymbol{\Delta}_s^{(m)}\right\|_2 \ll \lambda_{\max}^{\star 1/3}$ in the last inequality. Combining (94) and (95), and summing over $s\in[r]$, we obtain

$$\begin{aligned}&\left\|\boldsymbol{e}_m^\top\left[\mathcal{P}_\Omega(\boldsymbol{E})\times_1 \boldsymbol{u}_s^{(m)}\times_2 \boldsymbol{u}_s^{(m)} - \mathcal{P}_\Omega(\boldsymbol{E})\times_1 \boldsymbol{u}_s^{\star}\times_2 \boldsymbol{u}_s^{\star}\right]_s\right\|_2\\ &\lesssim \sigma_{\max}\lambda_{\max}^{\star 1/3}\cdot\left\{\sqrt{\frac{\mu}{d}}\log^2 d\sqrt{r}\left\|\boldsymbol{\Delta}_s^{(m)}\right\|_\infty + \sqrt{p\log d}\left\|\boldsymbol{U}^{(m)}-\boldsymbol{U}^\star\right\|_{\mathrm{F}}\right\}\\ &\lesssim \sigma_{\max}\lambda_{\max}^{\star 1/3}\cdot\left\{\sqrt{\frac{\mu}{d}}\log^2 d\cdot\frac{\sigma_{\max}}{\lambda_{\min}^\star}\sqrt{\frac{\mu r^2\log d}{p}}\lambda_{\max}^{\star 1/3} + \sqrt{p\log d}\cdot\frac{\sigma_{\max}}{\lambda_{\min}^\star}\sqrt{\frac{rd\log d}{p}}\lambda_{\max}^{\star 1/3}\right\}\\ &\asymp \frac{\sigma_{\max}\sqrt{p}}{\lambda_{\min}^{\star 1/3}}\sigma_{\max}\sqrt{\frac{rd\log^2 d}{p}},\end{aligned} \tag{96}$$

  where the last step holds as long as $p\gtrsim \mu^2 rd^{-2}\log^3 d$ and $\kappa\asymp 1$.

- Turning to γ_2, we can decompose

$$\begin{aligned}&\left(\mathcal{P}_\Omega(\boldsymbol{E})\times_1 \boldsymbol{u}_s\times_2 \boldsymbol{u}_s - \mathcal{P}_\Omega(\boldsymbol{E})\times_1 \boldsymbol{u}_s^{(m)}\times_2 \boldsymbol{u}_s^{(m)}\right)_m\\ &= 2\left(\boldsymbol{u}_s-\boldsymbol{u}_s^{(m)}\right)^\top(\mathcal{P}_\Omega(\boldsymbol{E}))_{:,:,m}\boldsymbol{u}_s^{(m)}\\ &\quad + \left(\boldsymbol{u}_s-\boldsymbol{u}_s^{(m)}\right)^\top(\mathcal{P}_\Omega(\boldsymbol{E}))_{:,:,m}\left(\boldsymbol{u}_s-\boldsymbol{u}_s^{(m)}\right).\end{aligned} \tag{97}$$

  For the first term, we use the Cauchy-Schwartz to derive

$$\begin{aligned}&\left|\left(\boldsymbol{u}_s-\boldsymbol{u}_s^{(m)}\right)^\top(\mathcal{P}_\Omega(\boldsymbol{E}))_{:,:,m}\boldsymbol{u}_s^{(m)}\right|\\ &\le \left\|\boldsymbol{u}_s-\boldsymbol{u}_s^{(m)}\right\|_2\left\|(\mathcal{P}_\Omega(\boldsymbol{E}))_{:,:,m}\boldsymbol{u}_s^{(m)}\right\|_2.\end{aligned}$$

  This motivates us to bound the ℓ_2 norm of $(\mathcal{P}_\Omega(\boldsymbol{E}))_{:,:,m}\boldsymbol{u}_s^{(m)} = \sum_{1\le i,j\le d}E_{i,j,m}\chi_{i,j,m}\boldsymbol{e}_i\boldsymbol{e}_j^\top\boldsymbol{u}_s^{(m)}$ — which is a sum of independent random zero-mean matrices. By Lemma 13, it is straightforward to calculate

$$\begin{aligned}B_2 &:= \max_{1\le i,j\le d}\left\|E_{i,j,m}\chi_{i,j,m}\boldsymbol{e}_i\boldsymbol{e}_j^\top\boldsymbol{u}_s^{(m)}\right\|_{\psi_1}\\ &\lesssim \sigma_{\max}\left\|\boldsymbol{u}_s^{(m)}\right\|_\infty \le \sigma_{\max}\lambda_{\max}^{\star 1/3}\sqrt{\frac{\mu}{d}};\\ V_2 &:= \sum_{1\le i,j\le d}\mathbb{E}\left[E_{i,j,m}^2\chi_{i,j,m}^2\right]\left(\boldsymbol{u}_{s,j}^{(m)}\right)^2\\ &\le \sigma_{\max}^2 dp\left\|\boldsymbol{u}_s^{(m)}\right\|_2^2 \lesssim \sigma_{\max}^2\lambda_{\max}^{\star 2/3}dp.\end{aligned}$$

  In view of the matrix Bernstein inequality, we show that with probability exceeding $1-O\left(d^{-20}\right)$,

$$\begin{aligned}\left\|(\mathcal{P}_\Omega(\boldsymbol{E}))_{:,:,m}\boldsymbol{u}_s^{(m)}\right\|_2 &\lesssim B_2\log^2 d + \sqrt{V_2\log d}\\ &\lesssim \sigma_{\max}\lambda_{\max}^{\star 1/3}\left\{\sqrt{\frac{\mu}{d}}\log^2 d + \sqrt{dp\log d}\right\}\\ &\asymp \sigma_{\max}\lambda_{\max}^{\star 1/3}\sqrt{dp\log d},\end{aligned}$$

  where the last step holds as long as $p\gg \mu d^{-2}\log^3 d$. Consequently, we reach

$$\left|\left(\boldsymbol{u}_s-\boldsymbol{u}_s^{(m)}\right)^\top(\mathcal{P}_\Omega(\boldsymbol{E}))_{:,:,m}\boldsymbol{u}_s^{(m)}\right|$$

$$\lesssim \sigma_{\max}\lambda_{\max}^{\star 1/3}\sqrt{dp\log d}\,\big\|\boldsymbol{u}_s-\boldsymbol{u}_s^{(m)}\big\|_2. \tag{98}$$

For the second term of (97), invoke [44, Lemma 11] to demonstrate that: with probability at least $1-O\left(d^{-20}\right)$,

$$\max_{1\le m\le d}\big\|\big(\mathcal{P}_{\Omega}(\boldsymbol{E})\big)_{:,:,m}\big\|\lesssim \sigma_{\max}\big(\sqrt{dp}+\log d\big). \tag{99}$$

This enables us to bound

$$\begin{aligned}
&\big|\big(\boldsymbol{u}_s-\boldsymbol{u}_s^{(m)}\big)^{\top}\left(\mathcal{P}_{\Omega}\left(\boldsymbol{E}\right)\right)_{:,:,m}\left(\boldsymbol{u}_s-\boldsymbol{u}_s^{(m)}\right)\big|\\
&\quad\le \big\|\big(\mathcal{P}_{\Omega}(\boldsymbol{E})\big)_{:,:,m}\big\|\big\|\boldsymbol{u}_s-\boldsymbol{u}_s^{(m)}\big\|_2^2\\
&\quad\lesssim \sigma_{\max}\big(\sqrt{dp}+\log d\big)\big\|\boldsymbol{u}_s-\boldsymbol{u}_s^{(m)}\big\|_2^2. \qquad (100)
\end{aligned}$$

Taking together the bounds (98), (100) and summing over $s\in[r]$, we obtain

$$\begin{aligned}
&\Big\|\boldsymbol{e}_m^{\top}\big[\mathcal{P}_{\Omega}(\boldsymbol{E})\times_1\boldsymbol{u}_s\times_2\boldsymbol{u}_s-\mathcal{P}_{\Omega}(\boldsymbol{E})\times_1\boldsymbol{u}_s^{(m)}\times_2\boldsymbol{u}_s^{(m)}\big]_s\Big\|_2\\
&\lesssim \sigma_{\max}\lambda_{\max}^{\star 1/3}\sqrt{dp\log d}\,\big\|\boldsymbol{U}-\boldsymbol{U}^{(m)}\big\|_{\mathrm{F}}\\
&\quad+\sigma_{\max}\big(\sqrt{dp}+\log d\big)\big\|\boldsymbol{U}-\boldsymbol{U}^{(m)}\big\|_{\mathrm{F}}^2\\
&\lesssim \sigma_{\max}\lambda_{\max}^{\star 1/3}\sqrt{dp\log d}\cdot\frac{\sigma_{\max}}{\lambda_{\min}^{\star}}\sqrt{\frac{\mu r\log d}{p}}\lambda_{\max}^{\star 1/3}\\
&\quad+\sigma_{\max}\big(\sqrt{dp}+\log d\big)\cdot\frac{\sigma_{\max}^2}{\lambda_{\min}^{\star 2}}\frac{\mu r\log d}{p}\lambda_{\max}^{\star 2/3}\\
&\asymp\frac{\sigma_{\max}\sqrt{p}}{\lambda_{\min}^{\star 1/3}}\sigma_{\max}\sqrt{\frac{\mu r d\log^2 d}{p}}, \qquad (101)
\end{aligned}$$

where we have used the conditions that $\sigma_{\max}/\lambda_{\min}^{\star}\ll\sqrt{p}/d^{3/4}$, $p\gg\mu r d^{-3/2}\log^2 d$ and $\kappa\asymp 1$.

- Putting (100) and (101) together and substituting them into (93) yield

$$\begin{aligned}
&\big\|\boldsymbol{e}_m^{\top}\mathsf{unfold}\big(p^{-1}\mathcal{P}_{\Omega}(\boldsymbol{E})\big)\big(\widetilde{\boldsymbol{U}}-\widetilde{\boldsymbol{U}}^{\star}\big)\big(\widetilde{\boldsymbol{U}}^{\top}\widetilde{\boldsymbol{U}}\big)^{-1}\big\|_2\\
&\quad\lesssim\frac{\sigma_{\max}\sqrt{p}}{\lambda_{\min}^{\star 2/3}}\frac{\sigma_{\max}}{\lambda_{\min}^{\star}}\sqrt{\frac{\mu r d\log^2 d}{p}}. \qquad (102)
\end{aligned}$$

2) Controlling β_2: Recognizing that

$$\begin{aligned}
&\big\|\boldsymbol{e}_m^{\top}\mathsf{unfold}\big(p^{-1}\mathcal{P}_{\Omega}(\boldsymbol{E})\big)\widetilde{\boldsymbol{U}}^{\star}\big(\big(\widetilde{\boldsymbol{U}}^{\top}\widetilde{\boldsymbol{U}}\big)^{-1}-\big(\widetilde{\boldsymbol{U}}^{\star\top}\widetilde{\boldsymbol{U}}^{\star}\big)^{-1}\big)\big\|_2\\
&\le\big\|\boldsymbol{e}_m^{\top}\mathsf{unfold}\big(p^{-1}\mathcal{P}_{\Omega}(\boldsymbol{E})\big)\widetilde{\boldsymbol{U}}^{\star}\big\|_2\big\|\big(\widetilde{\boldsymbol{U}}^{\top}\widetilde{\boldsymbol{U}}\big)^{-1}-\big(\widetilde{\boldsymbol{U}}^{\star\top}\widetilde{\boldsymbol{U}}^{\star}\big)^{-1}\big\|,
\end{aligned}$$

it suffices to control the ℓ_2 norm of $\boldsymbol{e}_m^{\top}\mathsf{unfold}\big(\mathcal{P}_{\Omega}(\boldsymbol{E})\widetilde{\boldsymbol{U}}^{\star}\big)$. Let us express

$$\boldsymbol{e}_m^{\top}\mathsf{unfold}\left(\mathcal{P}_{\Omega}\left(\boldsymbol{E}\right)\right)\widetilde{\boldsymbol{U}}^{\star}=\sum_{1\le i,j\le d}E_{i,j,m}\chi_{i,j,m}\widetilde{\boldsymbol{U}}^{\star}_{(i,j),:}$$

as a sum of independent zero-mean random vectors in $\mathbb{R}^r$. By (63a), it is straightforward to compute that

$$\begin{aligned}
B_3&\coloneqq\max_{1\le i,j\le d}\big\|E_{i,j,m}\chi_{i,j,m}\widetilde{\boldsymbol{U}}^{\star}_{(i,j),:}\big\|_{\psi_1}\lesssim\sigma_{\max}\,\big\|\widetilde{\boldsymbol{U}}^{\star}\big\|_{2,\infty}\\
&\lesssim\sigma_{\max}\lambda_{\max}^{\star 2/3}\frac{\mu\sqrt{r}}{d},\\
V_3&\coloneqq\sum_{1\le i,j\le d}\mathbb{E}\big[E_{i,j,m}^2\chi_{i,j,m}\big]\big\|\widetilde{\boldsymbol{U}}^{\star}_{(i,j),:}\big\|_2^2\lesssim\sigma_{\max}^2 p\,\big\|\widetilde{\boldsymbol{U}}^{\star}\big\|_{\mathrm{F}}^2\\
&\lesssim\sigma_{\max}^2\lambda_{\max}^{\star 4/3}rp,
\end{aligned}$$

where $\|\cdot\|_{\psi_1}$ denotes the sub-exponential norm. Applying the matrix Bernstein inequality yields that

$$\begin{aligned}
&\big\|\boldsymbol{e}_m^{\top}\mathsf{unfold}\big(p^{-1}\mathcal{P}_{\Omega}(\boldsymbol{E})\big)\widetilde{\boldsymbol{U}}^{\star}\big\|\lesssim B_3\log^2 d+\sqrt{V_3\log d}\\
&\quad\lesssim\sigma_{\max}\lambda_{\max}^{\star 2/3}\left\{\frac{\mu\sqrt{r}\,\log^2 d}{d}+\sqrt{rp\log d}\right\}\\
&\quad\asymp\sigma_{\max}\lambda_{\max}^{\star 2/3}\sqrt{rp\log d}
\end{aligned}$$

with probability at least $1-O\left(d^{-20}\right)$, where the last line holds as long as $p\gtrsim\mu^2 d^{-2}\log^3 d$.

In addition, we can use (63b) and (63c) to upper bound

$$\begin{aligned}
&\big\|(\widetilde{\boldsymbol{U}}^{\top}\widetilde{\boldsymbol{U}})^{-1}-(\widetilde{\boldsymbol{U}}^{\star\top}\widetilde{\boldsymbol{U}}^{\star})^{-1}\big\|\\
&\le\big\|(\widetilde{\boldsymbol{U}}^{\top}\widetilde{\boldsymbol{U}})^{-1}\big\|\big\|\widetilde{\boldsymbol{U}}^{\top}(\widetilde{\boldsymbol{U}}-\widetilde{\boldsymbol{U}}^{\star})\\
&\quad+(\widetilde{\boldsymbol{U}}-\widetilde{\boldsymbol{U}}^{\star})^{\top}\widetilde{\boldsymbol{U}}^{\star}\big\|\big\|(\widetilde{\boldsymbol{U}}^{\star\top}\widetilde{\boldsymbol{U}}^{\star})^{-1}\big\|\\
&\le\big\|(\widetilde{\boldsymbol{U}}^{\top}\widetilde{\boldsymbol{U}})^{-1}\big\|\big\|(\widetilde{\boldsymbol{U}}^{\star\top}\widetilde{\boldsymbol{U}}^{\star})^{-1}\big\|\big\|\widetilde{\boldsymbol{U}}-\widetilde{\boldsymbol{U}}^{\star}\big\|\big(\|\widetilde{\boldsymbol{U}}^{\star}\|+\|\widetilde{\boldsymbol{U}}\|\big)\\
&\lesssim\frac{1}{\lambda_{\min}^{\star 2}}\big\|\widetilde{\boldsymbol{U}}-\widetilde{\boldsymbol{U}}^{\star}\big\|\lesssim\frac{\sigma_{\max}}{\lambda_{\min}^{\star 3}}\sqrt{\frac{d}{p}}\,\lambda_{\max}^{\star 2/3}, \qquad (103)
\end{aligned}$$

where we use (81) in the last step.

Taking together the above bounds, we arrive at

$$\begin{aligned}
&\Big\|\boldsymbol{e}_m^{\top}\mathsf{unfold}\big(p^{-1}\mathcal{P}_{\Omega}(\boldsymbol{E})\big)\widetilde{\boldsymbol{U}}^{\star}\big(\big(\widetilde{\boldsymbol{U}}^{\top}\widetilde{\boldsymbol{U}}\big)^{-1}-\big(\widetilde{\boldsymbol{U}}^{\star\top}\widetilde{\boldsymbol{U}}^{\star}\big)^{-1}\big)\Big\|_2\\
&\quad\lesssim\frac{\sigma_{\max}\sqrt{p}}{\lambda_{\min}^{\star 2/3}}\frac{\sigma_{\max}}{\lambda_{\min}^{\star}}\sqrt{\frac{rd\log d}{p}}. \qquad (104)
\end{aligned}$$

3) Combining β_1 and β_2: Putting (102) and (104) together, we conclude that with probability at least $1-O\left(d^{-20}\right)$, one has

$$\begin{aligned}
&\big\|\mathsf{unfold}\big(p^{-1}\mathcal{P}_{\Omega}(\boldsymbol{E})\big)\big(\widetilde{\boldsymbol{U}}\big(\widetilde{\boldsymbol{U}}^{\top}\widetilde{\boldsymbol{U}}\big)^{-1}-\widetilde{\boldsymbol{U}}^{\star}\big(\widetilde{\boldsymbol{U}}^{\star\top}\widetilde{\boldsymbol{U}}^{\star}\big)^{-1}\big)\big\|_{2,\infty}\\
&\quad\lesssim\frac{\sigma_{\max}}{\lambda_{\min}^{\star 2/3}\sqrt{p}}\frac{\sigma_{\max}}{\lambda_{\min}^{\star}}\sqrt{\frac{\mu r d\log d}{p}}.
\end{aligned}$$

E. Proof of Lemma 5

Without loss of generality, assume that $\boldsymbol{\Pi}=\boldsymbol{I}_r$ for simplicity of presentation. Fix an arbitrary $1\le m\le d$. From (63c), we can upper bound

$$\begin{aligned}
&\big\|\boldsymbol{e}_m^{\top}\mathsf{unfold}\big((p^{-1}\mathcal{P}_{\Omega}-\mathcal{I})(\boldsymbol{T}-\boldsymbol{T}^{\star})\big)\widetilde{\boldsymbol{U}}\big(\widetilde{\boldsymbol{U}}^{\top}\widetilde{\boldsymbol{U}}\big)^{-1}\big\|_2\\
&\quad\le\big\|\boldsymbol{e}_m^{\top}\mathsf{unfold}\big((p^{-1}\mathcal{P}_{\Omega}-\mathcal{I})(\boldsymbol{T}-\boldsymbol{T}^{\star})\big)\widetilde{\boldsymbol{U}}\big\|_2\big\|\big(\widetilde{\boldsymbol{U}}^{\top}\widetilde{\boldsymbol{U}}\big)^{-1}\big\|\\
&\quad\asymp\frac{1}{\lambda_{\min}^{\star 4/3}}\big\|\boldsymbol{e}_m^{\top}\mathsf{unfold}\big((p^{-1}\mathcal{P}_{\Omega}-\mathcal{I})(\boldsymbol{T}-\boldsymbol{T}^{\star})\big)\widetilde{\boldsymbol{U}}\big\|_2. \qquad (105)
\end{aligned}$$

As a result, it suffices to upper bound the ℓ_2 norm of the m-th row of $\mathsf{unfold}\big((p^{-1}\mathcal{P}_{\Omega}-\mathcal{I})(\boldsymbol{T}-\boldsymbol{T}^{\star})\big)\widetilde{\boldsymbol{U}}$. Observe that this matrix can be decomposed as follows

$$\begin{aligned}
&\mathsf{unfold}\big((p^{-1}\mathcal{P}_{\Omega}-\mathcal{I})(\boldsymbol{T}-\boldsymbol{T}^{\star})\big)\widetilde{\boldsymbol{U}}\\
&\quad=\underbrace{\mathsf{unfold}\big((p^{-1}\mathcal{P}_{\Omega}-\mathcal{I})(\boldsymbol{T}^{(m)}-\boldsymbol{T}^{\star})\big)\widetilde{\boldsymbol{U}}^{(m)}}_{=:\beta_1}\\
&\quad\quad+\underbrace{\mathsf{unfold}\big((p^{-1}\mathcal{P}_{\Omega}-\mathcal{I})(\boldsymbol{T}-\boldsymbol{T}^{\star})\big)\big(\widetilde{\boldsymbol{U}}-\widetilde{\boldsymbol{U}}^{(m)}\big)}_{=:\beta_2}
\end{aligned}$$

$$+\underbrace{\mathsf{unfold}\big((p^{-1}\mathcal{P}_\Omega-\mathcal{I})(\boldsymbol{T}-\boldsymbol{T}^{(m)})\big)\widetilde{\boldsymbol{U}}^{(m)}}_{=:\,\beta_3},$$

where $\widetilde{\boldsymbol{U}}^{(m)}:=\big[\boldsymbol{u}_l^{(m)}\otimes\boldsymbol{u}_l^{(m)}\big]_{1\le l\le r}$. In what follows, we shall control these three terms separately. To simplify presentation, let us define

$$\boldsymbol{\xi}_s^{(m)}:=\boldsymbol{u}_s-\boldsymbol{u}_s^{(m)},\qquad 1\le s\le r. \tag{106}$$

1) Controlling β_1: By construction, we can express

$$\begin{aligned}&\boldsymbol{e}_m^\top\mathsf{unfold}\big((p^{-1}\mathcal{P}_\Omega-\mathcal{I})(\boldsymbol{T}^{(m)}-\boldsymbol{T}^\star)\big)\widetilde{\boldsymbol{U}}^{(m)}\\&\quad=\sum_{1\le i,j\le d}\big(\boldsymbol{T}^{(m)}-\boldsymbol{T}^\star\big)_{i,j,m}(p^{-1}\chi_{i,j,m}-1)\widetilde{\boldsymbol{U}}^{(m)}_{(i,j),:}\end{aligned}$$

as a sum of independent zero-mean random vectors. We claim for the moment that

$$\sum_{1\le i,j\le d}\big(\boldsymbol{T}^{(m)}-\boldsymbol{T}^\star\big)^2_{i,j,m}\lesssim\frac{\sigma_{\max}^2}{\lambda_{\min}^{\star2}}\frac{\mu r\log d}{p}\lambda_{\max}^{\star2}. \tag{107}$$

Combined with (61d) and Lemma 14, it is straightforward to compute that

$$\begin{aligned}B&:=\max_{1\le i,j\le d}\big\|\big(\boldsymbol{T}^{(m)}-\boldsymbol{T}^\star\big)_{i,j,m}(p^{-1}\chi_{ijm}-1)\widetilde{\boldsymbol{U}}^{(m)}_{(i,j),:}\big\|_{\psi_1}\\&\lesssim\frac{1}{p}\big\|\boldsymbol{T}^{(m)}-\boldsymbol{T}^\star\big\|_\infty\big\|\widetilde{\boldsymbol{U}}^{(m)}\big\|_{2,\infty}\\&\lesssim\frac{1}{p}\cdot\frac{\sigma_{\max}}{\lambda_{\min}^\star}\sqrt{\frac{\mu^3r^2\log d}{d^2p}}\lambda_{\max}^\star\cdot\frac{\mu\sqrt{r}\,\lambda_{\max}^{\star2/3}}{d},\end{aligned}$$

and

$$\begin{aligned}V&:=\sum_{1\le i,j\le d}\mathbb{E}\big[(\boldsymbol{T}^{(m)}-\boldsymbol{T}^\star)^2_{i,j,m}(p^{-1}\chi_{i,j,m}-1)^2\big]\big\|\widetilde{\boldsymbol{U}}^{(m)}_{(i,j),:}\big\|_2^2\\&\le\frac{1}{p}\big\|\widetilde{\boldsymbol{U}}^{(m)}\big\|_{2,\infty}^2\sum_{1\le i,j\le d}\big(\boldsymbol{T}^{(m)}-\boldsymbol{T}^\star\big)^2_{i,j,m}\\&\lesssim\frac{1}{p}\cdot\frac{\mu^2r\lambda_{\max}^{\star4/3}}{d^2}\cdot\frac{\sigma_{\max}^2}{\lambda_{\min}^{\star2}}\frac{\mu r\log d}{p}\lambda_{\max}^{\star2}.\end{aligned}$$

In view of the matrix Bernstein inequality, one has with probability at least $1-O\big(d^{-20}\big)$,

$$\begin{aligned}&\big\|\boldsymbol{e}_m^\top\mathsf{unfold}\big(p^{-1}\mathcal{P}_\Omega(\boldsymbol{T}^{(m)}-\boldsymbol{T}^\star)-(\boldsymbol{T}^{(m)}-\boldsymbol{T}^\star)\big)\widetilde{\boldsymbol{U}}^{(m)}\big\|_2\\&\quad\lesssim B\log^2 d+\sqrt{V\log d}\\&\quad\lesssim\frac{\sigma_{\max}\lambda_{\max}^{\star5/3}}{\lambda_{\min}^\star\sqrt{p}}\left\{\sqrt{\frac{\mu^3r^2}{d^2p}}\frac{\mu\sqrt{r}\,\log^{5/2}d}{d\sqrt{p}}+\sqrt{\frac{\mu^3r^2\log^2 d}{d^2p}}\right\}\\&\quad\asymp\frac{\sigma_{\max}\lambda_{\max}^{\star2/3}}{\sqrt{p}}\sqrt{\frac{\mu^3r^2\log^2 d}{d^2p}},\end{aligned} \tag{108}$$

where the last step holds as long as $p\gtrsim\mu^2rd^{-2}\log^3 d$ and $\kappa\asymp1$.

Now we are left with justifying the claim (107). Observe that

$$\begin{aligned}&\sum_{1\le i,j\le d}\big(\boldsymbol{T}^{(m)}-\boldsymbol{T}^\star\big)^2_{i,j,m}\\&\quad=\big\|\boldsymbol{U}^{(m)}\boldsymbol{F}^{(m)}\boldsymbol{U}^{(m)\top}-\boldsymbol{U}^\star\boldsymbol{F}^\star\boldsymbol{U}^{\star\top}\big\|_{\mathrm{F}}^2,\end{aligned}$$

where $\boldsymbol{F}^{(m)}$ and $\boldsymbol{F}^\star$ are diagonal matrices in $\mathbb{R}^{r\times r}$ with entries $F^{(m)}_{i,i}=\big(\boldsymbol{u}_i^{(m)}\big)_m$ and $F^\star_{i,i}=(\boldsymbol{u}_i^\star)_m$. Note that $\|\boldsymbol{F}^{(m)}\|\le\max_{1\le i\le r}\|\boldsymbol{u}_i^{(m)}\|_\infty$, $\|\boldsymbol{F}^\star\|\le\max_{1\le i\le r}\|\boldsymbol{u}_i^\star\|_\infty$and $\big\|\boldsymbol{F}^{(m)}-\boldsymbol{F}^\star\big\|_{\mathrm{F}}\le\big\|\boldsymbol{U}^{(m)}-\boldsymbol{U}^\star\big\|_{2,\infty}$. By the triangle inequality, it is straightforward to bound

$$\begin{aligned}&\big\|\boldsymbol{U}^{(m)}\boldsymbol{F}^{(m)}\boldsymbol{U}^{(m)\top}-\boldsymbol{U}^\star\boldsymbol{F}^\star\boldsymbol{U}^{\star\top}\big\|_{\mathrm{F}}\\&\le\big\|(\boldsymbol{U}^{(m)}-\boldsymbol{U}^\star)\boldsymbol{F}^{(m)}\boldsymbol{U}^{(m)\top}\big\|_{\mathrm{F}}\\&\quad+\big\|\boldsymbol{U}^\star(\boldsymbol{F}^{(m)}-\boldsymbol{F}^\star)\boldsymbol{U}^{(m)\top}\big\|_{\mathrm{F}}\\&\quad+\big\|\boldsymbol{U}^\star\boldsymbol{F}^\star(\boldsymbol{U}^{(m)}-\boldsymbol{U}^\star)^\top\big\|_{\mathrm{F}}\\&\le\big\|\boldsymbol{U}^{(m)}-\boldsymbol{U}^\star\big\|_{\mathrm{F}}\big\|\boldsymbol{F}^{(m)}\big\|\big\|\boldsymbol{U}^{(m)}\big\|\\&\quad+\|\boldsymbol{U}^\star\|\,\big\|\boldsymbol{F}^{(m)}-\boldsymbol{F}^\star\big\|_{\mathrm{F}}\big\|\boldsymbol{U}^{(m)}\big\|\\&\quad+\|\boldsymbol{U}^\star\|\,\|\boldsymbol{F}^\star\|\,\big\|\boldsymbol{U}^{(m)}-\boldsymbol{U}^\star\big\|_{\mathrm{F}}\\&\le\big\|\boldsymbol{U}^{(m)}-\boldsymbol{U}^\star\big\|_{\mathrm{F}}\max_{1\le i\le r}\|\boldsymbol{u}_i^{(m)}\|_\infty\big\|\boldsymbol{U}^{(m)}\big\|\\&\quad+\|\boldsymbol{U}^\star\|\,\big\|\boldsymbol{U}^{(m)}-\boldsymbol{U}^\star\big\|_{2,\infty}\big\|\boldsymbol{U}^{(m)}\big\|\\&\quad+\|\boldsymbol{U}^\star\|\max_{1\le i\le r}\|\boldsymbol{u}_i^\star\|_\infty\,\big\|\boldsymbol{U}^{(m)}-\boldsymbol{U}^\star\big\|_{\mathrm{F}}\\&\overset{\text{(i)}}{\lesssim}\frac{\sigma_{\max}}{\lambda_{\min}^\star}\sqrt{\frac{rd\log d}{p}}\,\lambda_{\max}^{\star1/3}\cdot\sqrt{\frac{\mu}{d}}\,\lambda_{\max}^{\star1/3}\cdot\lambda_{\max}^{\star1/3}\\&\quad+\lambda_{\max}^{\star2/3}\cdot\frac{\sigma_{\max}}{\lambda_{\min}^\star}\sqrt{\frac{\mu r\log d}{p}}\,\lambda_{\max}^{\star1/3}\\&\quad+\lambda_{\max}^{\star1/3}\cdot\sqrt{\frac{\mu}{d}}\,\lambda_{\max}^{\star1/3}\cdot\frac{\sigma_{\max}}{\lambda_{\min}^\star}\sqrt{\frac{rd\log d}{p}}\,\lambda_{\max}^{\star1/3}\\&\lesssim\frac{\sigma_{\max}}{\lambda_{\min}^\star}\sqrt{\frac{\mu r\log d}{p}}\,\lambda_{\max}^\star,\end{aligned}$$

where (i) results from Lemmas 11-13.

2) Controlling β_2: For each $s\in[r]$, we can further decompose

$$\begin{aligned}&\big(\mathsf{unfold}\big((p^{-1}\mathcal{P}_\Omega-\mathcal{I})(\boldsymbol{T}-\boldsymbol{T}^\star)\big)\big(\widetilde{\boldsymbol{U}}-\widetilde{\boldsymbol{U}}^{(m)}\big)\big)_{m,s}\\&\quad=\boldsymbol{u}_s^\top\big((p^{-1}\mathcal{P}_\Omega-\mathcal{I})(\boldsymbol{T}-\boldsymbol{T}^\star)\big)_{:,:,m}\boldsymbol{u}_s\\&\qquad-\boldsymbol{u}_s^{(m)\top}\big((p^{-1}\mathcal{P}_\Omega-\mathcal{I})(\boldsymbol{T}-\boldsymbol{T}^\star)\big)_{:,:,m}\boldsymbol{u}_s^{(m)}\\&\quad=\boldsymbol{\xi}_s^{(m)\top}\big((p^{-1}\mathcal{P}_\Omega-\mathcal{I})(\boldsymbol{T}-\boldsymbol{T}^\star)\big)_{:,:,m}\boldsymbol{u}_s\\&\qquad+\boldsymbol{u}_s^{(m)\top}\big((p^{-1}\mathcal{P}_\Omega-\mathcal{I})(\boldsymbol{T}-\boldsymbol{T}^\star)\big)_{:,:,m}\boldsymbol{\xi}_s^{(m)},\end{aligned}$$

where we recall that $\boldsymbol{\xi}_s^{(m)}:=\boldsymbol{u}_s-\boldsymbol{u}_s^{(m)}$. Let us first upper bound the spectral norm of the m-th mode-3 slice of $(p^{-1}\mathcal{P}_\Omega-\mathcal{I})(\boldsymbol{T}-\boldsymbol{T}^\star)$. Recalling the notation $\boldsymbol{\Delta}:=\boldsymbol{U}-\boldsymbol{U}^\star$ and $\boldsymbol{\Delta}_s=\boldsymbol{u}_s-\boldsymbol{u}_s^\star,1\le s\le r$, one can express

$$\boldsymbol{T}-\boldsymbol{T}^\star=\sum_{1\le s\le r}\big(\boldsymbol{\Delta}_s\otimes\boldsymbol{u}_s^{\otimes2}+\boldsymbol{u}_s^\star\otimes\boldsymbol{\Delta}_s\otimes\boldsymbol{u}_s+\boldsymbol{u}_s^{\star\otimes2}\otimes\boldsymbol{\Delta}_s\big),$$

and consequently,

$$(\boldsymbol{T}-\boldsymbol{T}^\star)_{:,:,m}=\boldsymbol{\Delta}\boldsymbol{F}\boldsymbol{U}^\top+\boldsymbol{U}^\star\boldsymbol{F}\boldsymbol{\Delta}^\top+\boldsymbol{U}^\star\boldsymbol{F}^\star\boldsymbol{\Delta}^\top,$$

where $\boldsymbol{F}$ and $\boldsymbol{F}^\star$ are diagonal matrices in $\mathbb{R}^{r\times r}$ with entries $F_{i,i}=(\boldsymbol{u}_i)_m$ and $F^\star_{i,i}=(\boldsymbol{u}_i^\star)_m$. Applying [94, Lemma 4.5] yields

$$\big\|\big((p^{-1}\mathcal{P}_\Omega-\mathcal{I})(\boldsymbol{T}-\boldsymbol{T}^\star)\big)_{:,:,m}\big\|$$

$$\le \left\| \left(p^{-1}\mathcal{P}_\Omega(\mathbf{1}^{\otimes 3}) - \mathbf{1}^{\otimes 3}\right)_{:,:,m} \right\| \|\boldsymbol{\Delta}\|_{2,\infty} \cdot \left(\|\boldsymbol{U}\boldsymbol{F}\|_{2,\infty} + \|\boldsymbol{U}^\star \boldsymbol{F}\|_{2,\infty} + \|\boldsymbol{U}^\star \boldsymbol{F}^\star\|_{2,\infty} \right).$$

The matrix Bernstein inequality then reveals that with probability at least $1 - O\left(d^{-20}\right)$,

$$\left\| \left(p^{-1}\mathcal{P}_\Omega(\mathbf{1}^{\otimes 3}) - \mathbf{1}^{\otimes 3}\right)_{:,:,m} \right\| \lesssim \frac{\log d}{p} + \sqrt{\frac{d \log d}{p}}.$$

Moreover, we know from (12b), (62a) and (62f) that

$$\|\boldsymbol{U}\boldsymbol{F}\|_{2,\infty} + \|\boldsymbol{U}^\star \boldsymbol{F}\|_{2,\infty} + \|\boldsymbol{U}^\star \boldsymbol{F}^\star\|_{2,\infty} \lesssim \frac{\mu\sqrt{r}}{d} \lambda_{\max}^{\star 2/3}.$$

Combining this with (59b), we demonstrate that

$$\begin{aligned} &\left\| \left((p^{-1}\mathcal{P}_\Omega - \mathcal{I})(\boldsymbol{T} - \boldsymbol{T}^\star)\right)_{:,:,m} \right\| \\ &\lesssim \left\{ \frac{\log d}{p} + \sqrt{\frac{d \log d}{p}} \right\} \cdot \frac{\sigma_{\max}}{\lambda_{\min}^\star} \sqrt{\frac{\mu r \log d}{p}}\, \lambda_{\max}^{\star 1/3} \frac{\mu\sqrt{r}}{d} \lambda_{\max}^{\star 2/3} \\ &\lesssim \frac{\sigma_{\max}}{\lambda_{\min}^\star} \left\{ \frac{\log d}{p} + \sqrt{\frac{d \log d}{p}} \right\} \sqrt{\frac{\mu^3 r^2 \log d}{d^2 p}}\, \lambda_{\max}^\star. \end{aligned}$$

As a result, one can use (61a) and (62e) to bound

$$\begin{aligned} &\left| \boldsymbol{\xi}_s^{(m)\top} \left((p^{-1}\mathcal{P}_\Omega - \mathcal{I})(\boldsymbol{T} - \boldsymbol{T}^\star)\right)_{:,:,m} \boldsymbol{u}_s \right| \\ &\le \left\| \left((p^{-1}\mathcal{P}_\Omega - \mathcal{I})(\boldsymbol{T} - \boldsymbol{T}^\star)\right)_{:,:,m} \right\| \left\| \boldsymbol{\xi}_s^{(m)} \right\|_2 \|\boldsymbol{u}_s\|_2 \\ &\lesssim \sigma_{\max} \lambda_{\max}^{\star 1/3} \sqrt{\frac{\mu^3 r^2 \log d}{d^2 p}} \left\{ \frac{\log d}{p} + \sqrt{\frac{d \log d}{p}} \right\} \left\| \boldsymbol{\xi}_s^{(m)} \right\|_2. \end{aligned}$$

Clearly, the upper bound also holds for $\boldsymbol{u}_s^{(m)\top}\left((p^{-1}\mathcal{P}_\Omega - \mathcal{I})\left(\boldsymbol{T} - \boldsymbol{T}^\star\right)\right)_{:,:,m} \boldsymbol{\xi}_s^{(m)}$. Summing over $s \in [r]$, we conclude that

$$\begin{aligned} &\left\| \boldsymbol{e}_m^\top \mathsf{unfold}\left((p^{-1}\mathcal{P}_\Omega - \mathcal{I})(\boldsymbol{T} - \boldsymbol{T}^\star)\right) \left(\widetilde{\boldsymbol{U}} - \widetilde{\boldsymbol{U}}^{(m)}\right) \right\|_2 \\ &\lesssim \sigma_{\max} \lambda_{\max}^{\star 1/3} \sqrt{\frac{\mu^3 r^2 \log d}{d^2 p}} \left\{ \frac{\log d}{p} + \sqrt{\frac{d \log d}{p}} \right\} \left\| \boldsymbol{U} - \boldsymbol{U}^{(m)} \right\|_{\mathrm{F}} \\ &\lesssim \sigma_{\max} \lambda_{\max}^{\star 1/3} \sqrt{\frac{\mu^3 r^2 \log d}{d^2 p}} \left\{ \frac{\log d}{p} + \sqrt{\frac{d \log d}{p}} \right\} \\ &\quad \cdot \frac{\sigma_{\max}}{\lambda_{\min}^\star} \sqrt{\frac{\mu r \log d}{p}} \lambda_{\max}^{\star 1/3} \\ &\ll \frac{\sigma_{\max} \lambda_{\max}^{\star 2/3}}{\sqrt{p}} \sqrt{\frac{\mu^3 r^2 \log d}{d^2 p}}. \end{aligned} \tag{109}$$

Here, the last inequality holds due to our assumptions $p \gg \mu^2 d^{-3/2} \log^3 d$ and $\sigma_{\max}/\lambda_{\min}^\star \ll \sqrt{p}/d^{3/4}$ and $\kappa \asymp 1$.

3) Controlling β_3: For each $s \in [r]$, we have

$$\begin{aligned} &\left(\mathsf{unfold}\left((p^{-1}\mathcal{P}_\Omega - \mathcal{I})\left(\boldsymbol{T} - \boldsymbol{T}^{(m)}\right)\right) \widetilde{\boldsymbol{U}}^{(m)}\right)_{m,s} \\ &\quad = \boldsymbol{u}_s^{(m)\top} \left((p^{-1}\mathcal{P}_\Omega - \mathcal{I})\left(\boldsymbol{T} - \boldsymbol{T}^{(m)}\right)\right)_{:,:,m} \boldsymbol{u}_s^{(m)}. \end{aligned}$$

Recall the definition of $\boldsymbol{\xi}_s^{(m)}$(cf. (106)), we decompose

$$\begin{aligned} \boldsymbol{T} - \boldsymbol{T}^{(m)} &= \sum_{1 \le s \le r} \left(\left(\boldsymbol{u}_s^{(m)} + \boldsymbol{\xi}_s^{(m)}\right)^{\otimes 3} - \left(\boldsymbol{u}_s^{(m)}\right)^{\otimes 3} \right) \\ &= \sum_{1 \le s \le r} \Big(\boldsymbol{\xi}_s^{(m)} \otimes \boldsymbol{u}_s^{(m)} \otimes \boldsymbol{u}_s^{(m)} + \boldsymbol{u}_s^{(m)} \otimes \boldsymbol{\xi}_s^{(m)} \otimes \boldsymbol{u}_s^{(m)} \\ &\quad + \boldsymbol{u}_s^{(m)} \otimes \boldsymbol{u}_s^{(m)} \otimes \boldsymbol{\xi}_s^{(m)} + \boldsymbol{\xi}_s^{(m)} \otimes \boldsymbol{\xi}_s^{(m)} \otimes \boldsymbol{u}_s^{(m)} \\ &\quad + \boldsymbol{\xi}_s^{(m)} \otimes \boldsymbol{u}_s^{(m)} \otimes \boldsymbol{\xi}_s^{(m)} + \boldsymbol{u}_s^{(m)} \otimes \boldsymbol{\xi}_s^{(m)} \otimes \boldsymbol{\xi}_s^{(m)} \\ &\quad + \boldsymbol{\xi}_s^{(m)} \otimes \boldsymbol{\xi}_s^{(m)} \otimes \boldsymbol{\xi}_s^{(m)} \Big). \end{aligned} \tag{110}$$

In view of the triangle inequality, it suffices to control these terms separately.

- Let us first consider the terms which are linear in terms of $\boldsymbol{\xi}_s^{(m)}$. For $\boldsymbol{M}_1 \coloneqq \sum_{1 \le s \le r} \boldsymbol{\xi}_s^{(m)} \otimes \boldsymbol{u}_s^{(m)} \otimes \boldsymbol{u}_s^{(m)}$, one can write

$$\begin{aligned} &\boldsymbol{u}_s^{(m)\top} \left\{ \left(p^{-1}\mathcal{P}_\Omega - \mathcal{I}\right) \boldsymbol{M}_1 \right\}_{:,:,m} \boldsymbol{u}_s^{(m)} \\ &= \sum_{1 \le s \le r} \Big\{ \left(\boldsymbol{u}_s^{(m)}\right)_m \\ &\quad \cdot \sum_{1 \le i,j \le d} \left(\boldsymbol{\xi}_s^{(m)}\right)_i \left(\boldsymbol{u}_s^{(m)}\right)_i \left(\boldsymbol{u}_s^{(m)}\right)_j^2 \left(p^{-1}\chi_{i,j,m} - 1\right) \Big) \Big\}. \end{aligned}$$

We shall use the Cauchy-Schwartz inequality to upper bound the absolute value of the quantity above. By construction, $\boldsymbol{u}_s^{(m)}$ is independent of $\{\chi_{i,j,m}\}_{1 \le i,j \le d}$. Applying a similar argument in [2, Lemma D.9], we know that with probability at least $1 - O\left(d^{-20}\right)$,

$$\begin{aligned} &\sum_{1 \le i \le d} \Big| \sum_{1 \le j \le d} \left(\boldsymbol{u}_s^{(m)}\right)_i \left(\boldsymbol{u}_s^{(m)}\right)_j^2 \left(p^{-1}\chi_{i,j,m} - 1\right) \Big|^2 \\ &\qquad \lesssim \frac{1}{p} \left\| \boldsymbol{u}_s^{(m)} \right\|_\infty^2 \left\| \boldsymbol{u}_s^{(m)} \right\|_2^4. \end{aligned}$$

This leads to the following upper bound

$$\begin{aligned} &\Big| \sum_{1 \le i \le d} \left(\boldsymbol{\xi}_s^{(m)}\right)_i \sum_{1 \le j \le d} \left(\boldsymbol{u}_s^{(m)}\right)_i \left(\boldsymbol{u}_s^{(m)}\right)_j^2 \left(p^{-1}\chi_{i,j,m} - 1\right) \Big|^2 \\ &\le \sum_{1 \le i \le d} \left(\boldsymbol{\xi}_s^{(m)}\right)_i^2 \\ &\quad \cdot \sum_{1 \le i \le d} \Big| \sum_{1 \le j \le d} \left(\boldsymbol{u}_s^{(m)}\right)_i^2 \left(\boldsymbol{u}_s^{(m)}\right)_j^2 \left(p^{-1}\chi_{i,j,m} - 1\right) \Big|^2 \\ &\lesssim \frac{1}{p} \left\| \boldsymbol{\xi}_s^{(m)} \right\|_2^2 \left\| \boldsymbol{u}_s^{(m)} \right\|_\infty^2 \left\| \boldsymbol{u}_s^{(m)} \right\|_2^4. \end{aligned}$$

It then follows that

$$\begin{aligned} &\left\| \left[\boldsymbol{u}_s^{(m)\top} \left\{ \left(p^{-1}\mathcal{P}_\Omega - \mathcal{I}\right) \boldsymbol{M}_1 \right\}_{:,:,m} \boldsymbol{u}_s^{(m)} \right]_s \right\|_2^2 \\ &\le \sum_{1 \le s \le r} \Big| \sum_{1 \le i,j \le d} \left(\boldsymbol{\xi}_s^{(m)}\right)_i \left(\boldsymbol{u}_s^{(m)}\right)_i \left(\boldsymbol{u}_s^{(m)}\right)_j^2 \left(p^{-1}\chi_{i,j,m} - 1\right) \Big|^2 \\ &\qquad \cdot \sum_{1 \le s \le r} \left(\boldsymbol{u}_s^{(m)}\right)_m^2 \\ &\lesssim \frac{1}{p} \sum_{1 \le s \le r} \left\| \boldsymbol{\xi}_s^{(m)} \right\|_2^2 \left\| \boldsymbol{u}_s^{(m)} \right\|_\infty^2 \left\| \boldsymbol{u}_s^{(m)} \right\|_2^4 \cdot \sum_{1 \le s \le r} \left(\boldsymbol{u}_s^{(m)}\right)_m^2 \\ &\le \frac{1}{p} \left\| \boldsymbol{U}^{(m)} \right\|_{2,\infty}^2 \max_{1 \le s \le r} \left\| \boldsymbol{u}_s^{(m)} \right\|_\infty^2 \left\| \boldsymbol{u}_s^{(m)} \right\|_2^4 \left\| \boldsymbol{U} - \boldsymbol{U}^{(m)} \right\|_{\mathrm{F}}^2 \\ &\lesssim \frac{1}{p} \cdot \frac{\mu r \lambda_{\max}^{\star 2/3}}{d} \cdot \frac{\mu \lambda_{\max}^{\star 2/3}}{d} \cdot \lambda_{\max}^{\star 4/3} \cdot \frac{\sigma_{\max}^2}{\lambda_{\min}^{\star 2}} \frac{\mu r \log d\, \lambda_{\max}^{\star 2/3}}{p} \\ &\lesssim \frac{\sigma_{\max}^2 \lambda_{\max}^{\star 10/3}}{\lambda_{\min}^{\star 2} p} \frac{\mu^3 r^2 \log d}{d^2 p}, \end{aligned} \tag{111}$$

where we use Lemmas 12 and 13 in (12b). Clearly, the upper bound is also valid for $\sum_{1 \le s \le r} \boldsymbol{u}_s^{(m)} \otimes \boldsymbol{\xi}_s^{(m)} \otimes \boldsymbol{u}_s^{(m)}$.

In an analogous manner, one can show that for $\boldsymbol{M}_2 := \sum_{1\le s\le r}\big(\boldsymbol{u}_s^{(m)}\big)\otimes\big(\boldsymbol{u}_s^{(m)}\big)\otimes\boldsymbol{\xi}_s^{(m)}$, with probability exceeding $1-O\big(d^{-20}\big)$,

$$\begin{aligned}
&\Big\|\Big[\boldsymbol{u}_s^{(m)\top}\big\{\big(p^{-1}\mathcal{P}_\Omega-\mathcal{I}\big)\boldsymbol{M}_2\big\}_{:,:,m}\boldsymbol{u}_s^{(m)}\Big]_s\Big\|_2^2\\
&\lesssim\sum_{1\le s\le r}\Big|\sum_{1\le i,j\le d}\big(\boldsymbol{u}_s^{(m)}\big)_i^2\big(\boldsymbol{u}_s^{(m)}\big)_j^2\big(p^{-1}\chi_{i,j,m}-1\big)\Big|^2\\
&\quad\cdot\sum_{1\le s\le r}\big(\boldsymbol{\xi}_s^{(m)}\big)_m^2\\
&\lesssim\frac{1}{p}\sum_{1\le s\le r}\big\|\boldsymbol{u}_s^{(m)}\big\|_\infty^4\big\|\boldsymbol{u}_s^{(m)}\big\|_2^4\cdot\sum_{1\le s\le r}\big(\boldsymbol{\xi}_s^{(m)}\big)_m^2\\
&\le\frac{1}{p}\big\|\boldsymbol{U}-\boldsymbol{U}^{(m)}\big\|_{\mathrm{F}}^2\max_{1\le s\le r}\big\|\boldsymbol{u}_s^{(m)}\big\|_\infty^4\big\|\boldsymbol{u}_s^{(m)}\big\|_2^2\big\|\boldsymbol{U}^{(m)}\big\|_{\mathrm{F}}^2\\
&\lesssim\frac{1}{p}\cdot\frac{\sigma_{\max}^2}{\lambda_{\min}^{\star 2}}\frac{\mu r\log d\lambda_{\max}^{\star 2/3}}{p}\cdot\frac{\mu^2\lambda_{\max}^{\star 4/3}}{d^2}\cdot\lambda_{\max}^{\star 2/3}\cdot r\lambda_{\max}^{\star 4/3}\\
&\lesssim\frac{\sigma_{\max}^2\lambda_{\max}^{\star 10/3}}{\lambda_{\min}^{\star 2}p}\frac{\mu^3r^2\log d}{d^2p},
\end{aligned}\tag{112}$$

where the last step arises from use (61) and (62).

- Next, we turn to the quadratic terms with respect to $\boldsymbol{\xi}_s^{(m)}$ in (110). For $\boldsymbol{M}_3 := \sum_{1\le s\le r}\boldsymbol{\xi}_s^{(m)}\otimes\boldsymbol{\xi}_s^{(m)}\otimes\boldsymbol{u}_s^{(m)}$, we can expand

$$\begin{aligned}
&\boldsymbol{u}_s^{(m)\top}\big\{\big(p^{-1}\mathcal{P}_\Omega-\mathcal{I}\big)\boldsymbol{M}_3\big\}_{:,:,m}\boldsymbol{u}_s^{(m)}\\
&=\sum_{1\le s\le r}\Big\{\big(\boldsymbol{u}_s^{(m)}\big)_m\sum_{1\le i,j\le d}\Big(\big(\boldsymbol{\xi}_s^{(m)}\big)_i\big(\boldsymbol{\xi}_s^{(m)}\big)_j\\
&\qquad\cdot\big(\boldsymbol{u}_s^{(m)}\big)_i\big(\boldsymbol{u}_s^{(m)}\big)_j\big(p^{-1}\chi_{i,j,m}-1\big)\Big)\Big\}.
\end{aligned}$$

Use the Cauchy-Schwartz inequality and the Bernstein inequality again to yield

$$\begin{aligned}
&\Big|\sum_{1\le i,j\le d}\big(\boldsymbol{\xi}_s^{(m)}\big)_i\big(\boldsymbol{\xi}_s^{(m)}\big)_j\big(\boldsymbol{u}_s^{(m)}\big)_i\big(\boldsymbol{u}_s^{(m)}\big)_j\big(p^{-1}\chi_{i,j,m}-1\big)\Big|^2\\
&\le\sum_{1\le i,j\le d}\big(\boldsymbol{\xi}_s^{(m)}\big)_i^2\big(\boldsymbol{\xi}_s^{(m)}\big)_j^2\\
&\quad\cdot\sum_{1\le i,j\le d}\big(\boldsymbol{u}_s^{(m)}\big)_i^2\big(\boldsymbol{u}_s^{(m)}\big)_j^2\big(p^{-1}\chi_{i,j,m}-1\big)^2\\
&\lesssim\frac{1}{p}\big\|\boldsymbol{\xi}_s^{(m)}\big\|_2^4\big\|\boldsymbol{u}_s^{(m)}\big\|_2^4
\end{aligned}$$

with probability at least $1-O\big(d^{-20}\big)$. Consequently, we find that

$$\begin{aligned}
&\Big\|\big[\boldsymbol{u}_s^{(m)\top}\big\{\big(p^{-1}\mathcal{P}_\Omega-\mathcal{I}\big)\boldsymbol{M}_3\big\}_{:,:,m}\boldsymbol{u}_s^{(m)}\big]_s\Big\|_2^2\\
&\lesssim\frac{1}{p}\sum_{1\le s\le r}\big(\boldsymbol{u}_s^{(m)}\big)_m^2\sum_{1\le s\le r}\big\|\boldsymbol{\xi}_s^{(m)}\big\|_2^4\big\|\boldsymbol{u}_s^{(m)}\big\|_2^4\\
&\lesssim\frac{1}{p}\big\|\boldsymbol{U}^{(m)}\big\|_{2,\infty}^2\max_{1\le s\le r}\big\|\boldsymbol{u}_s^{(m)}\big\|_2^4\\
&\quad\cdot\max_{1\le s\le r}\big\|\boldsymbol{u}_s-\boldsymbol{u}_s^{(m)}\big\|_\infty^2\big\|\boldsymbol{U}-\boldsymbol{U}^{(m)}\big\|_{\mathrm{F}}^2\\
&\lesssim\frac{1}{p}\cdot\frac{\mu r\lambda_{\max}^{\star 2/3}}{d}\cdot\lambda_{\max}^{\star 4/3}\cdot\frac{\mu\lambda_{\max}^{\star 2/3}}{d}\cdot\frac{\sigma_{\max}^2}{\lambda_{\min}^{\star 2}}\frac{\mu r\log d\lambda_{\max}^{\star 2/3}}{p}\\
&\lesssim\frac{\sigma_{\max}^2\lambda_{\max}^{\star 10/3}}{\lambda_{\min}^{\star 2}p}\frac{\mu^3r^2\log d}{d^2p},
\end{aligned}\tag{113}$$

where the last line holds true due to Lemmas 12 and 13, and $\big\|\boldsymbol{u}_s-\boldsymbol{u}_s^{(m)}\big\|_\infty\le\|\boldsymbol{u}_s\|_\infty+\|\boldsymbol{u}_s^{(m)}\|_\infty\lesssim\sqrt{\mu/d}\,\lambda_{\max}^{\star 1/3}$. Using a similar argument for (111), one can verify that for $\boldsymbol{M}_4 := \sum_{1\le s\le r}\boldsymbol{\xi}_s^{(m)}\otimes\boldsymbol{u}_s^{(m)}\otimes\boldsymbol{\xi}_s^{(m)}$ with probability at least $1-O\big(d^{-20}\big)$,

$$\begin{aligned}
&\Big\|\big[\boldsymbol{u}_s^{(m)\top}\big\{\big(p^{-1}\mathcal{P}_\Omega-\mathcal{I}\big)\boldsymbol{M}_4\big\}_{:,:,m}\boldsymbol{u}_s^{(m)}\big]_s\Big\|_2^2\\
&\lesssim\frac{1}{p}\big\|\boldsymbol{U}-\boldsymbol{U}^{(m)}\big\|_{2,\infty}^2\max_{1\le s\le r}\big\|\boldsymbol{u}_s^{(m)}\big\|_\infty^2\\
&\quad\cdot\max_{1\le s\le r}\big\|\boldsymbol{u}_s^{(m)}\big\|_2^4\big\|\boldsymbol{U}-\boldsymbol{U}^{(m)}\big\|_{\mathrm{F}}^2\\
&\lesssim\frac{1}{p}\cdot\frac{\mu r\lambda_{\max}^{\star 2/3}}{d}\cdot\frac{\mu\lambda_{\max}^{\star 2/3}}{d}\cdot\lambda_{\max}^{\star 4/3}\cdot\frac{\sigma_{\max}^2}{\lambda_{\min}^{\star 2}}\frac{\mu r\log d\lambda_{\max}^{\star 2/3}}{p}\\
&\lesssim\frac{\sigma_{\max}^2\lambda_{\max}^{\star 10/3}}{\lambda_{\min}^{\star 2}p}\frac{\mu^3r^2\log d}{d^2p},
\end{aligned}\tag{114}$$

where we use (61), (62) and $\|\boldsymbol{U}-\boldsymbol{U}^{(m)}\|_{2,\infty}\le\|\boldsymbol{U}\|_{2,\infty}+\|\boldsymbol{U}^{(m)}\|_{2,\infty}\lesssim\sqrt{\mu r/d}\lambda_{\max}^{\star 1/3}$ in the last step. Note that the same bound also holds for $\sum_{1\le s\le r}\boldsymbol{u}_s^{(m)}\otimes\boldsymbol{\xi}_s^{(m)}\otimes\boldsymbol{\xi}_s^{(m)}$.

- As for the cubic term $\boldsymbol{M}_5 := \sum_{1\le s\le r}\big(\boldsymbol{\xi}_s^{(m)}\big)^{\otimes 3}$ in (110), arguing similarly as in (113), we know that with probability greater than $1-O\big(d^{-20}\big)$,

$$\begin{aligned}
&\Big\|\big[\boldsymbol{u}_s^{(m)\top}\big\{\big(p^{-1}\mathcal{P}_\Omega-\mathcal{I}\big)\boldsymbol{M}_5\big\}_{:,:,m}\boldsymbol{u}_s^{(m)}\big]_s\Big\|_2^2\\
&\lesssim\frac{1}{p}\sum_{1\le s\le r}\big(\boldsymbol{\xi}_s^{(m)}\big)_m^2\sum_{1\le s\le r}\big\|\boldsymbol{\xi}_s^{(m)}\big\|_2^4\big\|\boldsymbol{u}_s^{(m)}\big\|_2^4\\
&\lesssim\frac{1}{p}\big\|\boldsymbol{U}-\boldsymbol{U}^{(m)}\big\|_{2,\infty}^2\max_{1\le s\le r}\big\|\boldsymbol{u}_s^{(m)}\big\|_2^4\\
&\quad\cdot\max_{1\le s\le r}\big\|\boldsymbol{\xi}_s^{(m)}\big\|_\infty^2\big\|\boldsymbol{U}-\boldsymbol{U}^{(m)}\big\|_{\mathrm{F}}^2\\
&\lesssim\frac{1}{p}\cdot\frac{\mu r\lambda_{\max}^{\star 2/3}}{d}\cdot\lambda_{\max}^{\star 4/3}\cdot\frac{\mu\lambda_{\max}^{\star 2/3}}{d}\cdot\frac{\sigma_{\max}^2}{\lambda_{\min}^{\star 2}}\frac{\mu r\log d\lambda_{\max}^{\star 2/3}}{p}\\
&\lesssim\frac{\sigma_{\max}^2\lambda_{\max}^{\star 10/3}}{\lambda_{\min}^{\star 2}p}\frac{\mu^3r^2\log d}{d^2p},
\end{aligned}\tag{115}$$

where we use (61), (62) and $\|\boldsymbol{U}-\boldsymbol{U}^{(m)}\|_{2,\infty}\le\|\boldsymbol{U}\|_{2,\infty}+\|\boldsymbol{U}^{(m)}\|_{2,\infty}\lesssim\sqrt{\mu r/d}\lambda_{\max}^{\star 1/3}$ in (115).

- Putting the results (111)-(115) together with the condition $\kappa\asymp 1$ reveals that: with probability at least $1-O\big(d^{-20}\big)$,

$$\begin{aligned}
&\big\|\boldsymbol{e}_m^\top\mathsf{unfold}\big((p^{-1}\mathcal{P}_\Omega-\mathcal{I})(\boldsymbol{T}-\boldsymbol{T}^{(m)})\big)\widetilde{\boldsymbol{U}}^{(m)}\big\|_2\\
&\lesssim\frac{\sigma_{\max}\lambda_{\max}^{\star 2/3}}{\sqrt{p}}\sqrt{\frac{\mu^3r^2\log^2 d}{d^2p}}.
\end{aligned}\tag{116}$$

4) Combining the Bounds on β_1, β_2 and β_3: Substituting (108), (109) and (116) into (105), and taking the union bound over $m\in[d]$, we conclude that

$$\begin{aligned}
&\big\|\mathsf{unfold}\big((p^{-1}\mathcal{P}_\Omega-\mathcal{I})(\boldsymbol{T}-\boldsymbol{T}^\star)\big)\widetilde{\boldsymbol{U}}\big(\widetilde{\boldsymbol{U}}^\top\widetilde{\boldsymbol{U}}\big)^{-1}\big\|_{2,\infty}\\
&\lesssim\frac{\sigma_{\max}}{\lambda_{\min}^{\star 2/3}\sqrt{p}}\sqrt{\frac{\mu^3r^2\log^2 d}{d^2p}}
\end{aligned}$$

with probability at least $1-O\big(d^{-10}\big)$, provided that $\kappa\asymp 1$.

F. Proof of Lemma 6

Without loss of generality, assume that $\boldsymbol{\Pi} = \boldsymbol{I}_r$. By (91), it is straightforward to invoke (63c) to obtain

$$\begin{aligned}\big\|\nabla g(\boldsymbol{U})(\widetilde{\boldsymbol{U}}^{\top}\widetilde{\boldsymbol{U}})^{-1}\big\|_{2,\infty} &\leq \big\|\nabla g(\boldsymbol{U})\big\|_{2,\infty}\big\|(\widetilde{\boldsymbol{U}}^{\top}\widetilde{\boldsymbol{U}})^{-1}\big\| \\ &\leq \big\|\nabla g(\boldsymbol{U})\big\|_{\mathrm{F}}\big\|(\widetilde{\boldsymbol{U}}^{\top}\widetilde{\boldsymbol{U}})^{-1}\big\| \\ &\lesssim \sigma_{\max}\lambda_{\max}^{\star 2/3}\sqrt{\frac{d}{p}}\,\frac{1}{d}\cdot\frac{1}{\lambda_{\min}^{\star 4/3}} \\ &\lesssim \frac{\sigma_{\max}}{\lambda_{\min}^{\star 2/3}}\sqrt{\frac{1}{p}}\,\frac{1}{\sqrt{d}},\end{aligned}$$

where the last step follows from the condition $\kappa \asymp 1$.

Appendix D
Proof of Auxiliary Lemmas: Distributional Theory for Tensor Entries

A. Proof of Lemma 7

Fix arbitrary $1 \leq m \leq n \leq l \leq d$. In what follows, we shall focus on the case $m < n < l$. The analysis naturally extends to the case where $m = n < l, m < n = l$ or $m = n = l$.

Before we embark on the proof, we remind the readers of the definitions of $\boldsymbol{z}$ (resp. $\boldsymbol{Z}$) in (66) (resp. (67)). While $\boldsymbol{Z}_{m,:}$, $\boldsymbol{Z}_{n,:}$ and $\boldsymbol{Z}_{l,:}$ are not mutually independent due to the symmetric sampling, we can show that the dependence between them are extremely weak. This in turn allows us to invoke the Berry-Esseen theorem to prove the advertised distributional guarantees.

We now begin to present our analysis. To decouple the weak dependence, we define the following auxiliary random vector:

$$\widehat{\boldsymbol{Z}}_{m,:} = \breve{\boldsymbol{Z}}_{m,:}\breve{\boldsymbol{\Sigma}}_m^{-1/2}\boldsymbol{\Sigma}_m^{\star 1/2} \tag{117}$$

with

$$\begin{aligned}\breve{\boldsymbol{Z}}_{m,:} &\coloneqq \sqrt{2}\sum_{i:\, i\neq n,l}\boldsymbol{z}_{i,i,m} + 2\sum_{i:\, i\neq m,n,l}\boldsymbol{z}_{i,m,m} \\ &\quad + 2\sum_{\substack{(i,j):\, i,j\neq m,n,l \\ 1\leq i<j\leq d}}\boldsymbol{z}_{i,j,m}, \\ \breve{\boldsymbol{\Sigma}}_m &\coloneqq \mathbb{E}\big[(\breve{\boldsymbol{Z}}_{m,:})^{\top}\breve{\boldsymbol{Z}}_{m,:}\big].\end{aligned}$$

The vectors $\widehat{\boldsymbol{Z}}_{n,:}$ and $\widehat{\boldsymbol{Z}}_{l,:}$ are defined in a similar manner. By construction, it is easy to verify that $\widehat{\boldsymbol{Z}}_{m,:}, \widehat{\boldsymbol{Z}}_{n,:}$ and $\widehat{\boldsymbol{Z}}_{l,:}$ are mutually independent. Moreover, Lemma 17 as stated below reveals that the constructed auxiliary vectors are sufficiently close to the original ones.

Lemma 17: Instate the assumptions and notation of Lemma 7. With probability at least $1 - O\left(d^{-13}\right)$, one has

$$\begin{aligned}\big\|(\boldsymbol{Z} - \widehat{\boldsymbol{Z}})_{m,:}\big\|_2 \lesssim \frac{\sigma_{\max}}{\lambda_{\min}^{\star 2/3}\sqrt{p}}\Bigg\{&\frac{\mu\sqrt{r}\log^2 d}{d\sqrt{p}} + \sqrt{\frac{\mu r\log d}{d}} \\ &+ \frac{\mu r^{3/2}\sqrt{\log d}}{d}\Bigg\}.\end{aligned}$$

In addition, the upper bound continues to hold for $(\boldsymbol{Z} - \widehat{\boldsymbol{Z}})_{n,:}$ and $(\boldsymbol{Z} - \widehat{\boldsymbol{Z}})_{l,:}$.

With these in place, we then define random variables

$$\begin{aligned}G_{m,n,l} &\coloneqq \big\langle\widehat{\boldsymbol{Z}}_{m,:}, \widetilde{\boldsymbol{U}}^{\star}_{(n,l),:}\big\rangle + \big\langle\widehat{\boldsymbol{Z}}_{n,:}, \widetilde{\boldsymbol{U}}^{\star}_{(m,l),:}\big\rangle \\ &\quad + \big\langle\widehat{\boldsymbol{Z}}_{l,:}, \widetilde{\boldsymbol{U}}^{\star}_{(m,n),:}\big\rangle; \\ H_{m,n,l} &\coloneqq Y_{m,n,l} - G_{m,n,l} \\ &= \big\langle(\boldsymbol{Z} - \widehat{\boldsymbol{Z}})_{m,:}, \widetilde{\boldsymbol{U}}^{\star}_{(n,l),:}\big\rangle + \big\langle(\boldsymbol{Z} - \widehat{\boldsymbol{Z}})_{n,:}, \widetilde{\boldsymbol{U}}^{\star}_{(m,l),:}\big\rangle \\ &\quad + \big\langle(\boldsymbol{Z} - \widehat{\boldsymbol{Z}})_{l,:}, \widetilde{\boldsymbol{U}}^{\star}_{(m,n),:}\big\rangle.\end{aligned}$$

By construction, $G_{m,n,l}$ is a sum of independent zero-mean random variables with variance $v^{\star}_{m,n,l}$ defined in (23). In the sequel, we shall apply the Berry-Esseen theorem [91, Theorem 1.1] (cf. Appendix B-C) to show that $G_{m,n,l}$ is close in distribution to a Gaussian random variable. As before, we need to control the quantity ρ defined in (65). From (63a) and (63b), it is straightforward to upper bound

$$\begin{aligned}&\frac{1}{p^3}\sum_{1\leq i,j\leq d}\mathbb{E}\big[|E_{i,j,k}|^3\chi_{i,j,k}\big]\big|\widetilde{\boldsymbol{U}}^{\star}_{(i,j),:}(\widetilde{\boldsymbol{U}}^{\star\top}\widetilde{\boldsymbol{U}}^{\star})^{-1}\big(\widetilde{\boldsymbol{U}}^{\star}_{(n,l),:}\big)^{\top}\big|^3 \\ &\lesssim \frac{\sigma_{\max}^3}{p^2}\max_{1\leq i,j\leq d}\big|\widetilde{\boldsymbol{U}}^{\star}_{(i,j),:}(\widetilde{\boldsymbol{U}}^{\star\top}\widetilde{\boldsymbol{U}}^{\star})^{-1}\big(\widetilde{\boldsymbol{U}}^{\star}_{(n,l),:}\big)^{\top}\big| \\ &\quad\cdot\widetilde{\boldsymbol{U}}^{\star}_{(n,l),:}(\widetilde{\boldsymbol{U}}^{\star\top}\widetilde{\boldsymbol{U}}^{\star})^{-1}\big(\widetilde{\boldsymbol{U}}^{\star}_{(n,l),:}\big)^{\top} \\ &\leq \frac{\sigma_{\max}^3}{p^2}\big\|\widetilde{\boldsymbol{U}}^{\star}\big\|_{2,\infty}\big\|(\widetilde{\boldsymbol{U}}^{\star\top}\widetilde{\boldsymbol{U}}^{\star})^{-1}\big\|^2\big\|\big(\widetilde{\boldsymbol{U}}^{\star}_{(n,l),:}\big)\big\|_2^3 \\ &\lesssim \frac{\sigma_{\max}^3}{p^2}\cdot\frac{\mu\sqrt{r}}{d}\lambda_{\max}^{\star 2/3}\cdot\left(\lambda_{\min}^{\star -4/3}\right)^2\big\|\big(\widetilde{\boldsymbol{U}}^{\star}_{(n,l),:}\big)\big\|_2^3 \\ &\lesssim \frac{\sigma_{\max}^3\lambda_{\max}^{\star 2/3}}{\lambda_{\min}^{\star 8/3}}\frac{\mu\sqrt{r}}{dp^2}\big\|\big(\widetilde{\boldsymbol{U}}^{\star}_{(n,l),:}\big)\big\|_2^3.\end{aligned}$$

In addition, we can use (70) to lower bound the variance as follows

$$\begin{aligned}v^{\star}_{m,n,l} &\geq \lambda_{\min}\big(\boldsymbol{\Sigma}^{\star}_m\big)\big\|\widetilde{\boldsymbol{U}}^{\star}_{(n,l),:}\big\|_2^2 + \lambda_{\min}\big(\boldsymbol{\Sigma}^{\star}_n\big)\big\|\widetilde{\boldsymbol{U}}^{\star}_{(m,l),:}\big\|_2^2 \\ &\quad + \lambda_{\min}\big(\boldsymbol{\Sigma}^{\star}_l\big)\big\|\widetilde{\boldsymbol{U}}^{\star}_{(m,n),:}\big\|_2^2 \\ &\gtrsim \frac{\sigma_{\min}^2}{\lambda_{\max}^{\star 4/3}p}\Big(\big\|\widetilde{\boldsymbol{U}}^{\star}_{(n,l),:}\big\|_2^2 + \big\|\widetilde{\boldsymbol{U}}^{\star}_{(m,l),:}\big\|_2^2 + \big\|\widetilde{\boldsymbol{U}}^{\star}_{(m,n),:}\big\|_2^2\Big).\end{aligned} \tag{118}$$

Combining these two bounds, we arrive at

$$\begin{aligned}\rho &\lesssim (v^{\star}_{m,n,l})^{-3/2}\frac{\sigma_{\max}^3\lambda_{\max}^{\star 2/3}}{\lambda_{\min}^{\star 8/3}}\frac{\mu\sqrt{r}}{dp^2} \\ &\quad\cdot\Big(\big\|\widetilde{\boldsymbol{U}}^{\star}_{(n,l),:}\big\|_2^3 + \big\|\widetilde{\boldsymbol{U}}^{\star}_{(m,l),:}\big\|_2^3 + \big\|\widetilde{\boldsymbol{U}}^{\star}_{(m,n),:}\big\|_2^3\Big) \\ &\lesssim \frac{\mu\sqrt{r}}{d\sqrt{p}},\end{aligned}$$

using the conditions $\sigma_{\max}/\sigma_{\min} \asymp 1$ and $\kappa \asymp 1$. As a consequence, invoke the Berry-Esseen-type theorem in Appendix B-C to derive

$$\sup_{\tau\in\mathbb{R}}\left|\mathbb{P}\Big\{G_{m,n,l} \leq \tau\sqrt{v^{\star}_{m,n,l}}\Big\} - \Phi(\tau)\right| \lesssim \frac{\mu\sqrt{r}}{d\sqrt{p}},$$

where $\Phi(\cdot)$ is the CDF of a standard Gaussian random variable.

We then move on to the residual term $H_{m,n,l}$. Recall the definition of $\omega_{m,n,l}$ in (33). By Lemma 17 and the Cauchy-Schwartz inequality, one can easily upper bound

$$
\begin{aligned}
&|H_{m,n,l}| \\
&\lesssim \omega_{m,n,l}\Big(\big\|(\boldsymbol{Z}-\widehat{\boldsymbol{Z}})_{m,:}\big\|_2+\big\|(\boldsymbol{Z}-\widehat{\boldsymbol{Z}})_{n,:}\big\|_2+\big\|(\boldsymbol{Z}-\widehat{\boldsymbol{Z}})_{l,:}\big\|_2\Big) \\
&\lesssim \left\{\frac{\mu\sqrt{r}\log^2 d}{d\sqrt{p}}+\sqrt{\frac{\mu r\log d}{d}}+\frac{\mu r^{3/2}\sqrt{\log d}}{d}\right\}\frac{\sigma_{\max}\omega_{m,n,l}}{\lambda_{\min}^{\star 2/3}\sqrt{p}} \\
&\lesssim \left\{\frac{\mu\sqrt{r}\log^2 d}{d\sqrt{p}}+\sqrt{\frac{\mu r\log d}{d}}+\frac{\mu r^{3/2}\sqrt{\log d}}{d}\right\}\sqrt{v_{m,n,l}^{\star}},
\end{aligned}
$$

where the last step arises from (118) and the conditions $\sigma_{\max}/\sigma_{\min}\asymp 1$ and $\kappa\asymp 1$.

1) Proof of Lemma 17: By the triangle inequality, we have

$$
\begin{aligned}
\big\|(\boldsymbol{Z}-\widehat{\boldsymbol{Z}})_{m,:}\big\|_2 \le{}& \underbrace{\big\|\big(\boldsymbol{Z}-\breve{\boldsymbol{Z}}\big)_{m,:}\big\|_2}_{=:\,\beta_1} \\
&+\underbrace{\big\|\breve{\boldsymbol{Z}}_{m,:}\breve{\boldsymbol{\Sigma}}_m^{-1/2}\big(\breve{\boldsymbol{\Sigma}}_m^{1/2}-\boldsymbol{\Sigma}_m^{\star 1/2}\big)\big\|_2}_{=:\,\beta_2}.
\end{aligned}
$$

We shall upper bound these two terms separately.

1) For β_1, observe that $(\boldsymbol{Z}-\breve{\boldsymbol{Z}})_{m,:}=\sqrt{2}\,\boldsymbol{z}_{n,n,m}+\sqrt{2}\,\boldsymbol{z}_{l,l,m}+2\sum_{i:i\neq n,l}\left(\boldsymbol{z}_{i,n,m}+\boldsymbol{z}_{i,l,m}\right)$ is a sum of independent zero-mean random vectors. From (63a) and (63b), straightforward computation gives

$$
\begin{aligned}
B &:= \max_{1\le i,j\le d}\big\|p^{-1}E_{i,j,k}\chi_{i,j,k}\widetilde{\boldsymbol{U}}^{\star}_{(i,j),:}(\widetilde{\boldsymbol{U}}^{\star\top}\widetilde{\boldsymbol{U}}^{\star})^{-1}\big\|_{\psi_1} \\
&\lesssim \frac{\sigma_{\max}}{p}\big\|\widetilde{\boldsymbol{U}}^{\star}\big\|_{2,\infty}\big\|(\widetilde{\boldsymbol{U}}^{\star\top}\widetilde{\boldsymbol{U}}^{\star})^{-1}\big\| \\
&\lesssim \frac{\sigma_{\max}}{p}\cdot\frac{\mu\sqrt{r}}{d}\lambda_{\max}^{\star 2/3}\cdot\frac{1}{\lambda_{\min}^{\star 4/3}}
\end{aligned}
$$

and

$$
\begin{aligned}
V &:= \sum_{1\le i\le d}p^{-2}\mathbb{E}\big[E_{i,j,k}^2\chi_{i,j,k}\big]\big\|\widetilde{\boldsymbol{U}}^{\star}_{(i,j),:}(\widetilde{\boldsymbol{U}}^{\star\top}\widetilde{\boldsymbol{U}}^{\star})^{-1}\big\|_2^2 \\
&\le \frac{\sigma_{\max}^2}{p}\big\|(\widetilde{\boldsymbol{U}}^{\star\top}\widetilde{\boldsymbol{U}}^{\star})^{-1}\big\|^2\sum_{1\le i\le d}\big\|\widetilde{\boldsymbol{U}}^{\star}_{(i,j),:}\big\|_2^2 \\
&\le \frac{\sigma_{\max}^2}{p}\big\|(\widetilde{\boldsymbol{U}}^{\star\top}\widetilde{\boldsymbol{U}}^{\star})^{-1}\big\|^2\max_{1\le s\le r}\|\boldsymbol{u}_s^{\star}\|_2^2\,\|\boldsymbol{U}^{\star}\|_{2,\infty}^2 \\
&\lesssim \frac{\sigma_{\max}^2}{p}\cdot\frac{1}{\lambda_{\min}^{\star 8/3}}\cdot\lambda_{\max}^{\star 2/3}\cdot\frac{\mu r}{d}\lambda_{\max}^{\star 2/3}.
\end{aligned}
$$

Applying the matrix Bernstein inequality reveals that with probability exceeding $1-O\left(d^{-20}\right)$,

$$
\begin{aligned}
\big\|\big(\boldsymbol{Z}-\breve{\boldsymbol{Z}}\big)_{m,:}\big\|_2 &\lesssim B\log^2 d+\sqrt{V\log d} \\
&\lesssim \frac{\sigma_{\max}\lambda_{\max}^{\star 2/3}}{\lambda_{\min}^{\star 4/3}\sqrt{p}}\left\{\frac{\mu\sqrt{r}\log^2 d}{d\sqrt{p}}+\sqrt{\frac{\mu r\log d}{d}}\right\}.
\end{aligned}
$$

In particular, we can combine it with (72) and $\kappa\asymp 1$ to obtain

$$
\begin{aligned}
\big\|\breve{\boldsymbol{Z}}_{m,:}\big\|_2 &\le \|\boldsymbol{Z}_{m,:}\|_2+\big\|\big(\boldsymbol{Z}-\breve{\boldsymbol{Z}}\big)_{m,:}\big\|_2 \\
&\lesssim \frac{\sigma_{\max}}{\lambda_{\min}^{\star 2/3}\sqrt{p}}\left\{\frac{\mu\sqrt{r}\log^2 d}{d\sqrt{p}}+\sqrt{\frac{\mu r\log d}{d}}+\sqrt{r\log d}\right\} \\
&\asymp \frac{\sigma_{\max}\sqrt{r\log d}}{\lambda_{\min}^{\star 2/3}\sqrt{p}},
\end{aligned} \tag{119}
$$

with the proviso that $p\gtrsim \mu^2 d^{-2}\log^3 d$ and $\mu\lesssim d$.

2) Turning to β_2, we invoke the independence between $(\boldsymbol{Z}-\breve{\boldsymbol{Z}})_{m,:}$ and $\breve{\boldsymbol{Z}}_{m,:}$ to derive

$$
\begin{aligned}
\boldsymbol{\Sigma}_m^{\star}-\breve{\boldsymbol{\Sigma}}_m &= \mathbb{E}\big[\boldsymbol{Z}_{m,:}\boldsymbol{Z}_{m,:}^{\top}\big]-\mathbb{E}\big[(\breve{\boldsymbol{Z}}_{m,:})^{\top}\breve{\boldsymbol{Z}}_{m,:}\big] \\
&= \mathbb{E}\big[(\boldsymbol{Z}_{m,:}-\breve{\boldsymbol{Z}}_{m,:})^{\top}(\boldsymbol{Z}_{m,:}-\breve{\boldsymbol{Z}}_{m,:})\big] \\
&= \frac{4}{p}\sum_{i:i\neq n,l}\sigma_{i,n,m}^2(\widetilde{\boldsymbol{U}}^{\star\top}\widetilde{\boldsymbol{U}}^{\star})^{-1}(\widetilde{\boldsymbol{U}}^{\star}_{(i,n),:})^{\top}\widetilde{\boldsymbol{U}}^{\star}_{(i,n),:}(\widetilde{\boldsymbol{U}}^{\star\top}\widetilde{\boldsymbol{U}}^{\star})^{-1} \\
&\quad+\frac{4}{p}\sum_{i:i\neq n,l}\sigma_{i,l,m}^2(\widetilde{\boldsymbol{U}}^{\star\top}\widetilde{\boldsymbol{U}}^{\star})^{-1}(\widetilde{\boldsymbol{U}}^{\star}_{(i,l),:})^{\top}\widetilde{\boldsymbol{U}}^{\star}_{(i,l),:}(\widetilde{\boldsymbol{U}}^{\star\top}\widetilde{\boldsymbol{U}}^{\star})^{-1} \\
&\quad+\frac{2}{p}\sigma_{n,n,m}^2(\widetilde{\boldsymbol{U}}^{\star\top}\widetilde{\boldsymbol{U}}^{\star})^{-1}(\widetilde{\boldsymbol{U}}^{\star}_{(n,n),:})^{\top}\widetilde{\boldsymbol{U}}^{\star}_{(n,n),:}(\widetilde{\boldsymbol{U}}^{\star\top}\widetilde{\boldsymbol{U}}^{\star})^{-1} \\
&\quad+\frac{2}{p}\sigma_{l,l,m}^2(\widetilde{\boldsymbol{U}}^{\star\top}\widetilde{\boldsymbol{U}}^{\star})^{-1}(\widetilde{\boldsymbol{U}}^{\star}_{(l,l),:})^{\top}\widetilde{\boldsymbol{U}}^{\star}_{(l,l),:}(\widetilde{\boldsymbol{U}}^{\star\top}\widetilde{\boldsymbol{U}}^{\star})^{-1}.
\end{aligned}
$$

One can then use (63b) to upper bound

$$
\begin{aligned}
\big\|\boldsymbol{\Sigma}_m^{\star}-\breve{\boldsymbol{\Sigma}}_m\big\| &\lesssim \frac{\sigma_{\max}^2}{p}\big\|(\widetilde{\boldsymbol{U}}^{\star\top}\widetilde{\boldsymbol{U}}^{\star})^{-1}\big\|^2 \\
&\quad\cdot\Big\|\sum_{i:i\neq l}(\widetilde{\boldsymbol{U}}^{\star}_{(i,n),:})^{\top}\widetilde{\boldsymbol{U}}^{\star}_{(i,n),:}+\sum_{i:i\neq n}(\widetilde{\boldsymbol{U}}^{\star}_{(i,l),:})^{\top}\widetilde{\boldsymbol{U}}^{\star}_{(i,l),:}\Big\| \\
&\le \frac{\sigma_{\max}^2}{p}\big\|(\widetilde{\boldsymbol{U}}^{\star\top}\widetilde{\boldsymbol{U}}^{\star})^{-1}\big\|^2\sum_{1\le i\le d}\Big(\big\|\widetilde{\boldsymbol{U}}^{\star}_{(i,n),:}\big\|_2^2+\big\|\widetilde{\boldsymbol{U}}^{\star}_{(i,l),:}\big\|_2^2\Big) \\
&\lesssim \frac{\sigma_{\max}^2}{p}\cdot\frac{1}{\lambda_{\min}^{\star 8/3}}\cdot\frac{\mu r}{d}\lambda_{\max}^{\star 4/3} \\
&\lesssim \frac{\sigma_{\max}^2}{\lambda_{\min}^{\star 4/3}p}\frac{\mu r}{d}\ll\frac{\sigma_{\min}^2}{\lambda_{\max}^{\star 4/3}p},
\end{aligned}
$$

where the last line arises from the assumption that $\sigma_{\max}/\sigma_{\min}\asymp 1$, $\kappa\asymp 1$ and $r\ll d/\mu$. Combining this with (70) and Weyl's inequality, one arrives at

$$
\lambda_{\max}\big(\breve{\boldsymbol{\Sigma}}_m\big)\lesssim\frac{\sigma_{\max}^2}{\lambda_{\min}^{\star 4/3}p}\quad\text{and}\quad\lambda_{\min}\big(\breve{\boldsymbol{\Sigma}}_m\big)\gtrsim\frac{\sigma_{\min}^2}{\lambda_{\max}^{\star 4/3}p}.
$$

Applying the perturbation bound for matrix square roots [93, Lemma 2.1] yields

$$
\begin{aligned}
\big\|\breve{\boldsymbol{\Sigma}}_m^{1/2}-\boldsymbol{\Sigma}_m^{\star 1/2}\big\| &\lesssim \frac{\big\|\breve{\boldsymbol{\Sigma}}_m-\boldsymbol{\Sigma}_m^{\star}\big\|}{\lambda_{\min}^{1/2}(\boldsymbol{\Sigma}_m^{\star})+\lambda_{\min}^{1/2}(\breve{\boldsymbol{\Sigma}}_m)} \\
&\lesssim \frac{\lambda_{\max}^{\star 2/3}\sqrt{p}}{\sigma_{\min}}\cdot\frac{\sigma_{\max}^2}{\lambda_{\min}^{\star 4/3}p}\frac{\mu r}{d}.
\end{aligned}
$$

This taken collectively with (119) implies that

$$
\begin{aligned}
&\big\|\breve{\boldsymbol{Z}}_{m,:}\breve{\boldsymbol{\Sigma}}_m^{-1/2}\big(\breve{\boldsymbol{\Sigma}}_m^{1/2}-\boldsymbol{\Sigma}_m^{\star 1/2}\big)\big\|_2 \\
&\lesssim \big\|\breve{\boldsymbol{Z}}_{m,:}\big\|_2\big\|\breve{\boldsymbol{\Sigma}}_m^{-1/2}\big\|\big\|\breve{\boldsymbol{\Sigma}}_m^{1/2}-\boldsymbol{\Sigma}_m^{\star 1/2}\big\| \\
&\lesssim \frac{\sigma_{\max}\sqrt{r\log d}}{\lambda_{\min}^{\star 2/3}\sqrt{p}}\cdot\frac{\lambda_{\max}^{\star 2/3}\sqrt{p}}{\sigma_{\min}}\cdot\frac{\lambda_{\max}^{\star 2/3}\sqrt{p}}{\sigma_{\min}}\frac{\sigma_{\max}^2}{\lambda_{\min}^{\star 4/3}p}\frac{\mu r}{d} \\
&\lesssim \frac{\sigma_{\max}}{\lambda_{\min}^{\star 2/3}\sqrt{p}}\frac{\mu r^{3/2}\sqrt{\log d}}{d},
\end{aligned}
$$

where the last line follows from the conditions $\sigma_{\max}/\sigma_{\min}\asymp 1$ and $\kappa\asymp 1$.

3) Putting the above bounds together yields the advertised bound in Lemma 17.

B. Proof of Lemma 8

We shall bound the terms in (47) separately, followed by the triangle inequality. In what follows, we denote $u^\star_{s,i} := (\boldsymbol{u}_s)_i$ and $\Delta_{s,i} := (\boldsymbol{\Delta}_s)_i$ for any $1 \le s \le r$ and $1 \le i \le d$.

1) Regarding the first three terms involving $\boldsymbol{W}$, combine the Cauchy-Schwartz with (45) to show that
$$\begin{aligned}&\left|\langle \boldsymbol{W}_{m,:}, \widetilde{\boldsymbol{U}}^\star_{(n,l),:}\rangle + \langle \boldsymbol{W}_{n,:}, \widetilde{\boldsymbol{U}}^\star_{(m,l),:}\rangle + \langle \boldsymbol{W}_{l,:}, \widetilde{\boldsymbol{U}}^\star_{(m,n),:}\rangle\right| \\ &\quad \le \|\boldsymbol{W}\|_{2,\infty}\,\omega_{m,n,l} \lesssim \zeta \frac{\sigma_{\max}}{\lambda_{\min}^{\star 2/3}\sqrt{p}}\omega_{m,n,l}, \qquad (120)\end{aligned}$$
where we recall the definition of ζ (resp. $\omega_{m,n,l}$) in (43) (resp. (33)).

2) We now turn to $\langle \boldsymbol{U}^\star_{m,:}, \widetilde{\boldsymbol{\Delta}}_{(n,l),:}\rangle = \sum_{1\le s\le r} u^\star_{s,m}\Delta_{s,n}\Delta_{s,l}$. By virtue of (12b) and (59b), one can invoke Cauchy-Schwartz to bound
$$\begin{aligned}\Big|\sum_{1\le s\le r} u^\star_{s,m}\Delta_{s,n}\Delta_{s,l}\Big| &\le \sqrt{\sum_{1\le s\le r}\Delta_{s,l}^2}\sqrt{\sum_{1\le s\le r}u^{\star 2}_{s,m}\Delta^2_{s,n}} \\ &\le \|\boldsymbol{\Delta}\|^2_{2,\infty}\max_{1\le s\le r}\|\boldsymbol{u}^\star_s\|_\infty \\ &\lesssim \left(\frac{\sigma_{\max}}{\lambda^\star_{\min}}\sqrt{\frac{\mu r\log d}{p}}\lambda_{\max}^{\star 1/3}\right)^2\sqrt{\frac{\mu}{d}}\,\lambda_{\max}^{\star 1/3}. \qquad (121)\end{aligned}$$
Clearly, this upper bound also holds for both $\langle \boldsymbol{U}^\star_{n,:}, \widetilde{\boldsymbol{\Delta}}_{(m,l),:}\rangle$ and $\langle \boldsymbol{U}^\star_{l,:}, \widetilde{\boldsymbol{\Delta}}_{(m,n),:}\rangle$.

3) As for the last term $\langle \boldsymbol{\Delta}_{m,:}, \widetilde{\boldsymbol{\Delta}}_{(n,l),:}\rangle$, we know from (59b) and (62d) that
$$\begin{aligned}\Big|\sum_{1\le s\le r}\Delta_{s,m}\Delta_{s,n}\Delta_{s,l}\Big| &\le \|\boldsymbol{\Delta}\|^2_{2,\infty}\max_{1\le s\le r}\|\boldsymbol{\Delta}_s\|_\infty \\ &\ll \left(\frac{\sigma_{\max}}{\lambda^\star_{\min}}\sqrt{\frac{\mu r\log d}{p}}\lambda_{\max}^{\star 1/3}\right)^2\sqrt{\frac{\mu}{d}}\,\lambda_{\max}^{\star 1/3}. \qquad (122)\end{aligned}$$

4) Combining (120), (121) and (122), we arrive at the advertised bound
$$\begin{aligned}|R_{m,n,l}| &\lesssim \zeta\frac{\sigma_{\max}}{\lambda_{\min}^{\star 2/3}\sqrt{p}}\omega_{m,n,l} \\ &\quad + \left(\frac{\sigma_{\max}}{\lambda^\star_{\min}}\sqrt{\frac{\mu r\log d}{p}}\lambda_{\max}^{\star 1/3}\right)^2\sqrt{\frac{\mu}{d}}\,\lambda_{\max}^{\star 1/3} \\ &\overset{\text{(i)}}{\lesssim} \left(\zeta + \frac{\sigma_{\max}}{\lambda_{\min}^{\star 1/3}}\frac{\mu^{3/2}r\log d}{\sqrt{dp}}\frac{1}{\omega_{m,n,l}}\right)\sqrt{v^\star_{mnl}} \\ &\overset{\text{(ii)}}{=} o(1)\sqrt{v^\star_{mnl}}.\end{aligned}$$
Here, (i) arises from the lower bound on $v^\star_{m,n,l}$ (cf. (118)) and the conditions $\sigma_{\max}/\sigma_{\min} \asymp 1$ and $\kappa \asymp 1$, whereas (ii) makes use of the assumptions (32) and (34).

Appendix E
Proof of Auxiliary Lemmas: Confidence Intervals

A. Proof of Lemma 9

Fix arbitrary $1 \le l \le r$ and $1 \le k \le d$. Before proceeding, we pause to introduce some notation for simplicity of presentation. Recalling the notation $\widetilde{\boldsymbol{U}}^\star := [\boldsymbol{u}_l^{\star\otimes 2}]_{1\le l\le r} \in \mathbb{R}^{d^2\times r}$ and $\widetilde{\boldsymbol{U}} := [\boldsymbol{u}_l^{\otimes 2}]_{1\le l\le r} \in \mathbb{R}^{d^2\times r}$, we define two $d^2\times r$ matrices as follows
$$\boldsymbol{V}^\star := \widetilde{\boldsymbol{U}}^\star(\widetilde{\boldsymbol{U}}^{\star\top}\widetilde{\boldsymbol{U}}^\star)^{-1}, \qquad \boldsymbol{V} := \widetilde{\boldsymbol{U}}(\widetilde{\boldsymbol{U}}^\top\widetilde{\boldsymbol{U}})^{-1}. \qquad (123)$$
These allow us to express the covariance matrix as $\boldsymbol{\Sigma}^\star_k = \boldsymbol{V}^\star\boldsymbol{D}^\star_k\boldsymbol{V}^\star$ (resp. $\boldsymbol{\Sigma}_k = \boldsymbol{V}\boldsymbol{D}_k\boldsymbol{V}$), where $\boldsymbol{D}^\star_k$ (resp. $\boldsymbol{D}_k$) is defined in (18) (resp. (25)). In addition, let us define
$$s^\star_{l,k} := \sqrt{(\boldsymbol{\Sigma}^\star_k)_{l,l}} \quad \text{and} \quad s_{l,k} := \sqrt{(\boldsymbol{\Sigma}_k)_{l,l}}.$$
Lemma 18 below collects several useful properties regrading $\boldsymbol{V}_{:,l}$ and $\boldsymbol{V}^\star_{:,l}$; the proof is deferred to the end of this section.

Lemma 18: Instate the assumptions and notation of Lemma 9. For each $1 \le l \le r$, one has
$$\left\|\boldsymbol{V}^\star_{:,l}\right\|_2 = \frac{1+o(1)}{\|\boldsymbol{u}^\star_l\|_2^2}, \qquad (124)$$
$$\left\|\boldsymbol{V}^\star_{:,l}\right\|_\infty \lesssim \frac{\mu\sqrt{r}}{d}\frac{1}{\lambda_{\min}^{\star 2/3}}, \qquad (125)$$
$$\sum_{1\le k\le d} V^{\star 2}_{(i,k),l} \lesssim \frac{\mu r}{d}\frac{1}{\lambda_{\min}^{\star 4/3}}; \qquad (126)$$
$$\left\|\boldsymbol{V}_{:,l} - \boldsymbol{V}^\star_{:,l}\right\|_2 \lesssim \frac{\sigma_{\max}}{\lambda^\star_{\min}}\sqrt{\frac{d}{p}}\frac{1}{\lambda_{\min}^{\star 2/3}}, \qquad (127)$$
$$\left\|\boldsymbol{V}_{:,l} - \boldsymbol{V}^\star_{:,l}\right\|_\infty \lesssim \frac{\sigma_{\max}}{\lambda^\star_{\min}}\sqrt{\frac{\mu^2 r\log d}{dp}}\frac{1}{\lambda_{\min}^{\star 2/3}}. \qquad (128)$$

With these in place, we are ready to control $J_{l,k}$, which can be expressed as
$$\begin{aligned}J_{l,k} &= \frac{u_{l,k}-u^\star_{l,k}}{s_{l,k}} - \frac{u_{l,k}-u^\star_{l,k}}{s^\star_{l,k}} \\ &= \left(u_{l,k}-u^\star_{l,k}\right)\frac{s^{\star 2}_{l,k}-s^2_{l,k}}{s_{l,k}s^\star_{l,k}}\frac{1}{s^\star_{l,k}+s_{l,k}}.\end{aligned}$$
This suggest that we control both $u_{l,k}-u^\star_{l,k}$ and $s^{\star 2}_{l,k}-s^2_{l,k}$.

- Regarding the estimation error of $u_{l,k}$, combining (124) with the assumptions $\sigma_{\max}/\sigma_{\min} \asymp 1$ and $\kappa \asymp 1$ allows us to lower bound
$$\begin{aligned}s^{\star 2}_{l,k} &= \frac{1}{p}(\boldsymbol{V}^\star_{:,l})^\top\boldsymbol{D}^\star_k\boldsymbol{V}^\star_{:,l} \ge \frac{1}{p}\lambda_{\min}(\boldsymbol{D}^\star_k)\left\|\boldsymbol{V}^\star_{:,l}\right\|_2^2 \\ &\gtrsim \frac{\sigma^2_{\min}}{p\|\boldsymbol{u}^\star_l\|_2^4}. \qquad (129)\end{aligned}$$
Hence, we know from (59b) and the conditions $\sigma_{\max}/\sigma_{\min} \asymp 1$ and $\kappa \asymp 1$ that
$$\begin{aligned}\left|u_{l,k}-u^\star_{l,k}\right| &\le \left\|\boldsymbol{U}-\boldsymbol{U}^\star\right\|_{2,\infty} \lesssim \frac{\sigma_{\max}}{\lambda^\star_{\min}}\sqrt{\frac{\mu r\log d}{p}}\lambda_{\max}^{\star 1/3} \\ &\lesssim s^\star_{l,k}\sqrt{\mu r\log d}.\end{aligned}$$
- Next, we claim that
$$\begin{aligned}\left|s^{\star 2}_{l,k}-s^2_{l,k}\right| &\lesssim \left\{\sqrt{\frac{\mu^3 r^2\log^2 d}{d^2 p}} + \frac{\sigma_{\max}}{\lambda^\star_{\min}}\sqrt{\frac{\mu^2 r d\log d}{p}}\right\}s^{\star 2}_{l,k} \\ &\ll s^{\star 2}_{l,k}; \qquad (130)\end{aligned}$$

if this were true, then one would further obtain

$$s_{l,k} \geq s^{\star}_{l,k} - |s_{l,k} - s^{\star}_{l,k}| \gtrsim s^{\star}_{l,k}.$$

- Putting the above bounds together reveals that

$$\begin{aligned}|J_{l,k}| &\lesssim \frac{\left|u_{l,k} - u^{\star}_{l,k}\right| \left|s^{\star 2}_{l,k} - s^{2}_{l,k}\right|}{s^{\star 3}_{l,k}} \\ &\lesssim \sqrt{\mu r \log d}\left\{\sqrt{\frac{\mu^3 r^2 \log^2 d}{d^2 p}} + \frac{\sigma_{\max}}{\lambda^{\star}_{\min} p}\sqrt{\frac{\mu^2 r d \log d}{p}}\right\}\end{aligned}$$

as claimed.

Hence, the remainder of the proof boils down to establishing the claim (130). Towards this, our starting point is the following decomposition

$$\begin{aligned}\frac{p}{2}\left(s^{2}_{l,k} - s^{\star 2}_{l,k}\right) &= (\boldsymbol{V}_{:,l})^{\top}\boldsymbol{D}_k\boldsymbol{V}_{:,l} - (\boldsymbol{V}^{\star}_{:,l})^{\top}\boldsymbol{D}^{\star}_k\boldsymbol{V}^{\star}_{:,l} \\ &= \underbrace{(\boldsymbol{V}_{:,l})^{\top}(\boldsymbol{D}_k - \widehat{\boldsymbol{D}}_k)\boldsymbol{V}_{:,l}}_{=:\,\beta_1} \\ &\quad + \underbrace{(\boldsymbol{V}_{:,l})^{\top}\widehat{\boldsymbol{D}}_k\boldsymbol{V}_{:,l} - (\boldsymbol{V}^{\star}_{:,l})^{\top}\boldsymbol{D}^{\star}_k\boldsymbol{V}^{\star}_{:,l}}_{=:\,\beta_2},\end{aligned} \tag{131}$$

where $\widehat{\boldsymbol{D}}_k \in \mathbb{R}^{d^2 \times d^2}$ is a diagonal matrix with entries given by

$$(\widehat{\boldsymbol{D}}_k)_{(i,j),(i,j)} = p^{-1}E^{2}_{i,j,k}\chi_{i,j,k}, \qquad 1 \leq i,j \leq d. \tag{132}$$

In what follows, we shall control β_1 and β_2 separately.

1) Bounding β_1: To begin with, let us decompose

$$\begin{aligned}\beta_1 = &\underbrace{(\boldsymbol{V}^{\star}_{:,l})^{\top}(\boldsymbol{D}_k - \widehat{\boldsymbol{D}}_k)\boldsymbol{V}^{\star}_{:,l}}_{=:\,\gamma_1} + 2\underbrace{(\boldsymbol{V}^{\star}_{:,l})^{\top}(\boldsymbol{D}_k - \widehat{\boldsymbol{D}}_k)(\boldsymbol{V}_{:,l} - \boldsymbol{V}^{\star}_{:,l})}_{=:\,\gamma_2} \\ &+ \underbrace{(\boldsymbol{V}_{:,l} - \boldsymbol{V}^{\star}_{:,l})^{\top}(\boldsymbol{D}_k - \widehat{\boldsymbol{D}}_k)(\boldsymbol{V}_{:,l} - \boldsymbol{V}^{\star}_{:,l})}_{=:\,\gamma_3}.\end{aligned}$$

- With respect to γ_1, the triangle inequality yields

$$\begin{aligned}|\gamma_1| &= \Big|\sum_{1\leq i,j\leq d}\big(\widehat{E}^{2}_{i,j,k} - E^{2}_{i,j,k}\big)p^{-1}\chi_{i,j,k}V^{\star 2}_{(i,j),l}\Big| \\ &\leq \max_{(i,j,k)\in\Omega}\big|\widehat{E}^{2}_{i,j,k} - E^{2}_{i,j,k}\big| \sum_{1\leq i,j\leq d} p^{-1}\chi_{i,j,k}V^{\star 2}_{(i,j),l}.\end{aligned}$$

From (51), we know that

$$\begin{aligned}\max_{(i,j,k)\in\Omega}\big|\widehat{E}_{ijk} - E_{ijk}\big| &\lesssim \frac{\sigma_{\max}}{\lambda^{\star}_{\min}}\sqrt{\frac{\mu^3 r^2 \log d}{d^2 p}}\,\lambda^{\star}_{\max} \\ &\ll \sigma_{\max}\sqrt{\log d},\end{aligned}$$

where the last inequality arises from the conditions $p \gg \mu^3 r^2 d^{-2}$ and $\kappa \asymp 1$. By the standard results of the sub-Gaussian random variables, one also has

$$\|\boldsymbol{E}\|_{\infty} \lesssim \sigma_{\max}\sqrt{\log d} \tag{133}$$

with probability at least $1 - O\left(d^{-20}\right)$. This reveals that

$$\begin{aligned}&\max_{(i,j,k)\in\Omega}\big|\widehat{E}^{2}_{i,j,k} - E^{2}_{i,j,k}\big| \\ &\quad\leq \max_{(i,j,k)\in\Omega}\big|(\widehat{E}_{i,j,k} - E_{i,j,k})(\widehat{E}_{i,j,k} + E_{i,j,k})\big| \\ &\quad\lesssim \sigma^{2}_{\max}\sqrt{\frac{\mu^3 r^2 \log^2 d}{d^2 p}}.\end{aligned} \tag{134}$$

In addition, apply the Bernstein inequality to find that with probability at least $1 - O\left(d^{-20}\right)$,

$$\begin{aligned}\sum_{1\leq i,j\leq d} p^{-1}\chi_{i,j,k}V^{\star 2}_{(i,j),l} &\lesssim \big\|\boldsymbol{V}^{\star}_{:,l}\big\|_2^2 + p^{-1}\log d\,\big\|\boldsymbol{V}^{\star}_{:,l}\big\|_{\infty}^2 \\ &\quad + \sqrt{p^{-1}\log d}\big\|\boldsymbol{V}^{\star}_{:,l}\big\|_{\infty}^2\big\|\boldsymbol{V}^{\star}_{:,l}\big\|_2^2 \\ &\overset{\text{(i)}}{\asymp} \big\|\boldsymbol{V}^{\star}_{:,l}\big\|_2^2 + p^{-1}\log d\,\big\|\boldsymbol{V}^{\star}_{:,l}\big\|_{\infty}^2 \\ &\overset{\text{(ii)}}{\lesssim} \frac{1}{\lambda^{\star 4/3}_{\min}}\left\{1 + \frac{\mu^2 r \log d}{d^2 p}\right\} \\ &\overset{\text{(iii)}}{\asymp} \frac{1}{\lambda^{\star 4/3}_{\min}},\end{aligned} \tag{135}$$

where (i) follows from the AM-GM inequality, (ii) makes use of (124) and (125), and (iii) holds true as long as $p \gg \mu^2 r d^{-2}\log d$. Combining the above bounds, we arrive at

$$|\gamma_1| \lesssim \frac{\sigma^{2}_{\max}}{\lambda^{\star 4/3}_{\min}}\sqrt{\frac{\mu^3 r^2 \log^2 d}{d^2 p}}$$

with probability at least $1 - O\left(d^{-20}\right)$.

- Turning to γ_2, one can use the triangle inequality and Cauchy-Schwartz to obtain

$$\begin{aligned}|\gamma_2| &= \Big|\sum_{1\leq i,j\leq d}\Big(\big(\widehat{E}^{2}_{i,j,k} - E^{2}_{i,j,k}\big)p^{-1}\chi_{i,j,k} \\ &\qquad \cdot V^{\star}_{(i,j),l}\big(V_{(i,j),l} - V^{\star}_{(i,j),l}\big)\Big)\Big| \\ &\leq \max_{(i,j,k)\in\Omega}\big|\widehat{E}^{2}_{i,j,k} - E^{2}_{i,j,k}\big|\,\big\|\boldsymbol{V}_{:,l} - \boldsymbol{V}^{\star}_{:,l}\big\|_{\infty} \\ &\qquad \cdot \sum_{1\leq i,j\leq d}\Big|p^{-1}\chi_{i,j,k}V^{\star}_{(i,j),l}\Big| \\ &\leq \max_{(i,j,k)\in\Omega}\big|\widehat{E}^{2}_{i,j,k} - E^{2}_{i,j,k}\big|\,\big\|\boldsymbol{V}_{:,l} - \boldsymbol{V}^{\star}_{:,l}\big\|_{\infty} \\ &\qquad \cdot \sqrt{\sum_{1\leq i,j\leq d} p^{-1}\chi_{i,j,k}V^{\star 2}_{(i,j),l}}\sqrt{\sum_{1\leq i,j\leq d} p^{-1}\chi_{i,j,k}}.\end{aligned}$$

It is straightforward to apply the Bernstein inequality to find that, with probability exceeding $1 - O\left(d^{-12}\right)$,

$$\begin{aligned}\sum_{1\leq i,j\leq d} p^{-1}\chi_{i,j,k} &\lesssim d^2 + p^{-1}\log d + \sqrt{d^2 p^{-1}\log d} \\ &\asymp d^2,\end{aligned} \tag{136}$$

with the proviso that $p \gg d^{-2}\log d$. Taking this with (128), (134) and (135) collectively, we conclude that

$$|\gamma_2| \lesssim \frac{\sigma^{2}_{\max}}{\lambda^{\star 4/3}_{\min}}\frac{\sigma_{\max}}{\lambda^{\star}_{\min}}\sqrt{\frac{\mu^5 r^3 \log^3 d}{d p^2}}.$$

- Regarding γ_3, we can develop an upper bound in an analogous manner:

$$\begin{aligned}|\gamma_3| &= \Big|\sum_{1\leq i,j\leq d}\Big(\big(\widehat{E}^{2}_{i,j,k} - E^{2}_{i,j,k}\big)p^{-1}\chi_{i,j,k} \\ &\qquad \cdot \big(V_{(i,j),l} - V^{\star}_{(i,j),l}\big)^2\Big)\Big|\end{aligned}$$

$$\leq \max_{(i,j,k)\in\Omega} \left|\widehat{E}_{i,j,k}^2 - E_{i,j,k}^2\right| \left\|\boldsymbol{V}_{:,l} - \boldsymbol{V}_{:,l}^\star\right\|_\infty^2 \sum_{1\leq i,j\leq d} p^{-1}\chi_{i,j,k}$$
$$\overset{\text{(i)}}{\lesssim} \sigma_{\max}^2 \sqrt{\frac{\mu^3 r^2 \log^2 d}{d^2 p}} \left(\frac{\sigma_{\max}}{\lambda_{\min}^\star}\sqrt{\frac{\mu^2 r \log d}{dp}}\frac{1}{\lambda_{\min}^{\star 2/3}}\right)^2 d^2$$
$$\overset{\text{(ii)}}{\ll} \frac{\sigma_{\max}^2}{\lambda_{\min}^{\star 4/3}} \frac{\sigma_{\max}}{\lambda_{\min}^\star}\sqrt{\frac{\mu^5 r^3 \log^3 d}{dp^2}}.$$

Here, (i) uses (128), (134) and (136), whereas (ii) holds as long as $\sigma_{\max}/\lambda_{\min}^\star \ll \sqrt{p/(\mu^2 r d \log d)}$.

- Taking these bounds together, we demonstrate that with probability at least $1 - O\left(d^{-12}\right)$,

$$|\beta_1| \lesssim \frac{\sigma_{\max}^2}{\lambda_{\min}^{\star 4/3}} \left\{\sqrt{\frac{\mu^3 r^2 \log^2 d}{d^2 p}} + \frac{\sigma_{\max}}{\lambda_{\min}^\star}\sqrt{\frac{\mu^5 r^3 \log^3 d}{dp^2}}\right\}$$
$$\asymp \frac{\sigma_{\max}^2}{\lambda_{\min}^{\star 4/3}} \sqrt{\frac{\mu^3 r^2 \log^2 d}{d^2 p}},$$

where the last step relies on the noise condition $\sigma_{\max}/\lambda_{\min}^\star \ll \sqrt{p/(\mu^2 r d^{3/2} \log d)}$.

2) Bounding β_2: Next, we move on to the term β_2 defined in (131), which admits the following decomposition

$$\beta_2 = (\boldsymbol{V}_{:,l})^\top \widehat{\boldsymbol{D}}_k \boldsymbol{V}_{:,l} - (\boldsymbol{V}_{:,l}^\star)^\top \widehat{\boldsymbol{D}}_k \boldsymbol{V}_{:,l}^\star$$
$$+ (\boldsymbol{V}_{:,l}^\star)^\top \widehat{\boldsymbol{D}}_k \boldsymbol{V}_{:,l}^\star - (\boldsymbol{V}_{:,l}^\star)^\top \boldsymbol{D}_k^\star \boldsymbol{V}_{:,l}^\star$$
$$= 2\underbrace{(\boldsymbol{V}_{:,l}^\star)^\top \widehat{\boldsymbol{D}}_k (\boldsymbol{V}_{:,l} - \boldsymbol{V}_{:,l}^\star)}_{=:\gamma_4} + \underbrace{(\boldsymbol{V}_{:,l} - \boldsymbol{V}_{:,l}^\star)^\top \widehat{\boldsymbol{D}}_k (\boldsymbol{V}_{:,l} - \boldsymbol{V}_{:,l}^\star)}_{=:\gamma_5}$$
$$+ \underbrace{(\boldsymbol{V}_{:,l}^\star)^\top \big(\widehat{\boldsymbol{D}}_k - \boldsymbol{D}_k^\star\big) \boldsymbol{V}_{:,l}^\star}_{=:\gamma_6}.$$

In the sequel, we shall upper bound each of these terms individually.

- To begin with, invoke the Cauchy-Schwartz inequality to bound

$$|\gamma_4| = \Big|\sum_{1\leq i,j\leq d} E_{i,j,k}^2 p^{-1}\chi_{i,j,k} V_{(i,j),l}^\star \big(V_{(i,j),l} - V_{(i,j),l}^\star\big)\Big|$$
$$\leq \left\|\boldsymbol{V}_{:,l} - \boldsymbol{V}_{:,l}^\star\right\|_\infty \cdot \sqrt{\sum_{1\leq i,j\leq d} E_{i,j,k}^4 p^{-1}\chi_{i,j,k} V_{(i,j),l}^{\star 2}}$$
$$\cdot \sqrt{\sum_{1\leq i,j\leq d} p^{-1}\chi_{i,j,k}}.$$

Applying the Bernstein inequality yields that, with probability at least $1 - O\left(d^{-13}\right)$,

$$\sum_{1\leq i,j\leq d} E_{i,j,k}^4 p^{-1}\chi_{i,j,k} V_{(i,j),l}^{\star 2}$$
$$\lesssim \sigma_{\max}^4 \Big\{\left\|\boldsymbol{V}_{:,l}^\star\right\|_2^2 + p^{-1}\log^2 d \left\|\boldsymbol{V}_{:,l}^\star\right\|_\infty^2$$
$$+ \sqrt{p^{-1}\log d}\left\|\boldsymbol{V}_{:,l}^\star\right\|_\infty \left\|\boldsymbol{V}_{:,l}^\star\right\|_2\Big\}$$
$$\overset{\text{(i)}}{\asymp} \sigma_{\max}^4 \left\{\left\|\boldsymbol{V}_{:,l}^\star\right\|_2^2 + p^{-1}\log^2 d \left\|\boldsymbol{V}_{:,l}^\star\right\|_\infty^2\right\}$$
$$\overset{\text{(ii)}}{\lesssim} \frac{\sigma_{\max}^4}{\lambda_{\min}^{\star 4/3}} \left\{1 + \frac{\mu^2 r \log^2 d}{d^2 p}\right\} \overset{\text{(iii)}}{\asymp} \frac{\sigma_{\max}^4}{\lambda_{\min}^{\star 4/3}},$$

where (i) is due to the AM-GM inequality, (ii) uses (124) and (125), and (iii) holds as long as $p \gtrsim \mu^2 r d^{-2} \log^2 d$. This combined with (128) and (136) further leads to

$$|\gamma_4| \lesssim \frac{\sigma_{\max}}{\lambda_{\min}^\star}\sqrt{\frac{\mu^2 r \log d}{dp}}\frac{1}{\lambda_{\min}^{\star 2/3}} \cdot d \cdot \frac{\sigma_{\max}^2}{\lambda_{\min}^{\star 2/3}}$$
$$= \frac{\sigma_{\max}^2}{\lambda_{\min}^{\star 4/3}}\frac{\sigma_{\max}}{\lambda_{\min}^\star}\sqrt{\frac{\mu^2 r d \log d}{p}}.$$

- Next, we turn to the term γ_5. By (128), (133) and (136), the following holds with probability at least $1 - O\left(d^{-13}\right)$,

$$\gamma_5 = \sum_{1\leq i,j\leq d} E_{i,j,k}^2 p^{-1}\chi_{i,j,k}\big(V_{(i,j),l} - V_{(i,j),l}^\star\big)^2$$
$$\leq \|\boldsymbol{E}\|_\infty^2 \left\|\boldsymbol{V}_{:,l} - \boldsymbol{V}_{:,l}^\star\right\|_\infty^2 \sum_{1\leq i,j\leq d} p^{-1}\chi_{i,j,k}$$
$$\lesssim \sigma_{\max}^2 \log d \cdot \left(\frac{\sigma_{\max}}{\lambda_{\min}^\star}\sqrt{\frac{\mu^2 r \log d}{dp}}\frac{1}{\lambda_{\min}^{\star 2/3}}\right)^2 \cdot d^2$$
$$\ll \frac{\sigma_{\max}^2}{\lambda_{\min}^{\star 4/3}}\frac{\sigma_{\max}}{\lambda_{\min}^\star}\sqrt{\frac{\mu^2 r d \log d}{p}},$$

where the last step holds as long as $\sigma_{\max}/\lambda_{\min}^\star \ll \sqrt{p/(\mu^2 r d \log^3 d)}$.

- As for γ_6, we can express $\gamma_6 = \sum_{1\leq i,j\leq d} V_{(i,j),l}^{\star 2}(p^{-1}E_{i,j,k}^2\chi_{i,j,k} - \sigma_{i,j,k}^2)$ as a sum of independent zero-mean random variables. With the assistance of (125), one derives

$$B := \max_{1\leq i,j\leq d}\left\|V_{(i,j),l}^{\star 2}(p^{-1}E_{i,j,k}^2\chi_{i,j,k} - \sigma_{i,j,k}^2)\right\|_{\psi_1}$$
$$\leq \frac{\sigma_{\max}^2}{p}\left\|\boldsymbol{V}_{:,l}^\star\right\|_\infty^2 \lesssim \frac{\sigma_{\max}^2}{\lambda_{\min}^{\star 4/3}}\frac{\mu^2 r}{d^2 p}.$$
$$V := \sum_{1\leq i,j\leq d} V_{(i,j),l}^{\star 4}\mathbb{E}\big[(p^{-1}E_{i,j,k}^2\chi_{i,j,k} - \sigma_{i,j,k}^2)^2\big]$$
$$\lesssim \frac{\sigma_{\max}^4}{p}\left\|\boldsymbol{V}_{:,l}^\star\right\|_\infty^2\left\|\boldsymbol{V}_{:,l}^\star\right\|_2^2 \lesssim \frac{\sigma_{\max}^4}{\lambda_{\min}^{\star 8/3}}\frac{\mu^2 r}{d^2 p}.$$

Applying the matrix Bernstein inequality, one has with probability at least $1 - O\left(d^{-20}\right)$,

$$|\gamma_6| \lesssim B\log^2 d + \sqrt{V\log d}$$
$$\lesssim \frac{\sigma_{\max}^2}{\lambda_{\min}^{\star 4/3}}\left\{\frac{\mu^2 r \log^2 d}{d^2 p} + \sqrt{\frac{\mu^2 r \log d}{d^2 p}}\right\}$$
$$\asymp \frac{\sigma_{\max}^2}{\lambda_{\min}^{\star 4/3}}\sqrt{\frac{\mu^2 r \log d}{d^2 p}},$$

where the last step holds as long as $p \gtrsim \mu^2 r d^{-2} \log^3 d$.

- Putting the bounds above together, we reach

$$|\beta_2| \lesssim \frac{\sigma_{\max}^2}{\lambda_{\min}^{\star 4/3}}\left\{\sqrt{\frac{\mu^2 r \log d}{d^2 p}} + \frac{\sigma_{\max}}{\lambda_{\min}^\star}\sqrt{\frac{\mu^2 r d \log d}{p}}\right\}.$$

3) Combining β_1 and β_2 to Establish the Claim (130): Taking the bounds on β_1 and β_2 collectively yields that, with probability exceeding $1 - O\left(d^{-10}\right)$,

$$\left|s_{l,k}^2 - s_{l,k}^{\star 2}\right| \lesssim \frac{\sigma_{\max}^2}{\lambda_{\min}^{\star 4/3} p}\left\{\sqrt{\frac{\mu^3 r^2 \log^2 d}{d^2 p}} + \frac{\sigma_{\max}}{\lambda_{\min}^{\star}}\sqrt{\frac{\mu^2 r d \log d}{p}}\right\}$$
$$\lesssim \left\{\sqrt{\frac{\mu^3 r^2 \log^2 d}{d^2 p}} + \frac{\sigma_{\max}}{\lambda_{\min}^{\star}}\sqrt{\frac{\mu^2 r d \log d}{p}}\right\} s_{l,k}^{\star 2},$$

where we have used the lower bound on $s_{l,k}^{\star 2}$ (cf. (129)) as well as the conditions $\sigma_{\max}/\sigma_{\min} \asymp 1$ and $\kappa \asymp 1$ in the last step.

4) Proof of Lemma 18:

1) We first consider the ℓ_2 norm of $\boldsymbol{V}_{:,l}^{\star}$. Let $\boldsymbol{\Lambda}^{\star} \in \mathbb{R}^{r\times r}$ be a diagonal matrix with entries $\Lambda_{i,i}^{\star} = \|\boldsymbol{u}_i^{\star}\|_2^3$ for all $1 \le i \le r$. We can then decompose

$$\boldsymbol{V}_{:,l}^{\star} = \left(\widetilde{\boldsymbol{U}}^{\star}\boldsymbol{\Lambda}^{\star -4/3}\right)_{:,l} + \widetilde{\boldsymbol{U}}^{\star}\left((\widetilde{\boldsymbol{U}}^{\star\top}\widetilde{\boldsymbol{U}}^{\star})^{-1} - \boldsymbol{\Lambda}^{\star -4/3}\right)_{:,l}$$
$$= \|\boldsymbol{u}_l^{\star}\|_2^{-4}\,\boldsymbol{u}_l^{\star\otimes 2} + \widetilde{\boldsymbol{U}}^{\star}\left((\widetilde{\boldsymbol{U}}^{\star\top}\widetilde{\boldsymbol{U}}^{\star})^{-1} - \boldsymbol{\Lambda}^{\star -4/3}\right)_{:,l}.$$

One can use the assumption (12b), as well as the conditions (63b) and $\kappa \asymp 1$, to bound the second term

$$\begin{aligned}
&\left\|\widetilde{\boldsymbol{U}}^{\star}\left((\widetilde{\boldsymbol{U}}^{\star\top}\widetilde{\boldsymbol{U}}^{\star})^{-1} - \boldsymbol{\Lambda}^{\star -4/3}\right)_{:,l}\right\|_2 \\
&\le \left\|\widetilde{\boldsymbol{U}}^{\star}\right\|\left\|(\widetilde{\boldsymbol{U}}^{\star\top}\widetilde{\boldsymbol{U}}^{\star})^{-1}\right\|\left\|\widetilde{\boldsymbol{U}}^{\star\top}\widetilde{\boldsymbol{U}}^{\star} - \boldsymbol{\Lambda}^{\star 4/3}\right\|\left\|\boldsymbol{\Lambda}^{\star -4/3}\right\| \\
&\lesssim \lambda_{\max}^{\star 2/3}\cdot\frac{1}{\lambda_{\min}^{\star 4/3}}\cdot r\max_{i\neq j}\left|\left\langle \boldsymbol{u}_i^{\star}, \boldsymbol{u}_j^{\star}\right\rangle\right|^2\cdot\frac{1}{\lambda_{\min}^{\star 4/3}} \\
&\lesssim \frac{1}{\lambda_{\min}^{\star 2/3}}\frac{\mu r}{d} = \frac{o(1)}{\|\boldsymbol{u}_l^{\star}\|_2^2},
\end{aligned} \tag{137}$$

where the last step arises from the condition $r = o(d/\mu)$. Therefore, we obtain that $\left\|\boldsymbol{V}_{:,l}^{\star}\right\|_2 = (1 + o(1))\,\|\boldsymbol{u}_l^{\star}\|_2^{-2}$.

2) Regarding the ℓ_∞ norm of $\boldsymbol{V}_{:,l}^{\star}$, we can use (63a) and (63b) to upper bound

$$\left\|\boldsymbol{V}_{:,l}^{\star}\right\|_\infty \le \left\|\widetilde{\boldsymbol{U}}^{\star}\right\|_{2,\infty}\left\|(\widetilde{\boldsymbol{U}}^{\star\top}\widetilde{\boldsymbol{U}}^{\star})^{-1}\right\|$$
$$\lesssim \frac{\mu\sqrt{r}\,\lambda_{\max}^{\star 2/3}}{d}\frac{1}{\lambda_{\min}^{\star 4/3}} \lesssim \frac{\mu\sqrt{r}}{d}\frac{1}{\lambda_{\min}^{\star 2/3}}.$$

3) Moreover, for any $1 \le i \le d$, one can apply (63a) and (63b) again to demonstrate that

$$\begin{aligned}
\sum_{1\le k\le d} V_{(i,k),l}^{\star 2} &= \sum_{1\le k\le d}\left(\widetilde{\boldsymbol{U}}_{(i,k),:}^{\star}(\widetilde{\boldsymbol{U}}^{\star\top}\widetilde{\boldsymbol{U}}^{\star})_{:,l}^{-1}\right)^2 \\
&\le \sum_{1\le k\le d}\left\|\widetilde{\boldsymbol{U}}_{(i,k),:}^{\star}\right\|_2^2\left\|(\widetilde{\boldsymbol{U}}^{\star\top}\widetilde{\boldsymbol{U}}^{\star})^{-1}\right\|^2 \\
&\le \frac{1}{\lambda_{\min}^{\star 8/3}}\sum_{1\le k\le d}\sum_{1\le s\le r}(\boldsymbol{u}_s^{\star})_i^2\,(\boldsymbol{u}_s^{\star})_k^2 \\
&\le \frac{1}{\lambda_{\min}^{\star 8/3}}\|\boldsymbol{U}^{\star}\|_{2,\infty}^2\max_{1\le s\le r}\|\boldsymbol{u}_s^{\star}\|_2^2 \\
&\lesssim \frac{1}{\lambda_{\min}^{\star 8/3}}\cdot\lambda_{\max}^{\star 2/3}\cdot\frac{\mu r}{d}\lambda_{\max}^{\star 2/3} \lesssim \frac{\mu r}{d}\frac{1}{\lambda_{\min}^{\star 4/3}},
\end{aligned}$$

where the last line holds due to $\kappa \asymp 1$.

4) Regarding the ℓ_2 loss, invoke (63c), (81) and (103) to upper bound

$$\begin{aligned}
\left\|(\boldsymbol{V}_{:,l} - \boldsymbol{V}_{:,l}^{\star})\right\|_2 &\le \left\|\widetilde{\boldsymbol{U}}(\widetilde{\boldsymbol{U}}^{\top}\widetilde{\boldsymbol{U}})^{-1} - \widetilde{\boldsymbol{U}}^{\star}(\widetilde{\boldsymbol{U}}^{\star\top}\widetilde{\boldsymbol{U}}^{\star})^{-1}\right\| \\
&\le \left\|\widetilde{\boldsymbol{U}} - \widetilde{\boldsymbol{U}}^{\star}\right\|\left\|(\widetilde{\boldsymbol{U}}^{\top}\widetilde{\boldsymbol{U}})^{-1}\right\| \\
&\quad + \left\|\widetilde{\boldsymbol{U}}^{\star}\right\|\left\|(\widetilde{\boldsymbol{U}}^{\top}\widetilde{\boldsymbol{U}})^{-1} - (\widetilde{\boldsymbol{U}}^{\star\top}\widetilde{\boldsymbol{U}}^{\star})^{-1}\right\| \\
&\lesssim \frac{\sigma_{\max}}{\lambda_{\min}^{\star}}\sqrt{\frac{d}{p}}\,\lambda_{\max}^{\star 2/3}\cdot\frac{1}{\lambda_{\min}^{\star 4/3}} \\
&\quad + \lambda_{\max}^{\star 2/3}\cdot\frac{\sigma_{\max}}{\lambda_{\min}^{\star 3}}\sqrt{\frac{d}{p}}\,\lambda_{\max}^{\star 2/3} \\
&\lesssim \frac{\sigma_{\max}}{\lambda_{\min}^{\star}}\sqrt{\frac{d}{p}}\,\frac{1}{\lambda_{\min}^{\star 2/3}},
\end{aligned}$$

which holds as long as $\kappa \asymp 1$.

5) Finally, combining (63e), (63c), (63a) and (103) allows us to upper bound the ℓ_∞ loss by

$$\begin{aligned}
\left\|(\boldsymbol{V}_{:,l} - \boldsymbol{V}_{:,l}^{\star})\right\|_\infty &\le \left\|\widetilde{\boldsymbol{U}} - \widetilde{\boldsymbol{U}}^{\star}\right\|_{2,\infty}\left\|(\widetilde{\boldsymbol{U}}^{\top}\widetilde{\boldsymbol{U}})^{-1}\right\| \\
&\quad + \left\|\widetilde{\boldsymbol{U}}^{\star}\right\|_{2,\infty}\left\|(\widetilde{\boldsymbol{U}}^{\top}\widetilde{\boldsymbol{U}})^{-1} - (\widetilde{\boldsymbol{U}}^{\star\top}\widetilde{\boldsymbol{U}}^{\star})^{-1}\right\| \\
&\lesssim \frac{\sigma_{\max}}{\lambda_{\min}^{\star}}\sqrt{\frac{\mu^2 r\log d}{dp}}\,\lambda_{\max}^{\star 2/3}\cdot\frac{1}{\lambda_{\min}^{\star 4/3}} \\
&\quad + \frac{\mu\sqrt{r}}{d}\lambda_{\max}^{\star 2/3}\cdot\frac{\sigma_{\max}}{\lambda_{\min}^{\star 3}}\sqrt{\frac{d}{p}}\,\lambda_{\max}^{\star 2/3} \\
&\asymp \frac{\sigma_{\max}}{\lambda_{\min}^{\star}}\sqrt{\frac{\mu^2 r\log d}{dp}}\,\frac{1}{\lambda_{\min}^{\star 2/3}},
\end{aligned}$$

where the last step arises from $\kappa \asymp 1$.

B. Proof of Lemma 10

Fix any $1 \le i \le j \le k \le d$. In order to control $K_{i,j,k}$, we will apply an almost identical argument as in Appendix E-A for Lemma 9. We omit some details of proof for the sake of conciseness.

By definition, one can express

$$K_{i,j,k} = \frac{T_{i,j,k} - T_{i,j,k}^{\star}}{v_{i,j,k}^{1/2}} - \frac{T_{i,j,k} - T_{i,j,k}^{\star}}{v_{i,j,k}^{\star 1/2}}$$
$$= \left(T_{i,j,k} - T_{i,j,k}^{\star}\right)\frac{v_{i,j,k}^{\star} - v_{i,j,k}}{\sqrt{v_{i,j,k}v_{i,j,k}^{\star}}}\frac{1}{\sqrt{v_{i,j,k}^{\star}} + \sqrt{v_{i,j,k}}}.$$

Recall the definitions $\boldsymbol{\Delta} \coloneqq \boldsymbol{U}\boldsymbol{\Pi} - \boldsymbol{U}^{\star}$, $\widetilde{\boldsymbol{\Delta}} = \left[\boldsymbol{\Delta}_l^{\otimes 2}\right]_{1\le l\le r} \in \mathbb{R}^{d^2\times r}$ and $\omega_{i,j,k} \coloneqq \left\|\widetilde{\boldsymbol{U}}_{(i,j),:}^{\star}\right\|_2 + \left\|\widetilde{\boldsymbol{U}}_{(i,k),:}^{\star}\right\|_2 + \left\|\widetilde{\boldsymbol{U}}_{(j,k),:}^{\star}\right\|_2$ in (33). In view of the decomposition in (46), it is straightforward to bound

$$\begin{aligned}
\left|T_{i,j,k} - T_{i,j,k}^{\star}\right| &\lesssim \|\boldsymbol{\Delta}\|_{2,\infty}\,\omega_{i,j,k} \\
&\quad + \left|\left\langle \boldsymbol{U}_{i,:}^{\star}, \widetilde{\boldsymbol{\Delta}}_{(j,k),:}\right\rangle\right| + \left|\left\langle \boldsymbol{U}_{j,:}^{\star}, \widetilde{\boldsymbol{\Delta}}_{(i,k),:}\right\rangle\right| \\
&\quad + \left|\left\langle \boldsymbol{U}_{k,:}^{\star}, \widetilde{\boldsymbol{\Delta}}_{(i,k),:}\right\rangle\right| + \left|\left\langle \boldsymbol{\Delta}_{i,:}, \widetilde{\boldsymbol{\Delta}}_{(j,k),:}\right\rangle\right| \\
&\overset{\text{(i)}}{\lesssim} \frac{\sigma_{\max}}{\lambda_{\min}^{\star}}\sqrt{\frac{\mu r\log d}{p}}\,\lambda_{\max}^{\star 1/3}\omega_{i,j,k}
\end{aligned}$$

$$
\begin{aligned}
&+\left(\frac{\sigma_{\max}}{\lambda_{\min}^{\star}}\sqrt{\frac{\mu r\log d}{p}}\lambda_{\max}^{\star 1/3}\right)^2\sqrt{\frac{\mu}{d}}\,\lambda_{\max}^{\star 1/3}\\
&\overset{\text{(ii)}}{\lesssim}\left\{\sqrt{\mu r\log d}+\frac{\sigma_{\max}}{\lambda_{\min}^{\star 1/3}}\frac{\mu^{3/2}r\log d}{\sqrt{dp}}\frac{1}{\omega_{i,j,k}}\right\}\sqrt{v_{i,j,k}^{\star}}\\
&\overset{\text{(iii)}}{\lesssim}\sqrt{\mu r\log d}\sqrt{v_{i,j,k}^{\star}},
\end{aligned}
$$

where (i) uses (59b), (121) and (122); (ii) follows from the lower bound of $v_{i,j,k}^{\star}$ in (118) and conditions $\sigma_{\max}/\sigma_{\min},\kappa\asymp 1$; and (iii) arises from the assumption (34).

Then the claim (138) would immediately follow as long as we could show that

$$
\begin{aligned}
\frac{\left|v_{i,j,k}-v_{i,j,k}^{\star}\right|}{v_{i,j,k}^{\star}}&\lesssim\sqrt{\frac{\mu^3r^2\log d}{d^2p}}+\frac{1}{\omega_{i,j,k}}\frac{\sigma_{\max}}{\lambda_{\min}^{\star 1/3}}\sqrt{\frac{\mu^4r^2\log d}{dp}}\\
&=o\left(1\right). \qquad (138)
\end{aligned}
$$

Indeed, one can apply the triangle inequality to show that $v_{i,j,k}\asymp v_{i,j,k}^{\star}$, and consequently obtain

$$
\begin{aligned}
\left|K_{i,j,k}\right|&\lesssim\frac{1}{v_{i,j,k}^{\star 3/2}}\left|T_{i,j,k}-T_{i,j,k}^{\star}\right|\left|v_{i,j,k}-v_{i,j,k}^{\star}\right|\\
&\lesssim\sqrt{\frac{\mu^4r^3\log^2 d}{d^2p}}+\frac{1}{\omega_{i,j,k}}\frac{\sigma_{\max}}{\lambda_{\min}^{\star 1/3}}\sqrt{\frac{\mu^5r^3\log^2 d}{dp}}
\end{aligned}
$$

as claimed.

Therefore, it remains to justify (138). For notational convenience, we define the following $d^2\times d^2$ matrices:

$$
\boldsymbol{P}\coloneqq\widetilde{\boldsymbol{U}}(\widetilde{\boldsymbol{U}}^{\top}\widetilde{\boldsymbol{U}})^{-1}\widetilde{\boldsymbol{U}}^{\top},\ \boldsymbol{P}^{\star}\coloneqq\widetilde{\boldsymbol{U}}^{\star}(\widetilde{\boldsymbol{U}}^{\star\top}\widetilde{\boldsymbol{U}}^{\star})^{-1}\widetilde{\boldsymbol{U}}^{\star\top}. \qquad (139)
$$

We can then express

$$
\begin{aligned}
v_{i,j,k}^{\star}&=\frac{2}{p}\Big(\boldsymbol{P}_{(i,j),:}^{\star}\boldsymbol{D}_k^{\star}\boldsymbol{P}_{:,(i,j)}^{\star}+\boldsymbol{P}_{(i,k),:}^{\star}\boldsymbol{D}_j^{\star}\boldsymbol{P}_{:,(i,k)}^{\star}\\
&\qquad+\boldsymbol{P}_{(j,k),:}^{\star}\boldsymbol{D}_i^{\star}\boldsymbol{P}_{:,(j,k)}^{\star}\Big), \qquad (140)\\
v_{i,j,k}&=\frac{2}{p}\Big(\boldsymbol{P}_{(i,j),:}\boldsymbol{D}_k\boldsymbol{P}_{:,(i,j)}+\boldsymbol{P}_{(i,k),:}\boldsymbol{D}_j\boldsymbol{P}_{:,(i,k)}\\
&\qquad+\boldsymbol{P}_{(j,k),:}\boldsymbol{D}_i\boldsymbol{P}_{:,(j,k)}\Big), \qquad (141)
\end{aligned}
$$

where $\boldsymbol{D}_k^{\star}$ (resp. $\boldsymbol{D}_k$) is defined in (18) (resp. (25)) for each $1\le k\le d$. Lemma 19 summarizes several bounds regarding $\boldsymbol{P}$ and $\boldsymbol{P}^{\star}$, whose proof can be found at the end of the section.

Lemma 19: Instate the assumptions and notations of Lemma 10. For any $1\le i,j\le d$, one has

$$
\begin{aligned}
\left\|\boldsymbol{P}_{(i,j),:}^{\star}\right\|_2&\lesssim\frac{1}{\lambda_{\min}^{\star 2/3}}\left\|\widetilde{\boldsymbol{U}}_{(i,j),:}^{\star}\right\|_2, &(142)\\
\left\|\boldsymbol{P}_{(i,j),:}^{\star}\right\|_\infty&\lesssim\frac{\mu\sqrt{r}}{d}\frac{1}{\lambda_{\min}^{\star 2/3}}\left\|\widetilde{\boldsymbol{U}}_{(i,j),:}^{\star}\right\|_2; &(143)\\
\left\|(\boldsymbol{P}-\boldsymbol{P}^{\star})_{(i,j),:}\right\|_2&\lesssim\frac{\sigma_{\max}}{\lambda_{\min}^{\star}}\sqrt{\frac{d\log d}{p}}\frac{\mu\sqrt{r}}{d}, &(144)\\
\left\|(\boldsymbol{P}-\boldsymbol{P}^{\star})_{(i,j),:}\right\|_\infty&\lesssim\frac{\sigma_{\max}}{\lambda_{\min}^{\star}}\sqrt{\frac{d\log d}{p}}\frac{\mu^2r}{d^2}. &(145)
\end{aligned}
$$

With these in mind, we are positioned to upper bound $v_{i,j,k}-v_{i,j,k}^{\star}$. By (140), (141) and the triangle inequality, we will show below how to upper bound $\boldsymbol{P}_{(i,j),:}\boldsymbol{D}_k\boldsymbol{P}_{:,(i,j)}-\boldsymbol{P}_{(i,j),:}^{\star}\boldsymbol{D}_k^{\star}\boldsymbol{P}_{:,(i,j)}^{\star}$. The other two terms can be controlled analogously.

Recall the auxiliary matrix $\widehat{\boldsymbol{D}}_k$ (cf. 132). One can then expand

$$
\begin{aligned}
&\boldsymbol{P}_{(i,j),:}\boldsymbol{D}_k\boldsymbol{P}_{:,(i,j)}-\boldsymbol{P}_{(i,j),:}^{\star}\boldsymbol{D}_k^{\star}\boldsymbol{P}_{:,(i,j)}^{\star}\\
&\quad=\underbrace{\boldsymbol{P}_{(i,j),:}(\boldsymbol{D}_k-\widehat{\boldsymbol{D}}_k)\boldsymbol{P}_{:,(i,j)}}_{=:\beta_1}\\
&\qquad+\underbrace{\boldsymbol{P}_{(i,j),:}\widehat{\boldsymbol{D}}_k\boldsymbol{P}_{:,(i,j)}-\boldsymbol{P}_{(i,j),:}^{\star}\boldsymbol{D}_k^{\star}\boldsymbol{P}_{:,(i,j)}^{\star}}_{=:\beta_2}.
\end{aligned}
$$

In what follows, we shall control β_1 and β_2 individually.

- For β_1, one decomposes it as follows

$$
\begin{aligned}
\beta_1&=\underbrace{\boldsymbol{P}_{(i,j),:}^{\star}(\boldsymbol{D}_k-\widehat{\boldsymbol{D}}_k)\boldsymbol{P}_{:,(i,j)}^{\star}}_{=:\gamma_1}\\
&\quad+2\underbrace{(\boldsymbol{P}-\boldsymbol{P}^{\star})_{:,(i,j)}(\boldsymbol{D}_k-\widehat{\boldsymbol{D}}_k)\boldsymbol{P}_{:,(i,j)}^{\star}}_{=:\gamma_2}\\
&\quad+\underbrace{(\boldsymbol{P}-\boldsymbol{P}^{\star})_{:,(i,j)}(\boldsymbol{D}_k-\widehat{\boldsymbol{D}}_k)(\boldsymbol{P}-\boldsymbol{P}^{\star})_{:,(i,j)}}_{=:\gamma_3}.
\end{aligned}
$$

The term γ_1 can be bounded by

$$
|\gamma_1|\le\max_{(s,l,k)\in\Omega}\left|\widehat{E}_{s,l,k}^2-E_{s,l,k}^2\right|\sum_{1\le s,l\le d}p^{-1}\chi_{s,l,k}P_{(i,j),(s,l)}^{\star 2}.
$$

Using (142) and (143), we know from the Bernstein inequality and the AM-GM inequality that with probability at least $1-O\left(d^{-13}\right)$,

$$
\begin{aligned}
&\sum_{1\le s,l\le d}p^{-1}\chi_{s,l,k}P_{(i,j),(s,l)}^{\star 2}\\
&\lesssim\left\|\boldsymbol{P}_{(i,j),:}^{\star}\right\|_2^2+p^{-1}\log d\left\|\boldsymbol{P}_{(i,j),:}^{\star}\right\|_\infty^2\\
&\quad+\sqrt{p^{-1}\log d}\left\|\boldsymbol{P}_{(i,j),:}^{\star}\right\|_\infty\left\|\boldsymbol{P}_{(i,j),:}^{\star}\right\|_2\\
&\asymp\left\|\boldsymbol{P}_{(i,j),:}^{\star}\right\|_2^2+p^{-1}\log d\left\|\boldsymbol{P}_{(i,j),:}^{\star}\right\|_\infty^2\\
&\lesssim\frac{1}{\lambda_{\min}^{\star 4/3}}\left\{1+\frac{\mu^2r\log d}{d^2p}\right\}\left\|\widetilde{\boldsymbol{U}}_{(i,j),:}^{\star}\right\|_2^2\\
&\asymp\frac{1}{\lambda_{\min}^{\star 4/3}}\left\|\widetilde{\boldsymbol{U}}_{(i,j),:}^{\star}\right\|_2^2, \qquad (146)
\end{aligned}
$$

where the last step arises from the condition $p\gtrsim\mu^2rd^{-2}\log d$. This combined with (134) leads to

$$
|\gamma_1|\lesssim\frac{\sigma_{\max}^2}{\lambda_{\min}^{\star 4/3}}\sqrt{\frac{\mu^3r^2\log^2 d}{d^2p}}\left\|\widetilde{\boldsymbol{U}}_{(i,j),:}^{\star}\right\|_2^2.
$$

As for γ_2, invoking Cauchy-Schwartz and applying (134) and (146) give

$$
\begin{aligned}
|\gamma_2|&\le\max_{(s,l,k)\in\Omega}\left|\widehat{E}_{s,l,k}^2-E_{s,l,k}^2\right|\left\|(\boldsymbol{P}-\boldsymbol{P}^{\star})_{(i,j),:}\right\|_\infty\\
&\quad\cdot\sqrt{\sum_{1\le s,l\le d}p^{-1}\chi_{s,l,k}P_{(i,j),(s,l)}^{\star 2}}\sqrt{\sum_{1\le s,l\le d}p^{-1}\chi_{s,l,k}}\\
&\lesssim\frac{\sigma_{\max}^2}{\lambda_{\min}^{\star 2/3}}\frac{\sigma_{\max}}{\lambda_{\min}^{\star}}\sqrt{\frac{d\log d}{p}}\frac{\mu^2r}{d}\left\|\widetilde{\boldsymbol{U}}_{(i,j),:}^{\star}\right\|_2,
\end{aligned}
$$

as long as $p \gtrsim \mu^3 r^2 d^{-2} \log^2 d$. Regarding γ_3, we can upper bound

$$\begin{aligned}|\gamma_3| &\le \max_{(s,l,k)\in\Omega} \left|\widehat{E}_{s,l,k}^2 - E_{s,l,k}^2\right| \left\|(\boldsymbol{P}-\boldsymbol{P}^\star)_{(i,j),:}\right\|_\infty^2 \\ &\quad \cdot \sum_{1\le s,l\le d} p^{-1}\chi_{s,l,k} \\ &\overset{\text{(i)}}{\lesssim} \sigma_{\max}^2 \sqrt{\frac{\mu^3 r^2 \log^2 d}{d^2 p}} \left(\frac{\sigma_{\max}}{\lambda_{\min}^\star}\sqrt{\frac{d\log d}{p}}\frac{\mu^2 r}{d^2}\right)^2 d^2 \\ &\overset{\text{(ii)}}{\lesssim} \sigma_{\max}^2 \left(\frac{\sigma_{\max}}{\lambda_{\min}^\star}\sqrt{\frac{d\log d}{p}}\frac{\mu^2 r}{d}\right)^2,\end{aligned}$$

where (i) uses (134), (136) and (145); (ii) holds as long as $p \gtrsim \mu^3 r^2 d^{-2} \log^2 d$. Taking the above bounds for γ_1, γ_2 and γ_3 together indicates that

$$\begin{aligned}|\beta_1| &\lesssim \frac{\sigma_{\max}^2}{\lambda_{\min}^{\star 4/3}} \sqrt{\frac{\mu^3 r^2 \log^2 d}{d^2 p}} \left\|\widetilde{\boldsymbol{U}}_{(i,j),:}^\star\right\|_2^2 \\ &\quad + \frac{\sigma_{\max}^2}{\lambda_{\min}^{\star 2/3}} \frac{\sigma_{\max}}{\lambda_{\min}^\star}\sqrt{\frac{d\log d}{p}}\frac{\mu^2 r}{d}\left\|\widetilde{\boldsymbol{U}}_{(i,j),:}^\star\right\|_2 \\ &\quad + \sigma_{\max}^2 \left(\frac{\sigma_{\max}}{\lambda_{\min}^\star}\sqrt{\frac{d\log d}{p}}\frac{\mu^2 r}{d}\right)^2.\end{aligned}$$

- Regarding β_2, we start by decomposing it as follows

$$\begin{aligned}\beta_2 &= \boldsymbol{P}_{(i,j),:}\widehat{\boldsymbol{D}}_k\boldsymbol{P}_{:,(i,j)} - \boldsymbol{P}_{(i,j),:}^\star\widehat{\boldsymbol{D}}_k\boldsymbol{P}_{:,(i,j)}^\star \\ &\quad + \boldsymbol{P}_{(i,j),:}^\star\widehat{\boldsymbol{D}}_k\boldsymbol{P}_{:,(i,j)}^\star - \boldsymbol{P}_{(i,j),:}^\star\boldsymbol{D}_k^\star\boldsymbol{P}_{:,(i,j)}^\star \\ &= 2\underbrace{\boldsymbol{P}_{(i,j),:}^\star\widehat{\boldsymbol{D}}_k(\boldsymbol{P}-\boldsymbol{P}^\star)_{:,(i,j)}}_{=:\gamma_4} \\ &\quad + \underbrace{(\boldsymbol{P}-\boldsymbol{P}^\star)_{(i,j),:}(\widehat{\boldsymbol{D}}_k - \boldsymbol{D}_k^\star)(\boldsymbol{P}-\boldsymbol{P}^\star)_{:,(i,j)}}_{=:\gamma_5} \\ &\quad + \underbrace{\boldsymbol{P}_{(i,j),:}^\star(\widehat{\boldsymbol{D}}_k - \boldsymbol{D}_k^\star)\boldsymbol{P}_{:,(i,j)}^\star}_{=:\gamma_6}.\end{aligned}$$

To bound γ_4, we can combine (142), (145) and (136) with the Cauchy-Schwartz inequality and the Bernstein inequality, to obtain with probability at least $1-O\left(d^{-13}\right)$,

$$\begin{aligned}|\gamma_4| &= \Big|\sum_{1\le s,l\le d}\Big(E_{s,l,k}^2 p^{-1}\chi_{s,l,k} \\ &\qquad \cdot P_{(i,j),(s,l)}^\star\big(P_{(i,j),(s,l)} - P_{(i,j),(s,l)}^\star\big)\Big)\Big| \\ &\le \left\|(\boldsymbol{P}-\boldsymbol{P}^\star)_{(i,j),:}\right\|_\infty \sqrt{\sum_{1\le s,l\le d} p^{-1}\chi_{s,l,k}} \\ &\quad \cdot \sqrt{\sum_{1\le s,l\le d} E_{s,l,k}^4 p^{-1}\chi_{s,l,k} P_{(i,j),(s,l)}^{\star 2}} \\ &\le \left\|(\boldsymbol{P}-\boldsymbol{P}^\star)_{(i,j),:}\right\|_\infty \cdot d \\ &\quad \cdot \sigma_{\max}^2\left\{\left\|\boldsymbol{P}_{(i,j),:}^\star\right\|_2 + p^{-1/2}\log d\left\|\boldsymbol{P}_{(i,j),:}^\star\right\|_\infty\right\} \\ &\lesssim \frac{\sigma_{\max}^2}{\lambda_{\min}^{\star 2/3}}\frac{\sigma_{\max}}{\lambda_{\min}^\star}\sqrt{\frac{d\log d}{p}}\frac{\mu^2 r}{d}\left\|\widetilde{\boldsymbol{U}}_{(i,j),:}^\star\right\|_2\end{aligned}$$

as long as $p \gtrsim \mu^2 r d^{-2}\log^2 d$. As for γ_5, combining (133), (136) and (145) shows that with probability at least $1-O\left(d^{-13}\right)$,

$$\begin{aligned}\gamma_5 &= \sum_{1\le s,l\le d} E_{s,l,k}^2 p^{-1}\chi_{s,l,k}\big(P_{(i,j),(s,l)} - P_{(i,j),(s,l)}^\star\big)^2 \\ &\le \|\boldsymbol{E}\|_\infty^2 \left\|(\boldsymbol{P}-\boldsymbol{P}^\star)_{(i,j),:}\right\|_\infty^2 \sum_{1\le s,l\le d} p^{-1}\chi_{s,l,k} \\ &\lesssim \sigma_{\max}^2\left(\frac{\sigma_{\max}}{\lambda_{\min}^\star}\sqrt{\frac{d\log^2 d}{p}}\frac{\mu^2 r}{d}\right)^2.\end{aligned}$$

Finally, observe that γ_6 is a sum of independent random variables. By (142) and (143), invoking the Bernstein inequality reveals that with probability at least $1-O\left(d^{-13}\right)$,

$$\begin{aligned}|\gamma_6| &\lesssim \sigma_{\max}^2\Bigg\{\frac{\log^2 d}{p}\left\|\boldsymbol{P}_{(i,j),:}^\star\right\|_\infty^2 \\ &\qquad + \sqrt{\frac{\log d}{p}}\left\|\boldsymbol{P}_{(i,j),:}^\star\right\|_\infty\left\|\boldsymbol{P}_{(i,j),:}^\star\right\|_2\Bigg\} \\ &\lesssim \frac{\sigma_{\max}^2}{\lambda_{\min}^{\star 4/3}}\left\{\frac{\mu^2 r\log^2 d}{d^2 p} + \sqrt{\frac{\mu^2 r\log d}{d^2 p}}\right\}\left\|\widetilde{\boldsymbol{U}}_{(i,j),:}^\star\right\|_2^2 \\ &\asymp \frac{\sigma_{\max}^2}{\lambda_{\min}^{\star 4/3}}\sqrt{\frac{\mu^2 r\log d}{d^2 p}}\left\|\widetilde{\boldsymbol{U}}_{(i,j),:}^\star\right\|_2^2,\end{aligned}$$

as long as $p \gtrsim \mu^2 r d^{-2}\log^3 d$. Therefore, we combine bounds for γ_4, γ_5 and γ_6 to conclude that

$$\begin{aligned}|\beta_2| &\lesssim \frac{\sigma_{\max}^2}{\lambda_{\min}^{\star 4/3}}\sqrt{\frac{\mu^2 r\log d}{d^2 p}}\left\|\widetilde{\boldsymbol{U}}_{(i,j),:}^\star\right\|_2^2 \\ &\quad + \frac{\sigma_{\max}^2}{\lambda_{\min}^{\star 2/3}}\frac{\sigma_{\max}}{\lambda_{\min}^\star}\sqrt{\frac{d\log d}{p}}\frac{\mu^2 r}{d}\left\|\widetilde{\boldsymbol{U}}_{(i,j),:}^\star\right\|_2 \\ &\quad + \sigma_{\max}^2\left(\frac{\sigma_{\max}}{\lambda_{\min}^\star}\sqrt{\frac{d\log^2 d}{p}}\frac{\mu^2 r}{d}\right)^2.\end{aligned}$$

- Putting the above bounds for β_1 and β_2 together reveals that

$$\begin{aligned}&\left|\boldsymbol{P}_{(i,j),:}\boldsymbol{D}_k\boldsymbol{P}_{:,(i,j)} - \boldsymbol{P}_{(i,j),:}^\star\boldsymbol{D}_k^\star\boldsymbol{P}_{:,(i,j)}^\star\right| \\ &\quad\lesssim \frac{\sigma_{\max}^2}{\lambda_{\min}^{\star 4/3}}\sqrt{\frac{\mu^3 r^2\log d}{d^2 p}}\left\|\widetilde{\boldsymbol{U}}_{(i,j),:}^\star\right\|_2^2 \\ &\quad\quad + \frac{\sigma_{\max}^2}{\lambda_{\min}^{\star 2/3}}\frac{\sigma_{\max}}{\lambda_{\min}^\star}\sqrt{\frac{d\log d}{p}}\frac{\mu^2 r}{d}\left\|\widetilde{\boldsymbol{U}}_{(i,j),:}^\star\right\|_2 \\ &\quad\quad + \sigma_{\max}^2\left(\frac{\sigma_{\max}}{\lambda_{\min}^\star}\sqrt{\frac{d\log^2 d}{p}}\frac{\mu^2 r}{d}\right)^2.\end{aligned}$$

- Clearly, we can apply an analogous argument to bound $\boldsymbol{P}_{(i,k),:}\boldsymbol{D}_j\boldsymbol{P}_{:,(i,k)} - \boldsymbol{P}_{(i,k),:}^\star\boldsymbol{D}_j^\star\boldsymbol{P}_{:,(i,k)}^\star$ and $\boldsymbol{P}_{(j,k),:}\boldsymbol{D}_i\boldsymbol{P}_{:,(j,k)} - \boldsymbol{P}_{(j,k),:}^\star\boldsymbol{D}_i^\star\boldsymbol{P}_{:,(j,k)}^\star$. Taken collectively with the lower bound of $v_{i,j,k}^\star$ (cf. (118))

and the conditions $\sigma_{\max}/\sigma_{\min} \asymp 1$ and $\kappa \asymp 1$, we obtain

$$\begin{aligned}\frac{\left|v_{i,j,k} - v^\star_{i,j,k}\right|}{v^\star_{i,j,k}} &\lesssim \sqrt{\frac{\mu^3 r^2 \log d}{d^2 p}} \\ &\quad + \frac{1}{\omega_{i,j,k}} \frac{\sigma_{\max}}{\lambda_{\min}^{\star 1/3}} \sqrt{\frac{d \log d}{p}} \frac{\mu^2 r}{d} \\ &\quad + \left(\frac{1}{\omega_{i,j,k}} \frac{\sigma_{\max}}{\lambda_{\min}^{\star 1/3}} \sqrt{\frac{d \log^2 d}{p}} \frac{\mu^2 r}{d}\right)^2. \\ &\asymp \sqrt{\frac{\mu^3 r^2 \log d}{d^2 p}} + \frac{1}{\omega_{i,j,k}} \frac{\sigma_{\max}}{\lambda_{\min}^{\star 1/3}} \sqrt{\frac{d \log d}{p}} \frac{\mu^2 r}{d} = o(1),\end{aligned}$$

where the last line holds due to the assumptions (34) and $p \gg \mu^3 r^2 d^{-2} \log^2 d$.

1) Proof of Lemma 19: Fix any $1 \leq i, j \leq d$.

- We start with the norms of the rows of $\boldsymbol{P}^\star$ (cf. 139). By (63b), it is straightforward to deduce that

$$\begin{aligned}\left\|\boldsymbol{P}^\star_{(i,j),:}\right\|_2 &\leq \left\|\widetilde{\boldsymbol{U}}^\star_{(i,j),:}\right\|_2 \left\|(\widetilde{\boldsymbol{U}}^{\star\top} \widetilde{\boldsymbol{U}}^\star)^{-1}\right\| \left\|\widetilde{\boldsymbol{U}}^\star\right\| \\ &\lesssim \frac{1}{\lambda_{\min}^{\star 2/3}} \left\|\widetilde{\boldsymbol{U}}^\star_{(i,j),:}\right\|_2; \\ \left\|\boldsymbol{P}^\star_{(i,j),:}\right\|_\infty &\leq \left\|\widetilde{\boldsymbol{U}}^\star_{(i,j),:}\right\|_2 \left\|(\widetilde{\boldsymbol{U}}^{\star\top} \widetilde{\boldsymbol{U}}^\star)^{-1}\right\| \left\|\widetilde{\boldsymbol{U}}^\star\right\|_{2,\infty} \\ &\lesssim \frac{\mu\sqrt{r}}{d} \frac{1}{\lambda_{\min}^{\star 2/3}} \left\|\widetilde{\boldsymbol{U}}^\star_{(i,j),:}\right\|_2.\end{aligned}$$

- Next, we move on to the ℓ_2 norm of $(\boldsymbol{P} - \boldsymbol{P}^\star)_{(i,j),:}$, which can be decomposed as

$$\begin{aligned}(\boldsymbol{P} - \boldsymbol{P}^\star)_{(i,j),:} &= \widetilde{\boldsymbol{U}}_{(i,j),:}(\widetilde{\boldsymbol{U}}^\top \widetilde{\boldsymbol{U}})^{-1} \widetilde{\boldsymbol{U}}^\top - \widetilde{\boldsymbol{U}}^\star_{(i,j),:}(\widetilde{\boldsymbol{U}}^{\star\top} \widetilde{\boldsymbol{U}}^\star)^{-1} \widetilde{\boldsymbol{U}}^{\star\top} \\ &= (\widetilde{\boldsymbol{U}} - \widetilde{\boldsymbol{U}}^\star)_{(i,j),:}(\widetilde{\boldsymbol{U}}^\top \widetilde{\boldsymbol{U}})^{-1} \widetilde{\boldsymbol{U}}^\top \\ &\quad + \widetilde{\boldsymbol{U}}^\star_{(i,j),:}\big((\widetilde{\boldsymbol{U}}^\top \widetilde{\boldsymbol{U}})^{-1} - (\widetilde{\boldsymbol{U}}^{\star\top} \widetilde{\boldsymbol{U}}^\star)^{-1}\big) \widetilde{\boldsymbol{U}}^\top \\ &\quad + \widetilde{\boldsymbol{U}}^\star_{(i,j),:}(\widetilde{\boldsymbol{U}}^{\star\top} \widetilde{\boldsymbol{U}}^\star)^{-1} (\widetilde{\boldsymbol{U}} - \widetilde{\boldsymbol{U}}^\star)^\top. \end{aligned} \tag{147}$$

By the triangle inequality, we shall control these three terms separately. From (63e), one has

$$\begin{aligned}&\left\|(\widetilde{\boldsymbol{U}} - \widetilde{\boldsymbol{U}}^\star)_{(i,j),:}(\widetilde{\boldsymbol{U}}^\top \widetilde{\boldsymbol{U}})^{-1} \widetilde{\boldsymbol{U}}^\top\right\|_2 \\ &\quad \leq \left\|(\widetilde{\boldsymbol{U}} - \widetilde{\boldsymbol{U}}^\star)_{(i,j),:}\right\|_2 \left\|(\widetilde{\boldsymbol{U}}^\top \widetilde{\boldsymbol{U}})^{-1}\right\| \left\|\widetilde{\boldsymbol{U}}\right\| \\ &\quad \lesssim \frac{\sigma_{\max}}{\lambda^\star_{\min}} \sqrt{\frac{\mu^2 r \log d}{dp}} \lambda_{\max}^{\star 2/3} \cdot \frac{1}{\lambda_{\min}^{\star 4/3}} \cdot \lambda_{\max}^{\star 2/3}.\end{aligned}$$

As for the remaining two terms, from (63a), (63c) and (103), one has

$$\begin{aligned}&\left\|\widetilde{\boldsymbol{U}}^\star_{(i,j),:}\right\|_2 \left\|(\widetilde{\boldsymbol{U}}^\top \widetilde{\boldsymbol{U}})^{-1} - (\widetilde{\boldsymbol{U}}^{\star\top} \widetilde{\boldsymbol{U}}^\star)^{-1}\right\| \left\|\widetilde{\boldsymbol{U}}\right\| \\ &\quad \lesssim \frac{\sigma_{\max}}{\lambda_{\min}^{\star 3}} \sqrt{\frac{d}{p}} \lambda_{\max}^{\star 2/3} \cdot \lambda_{\max}^{\star 2/3} \left\|\widetilde{\boldsymbol{U}}^\star_{(i,j),:}\right\|_2 \\ &\quad \lesssim \frac{\sigma_{\max}}{\lambda_{\min}^{\star 5/3}} \sqrt{\frac{d}{p}} \left\|\widetilde{\boldsymbol{U}}^\star_{(i,j),:}\right\|_2.\end{aligned}$$

Combining (63a), (63b) and (81) yields

$$\begin{aligned}&\left\|\widetilde{\boldsymbol{U}}^\star_{(i,j),:}\right\| \left\|(\widetilde{\boldsymbol{U}}^{\star\top} \widetilde{\boldsymbol{U}}^\star)^{-1}\right\| \left\|\widetilde{\boldsymbol{U}} - \widetilde{\boldsymbol{U}}^\star\right\| \\ &\quad \lesssim \frac{1}{\lambda_{\min}^{\star 4/3}} \cdot \frac{\sigma_{\max}}{\lambda^\star_{\min}} \sqrt{\frac{d}{p}} \lambda_{\max}^{\star 2/3} \left\|\widetilde{\boldsymbol{U}}^\star_{(i,j),:}\right\|_2 \\ &\quad \lesssim \frac{\sigma_{\max}}{\lambda_{\min}^{\star 5/3}} \sqrt{\frac{d}{p}} \left\|\widetilde{\boldsymbol{U}}^\star_{(i,j),:}\right\|_2.\end{aligned}$$

The above bounds taken collectively allow us to obtain

$$\begin{aligned}&\left\|(\boldsymbol{P} - \boldsymbol{P}^\star)_{(i,j),:}\right\|_2 \\ &\quad \lesssim \frac{\sigma_{\max}}{\lambda_{\min}^{\star 5/3}} \sqrt{\frac{d}{p}} \left(\frac{\mu\sqrt{r \log d}}{d} \lambda_{\max}^{\star 2/3} + \left\|\widetilde{\boldsymbol{U}}^\star_{(i,j),:}\right\|_2\right) \\ &\quad \asymp \frac{\sigma_{\max}}{\lambda^\star_{\min}} \sqrt{\frac{d \log d}{p}} \frac{\mu\sqrt{r}}{d},\end{aligned}$$

where the last step arises from the incoherence condition that $\|\widetilde{\boldsymbol{U}}^\star\|_{2,\infty} \lesssim \mu\sqrt{r}\lambda_{\max}^{\star 2/3}/d$ and $\kappa \asymp 1$.

- Finally, let us look at the ℓ_∞ norm of $(\boldsymbol{P} - \boldsymbol{P}^\star)_{(i,j),:}$. Armed with the decomposition in (147), we can bound

$$\begin{aligned}&\left\|(\boldsymbol{P} - \boldsymbol{P}^\star)_{(i,j),:}\right\|_\infty \\ &\quad \leq \left\|(\widetilde{\boldsymbol{U}} - \widetilde{\boldsymbol{U}}^\star)_{(i,j),:}\right\|_2 \left\|(\widetilde{\boldsymbol{U}}^\top \widetilde{\boldsymbol{U}})^{-1}\right\| \left\|\widetilde{\boldsymbol{U}}\right\|_{2,\infty} \\ &\qquad + \left\|\widetilde{\boldsymbol{U}}^\star_{(i,j),:}\right\|_2 \left\|(\widetilde{\boldsymbol{U}}^\top \widetilde{\boldsymbol{U}})^{-1} - (\widetilde{\boldsymbol{U}}^{\star\top} \widetilde{\boldsymbol{U}}^\star)^{-1}\right\| \left\|\widetilde{\boldsymbol{U}}\right\|_{2,\infty} \\ &\qquad + \left\|\widetilde{\boldsymbol{U}}^\star_{(i,j),:}\right\|_2 \left\|(\widetilde{\boldsymbol{U}}^{\star\top} \widetilde{\boldsymbol{U}}^\star)^{-1}\right\| \left\|\widetilde{\boldsymbol{U}} - \widetilde{\boldsymbol{U}}^\star\right\|_{2,\infty} \\ &\quad \overset{\text{(i)}}{\lesssim} \frac{\sigma_{\max}}{\lambda^\star_{\min}} \sqrt{\frac{\mu^2 r \log d}{dp}} \lambda_{\max}^{\star 2/3} \cdot \frac{1}{\lambda_{\min}^{\star 4/3}} \cdot \frac{\mu\sqrt{r}}{d} \lambda_{\max}^{\star 2/3} \\ &\qquad + \frac{\sigma_{\max}}{\lambda_{\min}^{\star 3}} \sqrt{\frac{d}{p}} \lambda_{\max}^{\star 2/3} \cdot \frac{\mu\sqrt{r}}{d} \lambda_{\max}^{\star 2/3} \left\|\widetilde{\boldsymbol{U}}^\star_{(i,j),:}\right\|_2 \\ &\qquad + \frac{1}{\lambda_{\min}^{\star 4/3}} \cdot \frac{\sigma_{\max}}{\lambda^\star_{\min}} \sqrt{\frac{\mu^2 r \log d}{dp}} \lambda_{\max}^{\star 2/3} \left\|\widetilde{\boldsymbol{U}}^\star_{(i,j),:}\right\|_2 \\ &\quad \overset{\text{(ii)}}{\lesssim} \frac{\sigma_{\max}}{\lambda_{\min}^{\star 5/3}} \sqrt{\frac{\mu^2 r \log d}{dp}} \left(\frac{\mu\sqrt{r}}{d} \lambda_{\max}^{\star 2/3} + \left\|\widetilde{\boldsymbol{U}}^\star_{(i,j),:}\right\|_2\right) \\ &\quad \overset{\text{(iii)}}{\lesssim} \frac{\sigma_{\max}}{\lambda^\star_{\min}} \sqrt{\frac{\mu^2 r \log d}{dp}} \frac{\mu\sqrt{r}}{d}.\end{aligned}$$

Here, (i) relies on (63) and (103), (ii) is due to the condition $\kappa \asymp 1$, whereas (iii) arises from (63a) and $\kappa \asymp 1$.

Appendix F
Proof of ℓ_2 Estimation Guarantees (Theorem 13)

As before, we assume that $\boldsymbol{\Pi} = \boldsymbol{I}_r$ for simplicity of notation throughout this section.

A. ℓ_2 Risk for Tensor Factor Estimation

Fix an arbitrary $1 \leq l \leq r$. Recalling the decomposition in (39), (40) and (44), we can write $\boldsymbol{u}_l - \boldsymbol{u}^\star_l = \boldsymbol{Z}_{:,l} + \boldsymbol{W}_{:,l}$. In what follows, we will first prove that $\boldsymbol{Z}_{:,l}$ converges to a Gaussian random vector in distribution. Then we can use the standard Gaussian concentration inequality to show that the ℓ_2 norm of the Gaussian random vector concentrates around its expectation. Combined with the observation that the ℓ_2 norm of $\boldsymbol{W}_{:,l}$ is negligible as shown in (45) (established in Lemmas 3-6), this implies the advertised bound on the ℓ_2 norm of $\boldsymbol{u}_l - \boldsymbol{u}_l^\star$.

Now we begin the proof. For convenience of presentation, we adopt the notation in (123) that $\boldsymbol{V}^\star := \widetilde{\boldsymbol{U}}^\star(\widetilde{\boldsymbol{U}}^{\star\top}\widetilde{\boldsymbol{U}}^\star)^{-1}$. Then we can express

$$\begin{aligned}\boldsymbol{Z}_{:,l} = \sqrt{2}\sum_{1\le i,k\le d} p^{-1}E_{i,i,k}\chi_{i,i,k}V^\star_{(i,i),l}\boldsymbol{e}_k \\ + \sum_{1\le i,j,k\le d} p^{-1}E_{i,j,k}\chi_{i,j,k}V^\star_{(i,j),l}\boldsymbol{e}_k\end{aligned}$$

as a sum of independent zero-mean random vectors in $\mathbb{R}^d$. Let us first compute the covariance matrix $\boldsymbol{S}_l^\star := \mathbb{E}\big[\boldsymbol{Z}_{:,l}(\boldsymbol{Z}_{:,l})^\top\big]$. Straightforward computation yields that for each $1\le i\le d$,

$$(\boldsymbol{S}_l^\star)_{i,i} = 2\sum_{1\le k_1,k_2\le d} p^{-1}\sigma^2_{i,k_1,k_2}V^{\star 2}_{(k_1,k_2),l}; \tag{148}$$

and for each $1\le i\neq j\le d$,

$$\begin{aligned}(\boldsymbol{S}_l^\star)_{i,j} &= 2\sqrt{2}\,p^{-1}\sigma^2_{i,i,j}V^\star_{(i,i),l}V^\star_{(i,j),l} \\ &\quad + 2\sqrt{2}\,p^{-1}\sigma^2_{i,j,j}V^\star_{(j,j),l}V^\star_{(i,j),l} \\ &\quad + \sum_{k:k\neq i,j} 4\,p^{-1}\sigma^2_{i,j,k}V^\star_{(i,k),l}V^\star_{(j,k),l} \\ &= 4\sum_{1\le k\le d} p^{-1}\sigma^2_{i,j,k}V^\star_{(i,k),l}V^\star_{(j,k),l} \\ &\quad - \big(4-2\sqrt{2}\big)p^{-1}\sigma^2_{i,i,j}V^\star_{(i,i),l}V^\star_{(i,j),l} \\ &\quad - \big(4-2\sqrt{2}\big)p^{-1}\sigma^2_{i,j,j}V^\star_{(j,j),l}V^\star_{(i,j),l}.\end{aligned} \tag{149}$$

Lemma 20 below collects several properties of $\boldsymbol{S}_l^\star$ and the proof is deferred to the end of this section.

Lemma 20: Instate the assumptions of Theorem 6. One has

$$\lambda_{\max}(\boldsymbol{S}_l^\star) \lesssim \frac{\sigma^2_{\max}}{p\,\|\boldsymbol{u}_l^\star\|_2^4}, \qquad \lambda_{\min}(\boldsymbol{S}_l^\star) \gtrsim \frac{\sigma^2_{\min}}{p\,\|\boldsymbol{u}_l^\star\|_2^4};$$
$$\mathsf{tr}(\boldsymbol{S}_l^\star) = \frac{(2+o(1))\,\sigma^2_{\max}d}{p\,\|\boldsymbol{u}_l^\star\|_2^4}, \qquad \|\boldsymbol{S}_l^\star\|_{\mathrm{F}} \lesssim \frac{\sigma^2_{\max}\sqrt{d}}{p\,\|\boldsymbol{u}_l^\star\|_2^4}.$$

Recall that we want to show $\boldsymbol{Z}_{:,l}$ converges to a Gaussian random vector $\boldsymbol{g}_l \sim \mathcal{N}(\boldsymbol{0},\boldsymbol{S}_l^\star)$ in distribution. By the Cramér–Wold theorem, it suffices to prove that for any $\boldsymbol{a} = (a_1,\cdots,a_d)^\top \in \mathbb{R}^d$, $\boldsymbol{a}^\top\boldsymbol{Z}_{:,l}$ converges to $\boldsymbol{a}^\top\boldsymbol{g}_l$ in distribution. Towards this, we apply the Berry-Esseen theorem [91, Theorem 1.1] (cf. Appendix B-C) again and upper bound ρ defined in (65). Without loss of generality, we assume $\|\boldsymbol{a}\|_2 = 1$. We can compute

$$\begin{aligned}&\sum_{1\le i,j,k\le d}\mathbb{E}\Big[\big|a_k p^{-1}E_{i,j,k}\chi_{i,j,k}V^\star_{(i,j),l}\big|^3\Big] \\ &\le \frac{1}{p^3}\|\boldsymbol{a}\|_\infty\big\|\boldsymbol{V}^\star_{:,l}\big\|_\infty\sum_{1\le i,j,k\le d} a_k^2\,\mathbb{E}\big[|E_{i,j,k}|^3\chi_{i,j,k}\big]V^{\star 2}_{(i,j),l} \\ &\overset{\text{(i)}}{\lesssim} \frac{\sigma^3_{\max}}{p^2}\|\boldsymbol{a}\|_\infty\|\boldsymbol{a}\|_2^2\big\|\boldsymbol{V}^\star_{:,l}\big\|_\infty\big\|\boldsymbol{V}^\star_{:,l}\big\|_2^2 \\ &\overset{\text{(ii)}}{\lesssim} \frac{\sigma^3_{\max}}{p^2}\cdot\frac{\mu\sqrt{r}}{d}\frac{1}{\lambda_{\min}^{\star 2/3}}\cdot\frac{1}{\lambda_{\min}^{\star 4/3}},\end{aligned}$$

where we use the property of sub-gaussian random variables in (i), and (ii) follows from (124), (125) and $\|\boldsymbol{a}\|_\infty \le \|\boldsymbol{a}\|_2 = 1$. Moreover, from Lemma 20, it is easy to see that

$$\begin{aligned}\mathsf{Var}\big(\boldsymbol{a}^\top\boldsymbol{Z}_{:,l}\big) &= \boldsymbol{a}^\top\boldsymbol{S}_l^\star\boldsymbol{a} \ge \lambda_{\min}(\boldsymbol{S}_l^\star)\,\|\boldsymbol{a}\|_2^2 \\ &\gtrsim \frac{\sigma^2_{\min}}{p\,\|\boldsymbol{u}_l^\star\|_2^4} \gtrsim \frac{\sigma^2_{\min}}{p\lambda_{\max}^{\star 4/3}}.\end{aligned}$$

One can then bound ρ

$$\begin{aligned}\rho &= \frac{1}{\mathsf{Var}^{3/2}(\boldsymbol{a}^\top\boldsymbol{Z}_{:,l})}\sum_{1\le i,j,k\le d}\mathbb{E}\Big[\big|a_k p^{-1}E_{i,j,k}\chi_{i,j,k}V^\star_{(i,j),l}\big|^3\Big] \\ &\lesssim \frac{p^{3/2}\lambda^{\star 2}_{\max}}{\sigma^3_{\min}}\cdot\frac{\sigma^3_{\max}}{\lambda^{\star 2}_{\min}}\frac{\mu\sqrt{r}}{dp^2} \lesssim \frac{\mu\sqrt{r}}{d\sqrt{p}} = o(1)\end{aligned}$$

where we use the condition that $\sigma_{\max}/\sigma_{\min},\kappa \asymp 1$ and $p \gg \mu^2 r d^{-3/2}$. Therefore, we justify the claimed distributional convergence of $\boldsymbol{a}^\top\boldsymbol{Z}_{:,l}$, which further implies the convergence of $\boldsymbol{Z}_{:,l}$ by the Cramér–Wold theorem.

Given that $\boldsymbol{Z}_{:,l}$ converges to $\boldsymbol{g}_l$ in distribution, we now apply the Gaussian concentration inequality [95, Proposition 1] to demonstrate the squared ℓ_2 norm of $\boldsymbol{g}_l$ is tightly concentrated around its mean with high probability. By Lemma 20, we can use the Gaussian concentration inequality [95, Proposition 1] to find that with probability at least $1 - O\big(d^{-11}\big)$,

$$\begin{aligned}\|\boldsymbol{g}_l\|_2^2 - \mathsf{tr}(\boldsymbol{S}_l^\star) &\lesssim \|\boldsymbol{S}_l^\star\|_{\mathrm{F}}\sqrt{\log d} + \|\boldsymbol{S}_l^\star\|\log d \\ &\lesssim \frac{\sigma^2_{\max}(\sqrt{d\log d}+\log d)}{p\,\|\boldsymbol{u}_l^\star\|_2^4} \\ &= o(1)\,\frac{\sigma^2_{\max}d}{p\,\|\boldsymbol{u}_l^\star\|_2^4},\end{aligned}$$

and consequently,

$$\|\boldsymbol{g}_l\|_2^2 \le \frac{2\,(1+o(1))\,\sigma^2_{\max}d}{p\,\|\boldsymbol{u}_l^\star\|_2^4}.$$

Moreover, we know from the continuous mapping theorem that $\|\boldsymbol{Z}_{:,l}\|_2^2$ converges to $\|\boldsymbol{g}_l\|_2^2$ in distribution because $\|\cdot\|_2^2$ is a continuous function. Therefore, we find that with probability at least $1-o(1)$,

$$\|\boldsymbol{Z}_{:,l}\|_2^2 \le \frac{2\,(1+o(1))\,\sigma^2_{\max}d}{p\,\|\boldsymbol{u}_l^\star\|_2^4}. \tag{150}$$

It remains to upper bound $\|\boldsymbol{W}_{:,l}\|_2$, which is easily accomplished with the help of (45). Indeed, it is straightforward to find that with probability at least $1 - O\big(d^{-11}\big)$,

$$\|\boldsymbol{W}_{:,l}\|_2^2 \le \sum_{1\le k\le d}\|\boldsymbol{W}_{k,:}\|_{2,\infty}^2 \le d\cdot\frac{o(1)\,\sigma^2_{\max}}{\lambda_{\min}^{\star 4/3}p} = \frac{o(1)\,\sigma^2_{\max}d}{p\,\|\boldsymbol{u}_l^\star\|_2^4},$$

where we use the assumption that $\kappa \asymp 1$ in the last step. Taken collectively with (150) finishes the proof.

1) Proof of Lemma 20: To begin with, let us consider the trace of $\boldsymbol{S}_l^\star$. From (148) and (124), it is straightforward to calculate

$$\mathsf{tr}(\boldsymbol{S}_l^\star) = \sum_{1\le i\le d} (\boldsymbol{S}_l^\star)_{i,i} \le 2 \sum_{1\le i,j,k\le d} p^{-1}\sigma_{i,j,k}^2 V_{(j,k),l}^{\star 2} = \frac{2\,\sigma_{\max}^2 d}{p}\left\|\boldsymbol{V}_{:,l}^\star\right\|_2^2 = \frac{2\,(1+o\,(1))\,\sigma_{\max}^2 d}{p\,\|\boldsymbol{u}_l^\star\|_2^4}.$$

As for the Frobenius norm of $\boldsymbol{S}_l^\star$, we note that it is an immediate consequence of the claim for the spectrum of $\boldsymbol{S}_l^\star$. Indeed, since $\boldsymbol{S}_l^\star$ is a positive semidefinite matrix, we know that

$$\|\boldsymbol{S}_l^\star\|_{\mathrm{F}}^2 = \mathsf{tr}(\boldsymbol{S}_l^{\star 2}) = \sum_{1\le i\le d}\lambda_i(\boldsymbol{S}_l^{\star 2}) = \sum_{1\le i\le d}\lambda_i^2(\boldsymbol{S}_l^\star) \le d\cdot\lambda_{\max}^2(\boldsymbol{S}_l^\star) \lesssim \frac{\sigma_{\max}^4 d}{p^2\,\|\boldsymbol{u}_l^\star\|_2^8}$$

as claimed.

Hence, the remainder of the proof amounts to controlling the eigenvalues of $\boldsymbol{S}_l^\star$. Let us decompose $\boldsymbol{S}_l^\star =: 2\widehat{\boldsymbol{S}}_l^\star - \breve{\boldsymbol{S}}_l^\star \in \mathbb{R}^{d\times d}$, where the entries of $\widehat{\boldsymbol{S}}_l^\star$ and $\breve{\boldsymbol{S}}_l^\star$ are given by

$$\big(\widehat{\boldsymbol{S}}_l^\star\big)_{i,j} = \begin{cases} \sum_{1\le k_1,k_2\le d} p^{-1}\sigma_{i,k_1,k_2}^2 V_{(k_1,k_2),l}^{\star 2}, & \text{if } i=j,\\ 2\sum_{1\le k\le d} p^{-1}\sigma_{i,j,k}^2 V_{(i,k),l}^\star V_{(j,k),l}^\star, & \text{if } i\ne j,\end{cases}$$

$$\big(\breve{\boldsymbol{S}}_l^\star\big)_{i,j} = \big(4-2\sqrt{2}\big)p^{-1}\Big(\sigma_{i,i,j}^2 V_{(i,i),l}^\star V_{(i,j),l}^\star + \sigma_{i,j,j}^2 V_{(j,j),l}^\star V_{(i,j),l}^\star\Big), \quad \forall 1\le i\ne j\le d.$$

and

$$\big(\breve{\boldsymbol{S}}_l^\star\big)_{i,i} = 0, \quad \forall 1\le i\le d.$$

Our proof strategy is to show that the spectrum of $\boldsymbol{S}_l^\star$ is mainly determined by $\widehat{\boldsymbol{S}}_l^\star$ (since $\|\breve{\boldsymbol{S}}_l^\star\|$ is a negligible term). One can then invoke Weyl's inequality to establish the conclusion.

Now we start the analysis. Note that by the symmetric sampling pattern, one equivalently express $\sigma_{i,j,k}^2 = s_i^2 s_j^2 s_k^2$ for each $1\le i,j,k\le d$ with $\max_{1\le i\le d} s_i \le \sigma_{\max}^{1/3}$. We then can decompose

$$\widehat{\boldsymbol{S}}_l^\star = 2\boldsymbol{A}\boldsymbol{A}^\top + \mathcal{P}_{\mathsf{diag}}\big(\widehat{\boldsymbol{S}}_l^\star - 2\boldsymbol{A}\boldsymbol{A}^\top\big),$$

where $\mathcal{P}_{\mathsf{diag}}(\boldsymbol{Z})$ extracts out the diagonal entries of a matrix $\boldsymbol{Z}$, and $\boldsymbol{A}\in\mathbb{R}^{d\times d}$ is a matrix with entries $A_{i,k} = \sqrt{1/p}\,s_i^2 s_k V_{(i,k),l}^\star$. Let us first control the spectral norm of $\mathcal{P}_{\mathsf{diag}}\big(\widehat{\boldsymbol{S}}_l^\star - 2\boldsymbol{A}\boldsymbol{A}^\top\big)$. From (126), it is easy to see that for all $1\le i\le d$,

$$\Big|\big(\widehat{\boldsymbol{S}}_l^\star - 2\boldsymbol{A}\boldsymbol{A}^\top\big)_{i,i}\Big| \le \frac{\sigma_{\min}^2}{p}\left\|\boldsymbol{V}_{:,l}^\star\right\|_2^2 + \frac{2\sigma_{\max}^2}{p}\sum_{1\le k\le d} V_{(i,k),l}^{\star 2} \le \frac{(1+o\,(1))\,\sigma_{\min}^2}{p\,\|\boldsymbol{u}_l^\star\|_2^4} + \frac{2\sigma_{\max}^2}{p}\frac{\mu r}{d}\frac{1}{\lambda_{\min}^{\star 4/3}} = \frac{(1+o\,(1))\,\sigma_{\max}^2}{p\,\|\boldsymbol{u}_l^\star\|_2^4}, \tag{151}$$

$$\big(\widehat{\boldsymbol{S}}_l^\star - 2\boldsymbol{A}\boldsymbol{A}^\top\big)_{i,i} \ge \frac{\sigma_{\min}^2}{p}\left\|\boldsymbol{V}_{:,l}^\star\right\|_2^2 - \frac{2\sigma_{\max}^2}{p}\sum_{1\le k\le d} V_{(i,k),l}^{\star 2} = \frac{(1-o\,(1))\,\sigma_{\min}^2}{p\,\|\boldsymbol{u}_l^\star\|_2^4}, \tag{152}$$

where we have used the condition that $\sigma_{\min}/\sigma_{\max}\asymp 1$, $\kappa\asymp 1$, and $r = o\,(d/\mu)$. It then suffices to focus on the spectrum of $\boldsymbol{A}\boldsymbol{A}^\top$, whose entries are given by

$$\Big(\boldsymbol{A}\boldsymbol{A}^\top\Big)_{i,j} = p^{-1}s_i^2 s_j^2 \sum_{1\le k\le d} s_k^2 V_{(i,k),l}^\star V_{(j,k),l}^\star, \quad 1\le i,j\le d.$$

Recalling the definitions $\boldsymbol{V}^\star := \widetilde{\boldsymbol{U}}^\star(\widetilde{\boldsymbol{U}}^{\star\top}\widetilde{\boldsymbol{U}}^\star)^{-1}$ and $u_{l,i}^\star := (\boldsymbol{u}_l^\star)_i$ for each $1\le l\le r, 1\le i\le d$, we can decompose

$$\begin{aligned} V_{(i,k),l}^\star &= \widetilde{\boldsymbol{U}}_{(i,k),:}^\star(\widetilde{\boldsymbol{U}}^{\star\top}\widetilde{\boldsymbol{U}}^\star)_{:,l}^{-1} \\ &= \widetilde{\boldsymbol{U}}_{(i,k),:}^\star\boldsymbol{\Lambda}_{:,l}^{\star -4/3} + \widetilde{\boldsymbol{U}}_{(i,k),:}^\star\big((\widetilde{\boldsymbol{U}}^{\star\top}\widetilde{\boldsymbol{U}}^\star)^{-1} - \boldsymbol{\Lambda}^{\star -4/3}\big)_{:,l} \\ &= \|\boldsymbol{u}_l^\star\|_2^{-4}\,\widetilde{U}_{(i,k),l}^\star + \widetilde{\boldsymbol{U}}_{(i,k),:}^\star\big((\widetilde{\boldsymbol{U}}^{\star\top}\widetilde{\boldsymbol{U}}^\star)^{-1} - \boldsymbol{\Lambda}^{\star -4/3}\big)_{:,l} \\ &= \|\boldsymbol{u}_l^\star\|_2^{-4}\,u_{l,i}^\star u_{l,k}^\star + \underbrace{\widetilde{\boldsymbol{U}}_{(i,k),:}^\star\big((\widetilde{\boldsymbol{U}}^{\star\top}\widetilde{\boldsymbol{U}}^\star)^{-1} - \boldsymbol{\Lambda}^{\star -4/3}\big)_{:,l}}_{=:\,\delta_{i,k}}. \end{aligned} \tag{153}$$

We will show shortly that $V_{(i,k),l}^\star$ is extremely close to $\|\boldsymbol{u}_l^\star\|_2^{-4}u_{l,i}^\star u_{l,k}^\star$. Then we obtain that for each $1\le i,j\le d$,

$$\begin{aligned} p\Big(\boldsymbol{A}\boldsymbol{A}^\top\Big)_{i,j} &= \|\boldsymbol{u}_l^\star\|_2^{-8}\Big(\sum_{1\le k\le d} s_k^2 u_{l,k}^{\star 2}\Big)\big(s_i^2 u_{l,i}^\star\big)\big(s_j^2 u_{l,j}^\star\big) \\ &\quad + \|\boldsymbol{u}_l^\star\|_2^{-4}\big(s_i^2 u_{l,i}^\star\big)\sum_{1\le k\le d} s_k^2 u_{l,k}^\star\big(s_j^2\delta_{j,k}\big) \\ &\quad + \|\boldsymbol{u}_l^\star\|_2^{-4}\big(s_j^2 u_{l,j}^\star\big)\sum_{1\le k\le d} s_k^2 u_{l,k}^\star\big(s_i^2\delta_{i,k}\big) \\ &\quad + \underbrace{\sum_{1\le k\le d} s_k^2\big(s_i^2\delta_{i,k}\big)\big(s_j^2\delta_{j,k}\big)}_{=:\,\Upsilon_{i,j}}, \end{aligned}$$

or equivalently,

$$p\boldsymbol{A}\boldsymbol{A}^\top = \boldsymbol{a}\boldsymbol{a}^\top\sum_{1\le k\le d} s_k^2 u_{l,k}^{\star 2} + \boldsymbol{a}\boldsymbol{b}^\top + \boldsymbol{b}\boldsymbol{a}^\top + \boldsymbol{\Upsilon},$$

where $\boldsymbol{a},\boldsymbol{b}\in\mathbb{R}^d$ with entries $a_i = s_i^2 u_{l,i}^\star/\|\boldsymbol{u}_l^\star\|_2^4$, $b_i = \sum_{1\le k\le d} s_k^2 u_{l,k}^\star\big(s_i^2\delta_{i,k}\big)$ and $\boldsymbol{\Upsilon} = [\Upsilon_{i,j}]_{1\le i,j\le d}\in\mathbb{R}^{d\times d}$. It is straightforward to see that $\boldsymbol{a}\boldsymbol{a}^\top\sum_{1\le k\le d} s_k^2 u_{l,k}^{\star 2}$ is a rank-1 matrix with the non-zero eigenvalue

$$\frac{\sigma_{\min}^2}{\|\boldsymbol{u}_l^\star\|_2^4} \le \lambda_{\max}\Big(\boldsymbol{a}\boldsymbol{a}^\top\sum_{1\le k\le d} s_k^2 u_{l,k}^{\star 2}\Big) \le \frac{\sigma_{\max}^2}{\|\boldsymbol{u}_l^\star\|_2^4}. \tag{154}$$

In addition, the remaining three terms are all small with respect to the spectral norm. Indeed, recalling the decomposition in (153), we can use (137), $\kappa\asymp 1$ and the condition $r = o(\sqrt{d/\mu})$ to show $\delta_{i,k}$ is sufficiently small, i.e.

$$\sum_{1\le i,k\le d}\delta_{i,k}^2 \le \big\|\widetilde{\boldsymbol{U}}^\star\big\|_{\mathrm{F}}^2\big\|(\widetilde{\boldsymbol{U}}^{\star\top}\widetilde{\boldsymbol{U}}^\star)^{-1} - \boldsymbol{\Lambda}^{\star -4/3}\big\|^2 \lesssim r\lambda_{\max}^{\star 4/3}\cdot\left(\frac{1}{\lambda_{\min}^{\star 8/3}}\frac{\mu r}{d}\lambda_{\max}^{\star 4/3}\right)^2 = \frac{o\,(1)}{\lambda_{\min}^{\star 4/3}}. \tag{155}$$

It then follows from the Cauchy-Schwartz inequality that

$$\|\boldsymbol{b}\|_2^2 = \sum_{1\le i\le d}\Big|\sum_{1\le k\le d} s_k^2 u_{l,k}^{\star}\left(s_i^2\delta_{i,k}\right)\Big|^2$$
$$\le \max_{1\le i\le d} s_i^8 \sum_{1\le k\le d} u_{l,k}^{\star 2}\sum_{1\le i,k\le d}\delta_{i,k}^2 \lesssim \frac{\sigma_{\max}^{8/3}}{\lambda_{\min}^{\star 4/3}}\frac{\mu^2 r^3}{d^2}\|\boldsymbol{u}_l^\star\|_2^2;$$
$$\|\boldsymbol{\Upsilon}\|_{\mathrm{F}}^2 = \sum_{1\le i,j\le d}\Big|\sum_{1\le k\le d} s_k^2\left(s_i^2\delta_{i,k}\right)\left(s_j^2\delta_{j,k}\right)\Big|^2$$
$$\le \max_{1\le i\le d} s_i^{12}\Big|\sum_{1\le i,k\le d}\delta_{i,k}^2\Big|^2 \lesssim \left(\frac{\sigma_{\max}^2}{\lambda_{\min}^{\star 4/3}}\frac{\mu^2 r^3}{d^2}\right)^2.$$

This combined with the condition $r = o\big(\sqrt{d/\mu}\big)$ and $\kappa \asymp 1$ reveals that

$$\big\|\boldsymbol{a}\boldsymbol{b}^\top + \boldsymbol{b}\boldsymbol{a}^\top + \boldsymbol{\Upsilon}\big\| \le 2\|\boldsymbol{a}\|_2\|\boldsymbol{b}\|_2 + \|\boldsymbol{\Upsilon}\|_{\mathrm{F}}$$
$$\lesssim \frac{\sigma_{\max}^{2/3}}{\|\boldsymbol{u}_l^\star\|_2^3}\frac{\sigma_{\max}^{4/3}}{\lambda_{\min}^{\star 2/3}}\frac{\mu r^{3/2}}{d}\|\boldsymbol{u}_l^\star\|_2 + \frac{\sigma_{\max}^2}{\lambda_{\min}^{\star 4/3}}\frac{\mu^2 r^3}{d^2}$$
$$= \frac{o(1)\,\sigma_{\max}^2}{\|\boldsymbol{u}_l^\star\|_2^4}.$$

Taken collectively with (151), (152) and (154), we conclude that

$$\lambda_{\max}(\widehat{\boldsymbol{S}}_l^\star) \lesssim \frac{\sigma_{\max}^2}{p\,\|\boldsymbol{u}_l^\star\|_2^4}, \qquad \lambda_{\min}(\widehat{\boldsymbol{S}}_l^\star) \gtrsim \frac{\sigma_{\min}^2}{p\,\|\boldsymbol{u}_l^\star\|_2^4}.$$

Applying a similar argument, one can easily show that

$$\|\breve{\boldsymbol{S}}_l^\star\| \lesssim \frac{\sigma_{\max}^2}{p}\sum_{1\le k\le d} V_{(k,k),l}^{\star 2} \lesssim \frac{\sigma_{\max}^2}{\lambda_{\min}^{\star 4/3} p}\frac{\mu r}{d} = \frac{o(1)\,\sigma_{\min}^2}{p\,\|\boldsymbol{u}_l^\star\|_2^4}$$
$$= o(1)\,\lambda_{\min}(\widehat{\boldsymbol{S}}_l^\star).$$

Therefore, the advertised bound of the eigenvalues of $\boldsymbol{S}_l^\star = 2\widehat{\boldsymbol{S}}_l^\star - \breve{\boldsymbol{S}}_l^\star$ immediately follows from Weyl's inequality.

B. ℓ_2 Risk for Tensor Estimation

To begin with, we recall the notation $\boldsymbol{\Delta}_l \coloneqq \boldsymbol{u}_l - \boldsymbol{u}_l^\star, 1 \le l \le r$, allowing us to expand

$$\boldsymbol{T} - \boldsymbol{T}^\star = \sum_{1\le l\le r}\boldsymbol{\Delta}_l\otimes\boldsymbol{u}_l^{\star\otimes 2} + \sum_{1\le l\le r}\boldsymbol{u}_l^\star\otimes\boldsymbol{\Delta}_l\otimes\boldsymbol{u}_l^\star$$
$$+ \sum_{1\le l\le r}\boldsymbol{u}_l^{\star\otimes 2}\otimes\boldsymbol{\Delta}_l + \sum_{1\le l\le r}\boldsymbol{u}_l^\star\otimes\boldsymbol{\Delta}_l^{\otimes 2}$$
$$+ \sum_{1\le l\le r}\boldsymbol{\Delta}_l\otimes\boldsymbol{u}_l^\star\otimes\boldsymbol{\Delta}_l + \sum_{1\le l\le r}\boldsymbol{\Delta}_l^{\otimes 2}\otimes\boldsymbol{u}_l^\star + \sum_{1\le l\le r}\boldsymbol{\Delta}_l^{\otimes 3}.$$

By symmetry, straightforward calculation yields

$$\|\boldsymbol{T} - \boldsymbol{T}^\star\|_{\mathrm{F}}^2 = 3\underbrace{\Big\|\sum_{1\le l\le r}\boldsymbol{\Delta}_l\otimes\boldsymbol{u}_l^{\star\otimes 2}\Big\|_{\mathrm{F}}^2}_{=:\beta_1}$$
$$+ 3\underbrace{\Big\|\sum_{1\le l\le r}\boldsymbol{u}_l^\star\otimes\boldsymbol{\Delta}_l^{\otimes 2}\Big\|_{\mathrm{F}}^2}_{=:\beta_2} + \underbrace{\Big\|\sum_{1\le l\le r}\boldsymbol{\Delta}_l^{\otimes 3}\Big\|_{\mathrm{F}}^2}_{=:\beta_4}$$
$$+ 6\underbrace{\Big\langle\sum_{1\le l\le r}\boldsymbol{\Delta}_l\otimes\boldsymbol{u}_l^{\star\otimes 2}, \sum_{1\le l\le r}\boldsymbol{u}_l^{\star\otimes 2}\otimes\boldsymbol{\Delta}_l\Big\rangle}_{=:\beta_4} + \beta_5,$$

where

$$\beta_5 \coloneqq 6\Big\langle\sum_{1\le l\le r}\boldsymbol{u}_l^\star\otimes\boldsymbol{\Delta}_l^{\otimes 2}, \sum_{1\le l\le r}\boldsymbol{\Delta}_l^{\otimes 2}\otimes\boldsymbol{u}_l^\star\Big\rangle$$
$$+ 6\Big\langle\sum_{1\le l\le r}\boldsymbol{u}_l^\star\otimes\boldsymbol{\Delta}_l^{\otimes 2}, \sum_{1\le l\le r}\boldsymbol{\Delta}_l^{\otimes 3}\Big\rangle$$
$$+ 6\Big\langle\sum_{1\le l\le r}\boldsymbol{\Delta}_l\otimes\boldsymbol{u}_l^{\star\otimes 2}, \sum_{1\le l\le r}\boldsymbol{u}_l^\star\otimes\boldsymbol{\Delta}_l^{\otimes 2}\Big\rangle$$
$$+ 6\Big\langle\sum_{1\le l\le r}\boldsymbol{\Delta}_l\otimes\boldsymbol{u}_l^{\star\otimes 2}, \sum_{1\le l\le r}\boldsymbol{\Delta}_l^{\otimes 3}\Big\rangle$$
$$+ 12\Big\langle\sum_{1\le l\le r}\boldsymbol{\Delta}_l\otimes\boldsymbol{u}_l^{\star\otimes 2}, \sum_{1\le l\le r}\boldsymbol{\Delta}_l^{\otimes 2}\otimes\boldsymbol{u}_l^\star\Big\rangle.$$

In what follows, we shall control the β_i's separately. In particular, we want to show that the ℓ_2 loss of interest is mainly controlled by β_1, with the remaining four terms being negligible with high probability.

1) We start with β_1. Recalling (39), (40) and (44) that $\boldsymbol{\Delta} := \boldsymbol{U} - \boldsymbol{U}^\star = \boldsymbol{Z} + \boldsymbol{W}$ as well as the notation $\widetilde{\boldsymbol{U}}^\star \coloneqq \big[\boldsymbol{u}_l^{\star\otimes 2}\big]_{1\le l\le r} \in \mathbb{R}^{d^2\times r}$, we can easily see that

$$\beta_1 = \big\|\boldsymbol{\Delta}\widetilde{\boldsymbol{U}}^{\star\top}\big\|_{\mathrm{F}}^2 = \big\|\boldsymbol{Z}\widetilde{\boldsymbol{U}}^{\star\top}\big\|_{\mathrm{F}}^2 + \big\|\boldsymbol{W}\widetilde{\boldsymbol{U}}^{\star\top}\big\|_{\mathrm{F}}^2$$
$$+ 2\big\langle\boldsymbol{Z}\widetilde{\boldsymbol{U}}^{\star\top}, \boldsymbol{W}\widetilde{\boldsymbol{U}}^{\star\top}\big\rangle.$$

One can apply an analogous argument as in Appendix F-A to show that the distribution of $\boldsymbol{Z}\widetilde{\boldsymbol{U}}^{\star\top}$ converges to a multivariate normal distribution, whose Euclidean norm concentrates around its expectation with high probability. We omit the detailed proof for conciseness. One can verify that with probability exceeding $1 - o(1)$,

$$\|\boldsymbol{Z}\widetilde{\boldsymbol{U}}^{\star\top}\|_{\mathrm{F}}^2 = \frac{(2+o(1))\,\sigma_{\max}^2 d}{p}\big\|\widetilde{\boldsymbol{U}}^\star(\widetilde{\boldsymbol{U}}^{\star\top}\widetilde{\boldsymbol{U}}^\star)^{-1}\widetilde{\boldsymbol{U}}^{\star\top}\big\|_{\mathrm{F}}^2$$
$$= \frac{(2+o(1))\,\sigma_{\max}^2 rd}{p}.$$

In addition, we know from (63b) and (45) that

$$\big\|\boldsymbol{W}\widetilde{\boldsymbol{U}}^{\star\top}\big\|_{\mathrm{F}}^2 \le \big\|\widetilde{\boldsymbol{U}}^\star\big\|^2\big\|\boldsymbol{W}\big\|_{\mathrm{F}}^2 \lesssim \lambda_{\max}^{\star 4/3}\cdot d\,\|\boldsymbol{W}\|_{2,\infty}^2$$
$$= o(1)\,\frac{\sigma_{\max}^2 d}{p},$$

which further implies that

$$\big|\big\langle\boldsymbol{Z}\widetilde{\boldsymbol{U}}^{\star\top}, \boldsymbol{W}\widetilde{\boldsymbol{U}}^{\star\top}\big\rangle\big| \le \big\|\boldsymbol{Z}\widetilde{\boldsymbol{U}}^{\star\top}\big\|_{\mathrm{F}}\big\|\boldsymbol{W}\widetilde{\boldsymbol{U}}^{\star\top}\big\|_{\mathrm{F}}$$
$$= o(1)\,\frac{\sigma_{\max}^2 rd}{p}.$$

As a result, we find that

$$\Big\|\sum_{1\le l\le r}\boldsymbol{\Delta}_l\otimes\boldsymbol{u}_l^{\star\otimes 2}\Big\|_{\mathrm{F}}^2 = \frac{(2+o(1))\,\sigma_{\max}^2 rd}{p}. \tag{156}$$

2) Next, let us look at β_2. We denote by $\widetilde{\boldsymbol{\Delta}} \coloneqq \big[\boldsymbol{\Delta}_l^{\otimes 2}\big]_{1\le l\le r} \in \mathbb{R}^{d^2\times r}$, whose Frobenius norm can be bounded by

$$\|\widetilde{\boldsymbol{\Delta}}\|_{\mathrm{F}}^2 = \sum_{1\le l\le r}\big\|\boldsymbol{\Delta}_l^{\otimes 2}\big\|_2^2 = \sum_{1\le l\le r}\|\boldsymbol{\Delta}_l\|_2^4$$
$$\le \max_{1\le l\le r}\big\|\boldsymbol{\Delta}_l\big\|_2^2\,\|\boldsymbol{\Delta}\|_{\mathrm{F}}^2$$

$$\leq \|\boldsymbol{U}-\boldsymbol{U}^\star\|_{2,\infty}^2 \|\boldsymbol{U}-\boldsymbol{U}^\star\|_{\mathrm{F}}^2 . \tag{157}$$

Consequently, we use (59a) and (59b) to obtain

$$\begin{aligned}\beta_2 &= \left\|\boldsymbol{U}^\star \widetilde{\boldsymbol{\Delta}}^\top\right\|_{\mathrm{F}}^2 \leq \|\boldsymbol{U}^\star\|^2 \left\|\widetilde{\boldsymbol{\Delta}}\right\|_{\mathrm{F}}^2 \\ &\lesssim \|\boldsymbol{U}^\star\|^2 \|\boldsymbol{U}-\boldsymbol{U}^\star\|_{2,\infty}^2 \|\boldsymbol{U}-\boldsymbol{U}^\star\|_{\mathrm{F}}^2 \\ &\lesssim \lambda_{\max}^{\star 2/3} \cdot \frac{\sigma_{\max}^2}{\lambda_{\min}^{\star 2}} \frac{\mu r \log d}{p} \lambda_{\max}^{\star 2/3} \cdot \frac{\sigma_{\max}^2}{\lambda_{\min}^{\star 2}} \frac{r d \log d}{p} \lambda_{\max}^{\star 2/3} \\ &= o(1) \frac{\sigma_{\max}^2 r d}{p},\end{aligned} \tag{158}$$

where the last step holds as long $\sigma_{\max}/\lambda_{\min}^\star \ll \sqrt{p/(\mu r \log^2 d)}$ and $\kappa \asymp 1$.

3) In a similar way, we can use (59) and (157) to upper bound β_3 as follows

$$\begin{aligned}\beta_3 &= \left\|\boldsymbol{\Delta} \widetilde{\boldsymbol{\Delta}}^\top\right\|_{\mathrm{F}}^2 \leq \|\boldsymbol{\Delta}\|^2 \|\widetilde{\boldsymbol{\Delta}}\|_{\mathrm{F}}^2 \\ &\lesssim \|\boldsymbol{U}-\boldsymbol{U}^\star\|^2 \|\boldsymbol{U}-\boldsymbol{U}^\star\|_{2,\infty}^2 \|\boldsymbol{U}-\boldsymbol{U}^\star\|_{\mathrm{F}}^2 \\ &\lesssim \frac{\sigma_{\max}^2}{\lambda_{\min}^{\star 2}} \frac{r d \log d}{p} \lambda_{\max}^{\star 2/3} \cdot \frac{\sigma_{\max}^2}{\lambda_{\min}^{\star 2}} \frac{\mu r \log d}{p} \lambda_{\max}^{\star 2/3} \\ &\quad \cdot \frac{\sigma_{\max}^2}{\lambda_{\min}^{\star 2}} \frac{r d \log d}{p} \lambda_{\max}^{\star 2/3} \\ &= o(1) \frac{\sigma_{\max}^2 r d}{p},\end{aligned} \tag{159}$$

where the last step follows from the conditions that $\sigma_{\max}/\lambda_{\min}^\star \ll \sqrt{p/(\mu r d \log^2 d)}$ and $\kappa \asymp 1$.

4) As for β_4, one can apply the triangle inequality and Cauchy-Schwartz to upper bound

$$\begin{aligned}|\beta_4| &= \Big| \sum_{1\leq l\leq r} \langle \boldsymbol{\Delta}_l, \boldsymbol{u}_l^\star\rangle^2 \|\boldsymbol{u}_l^\star\|_2^2 \Big| \\ &\quad + \Big| \sum_{1\leq l\neq s\leq r} \langle \boldsymbol{\Delta}_l, \boldsymbol{u}_s^\star\rangle \langle \boldsymbol{u}_l^\star, \boldsymbol{\Delta}_s\rangle \langle \boldsymbol{u}_l^\star, \boldsymbol{u}_s^\star\rangle \Big| \\ &\lesssim \sum_{1\leq l\leq r} \langle \boldsymbol{\Delta}_l, \boldsymbol{u}_l^\star\rangle^2 \|\boldsymbol{u}_l^\star\|_2^2 \\ &\quad + \max_{1\leq l\neq s\leq r} |\langle \boldsymbol{u}_l^\star, \boldsymbol{u}_s^\star\rangle| \Big(\sum_{1\leq l\leq r} \|\boldsymbol{\Delta}_l\|_2 \|\boldsymbol{u}_l^\star\|_2 \Big)^2 \\ &\leq \max_{1\leq l\leq r} \langle \boldsymbol{\Delta}_l, \boldsymbol{u}_l^\star\rangle^2 \|\boldsymbol{U}^\star\|_{\mathrm{F}}^2 \\ &\quad + \max_{1\leq l\neq s\leq r} |\langle \boldsymbol{u}_l^\star, \boldsymbol{u}_s^\star\rangle| \|\boldsymbol{U}-\boldsymbol{U}^\star\|_{\mathrm{F}}^2 \|\boldsymbol{U}^\star\|_{\mathrm{F}}^2 .\end{aligned}$$

We then use the incoherence condition (12c), Lemma 16 in Appendix C-C, (62a) and (59a) to find that

$$\begin{aligned}|\beta_4| &= o(1) \frac{\sigma^2}{\lambda_{\min}^{\star 2/3}} \frac{d}{p} \cdot r \lambda_{\max}^{\star 2/3} \\ &\quad + \sqrt{\frac{\mu}{d}} \lambda_{\max}^{\star 2/3} \cdot \frac{\sigma_{\max}^2}{\lambda_{\min}^{\star 2}} \frac{r d \log d}{p} \lambda_{\max}^{\star 2/3} \cdot r \lambda_{\max}^{\star 2/3} \\ &= o(1) \frac{\sigma_{\max}^2 r d}{p},\end{aligned} \tag{160}$$

where we use the assumption that $r = o\left(\sqrt{d/(r \log^2 d)}\right)$ and $\kappa \asymp 1$.

5) It remains to bound β_5. Given the Cauchy-Schwartz inequality $|\langle \boldsymbol{A}, \boldsymbol{B}\rangle| \leq \|\boldsymbol{A}\|_{\mathrm{F}} \|\boldsymbol{B}\|_{\mathrm{F}}$, it immediately follows from (156), (158) and (159) that

$$\begin{aligned}|\beta_5| &\lesssim \Big\| \sum_{1\leq l\leq r} \boldsymbol{u}_l^\star \otimes \boldsymbol{\Delta}_l^{\otimes 2} \Big\|_{\mathrm{F}}^2 \\ &\quad + \Big\| \sum_{1\leq l\leq r} \boldsymbol{u}_l^{\star\otimes 2} \otimes \boldsymbol{\Delta}_l \Big\|_{\mathrm{F}} \Big\| \sum_{1\leq l\leq r} \boldsymbol{\Delta}_l^{\otimes 3} \Big\|_{\mathrm{F}} \\ &\quad + \Big\| \sum_{1\leq l\leq r} \boldsymbol{u}_l^\star \otimes \boldsymbol{\Delta}_l^{\otimes 2} \Big\|_{\mathrm{F}} \Big\| \sum_{1\leq l\leq r} \boldsymbol{\Delta}_l^{\otimes 3} \Big\|_{\mathrm{F}} \\ &\quad + \Big\| \sum_{1\leq l\leq r} \boldsymbol{u}_l^\star \otimes \boldsymbol{\Delta}_l^{\otimes 2} \Big\|_{\mathrm{F}} \Big\| \sum_{1\leq l\leq r} \boldsymbol{u}_l^{\star\otimes 2} \otimes \boldsymbol{\Delta}_l \Big\|_{\mathrm{F}} \\ &= o(1) \frac{\sigma_{\max}^2 r d}{p}.\end{aligned}$$

This taken collectively with (160) finishes the proof.

Appendix G
Proof of Lower Bounds (Theorems 12 and 14)

In this section, we establish the lower bounds claimed in Theorems 12 and 14 (which subsume Theorems 5 and 7 as special cases, respectively). Recall the assumption that $\{E_{i,j,k}\}$ are independent Gaussians. For the sake of notational simplicity, we shall assume throughout this proof that $\sigma_{i,j,k}^2 \equiv \sigma_{\min}^2$ for all $1 \leq i, j, k \leq d$.

Given that the noise components $\{E_{i,j,k} \mid (i,j,k) \in \Omega, 1 \leq i \leq j \leq k \leq d\}$ are assumed to be independent Gaussian, the probability density function (conditional on Ω) can be computed as

$$f(\boldsymbol{T}^{\mathsf{obs}}) = c \prod_{\substack{1\leq i\leq j\leq k\leq d \\ (i,j,k)\in\Omega}} \exp\left(-\frac{\left(T_{i,j,k}^{\mathsf{obs}} - \sum_{l=1}^r u_{l,i}^\star u_{l,j}^\star u_{l,k}^\star\right)^2}{2\sigma_{\min}^2} \right)$$

for some normalization constant $c > 0$. Here, we abuse the notation $f(\cdot)$ to represent the probability density function whenever it is clear from the context. Denote by $\mathsf{vec}(\boldsymbol{U}^\star)$ the vectorization of $\boldsymbol{U}^\star = [\boldsymbol{u}_1^\star, \cdots, \boldsymbol{u}_r^\star]$, namely,

$$\mathsf{vec}(\boldsymbol{U}^\star) := \begin{bmatrix} \boldsymbol{u}_1^\star \\ \vdots \\ \boldsymbol{u}_r^\star \end{bmatrix} \in \mathbb{R}^{dr}.$$

By virtue of the Cramér-Rao lower bound, any unbiased estimator $\widehat{\boldsymbol{U}}$ for $\boldsymbol{U}^\star$ necessarily obeys

$$\mathsf{Cov}\big[\mathsf{vec}(\widehat{\boldsymbol{U}})\big] \succeq (\boldsymbol{\mathcal{I}}_\Omega)^{-1},$$

where $\boldsymbol{\mathcal{I}}_\Omega \in \mathbb{R}^{dr\times dr}$ denotes the corresponding Fisher information matrix (conditional on Ω) as follows

$$\begin{aligned}\boldsymbol{\mathcal{I}}_\Omega &:= \boldsymbol{\mathcal{I}}_\Omega(\boldsymbol{U}^\star) \\ &= \mathbb{E}\Big[\nabla_{\mathsf{vec}(\boldsymbol{U}^\star)} \log f(\boldsymbol{T}^{\mathsf{obs}}) \big(\nabla_{\mathsf{vec}(\boldsymbol{U}^\star)} \log f(\boldsymbol{T}^{\mathsf{obs}})\big)^\top\Big] \\ &= \Big[\mathbb{E}\Big[\nabla_{\boldsymbol{u}_{l_1}^\star} \log f(\boldsymbol{T}^{\mathsf{obs}}) \big(\nabla_{\boldsymbol{u}_{l_2}^\star} \log f(\boldsymbol{T}^{\mathsf{obs}})\big)^\top\Big]\Big]_{1\leq l_1, l_2\leq r} .\end{aligned} \tag{161}$$

It then suffices to compute the Fisher information matrix. Towards this end, we start by observing that

$$\frac{\partial T^\star_{i,j,k}}{\partial u^\star_{l,s}} = \sum_{\tau=1}^{r} \frac{\partial u^\star_{\tau,i} u^\star_{\tau,j} u^\star_{\tau,k}}{\partial u^\star_{l,s}} = u^\star_{l,j} u^\star_{l,k} \mathbb{1}_{\{i=s\}} + u^\star_{l,i} u^\star_{l,k} \mathbb{1}_{\{j=s\}} + u^\star_{l,i} u^\star_{l,j} \mathbb{1}_{\{k=s\}}, \tag{162}$$

and

$$\frac{\partial \log f(\boldsymbol{T}^{\mathsf{obs}})}{\partial u^\star_{l,s}} = \sum_{(i,j,k)\in\Omega,\, i\le j\le k} \bigg(\frac{E_{i,j,k}}{\sigma^2_{\min}} \big(u^\star_{l,j} u^\star_{l,k} \mathbb{1}_{\{i=s\}} + u^\star_{l,i} u^\star_{l,k} \mathbb{1}_{\{j=s\}} + u^\star_{l,i} u^\star_{l,j} \mathbb{1}_{\{k=s\}} \big) \bigg) \tag{163}$$

for any $1 \le l \le r$ and $1 \le s \le d$. In addition, let us define a collection of vectors $\{\boldsymbol{h}_{i,j,k}\}_{1\le i\le j\le k\le d}$ in $\mathbb{R}^{dr}$ with entries

$$h_{i,j,k}(l,s) := h_{i,j,k}\big((l-1)\times r + s\big) := u^\star_{l,j} u^\star_{l,k} \mathbb{1}_{\{i=s\}} + u^\star_{l,i} u^\star_{l,k} \mathbb{1}_{\{j=s\}} + u^\star_{l,i} u^\star_{l,j} \mathbb{1}_{\{k=s\}} \tag{164}$$

for any $1 \le l \le r$ and $1 \le s \le d$. One can then express

$$\boldsymbol{\mathcal{I}}_\Omega = \frac{1}{\sigma^4_{\min}} \sum_{(i,j,k)\in\Omega,\, i\le j\le k} \mathbb{E}\left[E^2_{i,j,k}\right] \boldsymbol{h}_{i,j,k} \boldsymbol{h}^\top_{i,j,k} = \frac{1}{\sigma^2_{\min}} \sum_{1\le i\le j\le k\le d} \chi_{i,j,k} \boldsymbol{h}_{i,j,k} \boldsymbol{h}^\top_{i,j,k}, \tag{165}$$

where we recall the notation $\chi_{i,j,k} := \mathbb{1}_{\{(i,j,k)\in\Omega\}}$. Let us further define

$$\boldsymbol{\mathcal{I}} := \mathbb{E}_\Omega\big[\boldsymbol{\mathcal{I}}_\Omega\big] = \frac{p}{\sigma^2_{\min}} \sum_{1\le i\le j\le k\le d} \boldsymbol{h}_{i,j,k} \boldsymbol{h}^\top_{i,j,k} \tag{166}$$

where the expectation is taken over randomness of $\{\chi_{i,j,k}\}_{1\le i,j,k\le d}$. In what follows, we shall compute the spectrum of $\boldsymbol{\mathcal{I}}$, and show that $\boldsymbol{\mathcal{I}}_\Omega$ is, with high probability, sufficiently close to $\boldsymbol{\mathcal{I}}$ in the spectral norm. In addition, denote $(l,s) := (l-1)\times r + d$ for all $1 \le l \le r, 1 \le s \le d$.

a) Spectrum of $\boldsymbol{\mathcal{I}}$: First, it is straightforward to calculate that for any $1 \le s \le d$ and any $1 \le l_1, l_2 \le d$,

$$\begin{aligned}\frac{\sigma^2_{\min}}{p} \mathcal{I}_{(l_1,s),(l_2,s)} &= \sum_{(i,j):i\le j} u^\star_{l_1,i} u^\star_{l_2,i} u^\star_{l_1,j} u^\star_{l_2,j} + 3\, u^\star_{l_1,s} u^\star_{l_2,s} \sum_{1\le i\le d} u^\star_{l_1,i} u^\star_{l_2,i} + 5\, u^{\star 2}_{l_1,s} u^{\star 2}_{l_2,s} \\ &= \frac{1}{2} \langle \boldsymbol{u}^\star_{l_1}, \boldsymbol{u}^\star_{l_2} \rangle^2 + \frac{1}{2} \sum_{1\le i\le d} u^{\star 2}_{l_1,i} u^{\star 2}_{l_2,i} + 3\, u^\star_{l_1,s} u^\star_{l_2,s} \langle \boldsymbol{u}^\star_{l_1}, \boldsymbol{u}^\star_{l_2} \rangle + 5\, u^{\star 2}_{l_1,s} u^{\star 2}_{l_2,s},\end{aligned}$$

and for any $1 \le s_1 \ne s_2 \le d$, $1 \le l_1, l_2 \le d$,

$$\frac{\sigma^2_{\min}}{p} \mathcal{I}_{(l_1,s_1),(l_2,s_2)} = u^\star_{l_1,s_2} u^\star_{l_2,s_1} \langle \boldsymbol{u}^\star_{l_1}, \boldsymbol{u}^\star_{l_2} \rangle.$$

As a consequence, one can express

$$\frac{\sigma^2_{\min}}{p} \boldsymbol{\mathcal{I}} = \boldsymbol{J}_1 + \boldsymbol{\Psi} + \underbrace{\begin{bmatrix} \boldsymbol{0} & \cdots & \langle \boldsymbol{u}^\star_r, \boldsymbol{u}^\star_1 \rangle \boldsymbol{u}^\star_r \boldsymbol{u}^{\star\top}_1 \\ \vdots & \ddots & \vdots \\ \langle \boldsymbol{u}^\star_1, \boldsymbol{u}^\star_r \rangle \boldsymbol{u}^\star_1 \boldsymbol{u}^{\star\top}_r & \cdots & \boldsymbol{0} \end{bmatrix}}_{=:\boldsymbol{\Phi}} \tag{167}$$

where $\boldsymbol{J}_1$ is a block diagonal matrix in $\mathbb{R}^{dr\times dr}$ whose i-th block is equal to

$$\frac{1}{2} \left\|\boldsymbol{u}^\star_i\right\|^4_2 \boldsymbol{I}_d + \left\|\boldsymbol{u}^\star_i\right\|^2_2 \boldsymbol{u}^\star_i \boldsymbol{u}^{\star\top}_i, \quad 1 \le i \le r; \tag{168}$$

$\boldsymbol{\Phi}$ is a matrix in $\mathbb{R}^{dr\times dr}$ with all-zero diagonal blocks and (l_1,l_2)-th block equal to $\langle \boldsymbol{u}^\star_{l_2}, \boldsymbol{u}^\star_{l_1} \rangle \boldsymbol{u}^\star_{l_2} \boldsymbol{u}^{\star\top}_{l_1}$; $\boldsymbol{\Psi}$ is a matrix in $\mathbb{R}^{dr\times dr}$ with entries as follows: if $l_1 = l_2 = l, s_1 = s_2 = s$,

$$\Psi_{(l_1,s_1),(l_2,s_2)} = \frac{1}{2} \sum_{1\le i\le d} u^{\star 4}_{l,i} + 2\, u^{\star 2}_{l,s} \|\boldsymbol{u}^\star_l\|^2_2 + 5 u^{\star 4}_{l,s};$$

if $l_1 \ne l_2, s_1 = s_2 = s$,

$$\Psi_{(l_1,s),(l_2,s)} = \frac{1}{2} \langle \boldsymbol{u}^\star_{l_1}, \boldsymbol{u}^\star_{l_2} \rangle^2 + \frac{1}{2} \sum_{1\le i\le d} u^{\star 2}_{l_1,i} u^{\star 2}_{l_2,i} + 2\, u^\star_{l_1,s} u^\star_{l_2,s} \langle \boldsymbol{u}^\star_{l_1}, \boldsymbol{u}^\star_{l_2} \rangle + 5\, u^{\star 2}_{l_1,s} u^{\star 2}_{l_2,s};$$

otherwise,

$$\Psi_{(l_1,s_1),(l_2,s_2)} = 0.$$

In the sequel, we shall control the spectral norm of $\boldsymbol{\Phi}$ and $\boldsymbol{\Psi}$ separately.

- For $\boldsymbol{\Phi}$, we can bound

$$\begin{aligned}\|\boldsymbol{\Phi}\| \le \|\boldsymbol{\Phi}\|_{\mathrm{F}} &\le \sqrt{\sum_{1\le l_1\ne l_2\le r} \langle \boldsymbol{u}^\star_{l_1}, \boldsymbol{u}^\star_{l_2} \rangle^2 \left\| \boldsymbol{u}^\star_{l_2} \boldsymbol{u}^{\star\top}_{l_1} \right\|^2_{\mathrm{F}}} \\ &\le \sqrt{\max_{1\le l_1\ne l_2\le r} \langle \boldsymbol{u}^\star_{l_1}, \boldsymbol{u}^\star_{l_2} \rangle^2 \sum_{1\le l_1,l_2\le r} \|\boldsymbol{u}^\star_{l_1}\|^2_2 \|\boldsymbol{u}^\star_{l_2}\|^2_2} \\ &\le r \sqrt{\frac{\mu}{d}} \max_{1\le l\le r} \|\boldsymbol{u}^\star_l\|^4_2 = o\Big(\min_{1\le l\le r} \|\boldsymbol{u}^\star_l\|^4_2 \Big),\end{aligned}$$

  where we have used the incoherence condition (12c) and the assumptions $r = o\big(\sqrt{d/\mu}\big)$ and $\kappa \asymp 1$.
- As for $\boldsymbol{\Psi}$, we note that each block of $\boldsymbol{\Psi}$ is a diagonal matrix. By the incoherence conditions, its entries can be bounded by

$$\big|\Psi_{(l,s),(l,s)}\big| \lesssim \|\boldsymbol{u}^\star_l\|^2_\infty \|\boldsymbol{u}^\star_l\|^2_2 \le \frac{\mu}{d} \max_{1\le l\le r} \|\boldsymbol{u}^\star_l\|^4_2,$$

  and

$$\begin{aligned}\big|\Psi_{(l_1,s),(l_2,s)}\big| &\lesssim \langle \boldsymbol{u}^\star_{l_1}, \boldsymbol{u}^\star_{l_2} \rangle^2 + \|\boldsymbol{u}^\star_{l_1}\|^2_\infty \|\boldsymbol{u}^\star_{l_2}\|^2_2 + 2\, \|\boldsymbol{u}^\star_{l_1}\|_\infty \|\boldsymbol{u}^\star_{l_2}\|_\infty \|\boldsymbol{u}^\star_{l_1}\|_2 \|\boldsymbol{u}^\star_{l_2}\|_2 \\ &\lesssim \frac{\mu}{d} \max_{1\le l\le r} \|\boldsymbol{u}^\star_l\|^4_2,\end{aligned}$$

  thus indicating that

$$\|\boldsymbol{\Psi}\|_\infty \lesssim \frac{\mu}{d} \max_{1\le l\le r} \|\boldsymbol{u}^\star_l\|^4_2.$$

  Given the special structure of $\boldsymbol{\Psi}$, one can easily permutation its columns and rows to arrive at another matrix $\widetilde{\boldsymbol{\Psi}} = [\widetilde{\boldsymbol{\Psi}}_{i,j}]_{1\le i,j\le r}$ such that (1) $\widetilde{\boldsymbol{\Psi}}$ is a block diagonal matrix; (2) $\widetilde{\boldsymbol{\Psi}}$ contains $d \times d$ blocks each of size $r \times r$; (3) each diagonal block $\widetilde{\boldsymbol{\Psi}}_{i,i}$ of $\widetilde{\boldsymbol{\Psi}}$ has spectral norm at most

$$\|\widetilde{\boldsymbol{\Psi}}_{i,i}\| \le \|\widetilde{\boldsymbol{\Psi}}_{i,i}\|_{\mathrm{F}} \le r\|\boldsymbol{\Psi}\|_{\infty} \lesssim \frac{\mu r}{d} \max_{1\le l\le r} \|\boldsymbol{u}_l^\star\|_2^4.$$

Consequently, one can derive

$$\begin{aligned}\|\boldsymbol{\Psi}\| = \|\widetilde{\boldsymbol{\Psi}}\| &\le \max_{1\le i\le r} \|\boldsymbol{\Psi}_{i,i}\| \lesssim \frac{\mu r}{d} \max_{1\le l\le r} \|\boldsymbol{u}_l^\star\|_2^4 \\ &= o\left(\min_{1\le l\le r} \|\boldsymbol{u}_l^\star\|_2^4\right),\end{aligned}$$

provided that $r = o(d/\mu)$ and $\kappa \asymp 1$.

- Putting the above two estimates together, we conclude that

$$\begin{aligned}\boldsymbol{\mathcal{I}} &\preceq \frac{p}{\sigma_{\min}^2}\boldsymbol{J}_1 + \frac{p}{\sigma_{\min}^2}\left(\|\boldsymbol{\Phi}\| + \|\boldsymbol{\Psi}\|\right)\boldsymbol{I}_{dr} \\ &= \big(1 + o\left(1\right)\big)\frac{p}{\sigma_{\min}^2}\boldsymbol{J}_1. \end{aligned} \tag{169}$$

b) Controlling $\|\boldsymbol{\mathcal{I}}_\Omega - \boldsymbol{\mathcal{I}}\|$: By construction, $\boldsymbol{\mathcal{I}}_\Omega - \boldsymbol{\mathcal{I}} = \frac{1}{\sigma_{\min}^2}\sum_{i\le j\le k}(\chi_{i,j,k} - p)\boldsymbol{h}_{i,j,k}\boldsymbol{h}_{i,j,k}^\top$ is a sum of independent zero-mean random matrix in $\mathbb{R}^{dr\times dr}$. By the incoherence conditions, it is straightforward to bound

$$\begin{aligned}B &:= \max_{1\le i\le j\le k\le d} \left\|(\chi_{i,j,k} - p)\boldsymbol{h}_{i,j,k}\boldsymbol{h}_{i,j,k}^\top\right\| \\ &\le \max_{1\le i\le j\le k\le d} \|\boldsymbol{h}_{i,j,k}\|_2^2 \\ &\overset{\text{(i)}}{\lesssim} r \max_{1\le l\le r} \|\boldsymbol{u}_l^\star\|_\infty^4 \le \frac{\mu^2 r}{d^2} \max_{1\le l\le r} \|\boldsymbol{u}_l^\star\|_2^4,\end{aligned}$$

where (i) arises from the definition of $\boldsymbol{h}_{i,j,k}$ in (164). In addition, we also have

$$\begin{aligned}V &:= \Big\| \sum_{i\le j\le k} \mathbb{E}\big[(\chi_{i,j,k} - p)^2\big] \|\boldsymbol{h}_{i,j,k}\|_2^2 \boldsymbol{h}_{i,j,k}\boldsymbol{h}_{i,j,k}^\top \Big\| \\ &\le p \max_{1\le i\le j\le k\le d} \|\boldsymbol{h}_{i,j,k}\|_2^2 \Big\| \sum_{1\le i\le j\le k\le d} \boldsymbol{h}_{i,j,k}\boldsymbol{h}_{i,j,k}^\top \Big\| \\ &\lesssim pr \max_{1\le l\le r} \|\boldsymbol{u}_l^\star\|_\infty^4 \max_{1\le l\le r} \|\boldsymbol{u}_l^\star\|_2^4 \\ &\lesssim \frac{\mu^2 rp}{d^2} \max_{1\le l\le r} \|\boldsymbol{u}_l^\star\|_2^8.\end{aligned}$$

Here, the second inequality arises from (164) and (169), whereas the third comes from our bound above for $\boldsymbol{\mathcal{I}}$. Invoking the matrix Bernstein inequality, we conclude that with probability at least $1 - O(d^{-10})$,

$$\begin{aligned}\|\boldsymbol{\mathcal{I}}_\Omega - \boldsymbol{\mathcal{I}}\| &\lesssim \frac{B\log d + \sqrt{V\log d}}{\sigma_{\min}^2} \\ &\lesssim \frac{p}{\sigma_{\min}^2} \max_{1\le l\le r} \|\boldsymbol{u}_l^\star\|_2^4 \left\{ \frac{\mu^2 r\log d}{d^2 p} + \frac{\mu\sqrt{r\log d}}{d\sqrt{p}} \right\} \\ &= o\left(1\right) \frac{p}{\sigma_{\min}^2} \min_{1\le l\le r} \|\boldsymbol{u}_l^\star\|_2^4, \end{aligned} \tag{170}$$

where the last step holds as long as $p \gg \mu^2 r d^{-2}\log^2 d$ and $\kappa \asymp 1$.

c) Combining the spectrum of $\boldsymbol{\mathcal{I}}_\Omega$ and the bound on $\|\boldsymbol{\mathcal{I}}_\Omega - \boldsymbol{\mathcal{I}}\|$: Combining (169) and (170) with Weyl's inequality reveals that with probability exceeding $1 - O(d^{-10})$,

$$\boldsymbol{\mathcal{I}}_\Omega \preceq \big(1 + o\left(1\right)\big)\frac{p}{\sigma_{\min}^2}\boldsymbol{J}_1,$$

where we recall the definition of $\boldsymbol{J}_1$ in (167) and (168). By the Woodbury matrix identity, it is straightforward to check

$$\left(\frac{1}{2}\|\boldsymbol{u}_i^\star\|_2^4\boldsymbol{I}_d + \|\boldsymbol{u}_i^\star\|_2^2\boldsymbol{u}_i^\star\boldsymbol{u}_i^{\star\top}\right)^{-1} = \frac{2}{\|\boldsymbol{u}_i^\star\|_2^4}\left(\boldsymbol{I}_d - \frac{2}{3}\frac{\boldsymbol{u}_i^\star\boldsymbol{u}_i^{\star\top}}{\|\boldsymbol{u}_i^\star\|_2^2}\right).$$

Hence, for any unbiased estimator $\widehat{\boldsymbol{U}}$ of $\boldsymbol{U}^\star$ we have

$$\mathsf{Cov}\big[\mathsf{vec}(\widehat{\boldsymbol{U}})\big] \succeq (\boldsymbol{\mathcal{I}}_\Omega)^{-1} \succeq \big(1 - o\left(1\right)\big)\frac{\sigma_{\min}^2}{p}\boldsymbol{J}_2. \tag{171}$$

where $\boldsymbol{J}_2$ is a block matrix in $\mathbb{R}^{dr\times dr}$ whose i-th block equals to

$$\frac{2}{\|\boldsymbol{u}_i^\star\|_2^4}\left(\boldsymbol{I}_d - \frac{2}{3}\frac{\boldsymbol{u}_i^\star\boldsymbol{u}_i^{\star\top}}{\|\boldsymbol{u}_i^\star\|_2^2}\right), \quad 1\le i\le r. \tag{172}$$

A few consequences from the above Cramér-Rao lower bound are in order.

- For each unbiased estimator $\widehat{\boldsymbol{u}}_l$ of $\boldsymbol{u}_l^\star$, one necessarily has

$$\begin{aligned}\mathbb{E}\big[(\widehat{u}_{l,k} - u_{l,k}^\star)^2\big] &\ge \big((\boldsymbol{\mathcal{I}}_\Omega)^{-1}\big)_{(l,k),(l,k)} \\ &\ge \big(1 - o\left(1\right)\big)\frac{2\sigma_{\min}^2}{p\,\|\boldsymbol{u}_l^\star\|_2^4}\left(1 - \frac{2}{3}\frac{u_{l,k}^{\star 2}}{\|\boldsymbol{u}_l^\star\|_2^2}\right) \\ &\ge \big(1 - o\left(1\right)\big)\frac{2\sigma_{\min}^2}{p\,\|\boldsymbol{u}_l^\star\|_2^4}\left(1 - \frac{2}{3}\frac{\|\boldsymbol{u}_l^\star\|_\infty^2}{\|\boldsymbol{u}_l^\star\|_2^2}\right) \\ &\overset{\text{(i)}}{\ge} \big(1 - o\left(1\right)\big)\frac{2\sigma_{\min}^2}{p\,\|\boldsymbol{u}_l^\star\|_2^4}\left(1 - \frac{2}{3}\frac{\mu}{d}\right) \\ &\overset{\text{(ii)}}{\ge} \big(1 - o\left(1\right)\big)\frac{2\sigma_{\min}^2}{p\,\|\boldsymbol{u}_l^\star\|_2^4} \\ &= \big(1 - o\left(1\right)\big)(\boldsymbol{\Sigma}_k^\star)_{l,l},\end{aligned}$$

  where (i) arises from the incoherence definition 1 and (ii) holds as long as $\mu = o(d)$. This further implies that

$$\begin{aligned}\mathbb{E}\big[\|\widehat{\boldsymbol{u}}_l - \boldsymbol{u}_l^\star\|_2^2\big] &= \sum_{k=1}^d \mathbb{E}\big[(\widehat{u}_{l,k} - u_{l,k}^\star)^2\big] \\ &\ge \big(1 - o\left(1\right)\big)\frac{2\sigma_{\min}^2 d}{p\,\|\boldsymbol{u}_l^\star\|_2^4}.\end{aligned}$$

- Any unbiased estimator $\widehat{T}_{i,j,k}$ of $T_{i,j,k}^\star$ necessarily obeys [96]

$$\begin{aligned}&\mathbb{E}\Big[\big(\widehat{T}_{i,j,k} - T_{i,j,k}^\star\big)^2\Big] \\ &\ge \left[\frac{\partial T_{i,j,k}^\star}{\partial \mathsf{vec}(\boldsymbol{U}^\star)}\right]^\top (\boldsymbol{\mathcal{I}}_\Omega)^{-1}\frac{\partial T_{i,j,k}^\star}{\partial \mathsf{vec}(\boldsymbol{U}^\star)} \\ &\overset{\text{(i)}}{\ge} \big(1 - o\left(1\right)\big)\sum_{1\le s\le d}\sum_{1\le l\le r}\frac{2\sigma_{\min}^2}{p\,\|\boldsymbol{u}_l^\star\|_2^4}\left(\frac{\partial T_{i,j,k}^\star}{\partial u_{l,s}^\star}\right)^2 \\ &\overset{\text{(ii)}}{=} \big(1 - o\left(1\right)\big)v_{i,j,k}^\star,\end{aligned}$$

where (i) uses (171), and (ii) follows from (162), (23) and direct algebraic manipulations.

- Any unbiased estimator $\widehat{\boldsymbol{T}}$ of $\boldsymbol{T}^\star$ necessarily satisfies

$$\begin{aligned}\mathbb{E}\big[\|\widehat{\boldsymbol{T}}-\boldsymbol{T}^\star\|_{\mathrm{F}}^2\big] &= \sum_{i,j,k}\mathbb{E}\Big[\big(\widehat{T}_{i,j,k}-T^\star_{i,j,k}\big)^2\Big]\\ &\geq \big(1-o\left(1\right)\big)\sum_{i,j,k} v^\star_{i,j,k}\\ &\overset{\text{(i)}}{\geq} \big(1-o\left(1\right)\big)\sum_{1\le i\le d}\ \sum_{1\le j,k\le d}\ \sum_{1\le l\le r}\frac{6\,\sigma_{\min}^2}{p\,\|\boldsymbol{u}_l^\star\|_2^4}\,u_{l,j}^{\star 2}u_{l,k}^{\star 2}\\ &= \big(1-o\left(1\right)\big)\frac{6\,\sigma_{\min}^2 dr}{p},\end{aligned}$$

where (i) arises from the definition of $v^\star_{i,j,k}$ in (23).

Appendix H
Proof of Auxiliary Lemmas: Preliminary Facts

A. *Proof of Lemma 13*

1) To begin with, by the incoherence condition (12b), it is easy to derive

$$\|\boldsymbol{U}^\star\|_{\mathrm{F}} \leq \sqrt{r}\max_{1\le l\le r}\|\boldsymbol{u}_l^\star\|_2 \leq \sqrt{r}\,\lambda_{\max}^{\star 1/3},$$
$$\|\boldsymbol{U}^\star\|_{2,\infty} \leq \sqrt{r}\max_{1\le l\le r}\|\boldsymbol{u}_l^\star\|_\infty \leq \sqrt{\frac{\mu r}{d}}\,\lambda_{\max}^{\star 1/3}.$$

2) Regarding the properties about the spectrum of $\boldsymbol{U}^\star$, we refer the reader to the proof of [2, Lemma D.1].

3) From Lemma 11, it is straightforward to show that: there exists a permutation $\pi(\cdot):[d]\mapsto[d]$ such that

$$\begin{aligned}\max_{1\le i\le r}\big\|\boldsymbol{u}_{\pi(i)}-\boldsymbol{u}_i^\star\big\|_2 &\leq \|\boldsymbol{U\Pi}-\boldsymbol{U}^\star\|_{\mathrm{F}}\\ &\lesssim \frac{\sigma}{\lambda_{\min}^\star}\sqrt{\frac{rd\log d}{p}}\,\lambda_{\max}^{\star 1/3} = o\left(\lambda_{\min}^{\star 1/3}\right),\\ \max_{1\le i\le r}\big\|\boldsymbol{u}_{\pi(i)}-\boldsymbol{u}_i^\star\big\|_\infty &\leq \|\boldsymbol{U\Pi}-\boldsymbol{U}^\star\|_{2,\infty}\\ &\lesssim \frac{\sigma}{\lambda_{\min}^\star}\sqrt{\frac{\mu r\log d}{p}}\,\lambda_{\max}^{\star 1/3} = o\left(\frac{\lambda_{\min}^{\star 1/3}}{\sqrt{d}}\right),\end{aligned}$$

where we have used the conditions that $\sigma_{\max}/\lambda_{\min}^\star \ll \sqrt{p/(\mu r d^{3/2}\log d)}$ and $\kappa\asymp 1$. Recognizing that $\lambda_{\min}^{\star 1/3} \leq \|\boldsymbol{u}_i^\star\|_2 \leq \lambda_{\max}^{\star 1/3}$ and that $\sqrt{1/d}\,\lambda_{\min}^{\star 1/3} \leq \|\boldsymbol{u}_i^\star\|_\infty \leq \sqrt{\mu/d}\,\lambda_{\max}^{\star 1/3}$ for all $1\le i\le r$, one immediately obtains (62c)-(62f) by invoking the triangle inequality.

4) For any $1\le i\neq j\le r$, applying the triangle inequality and the Cauchy-Schwartz inequality yields

$$\begin{aligned}|\langle\boldsymbol{u}_i,\boldsymbol{u}_j\rangle| &\leq \big|\langle\boldsymbol{u}_i^\star,\boldsymbol{u}_j^\star\rangle\big| + |\langle\boldsymbol{u}_i-\boldsymbol{u}_i^\star,\boldsymbol{u}_j\rangle|\\ &\quad + \big|\langle\boldsymbol{u}_i^\star,\boldsymbol{u}_j-\boldsymbol{u}_j^\star\rangle\big|\\ &\leq \big|\langle\boldsymbol{u}_i^\star,\boldsymbol{u}_j^\star\rangle\big| + \|\boldsymbol{u}_i-\boldsymbol{u}_i^\star\|_2\,\|\boldsymbol{u}_j\|_2\\ &\quad + \|\boldsymbol{u}_j-\boldsymbol{u}_j^\star\|_2\,\|\boldsymbol{u}_i^\star\|_2\\ &\lesssim \sqrt{\frac{\mu}{d}}\,\lambda_{\max}^{\star 2/3} + \frac{\sigma_{\max}}{\lambda_{\min}^\star}\sqrt{\frac{rd\log d}{p}}\,\lambda_{\max}^{\star 2/3}.\end{aligned}$$

5) Next, we move on to the spectrum of $\boldsymbol{U}$. In view of (59a) and the conditions that $\sigma_{\max}/\lambda_{\min}^\star \ll \sqrt{p/(rd^{3/2}\log d)}$ and $\kappa\asymp 1$, one can deduce that

$$\begin{aligned}\|\boldsymbol{U\Pi}-\boldsymbol{U}^\star\| \leq \|\boldsymbol{U\Pi}-\boldsymbol{U}^\star\|_{\mathrm{F}} &\lesssim \frac{\sigma_{\max}}{\lambda_{\min}^\star}\sqrt{\frac{rd\log d}{p}}\,\lambda_{\max}^{\star 1/3}\\ &= o\left(\lambda_{\min}^{\star 1/3}\right).\end{aligned}$$

Therefore, (62h) is an immediate consequence of Weyl's inequality and (62b).

6) Finally, we know from Lemma 12 that the estimation error bounds for $\|\boldsymbol{U\Pi}-\boldsymbol{U}^\star\|_{\mathrm{F}}$ and $\|\boldsymbol{U\Pi}-\boldsymbol{U}^\star\|_{2,\infty}$ continue to hold if we replace $\boldsymbol{U}$ with $\boldsymbol{U}^{(m)}$. Hence, the above results are also valid for $\boldsymbol{U}^{(m)}$ and $\boldsymbol{u}_l^{(m)}$ $(1\le l\le r)$.

B. *Proof of Lemma 14*

1) To begin with, it is straightforward to compute

$$\big\|\widetilde{\boldsymbol{U}}^\star\big\|_{\mathrm{F}}^2 = \sum_{1\le s\le r}\|\boldsymbol{u}_s^\star\otimes\boldsymbol{u}_s^\star\|_2^2 = \sum_{1\le s\le r}\|\boldsymbol{u}_s^\star\|_2^4 \leq r\lambda_{\max}^{\star 4/3}.$$

2) For any $1\le i,j\le d$, the incoherence condition (12b) yields

$$\big\|\widetilde{\boldsymbol{U}}^\star_{(i,j),:}\big\|_2^2 = \sum_{1\le s\le r}u_{s,i}^{\star 2}u_{s,j}^{\star 2} \leq \frac{\mu^2 r}{d^2}\lambda_{\max}^{\star 4/3}.$$

This leads to the claimed bound regarding $\|\widetilde{\boldsymbol{U}}^\star\|_{2,\infty}$.

3) Regarding the spectrum of $\widetilde{\boldsymbol{U}}^\star$ and $\widetilde{\boldsymbol{U}}$, we refer the reader to the proof of [2, Lemma 4.1 and Lemma D.1].

4) Next, we turn to $\|\widetilde{\boldsymbol{U}}\boldsymbol{\Pi}-\widetilde{\boldsymbol{U}}^\star\|_{\mathrm{F}}$. Without loss of generality, assume that $\boldsymbol{\Pi}=\boldsymbol{I}_r$. Using the fact that $\langle\boldsymbol{a}^{\otimes 2},\boldsymbol{b}^{\otimes 2}\rangle = \langle\boldsymbol{a},\boldsymbol{b}\rangle^2$ for any vectors $\boldsymbol{a},\boldsymbol{b}\in\mathbb{R}^d$, we can straightforwardly calculate that

$$\begin{aligned}&\big\|\boldsymbol{u}_s^{\otimes 2}-\boldsymbol{u}_s^{\star\otimes 2}\big\|_2^2\\ &\quad= \big\|\boldsymbol{u}_s^{\otimes 2}\big\|_2^2 + \big\|\boldsymbol{u}_s^{\star\otimes 2}\big\|_2^2 - 2\left\langle\boldsymbol{u}_s^{\otimes 2},\boldsymbol{u}_s^{\star\otimes 2}\right\rangle\\ &\quad= \|\boldsymbol{u}_s\|_2^4 + \|\boldsymbol{u}_s^\star\|_2^4 - 2\left\langle\boldsymbol{u}_s,\boldsymbol{u}_s^\star\right\rangle^2\\ &\quad= \|\boldsymbol{u}_s\|_2^4 + \|\boldsymbol{u}_s^\star\|_2^4 - \frac{1}{2}\big(\|\boldsymbol{u}_s\|_2^2 + \|\boldsymbol{u}_s^\star\|_2^2 - \|\boldsymbol{u}_s-\boldsymbol{u}_s^\star\|_2^2\big)^2\\ &\quad= \frac{1}{2}\big(\|\boldsymbol{u}_s\|_2^2 - \|\boldsymbol{u}_s^\star\|_2^2\big)^2\\ &\qquad + \big(\|\boldsymbol{u}_s\|_2^2 + \|\boldsymbol{u}_s^\star\|_2^2\big)\,\|\boldsymbol{u}_s-\boldsymbol{u}_s^\star\|_2^2\\ &\qquad - \frac{1}{2}\,\|\boldsymbol{u}_s-\boldsymbol{u}_s^\star\|_2^4.\end{aligned}$$

From the triangle inequality and the Cauchy-Schwartz inequality, we know that

$$\begin{aligned}&\big(\|\boldsymbol{u}_s\|_2^2 - \|\boldsymbol{u}_s^\star\|_2^2\big)^2\\ &\quad= \big(\|\boldsymbol{u}_s\|_2 + \|\boldsymbol{u}_s^\star\|_2\big)^2\big(\|\boldsymbol{u}_s\|_2 - \|\boldsymbol{u}_s^\star\|_2\big)^2\\ &\quad\leq 2\big(\|\boldsymbol{u}_s\|_2^2 + \|\boldsymbol{u}_s^\star\|_2^2\big)\,\|\boldsymbol{u}_s-\boldsymbol{u}_s^\star\|_2^2.\end{aligned}$$

The above two results taken together with (62) reveal that

$$\begin{aligned}\big\|\boldsymbol{u}_s^{\otimes 2}-\boldsymbol{u}_s^{\star\otimes 2}\big\|_2^2 &\leq 2\left(\|\boldsymbol{u}_s\|_2^2 + \|\boldsymbol{u}_s^\star\|_2^2\right)\|\boldsymbol{u}_s-\boldsymbol{u}_s^\star\|_2^2\\ &\lesssim \lambda_{\max}^{\star 2/3}\,\|\boldsymbol{u}_s-\boldsymbol{u}_s^\star\|_2^2,\end{aligned}$$

and consequently,

$$\begin{aligned}\|\widetilde{\boldsymbol{U}}-\widetilde{\boldsymbol{U}}^{\star}\|_{\mathrm{F}}^{2} &= \sum_{1\le s\le r}\left\|\boldsymbol{u}_s^{\otimes 2}-\boldsymbol{u}_s^{\star\otimes 2}\right\|_2^2\\ &\lesssim \lambda_{\max}^{\star 2/3}\sum_{1\le s\le r}\|\boldsymbol{u}_s-\boldsymbol{u}_s^{\star}\|_2^2\\ &= \lambda_{\max}^{\star 2/3}\|\boldsymbol{U}-\boldsymbol{U}^{\star}\|_{\mathrm{F}}^2 .\end{aligned}$$

Then the advertised bound on $\|\widetilde{\boldsymbol{U}}-\widetilde{\boldsymbol{U}}^{\star}\|_{\mathrm{F}}$ follows immediately from (59).

5) We proceed to the term $\|\widetilde{\boldsymbol{U}}\boldsymbol{\Pi}-\widetilde{\boldsymbol{U}}^{\star}\|_{2,\infty}$. Again, let us assume $\boldsymbol{\Pi}=\boldsymbol{I}_r$ and recall the notation $u_{s,i}:=(\boldsymbol{u}_s)_i$ and $u_{s,i}^{\star}:=(\boldsymbol{u}_s^{\star})_i$ for any $1\le s\le r, 1\le i\le d$. Then we can upper bound

$$\begin{aligned}&\sum_{1\le s\le r}\left(\boldsymbol{u}_s^{\otimes 2}-\boldsymbol{u}_s^{\star\otimes 2}\right)_{(i,j)}^2\\ &\quad=\sum_{1\le s\le r}\left(u_{s,i}u_{s,j}-u_{s,i}^{\star}u_{s,j}^{\star}\right)^2\\ &\quad\lesssim\sum_{1\le s\le r}\left(u_{s,i}-u_{s,i}^{\star}\right)^2u_{s,j}^2+\sum_{1\le s\le r}u_{s,i}^{\star 2}\left(u_{s,j}-u_{s,j}^{\star}\right)^2\\ &\quad\lesssim\max_{1\le s\le r}\|\boldsymbol{u}_s\|_\infty^2\|\boldsymbol{U}-\boldsymbol{U}^{\star}\|_{2,\infty}^2\end{aligned}$$

for any $1\le i,j\le d$. This taken collectively with (62f) and (59b) yields the claim.

6) Finally, we note that all bounds for $\boldsymbol{u}_l$ are also true for $\boldsymbol{u}_l^{(m)}$. Hence the above-mentioned results continue to hold for $\boldsymbol{U}^{(m)}$ and $\boldsymbol{u}_l^{(m)}$ $(1\le l\le r)$.

Appendix I
Other Auxiliary Lemmas

Lemma 21: Let $\boldsymbol{T}\in\mathbb{R}^{d\times d\times d}$ be an order-3 tensor with decomposition $\boldsymbol{T}=\sum_{i=1}^r\boldsymbol{u}_i\otimes\boldsymbol{v}_i\otimes\boldsymbol{w}_i$. Here, $\{\boldsymbol{u}_i,\boldsymbol{v}_i,\boldsymbol{w}_i\}_{i=1}^r$ is a collection of vectors in $\mathbb{R}^d$. Then for any index subset $\Omega\subset[d]^3$ and any $t\in\mathbb{R}$, one has

$$\begin{aligned}&\|\mathcal{P}_\Omega(\boldsymbol{T})-t\boldsymbol{T}\|\\ &\quad\le\left\|\mathcal{P}_\Omega\left(\boldsymbol{1}^{\otimes 3}\right)-t\boldsymbol{1}^{\otimes 3}\right\|\sum_{i=1}^r\|\boldsymbol{u}_i\|_\infty\|\boldsymbol{v}_i\|_\infty\|\boldsymbol{w}_i\|_\infty ,\end{aligned}$$

where $\boldsymbol{1}\in\mathbb{R}^d$ denotes the all-one vector. Here, Ω can be arbitrary.

Proof: Fix arbitrary vectors $\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}\in\mathbb{R}^d$ with $\|\boldsymbol{x}\|_2=\|\boldsymbol{y}\|_2=\|\boldsymbol{z}\|_2=1$. We have

$$\begin{aligned}&|\langle\mathcal{P}_\Omega(\boldsymbol{T})-t\boldsymbol{T},\,\boldsymbol{x}\otimes\boldsymbol{y}\otimes\boldsymbol{z}\rangle|\\ &\quad=\left|\left\langle\mathcal{P}_\Omega\left(\boldsymbol{1}^{\otimes 3}\right)-t\boldsymbol{1}^{\otimes 3},\boldsymbol{T}\odot(\boldsymbol{x}\otimes\boldsymbol{y}\otimes\boldsymbol{z})\right\rangle\right|\\ &\quad\le\left\|\mathcal{P}_\Omega\left(\boldsymbol{1}^{\otimes 3}\right)-t\boldsymbol{1}^{\otimes 3}\right\|\left\|\boldsymbol{T}\odot(\boldsymbol{x}\otimes\boldsymbol{y}\otimes\boldsymbol{z})\right\|_*,\end{aligned}$$

where we denote by $\|\cdot\|_*$ the tensor nuclear norm [28]. By the linearity of the Hadamard and tensor product, we can express

$$\begin{aligned}\boldsymbol{T}\odot(\boldsymbol{x}\otimes\boldsymbol{y}\otimes\boldsymbol{z})&=\Big(\sum_{i=1}^r\boldsymbol{u}_i\otimes\boldsymbol{v}_i\otimes\boldsymbol{w}_i\Big)\odot(\boldsymbol{x}\otimes\boldsymbol{y}\otimes\boldsymbol{z})\\ &=\sum_{i=1}^r(\boldsymbol{u}_i\otimes\boldsymbol{v}_i\otimes\boldsymbol{w}_i)\odot(\boldsymbol{x}\otimes\boldsymbol{y}\otimes\boldsymbol{z})\\ &=\sum_{i=1}^r(\boldsymbol{u}_i\odot\boldsymbol{x})\otimes(\boldsymbol{v}_i\odot\boldsymbol{y})\otimes(\boldsymbol{w}_i\odot\boldsymbol{z}) .\end{aligned}$$

From the triangle inequality, we can upper bound

$$\begin{aligned}&\left\|\boldsymbol{T}\odot(\boldsymbol{x}\otimes\boldsymbol{y}\otimes\boldsymbol{z})\right\|_*\\ &\quad\le\sum_{i=1}^r\|(\boldsymbol{u}_i\odot\boldsymbol{x})\otimes(\boldsymbol{v}_i\odot\boldsymbol{y})\otimes(\boldsymbol{w}_i\odot\boldsymbol{z})\|_*\\ &\quad\le\sum_{i=1}^r\|\boldsymbol{u}_i\odot\boldsymbol{x}\|_2\|\boldsymbol{v}_i\odot\boldsymbol{y}\|_2\|\boldsymbol{w}_i\odot\boldsymbol{z}\|_2\\ &\quad\le\sum_{i=1}^r\|\boldsymbol{u}_i\|_\infty\|\boldsymbol{v}_i\|_\infty\|\boldsymbol{w}_i\|_\infty .\end{aligned}$$

Here, the second inequality holds due to the fact that $\|\boldsymbol{a}\otimes\boldsymbol{b}\otimes\boldsymbol{c}\|=\|\boldsymbol{a}\otimes\boldsymbol{b}\otimes\boldsymbol{c}\|_*=\|\boldsymbol{a}\|_2\|\boldsymbol{b}\|_2\|\boldsymbol{c}\|_2$ for any vectors $\boldsymbol{a},\boldsymbol{b},\boldsymbol{c}\in\mathbb{R}^d$, whereas the last inequality follows by observing the following inequality

$$\|\boldsymbol{u}_i\odot\boldsymbol{x}\|_2^2=\sum_{j=1}^d(\boldsymbol{u}_i)_j^2x_j^2\le\|\boldsymbol{u}_i\|_\infty^2\sum_{j=1}^dx_j^2=\|\boldsymbol{u}_i\|_\infty^2$$

and similarly $\|\boldsymbol{v}_i\odot\boldsymbol{y}\|_2\le\|\boldsymbol{v}_i\|_\infty$ and $\|\boldsymbol{w}_i\odot\boldsymbol{z}\|_2\le\|\boldsymbol{w}_i\|_\infty$. Consequently, one arrives at

$$\begin{aligned}&|\langle\mathcal{P}_\Omega(\boldsymbol{T})-t\boldsymbol{T},\,\boldsymbol{x}\otimes\boldsymbol{y}\otimes z\rangle|\\ &\quad\le\left\|\mathcal{P}_\Omega\left(\boldsymbol{1}^{\otimes 3}\right)-t\boldsymbol{1}^{\otimes 3}\right\|\sum_{i=1}^r\|\boldsymbol{u}_i\|_\infty\|\boldsymbol{v}_i\|_\infty\|\boldsymbol{w}_i\|_\infty .\end{aligned}$$

Given that this holds for arbitrary $\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}\in\mathbb{R}^d$ with $\|\boldsymbol{x}\|_2=\|\boldsymbol{y}\|_2=\|\boldsymbol{z}\|_2=1$, we finish the proof by the definition of the spectral norm. □

References

[1] C. Cai, H. V. Poor, and Y. Chen, "Uncertainty quantification for nonconvex tensor completion: Confidence intervals, heteroscedasticity and optimality," in *Proc. 37th Int. Conf. Mach. Learn.*, 2020, pp. 1271–1282.

[2] C. Cai, G. Li, H. V. Poor, and Y. Chen, "Nonconvex low-rank tensor completion from noisy data," *Oper. Res.*, vol. 70, no. 2, pp. 1219–1237, 2022.

[3] T. G. Kolda and B. W. Bader, "Tensor decompositions and applications," *SIAM Rev.*, vol. 51, no. 3, pp. 455–500, Sep. 2009.

[4] N. D. Sidiropoulos, L. De Lathauwer, X. Fu, K. Huang, E. E. Papalexakis, and C. Faloutsos, "Tensor decomposition for signal processing and machine learning," *IEEE Trans. Signal Process.*, vol. 65, no. 13, pp. 3551–3582, Jul. 2017.

[5] A. Anandkumar, R. Ge, D. J. Hsu, S. M. Kakade, and M. Telgarsky, "Tensor decompositions for learning latent variable models," *J. Mach. Learn. Res.*, vol. 15, pp. 2773–2832, Aug. 2014.

[6] A. Zhang, "Cross: Efficient low-rank tensor completion," *Ann. Statist.*, vol. 47, no. 2, pp. 936–964, 2019.

[7] R. Han, R. Willett, and A. R. Zhang, "An optimal statistical and computational framework for generalized tensor estimation," 2020, *arXiv:2002.11255*.

[8] Y. Luo and A. R. Zhang, "Low-rank tensor estimation via Riemannian gauss-Newton: Statistical optimality and second-order convergence," 2021, *arXiv:2104.12031*.

[9] J.-F. Cai, J. Li, and D. Xia, "Generalized low-rank plus sparse tensor estimation by fast Riemannian optimization," 2021, *arXiv:2103.08895*.

[10] O. Semerci, N. Hao, M. E. Kilmer, and E. L. Miller, "Tensor-based formulation and nuclear norm regularization for multienergy computed tomography," *IEEE Trans. Image Process.*, vol. 23, no. 4, pp. 1678–1693, Apr. 2014.

[11] J. Liu, P. Musialski, P. Wonka, and J. Ye, "Tensor completion for estimating missing values in visual data," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 208–220, Jan. 2012.

[12] N. Kreimer, A. Stanton, and M. D. Sacchi, "Tensor completion based on nuclear norm minimization for 5D seismic data reconstruction," *Geophysics*, vol. 78, no. 6, pp. V273–V284, 2013.

[13] S. Gandy, B. Recht, and I. Yamada, "Tensor completion and low-n-rank tensor recovery via convex optimization," *Inverse Problems*, vol. 27, no. 2, p. 025010, 2011.

[14] B. Huang, C. Mu, D. Goldfarb, and J. Wright, "Provable models for robust low-rank tensor completion," *Pacific J. Optim.*, vol. 11, no. 2, pp. 339–364, 2015.

[15] B. Romera-Paredes and M. Pontil, "A new convex relaxation for tensor completion," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 2967–2975.

[16] A. Montanari and N. Sun, "Spectral algorithms for tensor completion," *Commun. Pure Appl. Math.*, vol. 71, no. 11, pp. 2381–2425, 2018.

[17] C. Cai, G. Li, Y. Chi, H. V. Poor, and Y. Chen, "Subspace estimation from unbalanced and incomplete data matrices: $\ell_{2,\infty}$ statistical guarantees," *Ann. Statist.*, vol. 49, no. 2, pp. 944–967, 2021.

[18] Y. Chen, Y. Chi, J. Fan, and C. Ma, "Spectral methods for data science: A statistical perspective," *Found. Trends Mach. Learn.*, vol. 14, no. 5, pp. 566–806, 2021.

[19] B. Barak and A. Moitra, "Noisy tensor completion via the sum-of-squares hierarchy," in *Proc. 29th Annu. Conf. Learn. Theory*, 2016, pp. 417–445.

[20] A. Potechin and D. Steurer, "Exact tensor completion with sum-of-squares," in *Proc. Conf. Learn. Theory*, 2017, pp. 1619–1673.

[21] P. Jain and S. Oh, "Provable tensor factorization with missing data," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 1431–1439.

[22] A. Liu and A. Moitra, "Tensor completion made practical," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 18905–18916.

[23] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Found. Comput. Math.*, vol. 9, no. 6, p. 717, Apr. 2019.

[24] E. J. Candès and Y. Plan, "Matrix completion with noise," *Proc. IEEE*, vol. 98, no. 6, pp. 925–936, Jun. 2010.

[25] D. Gross, "Recovering low-rank matrices from few coefficients in any basis," *IEEE Trans. Inf. Theory*, vol. 57, no. 3, pp. 1548–1566, Mar. 2011.

[26] Y. Chen, Y. Chi, J. Fan, C. Ma, and Y. Yan, "Noisy matrix completion: Understanding statistical guarantees for convex relaxation via nonconvex optimization," *SIAM J. Optim.*, vol. 30, no. 4, pp. 3098–3121, 2020.

[27] X. Li, "Compressed sensing and matrix completion with constant proportion of corruptions," *Constructive Approx.*, vol. 37, no. 1, pp. 73–99, Feb. 2013.

[28] M. Yuan and C.-H. Zhang, "On tensor completion via nuclear norm minimization," *Found. Comput. Math.*, vol. 16, no. 4, pp. 1031–1068, 2014.

[29] M. Yuan and C.-H. Zhang, "Incoherent tensor norms and their applications in higher order tensor completion," *IEEE Trans. Inf. Theory*, vol. 63, no. 10, pp. 6753–6766, Oct. 2017.

[30] R. Tomioka, K. Hayashi, and H. Kashima, "Estimation of low-rank tensors via convex optimization," 2010, *arXiv:1010.0789*.

[31] C. Mu, B. Huang, J. Wright, and D. Goldfarb, "Square deal: Lower bounds and improved relaxations for tensor recovery," in *Proc. 31st Int. Conf. Mach. Learn.*, 2014, pp. 73–81.

[32] R. Keshavan, A. Montanari, and S. Oh, "Matrix completion from a few entries," *IEEE Trans. Inf. Theory*, vol. 56, no. 6, pp. 2980–2998, Jun. 2009.

[33] Y. Chi, Y. M. Lu, and Y. Chen, "Nonconvex optimization meets low-rank matrix factorization: An overview," *IEEE Trans. Signal Process.*, vol. 67, no. 20, pp. 5239–5269, Oct. 2019.

[34] D. Xia, M. Yuan, and C.-H. Zhang, "Statistically optimal and computationally efficient low rank tensor completion from noisy entries," *Ann. Statist.*, vol. 49, no. 1, pp. 76–99, 2021.

[35] D. Xia and M. Yuan, "On polynomial time methods for exact low-rank tensor completion," *Found. Comput. Math.*, vol. 19, pp. 1265–1313, Jan. 2019.

[36] Z. Zhang and S. Aeron, "Exact tensor completion using t-SVD," *IEEE Trans. Signal Process.*, vol. 65, no. 6, pp. 1511–1526, Mar. 2016.

[37] T.-Y. Ji, T.-Z. Huang, X.-L. Zhao, T.-H. Ma, and G. Liu, "Tensor completion using total variation and low-rank matrix factorization," *Inf. Sci.*, vol. 326, pp. 243–257, Jan. 2016.

[38] S. Burer and R. D. Monteiro, "A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization," *Math. Program.*, vol. 95, no. 2, pp. 329–357, Feb. 2003.

[39] N. Srebro, "Learning with matrix factorizations," Ph.D. dissertation, Massachusetts Inst. Technol., Cambridge, MA, USA, 2004.

[40] R. H. Keshavan, A. Montanari, and S. Oh, "Matrix completion from noisy entries," *J. Mach. Learn. Res.*, vol. 11, pp. 2057–2078, Jul. 2010.

[41] P. Jain, P. Netrapalli, and S. Sanghavi, "Low-rank matrix completion using alternating minimization," in *Proc. 45th Annu. ACM Symp. Theory Comput.*, 2013, pp. 665–674.

[42] E. J. Candès, X. Li, and M. Soltanolkotabi, "Phase retrieval via Wirtinger flow: Theory and algorithms," *IEEE Trans. Inf. Theory*, vol. 61, no. 4, pp. 1985–2007, Apr. 2015.

[43] Y. Chen and E. J. Candès, "Solving random quadratic systems of equations is nearly as easy as solving linear systems," *Commun. Pure Appl. Math.*, vol. 70, pp. 822–883, May 2017.

[44] Y. Chen and M. J. Wainwright, "Fast low-rank estimation by projected gradient descent: General statistical and algorithmic guarantees," 2015, *arXiv:1509.03025*.

[45] C. Ma, K. Wang, Y. Chi, and Y. Chen, "Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval and matrix completion," *Found. Comput. Math.*, vol. 20, no. 3, pp. 451–633, 2020.

[46] Y. Chen and E. J. Candès, "The projected power method: An efficient algorithm for joint alignment from pairwise differences," *Commun. Pure Appl. Math.*, vol. 71, no. 8, pp. 1648–1714, Aug. 2018.

[47] Y. Chen, Y. Chi, J. Fan, and C. Ma, "Gradient descent with random initialization: Fast global convergence for nonconvex phase retrieval," *Math. Program.*, vol. 176, no. 1, pp. 5–37, Jul. 2018.

[48] P. Netrapalli, U. Niranjan, S. Sanghavi, A. Anandkumar, and P. Jain, "Non-convex robust PCA," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 1107–1115.

[49] B. Hao, A. Zhang, and G. Cheng, "Sparse and low-rank tensor estimation via cubic sketchings," *IEEE Trans. Inf. Theory*, vol. 66, no. 9, pp. 5927–5964, Sep. 2020.

[50] H. Zhang, Y. Chi, and Y. Liang, "Provable non-convex phase retrieval with outliers: Median truncated Wirtinger flow," in *Proc. 33rd Int. Conf. Mach. Learn.*, 2016, pp. 1022–1031.

[51] Y. Li, C. Ma, Y. Chen, and Y. Chi, "Nonconvex matrix factorization from rank-one measurements," *IEEE Trans. Inf. Theory*, vol. 67, no. 3, pp. 1928–1950, Mar. 2021.

[52] J.-F. Cai, H. Liu, and Y. Wang, "Fast rank one alternating minimization algorithm for phase retrieval," 2017, *arXiv:1708.08751*.

[53] J. Chen, D. Liu, and X. Li, "Nonconvex rectangular matrix completion via gradient descent without $\ell_{2,\infty}$ regularization," 2019, *arXiv:1901.06116*.

[54] G. Wang and G. Giannakis, "Solving random systems of quadratic equations via truncated generalized gradient flow," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 568–576.

[55] Y. Chen, J. Fan, C. Ma, and Y. Yan, "Bridging convex and nonconvex optimization in robust PCA: Noise, outliers and missing data," *Ann. Statist.*, vol. 49, no. 5, pp. 2948–2971, 2021.

[56] G. Li and Y. Wei, "A non-asymptotic framework for approximate message passing in spiked models," 2022, *arXiv:2208.03313*.

[57] J. Sun, Q. Qu, and J. Wright, "A geometric analysis of phase retrieval," *Found. Comput. Math.*, vol. 18, pp. 1131–1198, pp. 1–68, Oct. 2018.

[58] Q. Qu, Y. Zhang, Y. C. Eldar, and J. Wright, "Convolutional phase retrieval via gradient descent," *IEEE Trans. Inf. Theory*, vol. 66, no. 3, pp. 1785–1821, Mar. 2020.

[59] T. Tong, C. Ma, and Y. Chi, "Accelerating ill-conditioned low-rank matrix estimation via scaled gradient descent," 2020, *arXiv:2005.08898*.

[60] H. Zhang, Y. Zhou, Y. Liang, and Y. Chi, "A nonconvex approach for phase retrieval: Reshaped Wirtinger flow and incremental algorithms," *J. Mach. Learn. Res.*, vol. 18, no. 1, pp. 5164–5198, 2017.

[61] Y. Chen, J. Fan, B. Wang, and Y. Yan, "Convex and nonconvex optimization are both minimax-optimal for noisy blind deconvolution under random designs," *J. Amer. Stat. Assoc.*, to be published.

[62] S. Ling, "Near-optimal performance bounds for orthogonal and permutation group synchronization via spectral methods," *Appl. Comput. Harmon. Anal.*, vol. 60, pp. 20–52, Sep. 2022.

[63] N. El Karoui, "On the impact of predictor geometry on the performance on high-dimensional ridge-regularized generalized robust regression estimators," *Probab. Theory Rel. Fields*, vol. 170, pp. 95–175, Jan. 2018.

[64] N. El Karoui, D. Bean, P. J. Bickel, C. Lim, and B. Yu, "On robust regression with high-dimensional predictors," *Proc. Nat. Acad. Sci. USA*, vol. 110, no. 36, pp. 14557–14562, 2013.

[65] A. Javanmard and A. Montanari, "De-biasing the lasso: Optimal sample size for Gaussian designs," 2015, *arXiv:1508.02757*.

[66] G. Li, Y. Wei, Y. Chi, Y. Gu, and Y. Chen, "Breaking the sample size barrier in model-based reinforcement learning with a generative model," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 12861–12872.

[67] G. Li, L. Shi, Y. Chen, Y. Chi, and Y. Wei, "Settling the sample complexity of model-based offline reinforcement learning," 2022, *arXiv:2204.05275*.
[68] Y. Yan, Y. Chen, and J. Fan, "Inference for heteroskedastic PCA with missing data," 2021, *arXiv:2107.12365*.
[69] Y. Chen, J. Fan, C. Ma, and K. Wang, "Spectral method and regularized MLE are both optimal for top-K ranking," *Ann. Statist.*, vol. 47, no. 4, p. 2204, 2019.
[70] Y. Chen, J. Fan, C. Ma, and Y. Yan, "Inference and uncertainty quantification for noisy matrix completion," *Proc. Nat. Acad. Sci. USA*, vol. 116, no. 46, pp. 22931–22937, 2019.
[71] D. Xia and M. Yuan, "Statistical inferences of linear forms for noisy matrix completion," 2019, *arXiv:1909.00116*.
[72] C. Cheng, Y. Wei, and Y. Chen, "Tackling small eigen-gaps: Fine-grained eigenvector estimation and inference under heteroscedastic noise," *IEEE Trans. Inf. Theory*, vol. 67, no. 11, pp. 7380–7419, Nov. 2021.
[73] D. Xia, "Confidence region of singular subspaces for low-rank matrix regression," 2018, *arXiv:1805.09871*.
[74] D. Xia, "Normal approximation and confidence region of singular subspaces," 2019, *arXiv:1901.00304*.
[75] D. Xia, A. R. Zhang, and Y. Zhou, "Inference for low-rank tensors—No need to debias," 2020, *arXiv:2012.14844*.
[76] J. Huang, D. Z. Huang, Q. Yang, and G. Cheng, "Power iteration for tensor PCA," 2020, *arXiv:2012.13669*.
[77] C.-H. Zhang and S. S. Zhang, "Confidence intervals for low dimensional parameters in high dimensional linear models," *J. Roy. Stat. Soc., B, Stat. Methodol.*, vol. 76, no. 1, pp. 217–242, Jan. 2014.
[78] S. van de Geer, P. Bühlmann, Y. Ritov, and R. Dezeure, "On asymptotically optimal confidence regions and tests for high-dimensional models," *Ann. Statist.*, vol. 42, no. 3, pp. 1166–1202, 2014.
[79] A. Javanmard and A. Montanari, "Confidence intervals and hypothesis testing for high-dimensional regression," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 2869–2909, Jan. 2014.
[80] Z. Ren, T. Sun, C.-H. Zhang, and H. H. Zhou, "Asymptotic normality and optimalities in estimation of large Gaussian graphical models," *Ann. Statist.*, vol. 43, no. 3, pp. 991–1026, 2015.
[81] Y. Ning and H. Liu, "A general theory of hypothesis tests and confidence regions for sparse high dimensional models," *Ann. Statist.*, vol. 45, no. 1, pp. 158–195, 2017.
[82] M. Celentano, A. Montanari, and Y. Wei, "The lasso with general Gaussian designs with applications to hypothesis testing," 2020, *arXiv:2007.13716*.
[83] T. T. Cai and Z. Guo, "Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity," *Ann. Statist.*, vol. 45, no. 2, pp. 615–646, 2017.
[84] T. T. Cai, T. Liang, and A. Rakhlin, "Geometric inference for general high-dimensional linear inverse problems," *Ann. Statist.*, vol. 44, no. 4, pp. 1536–1563, 2016.
[85] C. Ma, J. Lu, and H. Liu, "Inter-subject analysis: A partial Gaussian graphical model approach," *J. Amer. Stat. Assoc.*, vol. 116, no. 534, pp. 746–755, 2021.
[86] P. Sur, Y. Chen, and E. J. Candès, "The likelihood ratio test in high-dimensional logistic regression is asymptotically a rescaled chi-square," *Probab. Theory Rel. Fields*, vol. 175, pp. 487–558, Jan. 2019.
[87] J. Janková and S. van de Geer, "De-biased sparse PCA: Inference and testing for eigenstructure of large covariance matrices," 2018, *arXiv:1801.10567*.
[88] L. Miolane and A. Montanari, "The distribution of the lasso: Uniform control over sparse balls and adaptive parameter tuning," 2018, *arXiv:1811.01212*.
[89] A. R. Zhang, T. Tony Cai, and Y. Wu, "Heteroskedastic PCA: Algorithm, optimality, and applications," 2018, *arXiv:1810.08316*.
[90] E. Richard and A. Montanari, "A statistical model for tensor PCA," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2897–2905.
[91] V. Bentkus, "A Lyapunov-type bound in R^d," *Theory Probab. Appl.*, vol. 49, no. 2, pp. 311–323, 2005.
[92] V. Koltchinskii, *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems*, vol. 2033. Berlin, Germany: Springer, 2011.
[93] B. A. Schmitt, "Perturbation bounds for matrix square roots and Pythagorean sums," *Linear Algebra Appl.*, vol. 174, pp. 215–227, Sep. 1992.
[94] J. Chen and X. Li, "Model-free nonconvex matrix completion: Local minima analysis and applications in memory-efficient kernel PCA," 2017, *arXiv:1711.01742*.
[95] D. Hsu, S. M. Kakade, and T. Zhang, "A tail inequality for quadratic forms of subGaussian random vectors," *Electron. Commun. Probab.*, vol. 17, no. 52, pp. 1–6, 2012.
[96] J. Shao, *Mathematical Statistics*. Berlin, Germany: Springer, 2003.

Changxiao Cai received the B.E. degree (Hons.) in electronic engineering from Tsinghua University in 2016, and the Ph.D. degree in electrical engineering from Princeton University in 2021. He is currently a Post-Doctoral Researcher with the Department of Biostatistics, University of Pennsylvania. His research interests include statistical machine learning, high-dimensional statistics, convex and nonconvex optimization, and information theory.

H. Vincent Poor (Life Fellow, IEEE) received the Ph.D. degree in EECS from Princeton University in 1977. From 1977 to 1990, he was on the Faculty of the University of Illinois at Urbana–Champaign. Since 1990, he has been on the Faculty at Princeton University, where he is currently the Michael Henry Strater University Professor. From 2006 to 2016, he worked as the Dean of the School of Engineering and Applied Science, Princeton University. He has also held visiting appointments at several other universities, including most recently at Berkeley and Cambridge. His research interests include information theory, machine learning and network science, and their applications in wireless networks, energy systems, and related fields. Among his publications in these areas is the recent book *Machine Learning and Wireless Communications* (Cambridge University Press, 2022). He is a member of the National Academy of Engineering and the National Academy of Sciences. He is a Foreign Member of the Chinese Academy of Sciences, the Royal Society, and other national and international academies. He received the IEEE Alexander Graham Bell Medal in 2017.

Yuxin Chen (Member, IEEE) received the B.E. degree (Hons.) from Tsinghua University in 2008, the M.S. degree in electrical and computer engineering from The University of Texas at Austin in 2010, and the Ph.D. degree in electrical engineering and the M.S. degree in statistics from Stanford University in January 2015. He is currently an Associate Professor of statistics and data science with The Wharton School, University of Pennsylvania. Prior to joining the University of Pennsylvania, he was an Assistant Professor of electrical and computer engineering at Princeton University from 2017 to 2021, and was a Post-Doctoral Scholar with the Department of Statistics, Stanford University, from 2015 to 2017. His research interests include high-dimensional data analysis, convex and nonconvex optimization, reinforcement learning, statistical learning, statistical signal processing, and information theory. He has received the Alfred P. Sloan Research Fellowship, the Google Research Scholar Award, the Princeton Graduate Mentoring Award, the AFOSR Young Investigator Award, the ARO Young Investigator Award, the ICCM Best Paper Award (Gold Medal), and the Princeton SEAS Junior Faculty Award, and was selected as the Finalist for the Best Paper Prize for Young Researchers in Continuous Optimization 2019.