Nonconvex Low-Rank Tensor Completion
from Noisy Data

Changxiao Cai*   Gen Li†   H. Vincent Poor*   Yuxin Chen*

November 11, 2020

Abstract

We study a noisy tensor completion problem of broad practical interest, namely, the reconstruction of a low-rank tensor from highly incomplete and randomly corrupted observations of its entries. While a variety of prior work has been dedicated to this problem, prior algorithms either are computationally too expensive for large-scale applications, or come with sub-optimal statistical guarantees. Focusing on "incoherent" and well-conditioned tensors of a constant CP rank, we propose a two-stage nonconvex algorithm — (vanilla) gradient descent following a rough initialization — that achieves the best of both worlds. Specifically, the proposed nonconvex algorithm faithfully completes the tensor and retrieves all individual tensor factors within nearly linear time, while at the same time enjoying near-optimal statistical guarantees (i.e. minimal sample complexity and optimal estimation accuracy). The estimation errors are evenly spread out across all entries, thus achieving optimal ℓ_∞ statistical accuracy. We have also discussed how to extend our approach to accommodate asymmetric tensors. The insight conveyed through our analysis of nonconvex optimization might have implications for other tensor estimation problems.

Keywords: tensor completion, nonconvex optimization, gradient descent, spectral methods, entrywise statistical guarantees, minimaxity

Contents

Corresponding author: Yuxin Chen. A short version of this work has appeared in NeurIPS 2019 [CLPC19].

*Department of Electrical Engineering, Princeton University, Princeton, NJ 08544, USA; Email: `{ccai,poor,yuxin.chen}@princeton.edu`.

†Department of Electronic Engineering, Tsinghua University, Beijing, 10084, China; Email: `g-li16@mails.tsinghua.edu.cn`.

1 Introduction and motivation

1.1 Tensor completion from noisy entries

Estimation of low-complexity models from highly incomplete observations is a fundamental task that spans a diverse array of science and engineering applications. Arguably one of the most extensively studied problems of this kind is matrix completion, where one wishes to recover a low-rank matrix given only partial entries [DR16, CC18]. Moving beyond matrix-type data, a natural higher-order generalization is *low-rank tensor completion*, which aims to reconstruct a low-rank tensor when the vast majority of its entries are unseen. There is certainly no shortage of applications that motivate the investigation of tensor completion (e.g. personalized medicine [SN16, Paw19], medical imaging [GRY11, SHKM14, CZA+17], seismic data analysis [KSS13, EAHK13], multi-dimensional harmonic retrieval [CC14, YLW+17]). One concrete example in operations research arises when learning the preference of individual customers for a collection of products on the basis of historical transactions [FL19, MP20]. Given the limited availability of transaction data (e.g. each customer might only have purchased very few products before), it is crucial to exploit multi-way customer-product interactions (e.g. users' browsing and searching histories) in order to better predict the likelihood of a customer purchasing a new product. Clearly, the presence of missing data and the need of exploiting multi-way structure result in the task of tensor completion. Additionally, tensor completion finds important applications in visual data in-painting [LMWY13, LYX17], where one wishes to reconstruct video data (or a sequence of images) from incomplete measurements. The video data consist of at least two spatial variables and one temporal variable, whose intrinsic connections are often modeled via certain low-complexity tensors.

For the sake of clarity, we phrase the problem formally before we proceed, focusing on a simple model that already captures the intrinsic difficulty of tensor completion in many aspects.[1] Imagine we are asked to estimate a symmetric order-three tensor[2] $\boldsymbol{T}^{\star} \in \mathbb{R}^{d \times d \times d}$ from a small number of noisy entries

$$T_{j,k,l} = T_{j,k,l}^{\star} + E_{j,k,l}, \qquad \forall (j,k,l) \in \Omega, \tag{1}$$

where $T_{j,k,l}$ is the observed noisy entry at location (j,k,l), $E_{j,k,l}$ stands for the associated noise, and $\Omega \subseteq \{1, \cdots, d\}^3$ is a symmetric index subset to sample from. For notational simplicity, we set $\boldsymbol{T} = [T_{j,k,l}]_{1 \leq j,k,l \leq d}$

---

[1]We focus on symmetric order-3 tensors primarily for simplicity of presentation. Many of our findings naturally extend to the more general case with asymmetric tensors of possibly higher order. Detailed discussions are deferred to Appendix E due to the space limits.

[2]Here, a tensor $\boldsymbol{T} \in \mathbb{R}^{d \times d \times d}$ is said to be symmetric if $T_{j,k,l} = T_{k,j,l} = T_{k,l,j} = T_{l,k,j} = T_{j,l,k} = T_{l,j,k}$ for all $1 \leq j,k,l \leq d$.

and $\boldsymbol{E} = [E_{j,k,l}]_{1\leq j,k,l\leq d}$, with $T_{j,k,l} = E_{j,k,l} = 0$ for any $(j,k,l) \notin \Omega$. We adopt a *random sampling* model such that each index (j,k,l) $(j \leq k \leq l)$ is included in Ω independently with probability p. In addition, we know *a priori* that the unknown tensor $\boldsymbol{T}^\star \in \mathbb{R}^{d\times d\times d}$ is a superposition of r rank-one tensors (often termed canonical polyadic (CP) decomposition if r is minimal)

$$\boldsymbol{T}^\star = \sum_{i=1}^{r} \boldsymbol{u}_i^\star \otimes \boldsymbol{u}_i^\star \otimes \boldsymbol{u}_i^\star, \qquad \text{or more concisely,} \qquad \boldsymbol{T}^\star = \sum_{i=1}^{r} \boldsymbol{u}_i^{\star\,\otimes 3}, \tag{2}$$

where each $\boldsymbol{u}_i^\star \in \mathbb{R}^d$ represents one of the r low-rank tensor components / factors. Here and throughout, for any vectors $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c} \in \mathbb{R}^d$, the tensor $\boldsymbol{a} \otimes \boldsymbol{b} \otimes \boldsymbol{c}$ is a $d \times d \times d$ array whose (j,k,l)-th entry is given by $a_j b_k c_l$. The primary question is this: can we hope to faithfully estimate $\boldsymbol{T}^\star$, as well as the individual tensor factors $\{\boldsymbol{u}_i^\star\}_{1\leq i\leq r}$, from the partially revealed entries (1), assuming that r is reasonably small?

1.2 Computational and statistical challenges

Even though tensor completion conceptually resembles matrix completion in various ways, it is considerably more challenging than the matrix counterpart. This is perhaps not surprising, given that a plethora of natural tensor problems (e.g. computing the spectral norm, finding the best low-rank approximation) are all notoriously hard [HL13]. As a notable example, while matrix completion is often efficiently solvable under nearly minimal sample complexity [CR09, Gro11], all polynomial-time algorithms developed so far for tensor completion — even in the noise-free case — require a sample size at least exceeding the order of $rd^{3/2}$, which is substantially larger than the degrees of freedom (i.e. rd) underlying the model (2). In fact, it is widely conjectured that there exists a large computational barrier away from the information-theoretic sampling limits [BM16].

With this fundamental gap in mind, the current paper focuses on the regime (in terms of the sample size) that enables reliable tensor completion in polynomial time. A variety of algorithms have been proposed that enjoy some sort of theoretical guarantees in (at least part of) this regime, including but not limited to spectral methods [MS18, CLC$^+$20], sum-of-squares hierarchy [BM16, PS17], nonconvex algorithms [JO14, XY17], and also convex relaxation (based on proper unfolding) [GRY11, HMGW15, RPP13, GQ14]. While these are all polynomial-time algorithms, most of the computational complexities supported by prior theory remain prohibitively high when dealing with large-scale tensor data — a point that we shall elaborate on later. The only exception is the unfolding-based spectral method, which, however, fails to achieve exact recovery as the noise vanishes. This leads to a critical question:

> **Q1:** *Is there any linear-time algorithm that is guaranteed to work for low-rank tensor completion?*

Going beyond such computational concerns, one might naturally wonder whether it is also possible for a fast algorithm to achieve a nearly un-improvable statistical accuracy in the presence of noise. Towards this end, intriguing stability guarantees have been established for sum-of-squares hierarchy in the noisy settings [BM16], although this paradigm is computationally expensive for large-scale data. The recent work [XYZ17] came up with a two-stage algorithm (i.e. a spectral method followed by tensor power iterations) for noisy tensor completion. Its estimation accuracy, however, falls short of achieving exact recovery in the absence of noise. This gives rise to another question of fundamental importance:

> **Q2:** *Can we achieve near-optimal statistical accuracy without compromising computational efficiency?*

In this paper, we aim to address the above two questions by developing a nonconvex algorithm that achieves optimal computational efficiency and statistical accuracy all at once.

2 Algorithm and main results

2.1 A two-stage nonconvex algorithm

To address the above-mentioned challenges, a first impulse is to resort to the following least squares problem:

$$\underset{\boldsymbol{u}_1,\cdots,\boldsymbol{u}_r\in\mathbb{R}^d}{\text{minimize}} \quad \sum_{j,k,l\in\Omega} \left(\left[\sum_{i=1}^{r} \boldsymbol{u}_i^{\otimes 3}\right]_{j,k,l} - T_{j,k,l} \right)^2, \tag{3}$$

or more concisely (up to proper re-scaling),

$$\underset{\boldsymbol{U}\in\mathbb{R}^{d\times r}}{\text{minimize}} \quad f(\boldsymbol{U}) := \frac{1}{6p} \left\| \mathcal{P}_\Omega \left(\sum_{i=1}^{r} \boldsymbol{u}_i^{\otimes 3} - \boldsymbol{T} \right) \right\|_{\mathrm{F}}^2 \tag{4}$$

if we take $\boldsymbol{U} := [\boldsymbol{u}_1,\ldots,\boldsymbol{u}_r] \in \mathbb{R}^{d\times r}$. Here, we denote by $\mathcal{P}_\Omega(\boldsymbol{T})$ the orthogonal projection of any tensor $\boldsymbol{T}$ onto the subspace of tensors which vanish outside of the index set Ω. This optimization problem, however, is highly nonconvex (which involves minimizing a degree-6 polynomial), thus resulting in computational intractability in general.

Fortunately, not all nonconvex problems are as daunting to solve as they may seem. For example, recent years have seen a flurry of activity in low-rank matrix factorization via nonconvex optimization, which provably achieves optimal statistical accuracy and computational efficiency at once; see [CLC19] for an overview of recent advances. Motivated by this strand of work, we propose to solve (4) via a two-stage nonconvex paradigm, presented below in reverse order. The whole procedure is summarized in Algorithms 1-3.

Gradient descent (GD). Arguably one of the simplest optimization algorithms is gradient descent, which adopts a gradient update rule

$$\boldsymbol{U}^{t+1} = \boldsymbol{U}^t - \eta_t \nabla f(\boldsymbol{U}^t), \qquad t = 0, 1, \cdots \tag{5}$$

where η_t is the learning rate or the stepsize, and $\boldsymbol{U}^t \in \mathbb{R}^{d\times r}$ is the estimate in the t-th iteration. The main computational burden in each iteration lies in gradient evaluation, which, in this case, can be performed in time proportional to that taken to read the data.

Despite the simplicity of this algorithm, two critical issues stand out and might significantly affect its efficiency, which we shall bear in mind throughout the algorithmic and theoretical development.

(i) Local stationary points and initialization. As is well known, GD is guaranteed to find an approximate local stationary point, provided that the learning rates do not exceed the inverse Lipschitz constant of the gradient [Bub15]. There exist, however, local stationary points (e.g. saddle points or spurious local minima) that might fall short of the desired statistical properties. This requires us to properly avoid such undesired points, while retaining computational efficiency. To address this issue, one strategy is to first identify a rough initial guess within a local region surrounding the global solution (which often helps rule out bad local minima), in order to guarantee proper convergence of subsequent optimization procedures [LT17, JO14]. As a side remark, while careful initialization might not be crucial for several matrix recovery cases [CCFM19, GBW18, TV19], it does seem to be critical in various tensor problems [RM14]. We shall elucidate this point in Section 3.3.

(ii) Learning rates and regularization. Learning rates play a pivotal role in determining the convergence properties of GD. The challenge, however, is that the loss function (4) is overall not sufficiently smooth (i.e. its gradient often has an exceedingly large Lipschitz constant), and hence generic optimization theory recommends a pessimistically slow update rule (i.e. an extremely small learning rate) so as to guard against over-shooting. This, however, slows down the algorithm significantly, thus destroying the main computational advantage of GD (i.e. low per-iteration cost). With this issue in mind, prior literature suggests carefully designed regularization steps (e.g. proper projection, regularized loss functions) in order to improve the geometry of the optimization landscape [XY17]. In contrast, we argue that one is allowed to take a constant learning rate — which is as aggressive as it can possibly be — even without enforcing any regularization procedures.

Algorithm 1 Gradient descent for nonconvex tensor completion

1: Generate an initial estimate $\boldsymbol{U}^0 \in \mathbb{R}^{d \times r}$ via Algorithm 2.

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

3:     $\boldsymbol{U}^{t+1} = \boldsymbol{U}^t - \eta_t \nabla f(\boldsymbol{U}^t) = \boldsymbol{U}^t - \frac{\eta_t}{p} \mathcal{P}_\Omega \left(\sum_{i=1}^r \left(\boldsymbol{u}_i^t \right)^{\otimes 3} - \boldsymbol{T} \right) \times_1^{\mathsf{seq}} \boldsymbol{U}^t \times_2^{\mathsf{seq}} \boldsymbol{U}^t$, where $\times_1^{\mathsf{seq}}$ and $\times_2^{\mathsf{seq}}$ are defined in Section 2.4.

Algorithm 2 Spectral initialization for nonconvex tensor completion

---

1: Let $\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top$ be the rank-r eigen-decomposition of

$$\boldsymbol{B} := \mathcal{P}_{\mathsf{off\text{-}diag}}(\boldsymbol{A}\boldsymbol{A}^\top), \tag{6}$$

where $\boldsymbol{A} = \mathsf{unfold}\left(p^{-1}\boldsymbol{T}\right)$ is the mode-1 matricization of $p^{-1}\boldsymbol{T}$, and $\mathcal{P}_{\mathsf{off\text{-}diag}}(\boldsymbol{Z})$ extracts out the off-diagonal entries of $\boldsymbol{Z}$.

2: **Output:** an initial estimate $\boldsymbol{U}^0 \in \mathbb{R}^{d \times r}$ on the basis of $\boldsymbol{U} \in \mathbb{R}^{d \times r}$ using Algorithm 3.

Initialization. Motivated by the above-mentioned issue (i), we develop a procedure that guarantees a reasonable initial estimate. In a nutshell, the proposed procedure consists of two steps:

(a) Estimate the subspace spanned by the r low-rank tensor factors $\{\boldsymbol{u}_i^\star\}_{1 \leq i \leq r}$ via a spectral method;

(b) Disentangle individual low-rank tensor factors from this subspace estimate.

As we shall see momentarily, the total computational complexity of the proposed initialization is $O(pd^3)$ when $r = O(1)$, $\kappa = O(1)$ and $p \geq 1/d^2$ (where κ is a sort of "condition number" defined later), which is a linear-time algorithm. Note, however, that these two steps in the initialization procedure are relatively more complicated to describe. To improve the flow of the current paper, we postpone the details to Section 3. The readers can catch a glimpse of these procedures in Algorithms 2-3.

2.2 Main results

Encouragingly, the proposed nonconvex algorithm provably achieves the best of both worlds — in terms of statistical accuracy and computational efficiency — for a class of low-rank, well-conditioned, and "incoherent" problem instances. This subsection summarizes our main findings.

Before continuing, we note that one cannot hope to recover an arbitrary tensor from highly sub-sampled and arbitrarily corrupted entries. In order to enable provably valid recovery, the present paper focuses on a tractable model by imposing the following assumptions.

Definition 2.1 (Incoherence and well-conditionedness). *Define the incoherence parameters and the condition number of $\boldsymbol{T}^\star$ as follows*

$$\mu_0 := \frac{d^3 \|\boldsymbol{T}^\star\|_\infty^2}{\|\boldsymbol{T}^\star\|_{\mathrm{F}}^2}, \tag{8a}$$

$$\mu_1 := \frac{d\|\boldsymbol{u}_i^\star\|_\infty^2}{\|\boldsymbol{u}_i^\star\|_2^2}, \tag{8b}$$

$$\mu_2 := \frac{d\langle \boldsymbol{u}_i^\star, \boldsymbol{u}_j^\star \rangle^2}{\|\boldsymbol{u}_i^\star\|_2^2 \|\boldsymbol{u}_j^\star\|_2^2}, \tag{8c}$$

$$\kappa := \frac{\max_i \|\boldsymbol{u}_i^\star\|_2}{\min_i \|\boldsymbol{u}_i^\star\|_2}. \tag{8d}$$

Remark 2.2. *Here, μ_0, μ_1 and μ_2 are termed the incoherence parameters. Definitions (8a)-(8c) can be viewed as some sort of incoherence conditions for the tensor. For instance, when μ_0, μ_1 and μ_2 are small, these conditions say that (1) the energy of tensor $\boldsymbol{T}^\star$ is (nearly) evenly spread across all entries; (2) each factor $\boldsymbol{u}_i^\star$ is de-localized; (3) the factors $\{\boldsymbol{u}_i^\star\}$ are nearly orthogonal to each other. Definition (8d) is concerned*

Algorithm 3 Retrieval of low-rank tensor factors from a given subspace estimate.

1: **Input:** number of restarts L, pruning threshold $\epsilon_{\sf th}$, subspace estimate $\boldsymbol{U} \in \mathbb{R}^{d \times r}$ given by Algorithm 2.
2: **for** $\tau = 1, \ldots, L$ **do**
3:      Generate an independent Gaussian vector $\boldsymbol{g}^\tau \sim \mathcal{N}(0, \boldsymbol{I}_d)$.
4:      $\left(\boldsymbol{\nu}^\tau, \lambda_\tau, {\sf spec\text{-}gap}_\tau\right) \leftarrow$ Retrieve-one-tensor-factor$(\boldsymbol{T}, p, \boldsymbol{U}, \boldsymbol{g}^\tau)$.
5: Generate tensor factor estimates $\left\{(\boldsymbol{w}^1, \lambda_1), \ldots, (\boldsymbol{w}^r, \lambda_r)\right\} \leftarrow$ Prune$\left(\left\{\left(\boldsymbol{\nu}^\tau, \lambda_\tau, {\sf spec\text{-}gap}_\tau\right)\right\}_{\tau=1}^L, \epsilon_{\sf th}\right)$.
6: **Output:** initial estimate $\boldsymbol{U}^0 = \left[\lambda_1^{1/3} \boldsymbol{w}^1, \ldots, \lambda_r^{1/3} \boldsymbol{w}^r\right]$.

1: **function** Retrieve-one-tensor-factor$(\boldsymbol{T}, p, \boldsymbol{U}, \boldsymbol{g})$
2:      Compute

$$\boldsymbol{\theta} = \boldsymbol{U}\boldsymbol{U}^\top \boldsymbol{g} =: \mathcal{P}_{\boldsymbol{U}}(\boldsymbol{g}), \tag{7a}$$

$$\boldsymbol{M} = p^{-1} \boldsymbol{T} \times_3 \boldsymbol{\theta}, \tag{7b}$$

where $\times_3$ is defined in Section 2.4.
3:      Let $\boldsymbol{\nu}$ be the leading singular vector of $\boldsymbol{M}$ obeying $\langle \boldsymbol{T}, \boldsymbol{\nu}^{\otimes 3} \rangle \geq 0$, and set $\lambda = \langle p^{-1} \boldsymbol{T}, \boldsymbol{\nu}^{\otimes 3} \rangle$.
4:      **return** $\left(\boldsymbol{\nu}, \lambda, \sigma_1(\boldsymbol{M}) - \sigma_2(\boldsymbol{M})\right)$.
---

with the "well-conditionedness" of the tensor, meaning that each rank-1 component is of roughly the same size. In particular, we note that an assumption on pairwise correlation (i.e. a constraint on μ_2) is often assumed in the literature of tensor decomposition/factorization (e.g. [AGJ14, SLLC17, HZC20]).

For notational simplicity, we shall set

$$\mu := \max\left\{\mu_0, \mu_1, \mu_2\right\}. \tag{9}$$

Note that our theory allows μ to grow with the problem dimension d (in fact, μ can be as large as $d/\text{poly}\log(d)$).

Assumption 2.3 (Random noise). *Suppose that $\boldsymbol{E}$ is a symmetric random tensor, where $\{E_{j,k,l}\}_{1 \leq j \leq k \leq l \leq d}$ (cf. (1)) are independently generated sub-Gaussian random variables with mean zero and variance $\mathsf{Var}(E_{j,k,l}) \leq \sigma^2$.*

In addition, recognizing that there is a global permutational ambiguity issue (namely, one cannot distinguish $\boldsymbol{u}_1^\star, \cdots, \boldsymbol{u}_r^\star$ from an arbitrary permutation of them), we introduce the following loss metrics to account for this ambiguity:

$$\mathsf{dist}_{\mathrm{F}}(\boldsymbol{U}, \boldsymbol{U}^\star) := \min_{\boldsymbol{\Pi} \in \mathsf{perm}_r} \|\boldsymbol{U}\boldsymbol{\Pi} - \boldsymbol{U}^\star\|_{\mathrm{F}}, \tag{10a}$$

$$\mathsf{dist}_\infty(\boldsymbol{U}, \boldsymbol{U}^\star) := \min_{\boldsymbol{\Pi} \in \mathsf{perm}_r} \|\boldsymbol{U}\boldsymbol{\Pi} - \boldsymbol{U}^\star\|_\infty, \tag{10b}$$

$$\mathsf{dist}_{2,\infty}(\boldsymbol{U}, \boldsymbol{U}^\star) := \min_{\boldsymbol{\Pi} \in \mathsf{perm}_r} \|\boldsymbol{U}\boldsymbol{\Pi} - \boldsymbol{U}^\star\|_{2,\infty}, \tag{10c}$$

where perm_r stands for the set of $r \times r$ permutation matrices. For notational simplicity, we also take

$$\lambda_{\min}^\star := \min_{1 \leq i \leq r} \|\boldsymbol{u}_i^\star\|_2^3 \qquad \text{and} \qquad \lambda_{\max}^\star := \max_{1 \leq i \leq r} \|\boldsymbol{u}_i^\star\|_2^3. \tag{11}$$

1: **function** Prune$\left(\left\{\left(\boldsymbol{\nu}^\tau, \lambda_\tau, {\sf spec\text{-}gap}_\tau\right)\right\}_{\tau=1}^L, \epsilon_{\sf th}\right)$
2:      Set $\Theta = \left\{\left(\boldsymbol{\nu}^\tau, \lambda_\tau, {\sf spec\text{-}gap}_\tau\right)\right\}_{\tau=1}^L$.
3:      **for** $i = 1, \ldots, r$ **do**
4:          Choose $\left(\boldsymbol{\nu}^\tau, \lambda_\tau, {\sf spec\text{-}gap}_\tau\right)$ from Θ with the largest ${\sf spec\text{-}gap}_\tau$; set $\boldsymbol{w}^i = \boldsymbol{\nu}^\tau$ and $\lambda_i = \lambda_\tau$.
5:          Update $\Theta \leftarrow \Theta \setminus \left\{\left(\boldsymbol{\nu}^\tau, \lambda_\tau, {\sf spec\text{-}gap}_\tau\right) \in \Theta : |\langle \boldsymbol{\nu}^\tau, \boldsymbol{w}^i \rangle| > 1 - \epsilon_{\sf th}\right\}$.
6:      **return** $\left\{(\boldsymbol{w}^1, \lambda_1), \ldots, (\boldsymbol{w}^r, \lambda_r)\right\}$.

With these notations in place, we are ready to present our main results. For simplicity of presentation, we shall start with the setting where $r, \mu, \kappa \asymp 1$.

Theorem 2.4. *Fix an arbitrary small constant $\delta > 0$. Suppose that $r, \kappa, \mu = O(1)$,*

$$p \geq c_0 \frac{\log^4 d}{d^{3/2}}, \qquad \frac{\sigma}{\lambda_{\min}^{\star}} \leq c_1 \frac{\sqrt{p}}{d^{3/4} \log^2 d},$$

$$L = c_2 \quad and \quad \epsilon_{\mathsf{th}} = c_3 \left(\frac{\log d}{d\sqrt{p}} + \frac{\sigma}{\lambda_{\min}^{\star}} \sqrt{\frac{d \log^2 d}{p}} + \sqrt{\frac{\log d}{d}} \right)$$

for some sufficiently large constants $c_0, c_2 > 0$ and some sufficiently small constants $c_1, c_3 > 0$. The learning rate $\eta_t \equiv \eta$ is taken to be a constant obeying $0 < \eta \leq \lambda_{\min}^{\star 4/3} / (32 \lambda_{\max}^{\star 8/3})$. Then with probability at least $1 - \delta$,

$$\mathsf{dist}_{\mathrm{F}}(\boldsymbol{U}^t, \boldsymbol{U}^{\star}) \leq \left(C_1 \rho^t + C_2 \frac{\sigma}{\lambda_{\min}^{\star}} \sqrt{\frac{d \log d}{p}} \right) \|\boldsymbol{U}^{\star}\|_{\mathrm{F}}, \tag{12a}$$

$$\mathsf{dist}_{\infty}(\boldsymbol{U}^t, \boldsymbol{U}^{\star}) \leq \mathsf{dist}_{2,\infty}(\boldsymbol{U}^t, \boldsymbol{U}^{\star}) \leq \left(C_3 \rho^t + C_4 \frac{\sigma}{\lambda_{\min}^{\star}} \sqrt{\frac{d \log d}{p}} \right) \|\boldsymbol{U}^{\star}\|_{2,\infty} \tag{12b}$$

hold simultaneously for all $0 \leq t \leq t_0 = d^5$. Here, $0 < C_1, C_3, \rho < 1$ and $C_2, C_4 > 0$ are some absolute constants.

Remark 2.5. *The theorem holds unchanged if d^5 is replaced by d^c for an arbitrarily large constant $c > 0$.*

Remark 2.6. *The upper bound t_0 on the iteration count arises from the leave-one-out analysis when handling noisy observations. In short, the leave-one-out argument can only provide high-probability bounds for each iteration, thus requiring an upper bound on the iteration count if we desire a uniform bound across iterations. Note that in the noiseless case, our results and analysis hold for an arbitrarily large number of iterations.*

As an immediate consequence of Theorem 2.4, we obtain appealing ℓ_{∞} statistical guarantees for estimating tensor entries, which are previously rarely available (see Table 1). Specifically, let our tensor estimate in the t-th iteration be

$$\boldsymbol{T}^t := \sum_{i=1}^{r} \boldsymbol{u}_i^t \otimes \boldsymbol{u}_i^t \otimes \boldsymbol{u}_i^t, \qquad \text{where } \boldsymbol{U}^t = [\boldsymbol{u}_1^t, \cdots, \boldsymbol{u}_r^t] \in \mathbb{R}^{d \times r}. \tag{13}$$

Then our result is this:

Corollary 2.7. *Fix an arbitrarily small constant $\delta > 0$. Instate the assumptions of Theorem 2.4. Then with probability at least $1 - \delta$,*

$$\left\| \boldsymbol{T}^t - \boldsymbol{T}^{\star} \right\|_{\mathrm{F}} \lesssim \left(C_1 \rho^t + C_2 \frac{\sigma}{\lambda_{\min}^{\star}} \sqrt{\frac{d \log d}{p}} \right) \|\boldsymbol{T}^{\star}\|_{\mathrm{F}}, \tag{14a}$$

$$\left\| \boldsymbol{T}^t - \boldsymbol{T}^{\star} \right\|_{\infty} \lesssim \left(C_3 \rho^t + C_4 \frac{\sigma}{\lambda_{\min}^{\star}} \sqrt{\frac{d \log d}{p}} \right) \|\boldsymbol{T}^{\star}\|_{\infty} \tag{14b}$$

hold simultaneously for all $0 \leq t \leq t_0 = d^5$. Here, $0 < C_1, C_3, \rho < 1$ and $C_2, C_4 > 0$ are some absolute constants.

Several important implications are provided as follows. The discussion below assumes $\lambda_{\max}^{\star} \asymp \lambda_{\min}^{\star} \asymp 1$ for notational simplicity.

1. *Linear convergence.* In the absence of noise, the proposed algorithm converges linearly, namely, it provably attains ε accuracy within $O(\log(1/\varepsilon))$ iterations. Given the inexpensiveness of each gradient iteration, this algorithm can be viewed as a linear-time algorithm, which can almost be implemented as long as we can read the data. In the noisy setting, the algorithm reaches an appealing statistical accuracy within a logarithmic number of iterations.

	algorithm	sample complexity	computational complexity	ℓ_2 error (noisy)	ℓ_∞ error (noisy)	recovery type (noiseless)
our theory	spectral method + (vanilla) GD	$d^{1.5}$	pd^3	$\sigma\sqrt{\frac{d}{p}}$	$\sigma\sqrt{\frac{1}{p}}$	exact
[XYZ17]	spectral initialization + tensor power method	$d^{1.5}$	pd^3	$(\|\boldsymbol{T}^\star\|_\infty + \sigma)\sqrt{\frac{d}{p}}$	n/a	approximate
[XY17]	spectral method + GD on manifold	$d^{1.5}$	$\mathrm{poly}(d)$	n/a	n/a	exact
[MS18]	spectral method	$d^{1.5}$	d^3	n/a	n/a	approximate
[BM16]	sum-of-squares	$d^{1.5}$	d^{15}	$\frac{\|\boldsymbol{T}^\star\|_{\mathrm{F}}}{\sqrt{pd^{1.5}}} + \sigma d^{1.5}$	n/a	approximate
[PS17]	sum-of-squares	$d^{1.5}$	d^{10}	n/a	n/a	exact
[YZ16] [YZ17]	tensor nuclear norm minimization	d	NP-hard	n/a	n/a	exact

Table 1: Comparison with prior theory for existing methods when $r, \mu, \kappa \asymp 1$ (neglecting logarithmic factors).

2. *Near-optimal sample complexity.* The fast convergence is guaranteed as soon as the sample size exceeds the order of $d^{3/2}\mathrm{poly}\log d$. This matches the minimal sample complexity — modulo some logarithmic factor — known so far for any polynomial-time algorithm.

3. *Near-optimal statistical accuracy.* The proposed algorithm converges geometrically fast to a point with Euclidean error $O\big(\sigma\sqrt{(d\log d)/p}\big)$. This matches the lower bound established in [XYZ17, Theorem 5] up to some logarithmic factor, thus justifying the statistical optimality of the proposed nonconvex algorithm.

4. *Entrywise estimation accuracy.* In addition to the Euclidean statistical guarantees, we have also established an entrywise error bound, which, to the best of our knowledge, has not been established in any of the prior work. When t is sufficiently large, the iterates reach an entrywise error bound $O\big(\sigma\sqrt{(\log d)/p}\big)$. This entrywise error bound is about an order of $\sqrt{d}$ times smaller than the above ℓ_2 error bound, thereby implying that the estimation errors are evenly spread out across all entries.

5. *Noise size.* The above theory operates in the regime where $\sigma \lesssim \sqrt{\frac{p}{d^{3/2}}}$ (modulo some log factor). Given that we have $\|\boldsymbol{T}^\star\|_\infty \asymp d^{-3/2}$ in this case, our noise size constraint can be equivalently written as (up to some log factor)

$$\frac{\sigma}{\|\boldsymbol{T}^\star\|_\infty} \lesssim \sqrt{pd^{3/2}}. \tag{15}$$

Since the sampling rate needs to satisfy $p \gg d^{-3/2}$, this condition essentially allows the typical size of each noise component to be considerably larger than the size of the corresponding entry of the truth, which covers a broad range of practical scenarios.

6. *Implicit regularization.* One appealing feature of our finding is the simplicity of the iterative refinement stage of the algorithm. All of the above statistical and computational benefits hold for vanilla gradient descent (when properly initialized). This should be contrasted with prior work (e.g. [XY17]) that relies on extra regularization terms to stabilize the optimization landscape. In principle, vanilla gradient descent implicitly constrains itself within a region of well-conditioned landscape, thus enabling fast convergence without explicit regularization.

7. *No need of sample splitting.* The theory developed herein does not require fresh samples in each iteration. We note that sample splitting has been frequently adopted in other context primarily to simplify mathematical analysis. Nevertheless, it typically does not exploit the data in an efficient manner (i.e. each data sample is used only once), thus resulting in the need of a much larger sample size in practice.

We shall take a moment to discuss the merits of our approach in comparison to prior work. One of the best-known polynomial-time algorithms is the degree-6 level of the sum-of-squares (SoS) hierarchy, which seems to

match the computationally feasible limit in terms of the sample complexity [BM16]. However, this approach has a well-documented limitation in that it involves solving a semidefinite program of dimensions $d^3 \times d^3$, which requires enormous storage and computation power. The work [MS18] alleviates this computational burden by resorting to a clever unfolding-based spectral algorithm; it is a nearly linear-time procedure that enables near-minimal sample complexity (among polynomial-time algorithms), although it does not achieve exact recovery even in the absence of noise. The two-stage algorithm developed by [XYZ17] — which is based on spectral initialization followed by tensor power methods — shares similar advantages and drawbacks as [MS18]. Further, the recent work [XY17] proposes a polynomial-time nonconvex algorithm based on gradient descent over Grassmann manifold (with a properly regularized objective function), which is an extension of the nonconvex matrix completion algorithm proposed by [KMO10a, KMO10b] to tensor data. The theory provided in [XY17], however, does not provide explicit computational complexities. The recent work [SY19] attempts tensor estimation via an interesting algorithm adapted from collaborative filtering and investigates both ℓ_2 and ℓ_∞ estimation accuracy. This approach, however, does not guarantee exact recovery in the absence of noise. We summarize and compare several prior results in Table 1 (omitting logarithmic factors).

Thus far, we have concentrated on the low-rank, well-conditioned, and incoherent case. Our main theory can be extended to cover a broader class of scenarios, as stated below.

Theorem 2.8. *Fix an arbitrary small constant $\delta > 0$. Suppose that $\kappa \asymp 1$,*

$$p \geq c_0 \frac{\mu^4 r^4 \log^4 d}{d^{3/2}}, \qquad \frac{\sigma}{\lambda_{\min}^\star} \leq c_1 \frac{\sqrt{p}}{\mu r^{3/2} d^{3/4} \log^2 d}, \quad r \leq c_2 \left(\frac{d}{\mu^6 \log^6 d} \right)^{1/6},$$

$$L = c_3 r^{2\kappa^2} \log^{3/2} r \qquad and \qquad \epsilon_{\mathsf{th}} = c_4 \left(\frac{\mu r \log d}{d\sqrt{p}} + \frac{\sigma}{\lambda_{\min}^\star} \sqrt{\frac{rd \log^2 d}{p}} + \sqrt{\frac{\mu r \log d}{d}} \right)$$

for some sufficiently large constants $c_0, c_3 > 0$ and some sufficiently small constants $c_1, c_2, c_4 > 0$. The learning rate $\eta_t \equiv \eta$ is taken to be a constant obeying $0 < \eta \leq \lambda_{\min}^{\star 4/3} / (32 \lambda_{\max}^{\star 8/3})$. Then with probability at least $1 - \delta$,

$$\mathsf{dist}_{\mathrm{F}}(\boldsymbol{U}^t, \boldsymbol{U}^\star) \leq \left(C_1 \rho^t + C_2 \frac{\sigma}{\lambda_{\min}^\star} \sqrt{\frac{d \log d}{p}} \right) \|\boldsymbol{U}^\star\|_{\mathrm{F}} \tag{16a}$$

$$\mathsf{dist}_\infty(\boldsymbol{U}^t, \boldsymbol{U}^\star) \leq \mathsf{dist}_{2,\infty}(\boldsymbol{U}^t, \boldsymbol{U}^\star) \leq \left(C_3 \rho^t + C_4 \frac{\sigma}{\lambda_{\min}^\star} \sqrt{\frac{d \log d}{p}} \right) \|\boldsymbol{U}^\star\|_{2,\infty} \tag{16b}$$

hold simultaneously for all $0 \leq t \leq t_0 = d^5$. Here, $0 < C_1, C_3, \rho < 1$ and $C_2, C_4 > 0$ are some absolute constants.

Corollary 2.9. *Fix an arbitrarily small constant $\delta > 0$. Instate the assumptions of Theorem 2.8. Then with probability at least $1 - \delta$,*

$$\left\| \boldsymbol{T}^t - \boldsymbol{T}^\star \right\|_{\mathrm{F}} \lesssim \left(C_1 \rho^t + C_2 \frac{\sigma}{\lambda_{\min}^\star} \sqrt{\frac{d \log d}{p}} \right) \|\boldsymbol{T}^\star\|_{\mathrm{F}}, \tag{17a}$$

$$\left\| \boldsymbol{T}^t - \boldsymbol{T}^\star \right\|_\infty \lesssim \left(C_3 \rho^t + C_4 \frac{\sigma}{\lambda_{\min}^\star} \sqrt{\frac{\mu^3 rd \log d}{p}} \right) \|\boldsymbol{T}^\star\|_\infty \tag{17b}$$

hold simultaneously for all $0 \leq t \leq t_0 = d^5$. Here, $0 < C_1, C_3, \rho < 1$ and $C_2, C_4 > 0$ are some absolute constants.

Remark 2.10. *Clearly, Theorem 2.8 and Corollary 2.9 subsume Theorem 2.4 and Corollary 2.7 as a special case respectively.*

Remark 2.11. *Our theorems require the rank r to not exceed $o(d^{1/6})$, which, we believe, is an artifact of the current nonconvex analysis (particularly for the initialization stage). For instance, our local convergence*

analysis is built upon strong convexity and smoothness, which holds only within a sufficiently small neighbor-hood surrounding the truth; given that the diameter of this neighborhood is no more than $o(1/r)$, our analysis requires an initial guess with higher accuracy than expected, thus leading to our rank constraint. It might be possible to improve the rank dependency via more refined analysis, and we leave it to future investigation.

In a nutshell, this theorem reveals intriguing theoretical support (including both ℓ_{F} and $\ell_{2,\infty}$ bounds) for more general settings. Assuming that the condition number $\kappa \asymp 1$, the nonconvex algorithm we propose is guaranteed to succeed in polynomial time. Note, however, that our theoretical dependency (including both sample and computational complexities) on the rank r and the incoherence parameter μ are likely loose and sub-optimal. In addition, if κ is allowed to grow with d, then the current theory requires a large number of restart attempts during the initialization stage, resulting in a very high computational burden. Improving these aspects, however, calls for a much more refined analysis framework, which we leave for future investigation.

2.3 Numerical experiments

We carry out a series of numerical experiments to corroborate our theoretical findings. Before proceeding, recall that Theorem 2.8 only guarantees successful recovery with probability $1 - \delta$ for some small constant δ; this means that we shall not anticipate a very high success rate (e.g. $1 - O(d^{-5})$) as in the matrix recovery case. As we shall make clear shortly, this happens mainly because the initialization stage works only with probability $1 - \delta$, where the uncertainty largely depends on the random vectors $\{\boldsymbol{g}^\tau\}_{1 \le \tau \le L}$. With this observation in mind, we recommend the following modification to improve the empirical success rate:

- Run Algorithm 2 independently for $t_{\mathrm{init}} = 5$ times to obtain multiple initial estimates (denoted by $\boldsymbol{U}^0_{[1]}, \cdots, \boldsymbol{U}^0_{[t_{\mathrm{init}}]}$); select the one achieving the smallest empirical loss, namely

$$\boldsymbol{U}^0_{\mathsf{best}} = \underset{\boldsymbol{U} \in \{\boldsymbol{U}^0_{[i]}\}_{1 \le i \le t_{\mathrm{init}}}}{\arg\min} f(\boldsymbol{U}). \tag{18}$$

- Run Algorithm 1 with the initial point $\boldsymbol{U}^0$ set to be $\boldsymbol{U}^0_{\mathsf{best}}$.

The final estimates for the low-rank factor and the whole tensor are denoted respectively by

$$\widehat{\boldsymbol{U}} = \boldsymbol{U}^{t_0} \qquad \text{and} \qquad \widehat{\boldsymbol{T}} = \sum_{i=1}^r \boldsymbol{u}^{t_0}_i \otimes \boldsymbol{u}^{t_0}_i \otimes \boldsymbol{u}^{t_0}_i, \tag{19}$$

where $\boldsymbol{U}^{t_0} = [\boldsymbol{u}^{t_0}_1, \cdots, \boldsymbol{u}^{t_0}_r] \in \mathbb{R}^{d \times r}$ is the iterate returned by Algorithm 1, with t_0 the total number of gradient iterations. In the sequel, we generate the true tensor $\boldsymbol{T}^\star = \sum_{1 \le i \le r} \boldsymbol{u}^{\star \otimes 3}_i$ randomly in such a way that $\boldsymbol{u}^\star_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$. The learning rates are taken to be $\eta_t \equiv 0.2$ unless otherwise noted.

We start with numerical convergence rates of our algorithm in the absence of noise. Set $d = 100$, $r = 4$, $p = 0.1$, $L = 16$ and $\epsilon_{\mathsf{th}} = 0.4$. Figure 1(a) the numerical estimation errors vs. iteration count t in a typical Monte Carlo trial. Here, four kinds of estimation errors are reported: (1) the relative Frobenius norm error $\frac{\mathsf{dist}_{\mathrm{F}}(\boldsymbol{U}^t, \boldsymbol{U}^\star)}{\|\boldsymbol{U}^\star\|_{\mathrm{F}}}$; (2) the relative $\|\cdot\|_{2,\infty}$ error $\frac{\mathsf{dist}_{2,\infty}(\boldsymbol{U}^t, \boldsymbol{U}^\star)}{\|\boldsymbol{U}^\star\|_{2,\infty}}$; (3) the relative Frobenius norm error $\frac{\|\boldsymbol{T}^t - \boldsymbol{T}^\star\|_{\mathrm{F}}}{\|\boldsymbol{T}^\star\|_{\mathrm{F}}}$; (4) the relative ℓ_∞ error $\frac{\|\boldsymbol{T}^t - \boldsymbol{T}^\star\|_\infty}{\|\boldsymbol{T}^\star\|_\infty}$. Here, $\boldsymbol{T}^t = \sum_{i=1}^r \boldsymbol{u}^t_i \otimes \boldsymbol{u}^t_i \otimes \boldsymbol{u}^t_i$ with $\boldsymbol{U}^t = [\boldsymbol{u}^t_1, \cdots, \boldsymbol{u}^t_r]$. For all these metrics, the numerical estimation errors decay geometrically fast.

Next, we study the phase transition (in terms of the success rates for exact recovery) in the noise-free settings. Set $d = 100$, $r = 4$, $L = 16$ and $\epsilon_{\mathsf{th}} = 0.4$. For the sake of comparisons, we also report the numerical performance of the tensor power method (TPM) followed by gradient descent. When running the tensor power method, we set both the number of iterations and the restart number to be 16. Each trial is claimed to succeed if the relative ℓ_2 error obeys $\frac{\mathsf{dist}_{\mathrm{F}}(\widehat{\boldsymbol{U}}, \boldsymbol{U}^\star)}{\|\boldsymbol{U}^\star\|_{\mathrm{F}}} \le 0.01$. Figure 1(b) plots the empirical success rates over 100 independent Monte Carlo trials. As can be seen, our initialization algorithm outperforms the tensor power method.

The third series of experiments is concerned with the dependence of the success rate on the rank r. Let us set $p = rd^{-3/2} \log^2 d$, $L = r^2$ and $\epsilon_{\mathsf{th}} = 0.4$, and the success recovery criterion is the same as above. Figure 1(c) depicts the empirical success rates (over 100 independent Monte Carlo trials) as the rank r varies. As
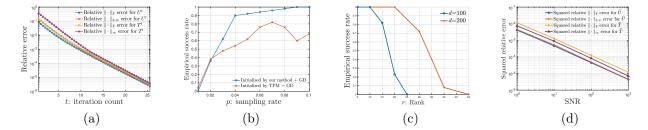
Figure 1: (a) relative errors of the estimates $\boldsymbol{U}^t$ and $\boldsymbol{T}^t$ vs. iteration count t for noiseless tensor completion, where $d = 100$, $r = 4$, $p = 0.1$; (b) empirical success rate vs. sampling rate, where $d = 100$, $r = 4$; (c) empirical success rate vs. rank, where $p = rd^{-3/2}\log^2 d$; (d) squared relative errors vs. SNR for noisy settings, where $d = 100$, $r = 4$ and $p = 0.1$. Each point in (b), (c) and (d) is averaged over 100 independent Monte Carlo trials.

can be seen from the plots, the proposed algorithm is able to achieve exact reconstruction as long as the rank r is sufficiently small compared to d. The plausible range of r, however, seems and seems to be larger than our theoretic requirement $r = o(d^{1/6})$. This, once again, suggests the need of future investigation to pin down the best possible dependency on r.

Finally, we consider the numerical estimation accuracy of our algorithm. Take $t_0 = 100$, $d = 100$, $r = 4$, $p = 0.1$, $L = 16$ and $\epsilon_{\mathsf{th}} = 0.4$. Define the signal-to-noise ratio (SNR) to be SNR $= \frac{\|\boldsymbol{T}^\star\|_{\mathrm{F}}^2/d^3}{\sigma^2}$. We report in Figure 1(d) three types of squared relative errors (namely, $\frac{\mathsf{dist}_{\mathrm{F}}^2(\widehat{\boldsymbol{U}}, \boldsymbol{U}^\star)}{\|\boldsymbol{U}^\star\|_{\mathrm{F}}^2}$, $\frac{\mathsf{dist}_{2,\infty}^2(\widehat{\boldsymbol{U}}, \boldsymbol{U}^\star)}{\|\boldsymbol{U}^\star\|_{2,\infty}^2}$ and $\frac{\|\widehat{\boldsymbol{T}} - \boldsymbol{T}^\star\|_\infty^2}{\|\boldsymbol{T}^\star\|_\infty^2}$) vs. SNR. Figure 1(d) illustrates that all three types of relative squared errors scale inversely proportional to the SNR (since the slope in the figure is roughly -1), which is consistent with our statistical guarantees.

2.4 Notation

Before proceeding, we gather a few notations that will be used throughout this paper. First of all, for any matrix $\boldsymbol{M} \in \mathbb{R}^{d \times d}$, we let $\|\boldsymbol{M}\|$ and $\|\boldsymbol{M}\|_{\mathrm{F}}$ denote the operator norm (or the spectral norm) and the Frobenius norm of $\boldsymbol{M}$, respectively, and let $\boldsymbol{M}_{i,:}$ and $\boldsymbol{M}_{:,i}$ denote the i-th row and i-th column, respectively. In addition, $\lambda_1(\boldsymbol{M}) \geq \lambda_2(\boldsymbol{M}) \geq \cdots \geq \lambda_d(\boldsymbol{M})$ denote the eigenvalues of $\boldsymbol{M}$ and $\sigma_1(\boldsymbol{M}) \geq \sigma_2(\boldsymbol{M}) \geq \cdots \geq \sigma_d(\boldsymbol{M})$ denote the singular values of $\boldsymbol{M}$.

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-slice with entries $(\boldsymbol{T}_{i,:,:})_{j,k} = T_{i,j,k}$, and $\boldsymbol{T}_{:,i,:}$ and $\boldsymbol{T}_{:,:,i}$ are defined in a similar way. For any tensors $\boldsymbol{T}, \boldsymbol{R} \in \mathbb{R}^{d \times d \times d}$, the inner product is defined as $\langle \boldsymbol{T}, \boldsymbol{R} \rangle := \sum_{j,k,l} T_{j,k,l} R_{j,k,l}$. The Frobenius norm of $\boldsymbol{T}$ is defined as $\|\boldsymbol{T}\|_{\mathrm{F}} := \sqrt{\langle \boldsymbol{T}, \boldsymbol{T} \rangle}$. For any vectors $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^d$, we define the vector products of a tensor $\boldsymbol{T} \in \mathbb{R}^{d \times d \times d}$ — denoted by $\boldsymbol{T} \times_3 \boldsymbol{u} \in \mathbb{R}^{d \times d}$ and $\boldsymbol{T} \times_1 \boldsymbol{u} \times_2 \boldsymbol{v} \in \mathbb{R}^d$ — such that

$$\left[\boldsymbol{T} \times_3 \boldsymbol{u}\right]_{ij} := \sum\nolimits_{1 \leq k \leq d} T_{i,j,k} u_k, \qquad 1 \leq i, j \leq d; \qquad (20\mathrm{a})$$

$$\left[\boldsymbol{T} \times_1 \boldsymbol{u} \times_2 \boldsymbol{v}\right]_k := \sum\nolimits_{1 \leq i,j \leq d} T_{i,j,k} u_i v_j, \qquad 1 \leq k \leq d. \qquad (20\mathrm{b})$$

The products $\boldsymbol{T} \times_2 \boldsymbol{u} \in \mathbb{R}^{d \times d}$, $\boldsymbol{T} \times_3 \boldsymbol{u} \in \mathbb{R}^{d \times d}$, $\boldsymbol{T} \times_1 \boldsymbol{u} \times_3 \boldsymbol{v} \in \mathbb{R}^d$ and $\boldsymbol{T} \times_2 \boldsymbol{u} \times_3 \boldsymbol{v} \in \mathbb{R}^d$ are defined in a similar manner. For any $\boldsymbol{U} = [\boldsymbol{u}_1, \cdots, \boldsymbol{u}_r] \in \mathbb{R}^{d \times r}$ and $\boldsymbol{V} = [\boldsymbol{v}_1, \cdots, \boldsymbol{v}_r] \in \mathbb{R}^{d \times r}$, we further define

$$\boldsymbol{T} \times_1^{\mathsf{seq}} \boldsymbol{U} \times_2^{\mathsf{seq}} \boldsymbol{V} := [\boldsymbol{T} \times_1 \boldsymbol{u}_i \times_2 \boldsymbol{v}_i]_{1 \leq i \leq r} \in \mathbb{R}^{d \times r}. \qquad (21)$$

In addition, the operator norm of $\boldsymbol{T}$ is defined as

$$\|\boldsymbol{T}\| := \sup_{\boldsymbol{u},\boldsymbol{v},\boldsymbol{w} \in \mathbb{S}^{d-1}} \langle \boldsymbol{T}, \boldsymbol{u} \otimes \boldsymbol{v} \otimes \boldsymbol{w} \rangle, \qquad (22)$$

where $\mathbb{S}^{d-1} := \{\boldsymbol{u} \in \mathbb{R}^d \mid \|\boldsymbol{u}\|_2 = 1\}$ indicates the unit sphere in $\mathbb{R}^d$.

12

Further, $f(n) \lesssim g(n)$ or $f(n) = O(g(n))$ means that $|f(n)/g(n)| \leq C_1$ for some constant $C_1 > 0$; $f(n) \gtrsim g(n)$ means that $|f(n)/g(n)| \geq C_2$ for some constant $C_2 > 0$; $f(n) \asymp g(n)$ means that $C_1 \leq |f(n)/g(n)| \leq C_2$ for some constants $C_1, C_2 > 0$; $f(n) = o(g(n))$ means that $\lim_{n\to\infty} f(n)/g(n) = 0$. In addition, $f(n) \ll g(n)$ means that $f(n) \leq c_1 g(n)$ for some sufficiently small constant $c_1 > 0$, and $f(n) \gg g(n)$ means that $f(n) \geq c_2 g(n)$ for some sufficiently large constant $c_2 > 0$.

3 Initialization

This section presents formal details of the proposed two-step initialization, accompanied by some intuition. Recall that the proposed initialization procedure consists of two steps, which we discuss separately.

3.1 Step 1: subspace estimation via a spectral method

The spectral algorithm is often applied in conjunction with simple "unfolding" (or "matricization") to estimate the *subspace* spanned by the r factors $\{u_i^\star\}_{1 \leq i \leq r}$. This strategy is partly motivated by prior approaches developed for covariance estimation with missing data [Lou14, MS18, CLC$^+$20]. We provide a brief introduction below.

Let

$$\boldsymbol{A} = \mathsf{unfold}^{1\times 2}\big(\tfrac{1}{p}\boldsymbol{T}\big) \in \mathbb{R}^{d\times d^2}, \quad \text{or more concisely} \quad \boldsymbol{A} = \mathsf{unfold}\big(\tfrac{1}{p}\boldsymbol{T}\big) \in \mathbb{R}^{d\times d^2} \tag{23}$$

be the mode-1 matricization of $p^{-1}\boldsymbol{T}$ (namely, $\frac{1}{p}T_{i,j,k} = A_{i,(j-1)d+k}$ for any $1 \leq i,j,k \leq d$) [KB09]. The rationale of this step is that: under our model, the unfolded matrix $\boldsymbol{A}$ obeys

$$\mathbb{E}[\boldsymbol{A}] = \mathsf{unfold}\big(\boldsymbol{T}^\star\big) = \sum_{i=1}^r \boldsymbol{u}_i^\star \left(\boldsymbol{u}_i^\star \otimes \boldsymbol{u}_i^\star\right)^\top =: \boldsymbol{A}^\star, \tag{24}$$

whose column space is precisely the span of $\{u^\star\}_{1 \leq i \leq r}$. This motivates one to estimate the r-dimensional column space of $\mathbb{E}[\boldsymbol{A}]$ from $\boldsymbol{A}$. Towards this, a natural strategy is to look at the principal subspace of $\boldsymbol{A}\boldsymbol{A}^\top$. However, the diagonal entries of $\boldsymbol{A}\boldsymbol{A}^\top$ bear too much influence on the principal directions and need to be properly down-weighed. The current paper chooses to work with the principal subspace of the following matrix that zeros out all diagonal components:

$$\boldsymbol{B} := \mathcal{P}_{\mathsf{off\text{-}diag}}(\boldsymbol{A}\boldsymbol{A}^\top), \tag{25}$$

where $\mathcal{P}_{\mathsf{off\text{-}diag}}(\boldsymbol{Z})$ extracts out the off-diagonal entries of a squared matrix $\boldsymbol{Z}$. If we let $\boldsymbol{U} \in \mathbb{R}^{d\times r}$ be an orthonormal matrix whose columns are the top-r eigenvectors of $\boldsymbol{B}$, then $\boldsymbol{U}$ serves as our subspace estimate. See Algorithm 2 for a summary of the procedure.

3.2 Step 2: retrieval of low-rank tensor factors from the subspace estimate

3.2.1 Procedure

As it turns out, it is possible to obtain rough (but reasonable) estimates of all individual low-rank tensor factors $\{u_i^\star\}_{1 \leq i \leq r}$ — up to global permutation — given a reliable subspace estimate $\boldsymbol{U}$. This is in stark contrast to the low-rank matrix recovery case, where there exists some global rotational ambiguity that prevents us from disentangling the r factors of interest.

We begin by describing how to retrieve *one* tensor factor from the subspace estimate — a procedure summarized in Retrieve-one-tensor-factor(). Let us generate a random vector from the provided subspace $\boldsymbol{U}$ (which has orthonormal columns), that is,

$$\boldsymbol{\theta} = \underbrace{\boldsymbol{U}\boldsymbol{U}^\top\boldsymbol{g}}_{\text{projection of } \boldsymbol{g} \text{ onto } \boldsymbol{U}}, \qquad \boldsymbol{g} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d). \tag{26}$$

13

The rescaled tensor data $p^{-1}\boldsymbol{T}$ is then transformed into a matrix via proper "projection" along this random direction $\boldsymbol{\theta}$, namely,

$$\boldsymbol{M} = \tfrac{1}{p}\boldsymbol{T} \times_3 \boldsymbol{\theta} \in \mathbb{R}^{d\times d}. \tag{27}$$

Our estimate for a tensor factor is then given by $\lambda^{1/3}\boldsymbol{\nu}$, where $\boldsymbol{\nu}$ is the leading singular vector of $\boldsymbol{M}$ obeying $\langle \boldsymbol{T}, \boldsymbol{\nu}^{\otimes 3}\rangle \geq 0$, and λ is taken as $\lambda = \langle p^{-1}\boldsymbol{T}, \boldsymbol{\nu}^{\otimes 3}\rangle$. Informally, $\boldsymbol{\nu}$ reflects the direction of the component $\boldsymbol{u}_i^\star$ that exhibits the largest correlation with the random direction $\boldsymbol{\theta}$, and λ forms an estimate of the corresponding size $\|\boldsymbol{u}_i^\star\|_2$.

A challenge remains, however, as there are oftentimes more than one tensor factors to estimate. To address this issue, we propose to re-run the aforementioned procedure multiple times, so as to ensure that we get to retrieve each tensor factor of interest at least once. We will then apply a careful pruning procedure (i.e. PRUNE()) to remove redundancy.

3.2.2 Intuition

To develop some intuition about the above procedure, consider the "heuristic" case where $\boldsymbol{\theta} = \boldsymbol{U}^\star(\boldsymbol{U}^{\star\top}\boldsymbol{U}^\star)^{-1}\boldsymbol{U}^{\star\top}\boldsymbol{g}$, namely, the idealistic scenario where the subspace estimate $\boldsymbol{U}$ is accurate. Averaging out the randomness in the sampling pattern and the noise, we see that the expected projected matrix (27) takes the following form:

$$\mathbb{E}\big[\boldsymbol{M} \mid \boldsymbol{\theta}\big] = \boldsymbol{T}^\star \times_3 \boldsymbol{\theta} = \sum_{i=1}^r \langle \boldsymbol{\theta}, \boldsymbol{u}_i^\star\rangle \boldsymbol{u}_i^\star \boldsymbol{u}_i^{\star\top}.$$

As a result, in the incoherent case where $\{\boldsymbol{u}_j^\star\}$ are nearly orthogonal to each other, the leading singular vector of $\mathbb{E}\big[\boldsymbol{M} \mid \boldsymbol{\theta}\big]$ — and hence that of $\boldsymbol{M}$ (i.e. $\boldsymbol{w}$) — is expected to be reasonably close to the factor $\boldsymbol{u}_i^\star$ that enjoys the largest projected coefficient. In other words, we expect

$$\boldsymbol{\nu} \approx \frac{1}{\|\boldsymbol{u}_i^\star\|_2}\boldsymbol{u}_i^\star, \qquad \text{where } i = \arg\max_{1\leq j\leq r}\big|\langle \boldsymbol{\theta}, \boldsymbol{u}_j^\star\rangle\big|. \tag{28}$$

In the mean time, armed with (28) and the incoherence assumption (such that $\boldsymbol{u}_i^\star$ and $\boldsymbol{u}_j^\star$ are nearly orthogonal for $i \neq j$), one might have

$$\lambda = \langle \boldsymbol{T}^\star, \boldsymbol{\nu}^{\otimes 3}\rangle \approx \frac{1}{\|\boldsymbol{u}_i^\star\|_2^3}\langle \boldsymbol{T}^\star, \boldsymbol{u}_i^{\star\otimes 3}\rangle \approx \frac{1}{\|\boldsymbol{u}_i^\star\|_2^3}\langle \boldsymbol{u}_i^{\star\otimes 3}, \boldsymbol{u}_i^{\star\otimes 3}\rangle = \|\boldsymbol{u}_i^\star\|_2^3, \tag{29}$$

thus explaining our choice of λ in the proposed procedure. These arguments hint at the ability of our procedure in retrieving one tensor factor in each round.

The above intuitive argument, however, does not explain why we need to first project a random vector $\boldsymbol{g}$ onto the (approximate) column space of $\boldsymbol{U}^\star$. While we won't go into detailed calculations here, we remark in passing a crucial high variability issue: without proper projection, the perturbation incurred by both the missing data and the noise might far exceed the strength of the true signal. As a result, it is advised to first project the data onto the desired subspace, in the hope of amplifying the signal-to-noise ratio.

3.3 Other alternatives?

The careful reader may naturally wonder whether a careful initialization is pivotal in achieving fast convergence. While a thorough answer to this has yet to be developed, we shall point out some alternatives that seem sub-optimal in both theory and practice. To simplify the presentation, the current subsection focuses on the rank-1 noiseless case, where

$$\boldsymbol{T}^\star = \boldsymbol{u}^{\star\otimes 3}, \qquad \boldsymbol{T} = \tfrac{1}{p}\mathcal{P}_\Omega(\boldsymbol{T}^\star), \qquad \|\boldsymbol{u}^\star\|_2 = 1. \tag{30}$$

Since the decision variable is now a d-dimensional vector, we shall employ the conventional notation $\boldsymbol{u}^t$ to represent $\boldsymbol{U}^t$.

Random initialization. We find it instrumental to begin with the population-level analysis, which corresponds to the scenario with no missing data and noise ($p = 1$ and $\sigma = 0$). A little calculation gives

$$\mathbb{E}\left[\boldsymbol{u}^1 \mid \boldsymbol{u}^0\right] = \mathbb{E}\left[\boldsymbol{u}^0 - \eta \nabla f(\boldsymbol{u}^0) \mid \boldsymbol{u}^0\right] = \left(1 - \eta\|\boldsymbol{u}^0\|_2^4\right)\boldsymbol{u}^0 + \eta\langle \boldsymbol{u}^0, \boldsymbol{u}^\star\rangle^2 \boldsymbol{u}^\star. \tag{31}$$

As an immediate consequence, the expected correlation between the next iterate and the truth obeys

$$\mathbb{E}\left[\langle \boldsymbol{u}^1, \boldsymbol{u}^\star\rangle \mid \boldsymbol{u}^0\right] = \left\{1 - \eta\|\boldsymbol{u}^0\|_2^4 + \eta\langle \boldsymbol{u}^0, \boldsymbol{u}^\star\rangle\|\boldsymbol{u}^\star\|_2^2\right\}\langle \boldsymbol{u}^0, \boldsymbol{u}^\star\rangle.$$

This means that if $\boldsymbol{u}^0$ and $\boldsymbol{u}^\star$ are positively correlated and if the initial guess $\boldsymbol{u}^0$ is sufficiently small,[3] then one has

$$\mathbb{E}\left[\langle \boldsymbol{u}^1, \boldsymbol{u}^\star\rangle \mid \boldsymbol{u}^0\right] \approx \left(1 + \eta\langle \boldsymbol{u}^0, \boldsymbol{u}^\star\rangle\|\boldsymbol{u}^\star\|_2^2\right)\langle \boldsymbol{u}^0, \boldsymbol{u}^\star\rangle; \tag{32}$$

a similar recursion holds for $\boldsymbol{u}^t$. As a result, the GD iterates are expected to get increasingly more aligned with the truth, at least at the population level. Caution needs to be exercised, however, that this population-level analysis alone fails to capture what is happening in the finite-sample case. In what follows, we point out potential issues with random initialization.

Consider the case where $\boldsymbol{u}^0$ is generated as a vector of i.i.d. Gaussian random variables. Suppose that $\boldsymbol{u}^0$ and $\boldsymbol{u}^\star$ are positively correlated and that $\|\boldsymbol{u}^0\|_2$ is sufficiently small. It is easily seen that, with high probability, the expected increment is on the order of (cf. (32))

$$\mathbb{E}\left[\langle \boldsymbol{u}^1, \boldsymbol{u}^\star\rangle \mid \boldsymbol{u}^0\right] - \langle \boldsymbol{u}^0, \boldsymbol{u}^\star\rangle \approx \eta\langle \boldsymbol{u}^0, \boldsymbol{u}^\star\rangle^2\|\boldsymbol{u}^\star\|_2^2 \lesssim \frac{\eta\, \mathrm{poly}\log(d)}{d}\|\boldsymbol{u}^\star\|_2^4\|\boldsymbol{u}^0\|_2^2, \tag{33}$$

which could be quite small as it depends quadratically on the current correlation $\langle \boldsymbol{u}^0, \boldsymbol{u}^\star\rangle$.

If we were to hope that the favorable population-level analysis captures more or less the finite-sample dynamics, we would need to ensure that the variability of the gradient update is well-controlled. Towards this, let us compute the variance of $\langle \boldsymbol{u}^1, \boldsymbol{u}^\star\rangle$, assuming that $\frac{\|\boldsymbol{u}^0\|_\infty}{\|\boldsymbol{u}^0\|_2} \asymp \frac{\|\boldsymbol{u}^\star\|_\infty}{\|\boldsymbol{u}^\star\|_2} \asymp \frac{\mathrm{poly}\log(d)}{\sqrt{d}}$:

$$\mathsf{Var}\left(\langle \boldsymbol{u}^1, \boldsymbol{u}^\star\rangle \mid \boldsymbol{u}^0\right) \asymp \mathsf{Var}\left(\frac{\eta}{p}\sum_{1 \le j,k,l \le d}\left(\chi_{jkl} - p\right)\left(u_j^0 u_k^0 u_l^0 - u_j^\star u_k^\star u_l^\star\right)u_j^0 u_k^0 u_l^\star\right)$$

$$\asymp \frac{\eta^2}{p}\sum_{1 \le j,k,l \le d}\left(u_j^0 u_k^0 u_l^0 + u_j^\star u_k^\star u_l^\star\right)^2 \left(u_j^0 u_k^0 u_l^\star\right)^2 \asymp \frac{\eta^2\mathrm{poly}\log(d)}{pd^3}\|\boldsymbol{u}^\star\|_2^8\|\boldsymbol{u}^0\|_2^4.$$

In other words, the typical size of the variability of $\langle \boldsymbol{u}^1, \boldsymbol{u}^\star\rangle$ is about the order of $\frac{\eta\mathrm{poly}\log(d)}{\sqrt{pd^3}}\|\boldsymbol{u}^\star\|_2^4\|\boldsymbol{u}^0\|_2^2$, which dominates (in fact, is order-of-magnitudes larger than) the mean increment (33) unless

$$p \gtrsim \frac{\mathrm{poly}\log(d)}{d}. \tag{34}$$

The sample size corresponding to (34) is, however, considerably larger than the computation limit $p \asymp \frac{\mathrm{poly}\log(d)}{d^{1.5}}$. The presence of a large variance implies highly volatile dynamics of randomly initialized GD, thus casting doubt on its efficiency in the most challenging sample-starved regime.

In summary, the main issue stems from the quadratic dependence of the expected increment (33) on the correlation $\langle \boldsymbol{u}^0, \boldsymbol{u}^\star\rangle$, which can be exceedingly small if $\boldsymbol{u}^0$ is randomly initialized.

Initialization via the tensor power method (TPM). Another alternative for initialization is the tensor power method, which has recently gained popularity in the context of learning latent-variable models [AGH+14, AGJ17]. Nevertheless, the TPM (with random initialization) suffers from the same high-volatility issue as randomly initialized GD. The argument for this would be nearly identical to the one presented

---

[3]In fact, if a random initialization $\boldsymbol{u}^0$ is not small, then one can easily show that, with high probability, the ℓ_2 norm of $\boldsymbol{u}^t$ is going to drop geometrically fast at the beginning.

above, and is hence omitted. Instead, we invoke a perturbation analysis result in [AGH⁺14, Theorem 5.1] to illustrate the insufficiency of the TPM.

Recall that $\frac{1}{p}\boldsymbol{T} = \boldsymbol{T}^\star + \left(\frac{1}{p}\boldsymbol{T} - \boldsymbol{T}^\star\right)$. A critical issue is that the perturbation bound in [AGH⁺14, Theorem 5.1] requires the tensor perturbation to be exceedingly small, namely,

$$\big\|\tfrac{1}{p}\boldsymbol{T} - \boldsymbol{T}^\star\big\| \lesssim 1/d. \tag{35}$$

This, however, cannot possibly hold if the sample size is merely $p \asymp \frac{\mathrm{poly}\log(d)}{d^{1.5}}$ (in which case one only expects a spectral norm bound on the order of $\|p^{-1}\boldsymbol{T} - \boldsymbol{T}^\star\| \lesssim \frac{1}{\mathrm{poly}\log(d)}$ shown in Corollary D.3 even in the absence of noise). In light of all this, existing stability analysis of the TPM does not imply either sample efficiency or computational efficiency.

4 Related work

One of the most natural ideas for solving tensor completion is to first unfold the tensor data into matrices, followed by proper convex relaxation commonly adopted for low-rank matrix completion. Given that there are more than one ways to matricize a tensor, several prior work has explored the design of matrix norms that can exploit the tensor structure more effectively [THK10, GRY11, LMWY13, RPP13, LFC⁺16, MHWG14]. Such algorithms have been robustified to enable reliable recovery against sparse outliers as well [GQ14]. For the most part, however, such unfolding-based convex relaxation necessarily incur loss of structural information, which is particularly severe when handling odd-order tensors. The sample complexity developed for this paradigm is often sub-optimal vis-a-vis the computational limits (namely, minimal sample complexity achievable by polynomial-time algorithms).

Motivated by the above sub-optimality issue, [YZ16, YZ17] proposed to minimize instead the tensor nuclear norm subject to data constraints, which provably allows for reduced sample complexity. The issue, however, is that computing the tensor nuclear norm itself is already computationally intractable, thus limiting its applicability to even moderate-dimensional problems. Similar findings have also been discovered for tensor atomic norm minimization [DBBG19]. When restricted to polynomial-time algorithms, the best statistical guarantees are often attained via convex relaxation tailored to the sum-of-squares hierarchy [BM16]; the resulting computational cost, however, remains prohibitively high for practical large-scale problems. Another matrix nuclear norm minimization algorithm has been proposed based on promoting certain structures on certain factor matrices [LSC⁺14]. Developing statistical guarantees is, however, not the focal point of this work.

Moving beyond convex relaxation, a number of prior papers have developed nonconvex algorithms for tensor completion, examples including iterative hard thresholding [RSS17], alternating minimization [JO14, WAA16, XHYS15], tensor SVD [ZA17], optimization on manifold [XY17, KM16, Ste16], proximal average algorithm with nonconvex regularizer [Yao18], and block coordinate decent [JHZ⁺16, XY13]. When it comes to the model considered herein, these algorithms either lack optimal statistical guarantees, or come with a computational cost that is significantly higher than a linear-time algorithm.

The algorithm and theory that we develop are largely inspired by the recent advances of nonconvex optimization algorithms for low-rank matrix recovery problems [KMO10a, KMO10b, CLS15, CC17, SL16, YPCC16, CW15]. The main theoretical tool — the leave-one-out analysis — is a powerful technique that has proved successful in various other statistical problems [EK15, CFMW19, AFWZ17, MWCC17, ZB18, CCFM19, CCF⁺19, LZT19, CFMY19, DC18, PW19]. There are several major differences between the analysis of nonconvex tensor completion and that of nonconvex matrix recovery. For instance, our initialization scheme is substantially more complicated than the matrix recovery counterpart, thus requiring much more sophisticated analysis; in addition, the local convergence stage of tensor completion does not suffer from rotational ambiguity (which often appears in nonconvex matrix completion), and hence we only need to handle permutational ambiguity.

In addition, the current paper focuses on non-adaptive uniform random sampling. If there is freedom in designing the sampling mechanism, then one can often expect improved performance; see [KS13, Zha19] as examples. Fundamental criteria that enable perfect low-CP-rank tensor completion have been studied in [AW17].

Tensor completion is simply a special example of the tensor recovery literature. There is a large body of results tackling various other tensor recovery and estimation problems, including but not limited to tensor decomposition [Kol01, KB09, AGH+14, AGJ14, TS15, KOKC13, HSSS16, GHJY15, ZKOM18, SDLF+17, SLLC17, GM17], tensor SVD and factorization [ZX18, KBHH13, ZA17], and tensor regression and sketching [RSS17, HZC20, CRY19, HWW+19]. The algorithmic ideas explored in this paper might have implications for these tensor-related problems as well.

5 Analysis

In this section, we outline the proof of Theorem 2.8. The proof of Corollary 2.9 is deferred to Appendix C. The analysis is divided into three parts:

- In Section 5.1, we show that given an initial estimate sufficiently close to the ground truth, vanilla gradient descent converges linearly. These are formalized in Lemmas 5.3 and 5.6.

- Sections 5.2-5.3 provide statistical guarantees for the two steps of the initialization procedure; see Theorems 5.9.

- Under the assumptions of Theorem 2.8, one can see that the initialization satisfies the requirement of linear convergence of vanilla gradient descent. Therefore, Theorem 2.8 immediately follows from the results in Sections 5.1-5.3.

5.1 Analysis for local convergence of GD

In this section, we demonstrate that: if the initialization is reasonably good, then vanilla gradient descent converges linearly to a solution with the desired statistical accuracy. We postpone the analysis for initialization to Sections 5.2-5.3 for convenience of presentation.

5.1.1 Preliminaries: gradient and Hessian calculation

First of all, using our notation $\times^{\mathsf{seq}}$ defined in (21), we can write

$$\nabla f\left(\boldsymbol{U}\right) = \frac{1}{p}\mathcal{P}_{\Omega}\Big(\sum\nolimits_{1\leq i\leq r}\boldsymbol{u}_i^{\otimes 3} - \boldsymbol{T}^{\star} - \boldsymbol{E}\Big) \times_1^{\mathsf{seq}} \boldsymbol{U} \times_2^{\mathsf{seq}} \boldsymbol{U}. \tag{36}$$

Next, we find it convenient to define an auxiliary loss function $f_{\mathsf{clean}}\left(\boldsymbol{U}\right) : \mathbb{R}^{d\times r} \to \mathbb{R}_+$ that corresponds to the noiseless case:

$$f_{\mathsf{clean}}\left(\boldsymbol{U}\right) = \frac{1}{6p}\Big\|\mathcal{P}_{\Omega}\Big(\sum\nolimits_{1\leq i\leq r}\boldsymbol{u}_i^{\otimes 3} - \boldsymbol{T}^{\star}\Big)\Big\|_{\mathrm{F}}^2. \tag{37}$$

The gradient of f_{clean} w.r.t. $\boldsymbol{u}_s$ $(1 \leq s \leq r)$ is thus given by

$$\nabla_{\boldsymbol{u}_s} f_{\mathsf{clean}}\left(\boldsymbol{U}\right) = \frac{1}{p}\mathcal{P}_{\Omega}\Big(\sum\nolimits_{1\leq i\leq r}\boldsymbol{u}_i^{\otimes 3} - \boldsymbol{T}^{\star}\Big) \times_1 \boldsymbol{u}_s \times_2 \boldsymbol{u}_s, \qquad 1 \leq s \leq r, \tag{38}$$

and hence one can write

$$\nabla f_{\mathsf{clean}}\left(\boldsymbol{U}\right) = \frac{1}{p}\mathcal{P}_{\Omega}\Big(\sum\nolimits_{1\leq i\leq r}\boldsymbol{u}_i^{\otimes 3} - \boldsymbol{T}^{\star}\Big) \times_1^{\mathsf{seq}} \boldsymbol{U} \times_2^{\mathsf{seq}} \boldsymbol{U}. \tag{39}$$

This clearly satisfies

$$\nabla f\left(\boldsymbol{U}\right) = \nabla f_{\mathsf{clean}}\left(\boldsymbol{U}\right) - \frac{1}{p}\mathcal{P}_{\Omega}(\boldsymbol{E}) \times_1^{\mathsf{seq}} \boldsymbol{U} \times_2^{\mathsf{seq}} \boldsymbol{U}. \tag{40}$$

Moreover, direct algebraic manipulations give that: for any matrix $\boldsymbol{V} = [\boldsymbol{v}_1, \dots, \boldsymbol{v}_r] \in \mathbb{R}^{d\times r}$,

$$\begin{aligned}
\mathsf{vec}\left(\boldsymbol{V}\right)^{\top} \nabla^2 f_{\mathsf{clean}}\left(\boldsymbol{U}\right) \mathsf{vec}\left(\boldsymbol{V}\right) = {} & \frac{1}{3p}\Big\|\mathcal{P}_{\Omega}\Big(\sum\nolimits_{1\leq s\leq r}\boldsymbol{u}_s \otimes \boldsymbol{u}_s \otimes \boldsymbol{v}_s + \boldsymbol{u}_s \otimes \boldsymbol{v}_s \otimes \boldsymbol{u}_s + \boldsymbol{v}_s \otimes \boldsymbol{u}_s \otimes \boldsymbol{u}_s\Big)\Big\|_{\mathrm{F}}^2 \\
& + \frac{2}{p}\Big\langle\mathcal{P}_{\Omega}\Big(\sum\nolimits_{s\in[r]}\boldsymbol{u}_s^{\otimes 3} - \boldsymbol{T}^{\star}\Big), \sum\nolimits_{s\in[r]}\boldsymbol{v}_s \otimes \boldsymbol{v}_s \otimes \boldsymbol{u}_s\Big\rangle,
\end{aligned} \tag{41}$$

where $\mathsf{vec}(\boldsymbol{V})$ denotes the vectorization of $\boldsymbol{V}$.

5.1.2   Local strong convexity and smoothness

At the heart of our analysis is a crucial geometric property of the objective function, that is, the noiseless loss function f_{clean} behaves like a locally strongly convex and smooth function. This fact, which is formally stated in the following lemma, is the key enabler of fast local convergence of vanilla GD.

Lemma 5.1 (**Local strong convexity and smoothness**). *Suppose that the sample complexity and the rank satisfy*

$$p \geq c_0 \max\left\{ \frac{\log^3 d}{d^{3/2}}, \frac{\mu^2 r^2 \log d}{d^2} \right\} \qquad r \leq c_1 \sqrt{\frac{d}{\mu}} \tag{42}$$

for some sufficiently large (resp. small) constant $c_0 > 0$ (resp. $c_1 > 0$). Then with probability greater than $1 - O(d^{-10})$,

$$\frac{1}{2} \lambda_{\min}^{\star 4/3} \|\boldsymbol{V}\|_{\mathrm{F}}^2 \leq \mathsf{vec}\left(\boldsymbol{V}\right)^\top \nabla^2 f_{\mathsf{clean}}\left(\boldsymbol{U}\right) \mathsf{vec}\left(\boldsymbol{V}\right) \leq 4 \lambda_{\max}^{\star 4/3} \|\boldsymbol{V}\|_{\mathrm{F}}^2 \tag{43}$$

holds simultaneously for all $\boldsymbol{V} \in \mathbb{R}^{d \times r}$ and all $\boldsymbol{U} \in \mathbb{R}^{d \times r}$ obeying

$$\|\boldsymbol{U} - \boldsymbol{U}^\star\|_{\mathrm{F}} \leq \delta \|\boldsymbol{U}^\star\|_{\mathrm{F}} \quad and \quad \|\boldsymbol{U} - \boldsymbol{U}^\star\|_{2,\infty} \leq \delta \|\boldsymbol{U}^\star\|_{2,\infty}. \tag{44}$$

Here, $\delta \leq c_2/(\mu^{3/2} r)$ for some sufficiently small constant $c_2 > 0$.

Proof. See Appendix A.1. ∎

In order to invoke Lemma 5.1, one needs to make sure that the decision matrix $\boldsymbol{U}$ of interest (e.g. $\boldsymbol{U}^t$ in the GD sequence) satisfies the condition (44). This, however, is a fairly stringent condition, as it requires $\boldsymbol{U}$ to be close to the truth in every single row.

5.1.3   Leave-one-out gradient descent sequences

Motivated by the analytical framework developed for low-rank matrix recovery [MWCC17, CLL19], we introduce the following leave-one-out sequences, which play a crucial role in guaranteeing that the entire trajectory $\{\boldsymbol{U}^t\}_{t \geq 0}$ satisfies the condition (44) as required in Lemma 5.1.

Specifically, we define for each $1 \leq m \leq d$ the following auxiliary loss function:

$$f^{(m)}\left(\boldsymbol{U}\right) \triangleq \frac{1}{6p} \left\| \mathcal{P}_{\Omega_{-m}}\left(\sum_{1 \leq s \leq r} \boldsymbol{u}_s^{\otimes 3} - \boldsymbol{T}^\star - \boldsymbol{E} \right) \right\|_{\mathrm{F}}^2 + \frac{1}{6} \left\| \mathcal{P}_m\left(\sum_{1 \leq s \leq r} \boldsymbol{u}_s^{\otimes 3} - \boldsymbol{T}^\star \right) \right\|_{\mathrm{F}}^2, \tag{45}$$

where

- $\mathcal{P}_{\Omega_m}$: the projection onto the subspace of tensors supported on $\{(i,j,k) \in \Omega : i = m \text{ or } j = m \text{ or } k = m\}$;

- $\mathcal{P}_{\Omega_{-m}}$: the projection onto the subspace of tensors supported on $\{(i,j,k) \in \Omega : i \neq m \text{ and } j \neq m \text{ and } k \neq m\}$;

- $\mathcal{P}_m$: the projection onto the subspace of tensors supported on $\{(i,j,k) \in [d]^3 : i = m \text{ or } j = m \text{ or } k = m\}$.

In words, this function is obtained by replacing all data at locations $\{(i,j,k) \in [d]^3 : i = m \text{ or } j = m \text{ or } k = m\}$ by their expected values, thus removing all randomness associated with this location subset. The gradient of $f^{(m)}(\boldsymbol{U})$ w.r.t. $\boldsymbol{u}_s$ $(1 \leq s \leq r)$ can be computed as:

$$\nabla_{\boldsymbol{u}_s} f^{(m)}(\boldsymbol{U}) = \frac{1}{p} \mathcal{P}_{\Omega_{-m}}\left(\sum_{1 \leq s \leq r} \boldsymbol{u}_s^{\otimes 3} - \boldsymbol{T}^\star - \boldsymbol{E} \right) \times_1 \boldsymbol{u}_s \times_2 \boldsymbol{u}_s + \mathcal{P}_m\left(\sum_{1 \leq s \leq r} \boldsymbol{u}_s^{\otimes 3} - \boldsymbol{T}^\star \right) \times_1 \boldsymbol{u}_s \times_2 \boldsymbol{u}_s. \tag{46}$$

We then denote by $\left\{ \boldsymbol{U}^{t,(m)} \right\}_{t \geq 0}$ the iterative sequence obtained by running gradient descent w.r.t. the leave-one-out loss $f^{(m)}(\cdot)$; see Algorithm 4. By construction, as long as $\boldsymbol{U}^{0,(m)}$ is independent of the sampling locations and the noise associated with the locations $\{(i,j,k) \in \Omega : i = m \text{ or } j = m \text{ or } k = m\}$ (which holds true as detailed momentarily), then the entire trajectory $\left\{ \boldsymbol{U}^{t,(m)} \right\}_{t \geq 0}$ becomes statistically independent of such randomness. This is a crucial property that allows us to decouple the complicated statistical dependency.

Algorithm 4 The m-th leave-one-out sequence

1: Generate an initial estimate $\boldsymbol{U}^{0,(m)}$ via Algorithm 5.
2: **for** $t = 0, 1, \ldots, t_0 - 1$ **do**
3:     $\boldsymbol{U}^{t+1,(m)} = \boldsymbol{U}^{t,(m)} - \eta_t \nabla f^{(m)}\big(\boldsymbol{U}^{t,(m)}\big).$

5.1.4 Key lemmas

The proof for local linear convergence of GD is inductive in nature, which proceeds on the basis of the following set of inductive hypotheses. As we shall see in Corollary 5.11 in Section 5.3, this set of inductive hypotheses — modulo some global permutation — is valid with high probability when $t = 0$. In order to simplify presentation, we remove the consideration of the global permutation factor throughout this section (namely, we assume that the following holds for $\boldsymbol{U}^0\boldsymbol{\Pi}^0$ with some permutation matrix $\boldsymbol{\Pi}^0 \in \mathbb{R}^{r \times r}$ obeying $\boldsymbol{\Pi}^0 = \boldsymbol{I}$. Our inductive hypotheses are summarized as follows:

Key hypotheses for the gradient update stage:

$$\left\| \boldsymbol{U}^t - \boldsymbol{U}^\star \right\|_{\mathrm{F}} \leq \left(C_1 \rho^t \mathcal{E}_{\mathsf{local}} + C_2 \frac{\sigma}{\lambda_{\min}^\star} \sqrt{\frac{d \log d}{p}} \right) \|\boldsymbol{U}^\star\|_{\mathrm{F}} \, ; \tag{47a}$$

$$\left\| \boldsymbol{U}^t - \boldsymbol{U}^\star \right\|_{2,\infty} \leq \left(C_3 \rho^t \mathcal{E}_{\mathsf{local}} + C_4 \frac{\sigma}{\lambda_{\min}^\star} \sqrt{\frac{d \log d}{p}} \right) \|\boldsymbol{U}^\star\|_{2,\infty} \, ; \tag{47b}$$

$$\left\| \boldsymbol{U}^t - \boldsymbol{U}^{t,(m)} \right\|_{\mathrm{F}} \leq \left(C_5 \rho^t \mathcal{E}_{\mathsf{local}} + C_6 \frac{\sigma}{\lambda_{\min}^\star} \sqrt{\frac{d \log d}{p}} \right) \|\boldsymbol{U}^\star\|_{2,\infty} \, ; \tag{47c}$$

$$\left\| \big(\boldsymbol{U}^{t,(m)} - \boldsymbol{U}^\star\big)_{m,:} \right\|_2 \leq \left(C_7 \rho^t \mathcal{E}_{\mathsf{local}} + C_8 \frac{\sigma}{\lambda_{\min}^\star} \sqrt{\frac{d \log d}{p}} \right) \|\boldsymbol{U}^\star\|_{2,\infty} \, ; \tag{47d}$$

for some quantity $\mathcal{E}_{\mathsf{local}} > 0$ (depending possibly on μ and r) and some constants $C_1, \cdots, C_8 > 0$. These exist a few straightforward consequences of the hypotheses (47), which we record in the following lemma.

Lemma 5.2. *Assume that the hypotheses* (47) *hold, then we have*

$$\left\| \boldsymbol{U}^{t,(m)} - \boldsymbol{U}^\star \right\|_{\mathrm{F}} \leq \left(2C_1 \rho^t \mathcal{E}_{\mathsf{local}} + 2C_2 \frac{\sigma}{\lambda_{\min}^\star} \sqrt{\frac{d \log d}{p}} \right) \|\boldsymbol{U}^\star\|_{\mathrm{F}} \, , \tag{48}$$

$$\left\| \boldsymbol{U}^{t,(m)} - \boldsymbol{U}^\star \right\|_{2,\infty} \leq \left((C_3 + C_5)\, \rho^t \mathcal{E}_{\mathsf{local}} + (C_4 + C_6) \frac{\sigma}{\lambda_{\min}^\star} \sqrt{\frac{d \log d}{p}} \right) \|\boldsymbol{U}^\star\|_{2,\infty} \, . \tag{49}$$

Proof. See Appendix A.2. ∎

Our proof for the hypotheses (47) is inductive in nature: we would like to show that if the hypotheses in (47) hold for the t-th iteration, then they continue to be valid for the $(t + 1)$-th iteration. We shall justify each of the above hypotheses inductively through the following lemmas.

Lemma 5.3. *Suppose that*

$$p \geq c_0 \frac{\mu^3 r^2 \log^3 d}{d^{3/2}}, \quad \frac{\sigma}{\lambda_{\min}^\star} \leq c_1 \frac{\sqrt{p}}{\mu^{3/2} r \sqrt{d \log d}}, \quad \text{and} \quad r \leq c_2 \sqrt{\frac{d}{\mu}}$$

for some sufficiently large constant $c_0 > 0$ and some sufficiently small constant $c_1, c_2 > 0$. Assume that the hypotheses (47) *hold for the t-th iteration and $\mathcal{E}_{\mathsf{local}} \leq c_3/\big(\mu^{3/2} r\big)$ for some sufficiently small constant $c_3 > 0$. Then with probability at least $1 - O(d^{-10})$,*

$$\left\| \boldsymbol{U}^{t+1} - \boldsymbol{U}^\star \right\|_{\mathrm{F}} \leq \left(C_1 \rho^{t+1} \mathcal{E}_{\mathsf{local}} + C_2 \frac{\sigma}{\lambda_{\min}^\star} \sqrt{\frac{d \log d}{p}} \right) \|\boldsymbol{U}^\star\|_{\mathrm{F}} \, , \tag{50}$$

provided that $0 < \eta \le \lambda_{\min}^{\star 4/3}/\left(32\lambda_{\max}^{\star 8/3}\right)$, $1 - \left(\lambda_{\min}^{\star 4/3}/5\right)\eta \le \rho < 1$, *and* C_2 *is sufficiently large.*

Proof. See Appendix A.3. ∎

Lemma 5.4. *Suppose that*

$$p \ge c_0 \frac{\mu^3 r^2 \log^3 d}{d^{3/2}}, \quad \frac{\sigma}{\lambda_{\min}^{\star}} \le c_1 \frac{\sqrt{p}}{\mu^{3/2} r \sqrt{d \log d}}, \quad \text{and} \quad r \le c_2 \sqrt{\frac{d}{\mu}}$$

for some sufficiently large constant $c_0 > 0$ *and some sufficiently small constant* $c_1, c_2 > 0$. *Assume that the hypotheses* (47) *hold for the t-th iteration and* $\mathcal{E}_{\mathsf{local}} \le c_3/\left(\mu^{3/2} r\right)$ *for some sufficiently small constant* $c_3 > 0$. *Then with probability at least* $1 - O(d^{-10})$, *one has*

$$\left\| \boldsymbol{U}^{t+1,(m)} - \boldsymbol{U}^{t+1} \right\|_{\mathrm{F}} \le \left(C_5 \rho^{t+1} \mathcal{E}_{\mathsf{local}} + C_6 \frac{\sigma}{\lambda_{\min}^{\star}} \sqrt{\frac{d \log d}{p}} \right) \left\| \boldsymbol{U}^{\star} \right\|_{2,\infty}, \tag{51}$$

provided that $0 < \eta \le \lambda_{\min}^{\star 4/3}/\left(32\lambda_{\max}^{\star 8/3}\right)$, $1 - \left(\lambda_{\min}^{\star 4/3}/5\right)\eta \le \rho < 1$ *and* C_6 *is sufficiently large.*

Proof. See Appendix A.4. ∎

Lemma 5.5. *Suppose that*

$$p \ge c_0 \frac{\mu^3 r^2 \log^3 d}{d^{3/2}}, \quad \frac{\sigma}{\lambda_{\min}^{\star}} \le c_1 \frac{\sqrt{p}}{\mu^{3/2} r \sqrt{d \log d}}, \quad \text{and} \quad r \le c_2 \sqrt{\frac{d}{\mu}}$$

for some sufficiently large constant $c_0 > 0$ *and some sufficiently small constant* $c_1, c_2 > 0$. *Assume that the hypotheses* (47) *hold for the t-th iteration and* $\mathcal{E}_{\mathsf{local}} \le c_3/\left(\mu^{3/2} r\right)$ *for some sufficiently small constant* $c_3 > 0$. *Then with probability at least* $1 - O(d^{-10})$, *one has*

$$\left\| \left(\boldsymbol{U}^{t+1,(m)} - \boldsymbol{U}^{\star}\right)_{m,:} \right\|_2 \le \left(C_7 \rho^{t+1} \mathcal{E}_{\mathsf{local}} + C_8 \frac{\sigma}{\lambda_{\min}^{\star}} \sqrt{\frac{d \log d}{p}} \right) \left\| \boldsymbol{U}^{\star} \right\|_{2,\infty}, \tag{52}$$

provided that $0 < \eta \le \lambda_{\min}^{\star 4/3}/\left(32\lambda_{\max}^{\star 8/3}\right)$, $1 - \left(\lambda_{\min}^{\star 4/3}/5\right)\eta \le \rho < 1$, C_7 *and* C_8 *are sufficiently large.*

Proof. See Appendix A.5. ∎

Lemma 5.6. *Suppose that*

$$p \ge c_0 \frac{\mu^3 r^2 \log^3 d}{d^{3/2}}, \quad \frac{\sigma}{\lambda_{\min}^{\star}} \le c_1 \frac{\sqrt{p}}{\mu^{3/2} r \sqrt{d \log d}}, \quad \text{and} \quad r \le c_2 \sqrt{\frac{d}{\mu}}$$

for some sufficiently large constant $c_0 > 0$ *and some sufficiently small constant* $c_1, c_2 > 0$. *Assume that the hypotheses* (47) *hold for the t-th iteration and* $\mathcal{E}_{\mathsf{local}} \le c_3/\left(\mu^{3/2} r\right)$ *for some sufficiently small constant* $c_3 > 0$. *Then with probability at least* $1 - O(d^{-10})$, *one has*

$$\left\| \boldsymbol{U}^{t+1} - \boldsymbol{U}^{\star} \right\|_{2,\infty} \le \left(C_3 \rho^{t+1} \mathcal{E}_{\mathsf{local}} + C_4 \frac{\sigma}{\lambda_{\min}^{\star}} \sqrt{\frac{d \log d}{p}} \right) \left\| \boldsymbol{U}^{\star} \right\|_{2,\infty}, \tag{53}$$

provided that $0 < \eta \le \lambda_{\min}^{\star 4/3}/\left(32\lambda_{\max}^{\star 8/3}\right)$, $1 - \left(\lambda_{\min}^{\star 4/3}/5\right)\eta \le \rho < 1$, $C_3/\left(C_5 + C_7\right)$ *and* $C_4/\left(C_6 + C_8\right)$ *are both sufficiently large.*

Proof. See Appendix A.6. ∎

The proofs of the above key lemmas are postponed to Appendix A.

5.2 Analysis for initialization: Part 1 (subspace estimation)

5.2.1 Key results

The aim of this subsection is to demonstrate that the subspace estimate U computed by Algorithm 2 is sufficiently close to the space spanned by the true tensor factors. Given that the columns of $U^\star = [u_1^\star, \cdots, u_r^\star]$ are in general not orthogonal to each other, we shall define $U_{\mathsf{orth}}^\star \in \mathbb{R}^{d \times r}$ as follows (obtained by proper orthonormalization) :

$$U_{\mathsf{orth}}^\star := U^\star (U^{\star\top} U^\star)^{-\frac{1}{2}}. \tag{54}$$

This matrix $U_{\mathsf{orth}}^\star$ reflects the rank-r principal subspace of $A^\star A^{\star\top} = \sum_i \|u_i^\star\|_2^4 u_i^\star u_i^{\star\top}$, where we recall that $A^\star \in \mathbb{R}^{d \times d^2}$ is the mode-1 matricization of $T^\star$. In addition, we define the rotation matrix

$$R := \underset{Q \in \mathcal{O}^{r \times r}}{\arg\min} \ \|UQ - U_{\mathsf{orth}}^\star\|_{\mathrm{F}}, \tag{55}$$

where $\mathcal{O}^{r \times r}$ stands for the set of $r \times r$ orthonormal matrices. This can be viewed as the global rotation matrix that best aligns the two subspaces represented by U and $U_{\mathsf{orth}}^\star$ respectively.

Equipped with the above notation, we can invoke [CLC+20, Corollary 1] to arrive at the following lemma, which upper bounds the distance between our subspace estimate U and the ground truth $U_{\mathsf{orth}}^\star$.

Lemma 5.7. *There exist some universal constants $c_0, c_1, c_2 > 0$ such that if*

$$p \geq c_0 \frac{\mu^2 r \log^2 d}{d^{3/2}}, \quad \frac{\sigma}{\lambda_{\min}^\star} \leq c_1 \frac{\sqrt{p}}{d^{3/4}\sqrt{\log d}}, \quad and \quad r \leq c_2 \sqrt{\frac{d}{\mu}},$$

then with probability $1 - O(d^{-10})$, the subspace estimate U computed by Algorithm 2 obeys

$$\|UR - U_{\mathsf{orth}}^\star\| \lesssim \mathcal{E}_{\mathsf{se}}, \tag{56a}$$

$$\|UR - U_{\mathsf{orth}}^\star\|_{2,\infty} \lesssim \mathcal{E}_{\mathsf{se}} \sqrt{\frac{\mu r}{d}}, \tag{56b}$$

where $U_{\mathsf{orth}}^\star$ and R are defined respectively in (54) and (55), and

$$\mathcal{E}_{\mathsf{se}} := \frac{\mu^2 r \log d}{d^{3/2} p} + \sqrt{\frac{\mu^2 r \log d}{d^2 p}} + \frac{\sigma^2}{\lambda_{\min}^{\star 2}} \frac{d^{3/2} \log d}{p} + \frac{\sigma}{\lambda_{\min}^\star} \sqrt{\frac{d \log d}{p}} + \frac{\mu r}{d}. \tag{57}$$

In a nutshell, Lemma 5.7 asserts that: under our sample size, noise and rank conditions, Algorithm 2 produces reliable estimates of the subspace spanned by the low-rank tensor factors $\{u_i^\star\}_{1 \leq i \leq r}$. The theorem quantifies the subspace distance in terms of both the spectral norm and $\|\cdot\|_{2,\infty}$, where the latter bound often reflects a considerably stronger sense of proximity compared to the former one.

As it turns out, in order to facilitate analysis for the subsequent stages, we need to introduce certain leave-one-out sequences as well, which we detail in the next subsection.

5.2.2 Leave-one-out sequences for subspace estimation

The key idea of the leave-one-out analysis is to create auxiliary leave-one-out sequences that are (1) independent of a small fraction of the data; (2) sufficiently close to the true estimates. We introduce the following auxiliary tensor and $d \times d^2$-dimensional matrix for each $1 \leq m \leq d$:

$$T^{(m)} := \mathcal{P}_{\Omega_{-m}}(T) + p\mathcal{P}_m(T^\star) \in \mathbb{R}^{d \times d \times d}, \tag{58}$$

$$A^{(m)} := \text{mode-1 matricization of } \frac{1}{p} T^{(m)}. \tag{59}$$

By construction, $T^{(m)}$ and $A^{(m)}$ are independent of $\mathcal{P}_{\Omega_m}(E)$, where we recall that

$$\Omega_{-m} := \{(i, j, k) \in \Omega : i \neq m \text{ and } j \neq m \text{ and } k \neq m\}, \tag{60}$$

$$\Omega_m := \{(i,j,k) \in \Omega \colon i = m \text{ or } j = m \text{ or } k = m\}. \tag{61}$$

We are now ready to introduce the auxiliary leave-one-out procedure for subspace estimation. Similar to the matrix $\boldsymbol{B}$ in Algorithm 2 (whose eigenspace serves as an estimate of the column space of $\boldsymbol{U}^\star$), we define an auxiliary matrix $\boldsymbol{B}^{(m)} \in \mathbb{R}^{d \times d}$ as follows:

$$\boldsymbol{B}^{(m)} = \mathcal{P}_{\mathsf{off\text{-}diag}}\big(\boldsymbol{A}^{(m)}\boldsymbol{A}^{(m)\top}\big), \tag{62}$$

where $\mathcal{P}_{\mathsf{off\text{-}diag}}(\cdot)$ (as already defined in Section 3.1) extracts out off-diagonal entries from a matrix. The rationale is simple: it can be easily verified that

$$\mathbb{E}\big[\boldsymbol{B}^{(m)}\big] = \boldsymbol{B}^\star - \mathcal{P}_{\mathsf{diag}}\big(\boldsymbol{B}^\star\big), \qquad \boldsymbol{B}^\star := \boldsymbol{A}^\star \boldsymbol{A}^{\star\top}, \tag{63}$$

where $\mathcal{P}_{\mathsf{diag}}(\cdot)$ extracts out the diagonal entries of the matrix. This gives hope that the eigenspace of $\boldsymbol{B}^{(m)}$ is also a reliable estimate of the column space of $\boldsymbol{U}^\star$, provided that the diagonal entries of $\boldsymbol{B}^\star$ are sufficiently small. Consequently, we shall compute $\boldsymbol{U}^{0,(m)} \in \mathbb{R}^{d \times r}$ — a matrix whose columns are the top-r leading eigenvectors of $\boldsymbol{B}^{(m)}$. The procedure is summarized in Algorithm 5.

Algorithm 5 The m-th leave-one-out sequence for spectral initialization

---

1: Let $\boldsymbol{U}^{(m)}\boldsymbol{\Lambda}^{(m)}\boldsymbol{U}^{(m)\top}$ be the rank-r eigen-decomposition of $\boldsymbol{B}^{(m)}$ defined in (62).
2: Generate the initial estimate $\boldsymbol{U}^{0,(m)} \in \mathbb{R}^{d \times r}$ from $\boldsymbol{U}^{(m)} \in \mathbb{R}^{d \times r}$ using Algorithm 6.

The following lemma plays a crucial role in our analysis, which formalizes the fact that the leave-one-out version $\boldsymbol{U}^{(m)}$ obtained by Algorithm 5 is extremely close to $\boldsymbol{U}$.

Lemma 5.8. *There exist some universal constants $c_0, c_1, c_2 > 0$ such that if*

$$p \geq c_0 \frac{\mu^2 r \log^2 d}{d^{3/2}}, \quad \frac{\sigma}{\lambda^\star_{\min}} \leq c_1 \frac{\sqrt{p}}{d^{3/4}\sqrt{\log d}}, \quad and \quad r \leq c_2\sqrt{\frac{d}{\mu}},$$

then with probability $1 - O\big(d^{-10}\big)$, the subspace estimate $\boldsymbol{U}^{(m)}$ computed by Algorithm 5 obeys

$$\big\|\boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{U}^{(m)}\boldsymbol{U}^{(m)\top}\big\|_{\mathrm{F}} \lesssim \mathcal{E}_{\mathsf{loo}}\sqrt{\frac{\mu r}{d}} \tag{64}$$

simultaneously for all $1 \leq m \leq d$, where

$$\mathcal{E}_{\mathsf{loo}} := \frac{\mu^2 r \log d}{d^{3/2}p} + \sqrt{\frac{\mu^2 r \log d}{d^2 p}} + \frac{\sigma^2}{\lambda^{\star 2}_{\min}}\frac{d^{3/2}\log d}{p} + \frac{\sigma}{\lambda^\star_{\min}}\sqrt{\frac{d \log d}{p}}. \tag{65}$$

Lemma 5.8 follows immediately from the analysis of [CLC+20, Lemma 4]. As a remark, the construction of the leave-one-out sequences herein is slightly different from the one in [CLC+20]. However, it is straightforward to adapt the proof of [CLC+20] to the case considered herein. We therefore omit the proof for the sake of brevity.

5.3 Analysis for initialization: Part 2 (retrieval of individual tensor factors)

5.3.1 Main results and leave-one-out sequences

This section justifies that the procedure presented in Algorithm 3 allows to disentangle the tensor factors. For notational simplicity, we let

$$\overline{\boldsymbol{u}}^\star_i := \boldsymbol{u}^\star_i / \|\boldsymbol{u}^\star_i\|_2, \qquad \lambda^\star_i := \|\boldsymbol{u}^\star_i\|^3_2, \qquad 1 \leq i \leq d. \tag{66}$$

Our result is this:

Theorem 5.9. *Fix any arbitrary small constant $\delta > 0$. Assume that*

$$p \geq c_0 \frac{\mu^2 r^4 \log^4 d}{d^{3/2}}, \quad \frac{\sigma}{\lambda^\star_{\min}} \leq c_1 \frac{\sqrt{p}}{r^{3/2} d^{3/4} \log^2 d}, \quad r \leq c_2 \left(\frac{d}{\mu^6 \log^6 d} \right)^{1/6},$$

$$L = c_3 r^{2\kappa^2} \log^{3/2} r, \quad \epsilon_{\mathsf{th}} = c_4 \left\{ \frac{\mu r \log d}{d\sqrt{p}} + \frac{\sigma}{\lambda^\star_{\min}} \sqrt{\frac{rd \log^2 d}{p}} + \sqrt{\frac{\mu r \log d}{d}} \right\} \tag{67}$$

for some sufficiently large universal constant $c_0, c_3 > 0$ and some sufficiently small universal constants $c_1, c_2, c_4 > 0$. Then with probability exceeding $1 - \delta$, there exists a permutation $\pi(\cdot) : [d] \mapsto [d]$ such that for all $1 \leq i \leq r$, the tensor factors $\{\boldsymbol{w}^i\}_{i=1}^r$ returned by Algorithm 3 satisfy

$$\left\| \boldsymbol{w}^i - \overline{\boldsymbol{u}}^\star_{\pi(i)} \right\|_2 \lesssim \frac{\mu r \log d}{d\sqrt{p}} + \frac{\sigma}{\lambda^\star_{\min}} \sqrt{\frac{rd \log^2 d}{p}} + \sqrt{\frac{\mu r \log d}{d}}; \tag{68a}$$

$$\left\| \boldsymbol{w}^i - \overline{\boldsymbol{u}}^\star_{\pi(i)} \right\|_\infty \lesssim \left\{ \frac{\mu^2 r \log^4 d}{d^{3/2} p} + \frac{\mu r \log^3 d}{d\sqrt{p}} + \frac{\sigma^2}{\lambda^{\star 2}_{\min}} \frac{d^{3/2} \log^4 d}{p} + \frac{\sigma}{\lambda^\star_{\min}} \sqrt{\frac{rd \log^6 d}{p}} + \sqrt{\frac{\mu r \log^2 d}{d}} \right\} \sqrt{\frac{\mu r}{d}}; \tag{68b}$$

$$\left| \lambda_i - \lambda^\star_{\pi(i)} \right| \lesssim \left\{ \frac{\mu r \log d}{d\sqrt{p}} + \frac{\sigma}{\lambda^\star_{\min}} \sqrt{\frac{rd \log^2 d}{p}} + \sqrt{\frac{\mu r \log d}{d}} \right\} \lambda^\star_{\pi(i)}. \tag{68c}$$

In short, this theorem asserts that the estimates returned by Algorithm 3 are — up to global permutation — reasonably close to the ground truth under our sample size and noise conditions. In order to establish this theorem and in order to provide initial guesses for the leave-one-out GD sequences, we need to produce a leave-one-out sequence tailored to this part of the algorithm. Such auxiliary sequences are generated in a similar spirit as the previous ones, and we summarize them in Algorithm 6. As usual, the resulting leave-one-out estimates $\{\lambda_i^{(m)}, \boldsymbol{w}^{i,(m)}\}_{i=1}^r$ are statistically independent of $\mathcal{P}_{\Omega_m}(\boldsymbol{E})$.

In what follows, we gather a few key properties of the leave-one-out estimates, which play a crucial role in the analysis.

Algorithm 6 The m-th leave-one-out sequence for retrieving individual tensor components

---

1: **Input:** restart number L, threshold ϵ_{th}, subspace estimate $\boldsymbol{U}^{(m)} \in \mathbb{R}^{d \times r}$ 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:     $\left(\boldsymbol{\nu}^{\tau,(m)}, \lambda_\tau^{(m)}, \mathsf{spec\text{-}gap}_\tau^{(m)} \right) \leftarrow \textsc{Retrieve-one-tensor-factor}(\boldsymbol{T}^{(m)}, p, \boldsymbol{U}^{(m)}, \boldsymbol{g}^\tau)$.
5: Generate tensor factor estimates

$$\left\{ (\boldsymbol{w}^{1,(m)}, \lambda_1^{(m)}), \ldots, (\boldsymbol{w}^{r,(m)}, \lambda_r^{(m)}) \right\} \leftarrow \textsc{Prune}(\{ (\boldsymbol{\nu}^{\tau,(m)}, \lambda_\tau^{(m)}, \mathsf{spec\text{-}gap}_\tau^{(m)}) \}_{\tau=1}^L, \epsilon_{\mathsf{th}}).$$

6: **Output:** an initial estimate $\boldsymbol{U}^{0,(m)} = \left[(\lambda_1^{(m)})^{1/3} \boldsymbol{w}^{1,(m)}, \ldots, (\lambda_r^{(m)})^{1/3} \boldsymbol{w}^{r,(m)} \right]$.

Theorem 5.10. *Fix any arbitrarily small constant $\delta > 0$. Instate the assumptions in Theorem 5.9. With probability exceeding $1 - \delta$, the permutation function stated in Theorem 5.9 obeys that: for all $1 \leq i \leq r$ and all $1 \leq m \leq d$:*

$$\left\| \boldsymbol{w}^i - \boldsymbol{w}^{i,(m)} \right\|_2 \lesssim \left\{ \frac{\mu^2 r \log^{3/2} d}{d^{3/2} p} + \frac{\mu \sqrt{r} \log d}{d\sqrt{p}} + \frac{\sigma^2}{\lambda^{\star 2}_{\min}} \frac{d^{3/2} \log^{3/2} d}{p} + \frac{\sigma}{\lambda^\star_{\min}} \sqrt{\frac{d \log^2 d}{p}} \right\} \sqrt{\frac{\mu r}{d}}; \tag{69a}$$

$$\left| \lambda_i - \lambda_i^{(m)} \right| \lesssim \left\{ \frac{\mu^2 r \log^{3/2} d}{d^{3/2} p} + \frac{\mu \sqrt{r} \log d}{d\sqrt{p}} + \frac{\sigma^2}{\lambda^{\star 2}_{\min}} \frac{d^{3/2} \log^{3/2} d}{p} + \frac{\sigma}{\lambda^\star_{\min}} \sqrt{\frac{d \log^2 d}{p}} \right\} \sqrt{\frac{\mu r}{d}} \lambda^\star_{\max}; \tag{69b}$$

$$\left|\left(\boldsymbol{w}^{i,(m)} - \overline{\boldsymbol{u}}_{\pi(i)}^{\star}\right)_m\right| \lesssim \left\{\frac{\sqrt{\mu r}\log^{7/2} d}{d^{3/2}p} + \frac{\mu r \log^3 d}{d\sqrt{p}} + \frac{\sigma}{\lambda_{\min}^{\star}}\frac{\log^4 d}{p} + \frac{\sigma}{\lambda_{\min}^{\star}}\sqrt{\frac{rd\log^6 d}{p}} + \sqrt{\frac{\mu r \log^2 d}{d}}\right\}\sqrt{\frac{\mu r}{d}}. \tag{69c}$$

With Theorems 5.9-5.10 in place, we can immediately establish a few desired properties (particularly those specified in Section 5.1) of our initial estimate, as asserted in the following corollary.

Corollary 5.11. *Fix any arbitrarily small constant $\delta > 0$. Instate the assumptions in Theorem 2.8. With probability exceeding $1 - \delta$, the estimates $\boldsymbol{U}^0$ and $\boldsymbol{U}^{0,(m)}$ returned by Algorithm 3 and Algorithm 6 respectively satisfy the hypotheses (47) for $t = 0$.*

Proof. See Appendix B.10. ∎

5.3.2 Analysis

Before we start with the proof, we first state the main idea. For the sake of clarify, we define

$$\boldsymbol{\theta}^{\tau} := \boldsymbol{U}\boldsymbol{U}^{\top}\boldsymbol{g}^{\tau}, \tag{70a}$$

$$\boldsymbol{\theta}^{\tau,(m)} := \boldsymbol{U}^{(m)}\boldsymbol{U}^{(m)\top}\boldsymbol{g}^{\tau}, \tag{70b}$$

$$\boldsymbol{M}^{\tau} := p^{-1}\boldsymbol{T} \times_3 \boldsymbol{\theta}^{\tau}, \tag{70c}$$

$$\boldsymbol{M}^{\tau,(m)} := p^{-1}\boldsymbol{T}^{(m)} \times_3 \boldsymbol{\theta}^{\tau,(m)}. \tag{70d}$$

In addition, let $\boldsymbol{\nu}^{\tau}$ be the top singular vector of $\boldsymbol{M}^{\tau}$ obeying $\langle \boldsymbol{T}, (\boldsymbol{\nu}^{\tau})^{\otimes 3}\rangle \geq 0$, and $\boldsymbol{\nu}^{\tau,(m)}$ the top singular vector of $\boldsymbol{M}^{\tau,(m)}$ obeying $\langle \boldsymbol{T}^{(m)}, (\boldsymbol{\nu}^{\tau,(m)})^{\otimes 3}\rangle \geq 0$. Set

$$\lambda_{\tau} := \langle p^{-1}\boldsymbol{T}, (\boldsymbol{\nu}^{\tau})^{\otimes 3}\rangle \qquad \text{and} \qquad \lambda_{\tau}^{(m)} := \langle p^{-1}\boldsymbol{T}^{(m)}, (\boldsymbol{\nu}^{\tau,(m)})^{\otimes 3}\rangle. \tag{71}$$

These are all computed in the function RETRIEVE-ONE-TENSOR-FACTOR() in the τ-th round.

1. We first show that for each $1 \leq i \leq r$, there exists at least one trial $1 \leq \tau \leq L$ such that the i-th tensor factor $\overline{\boldsymbol{u}}_i^{\star}$ is the top singular vector of the population version of $\boldsymbol{T} \times_3 \boldsymbol{\theta}^{\tau}$ (with respect to the missing data and noise). In addition, the spectral gap is large enough to guarantee accurate estimates.

2. Next, we prove that given this spectral gap, the top singular vector $\boldsymbol{\nu}^{\tau}$ of $\boldsymbol{T} \times_3 \boldsymbol{\theta}^{\tau}$ is close to $\overline{\boldsymbol{u}}_i^{\star}$ both in the $\|\cdot\|_2$ and $\|\cdot\|_{\infty}$ norm. This also enables us to accurately estimate the magnitude of $\boldsymbol{u}_i^{\star}$.

3. Finally, we need to show that one can find those reliable estimates among L random restarts. Combining the spectral gap information with the incoherence condition that tensor components are nearly orthogonal to each other, our selection procedure is guaranteed to recover all tensor factors.

Now we proceed to the proof. Without loss of generality, we prove the case for $i = 1$ in the sequel, i.e. there exists some $\tau \in [L]$ such that $\boldsymbol{\nu}^{\tau}$ accurately recovers $\boldsymbol{u}_1^{\star}$. Together with the union bound, this shows that we can find reliable estimates for all tensor factors. We then conclude the proof by showing that Algorithm 3 is able to find all of them without duplicates.

To this end, we find it convenient to introduce an auxiliary vector $\boldsymbol{\gamma}^{\star\tau} = [\gamma_1^{\star\tau}, \cdots, \gamma_r^{\star\tau}]^{\top} \in \mathbb{R}^r$ and its leave-one-out versions $\boldsymbol{\gamma}^{\star\tau,(m)} = [\gamma_1^{\star\tau,(m)}, \cdots, \gamma_r^{\star\tau,(m)}]^{\top}$ ($1 \leq m \leq d$) for each $1 \leq \tau \leq L$ as follows:

$$\gamma_i^{\star\tau} := \|\boldsymbol{u}_i^{\star}\|_2^2 \langle \boldsymbol{u}_i^{\star}, \boldsymbol{\theta}^{\tau}\rangle = \lambda_i^{\star}\langle \overline{\boldsymbol{u}}_i^{\star}, \boldsymbol{\theta}^{\tau}\rangle, \tag{72a}$$

$$\gamma_i^{\star\tau,(m)} := \|\boldsymbol{u}_i^{\star}\|_2^2 \langle \boldsymbol{u}_i^{\star}, \boldsymbol{\theta}^{\tau,(m)}\rangle = \lambda_i^{\star}\langle \overline{\boldsymbol{u}}_i^{\star}, \boldsymbol{\theta}^{\tau,(m)}\rangle, \tag{72b}$$

where $\overline{\boldsymbol{u}}_i^{\star}$ and $\lambda_i^{\star}$ are both defined in (66). The idea is to let $\boldsymbol{\gamma}^{\star\tau}$ approximate the singular values of $\boldsymbol{T}^{\star} \times_3 \boldsymbol{\theta}^{\tau}$; this can be seen, for instance, via the following calculation:

$$\boldsymbol{T}^{\star} \times_3 \boldsymbol{\theta}^{\tau} = \sum_{i=1}^r \langle \boldsymbol{u}_i^{\star}, \boldsymbol{\theta}^{\tau}\rangle\boldsymbol{u}_i^{\star}\boldsymbol{u}_i^{\star\top} = \sum_{i=1}^r \underbrace{\lambda_i^{\star}\langle \overline{\boldsymbol{u}}_i^{\star}, \boldsymbol{\theta}^{\tau}\rangle}_{=\gamma_i^{\star\tau}} \overline{\boldsymbol{u}}_i^{\star}\overline{\boldsymbol{u}}_i^{\star\top}, \tag{73}$$

24

where $\{\overline{\boldsymbol{u}}_i^\star\}_{i=1}^r$ — which are assumed to be incoherent (or nearly orthogonal to each other) — can be approximately viewed as the singular vectors of $\boldsymbol{T}^\star \times_3 \boldsymbol{\theta}^\tau$.

If we want our spectral estimate to be accurate, we would need to be assured that the two largest entries of $\boldsymbol{\gamma}^{\star\tau}$ (in magnitude) are sufficiently separated.

Lemma 5.12. *Instate the assumptions of Theorem 5.9. Define $\Delta_1^\tau := \gamma_1^{\star\tau} - \max_{1 < i \leq r} \left| \gamma_i^{\star\tau} \right|$ for each $1 \leq \tau \leq L$ and let $\Delta_1^{(1)} \geq \Delta_1^{(2)} \geq \cdots \geq \Delta_1^{(L)}$ denote the order statistics of $\left\{ \Delta_1^\tau \right\}_{\tau=1}^L$ (in descending order). Fix any arbitrary small constant $\delta > 0$. With probability greater than $1 - \delta/r$, one has*

$$\Delta_1^{(1)} \gtrsim \lambda_{\min}^\star, \tag{74a}$$

$$\Delta_1^{(1)} - \Delta_1^{(2)} \gtrsim \frac{\lambda_{\min}^\star}{r\sqrt{\log d}}, \tag{74b}$$

Additionally, for any fixed vector $\boldsymbol{v} \in \mathbb{R}^r$, with probability at least $1 - O\left(d^{-10}\right)$, for all $1 \leq \tau \leq L$, one has

$$\gamma_1^{\star\tau} \lesssim \sqrt{\log d}\, \lambda_{\max}^\star, \tag{75a}$$

$$\left\| \boldsymbol{\gamma}^{\star\tau} \right\|_2 \lesssim \sqrt{r \log d}\, \lambda_{\max}^\star, \tag{75b}$$

$$\left| \langle \boldsymbol{v}, \boldsymbol{\gamma}^{\star\tau} \rangle \right| \lesssim \| \boldsymbol{v} \|_2 \sqrt{\log d}\, \lambda_{\max}^\star. \tag{75c}$$

Proof. See Appendix B.1. ∎

Lemma 5.12 demonstrates that there exists some $\tau \in [L]$ such that $\gamma_1^{\star\tau} - \max_{1 < i \leq r} \left| \gamma_i^{\star\tau} \right| \gtrsim \lambda_{\min}^\star$. This means that $\overline{\boldsymbol{u}}_1^\star$ exhibits the largest correlation with the random projection θ, which further implies that $\overline{\boldsymbol{u}}_1^\star$ is the largest singular vector of $\boldsymbol{T}^\star \times_3 \boldsymbol{\theta}^\tau$ with a considerable spectral gap (as we will show shortly). With the desired spectral gap in place, we are ready to look at the eigenvectors / singular vectors of interest. To this end, we find it convenient to introduce another auxiliary vector $\overline{\boldsymbol{u}}^\tau$, defined as the leading singular vector of $\boldsymbol{M}^\tau$ (cf. (70c)) obeying

$$\langle \overline{\boldsymbol{u}}^\tau, \overline{\boldsymbol{u}}_1^\star \rangle \geq 0. \tag{76}$$

The careful reader would immediately notice the similarity between $\overline{\boldsymbol{u}}^\tau$ and $\boldsymbol{\nu}^\tau$ except for their global signs; namely, we determine the global sign of $\overline{\boldsymbol{u}}^\tau$ based on the ground truth information (76), but pick the global sign for $\boldsymbol{\nu}^\tau$ solely based on the observed data (cf. Algorithm 3). Fortunately, the vectors $\overline{\boldsymbol{u}}^\tau$ and $\boldsymbol{\nu}^\tau$ provably coincide, namely,

$$\overline{\boldsymbol{u}}^\tau = \boldsymbol{\nu}^\tau, \tag{77}$$

as we shall demonstrate momentarily in Lemma 5.16. In a similar way, we also denote by $\overline{\boldsymbol{u}}^{\tau,(m)}$ the leading singular vector of $\boldsymbol{M}^{\tau,(m)}$ defined in (70d) such that

$$\langle \overline{\boldsymbol{u}}^{\tau,(m)}, \overline{\boldsymbol{u}}_1^\star \rangle \geq 0. \tag{78}$$

Lemma 5.16 also shows that $\overline{\boldsymbol{u}}^{\tau,(m)} = \boldsymbol{\nu}^{\tau,(m)}$.

We shall now take a detour to look at $\overline{\boldsymbol{u}}^\tau$, which in turn would help us understand $\boldsymbol{\nu}^\tau$. We shall first demonstrate that $\overline{\boldsymbol{u}}^\tau$ (and hence $\boldsymbol{\nu}^\tau$) is sufficiently close to the corresponding true factor in the ℓ_2 sense.

Lemma 5.13. *Instate the assumptions of Theorem 5.9. Let $\overline{\boldsymbol{u}}^\tau$ and $\overline{\boldsymbol{u}}_1^\star$ be as defined in (76) and (66), respectively. Define $\mathcal{A}$ to be the event such that $\gamma_1^{\star\tau} - \max_{1 < i \leq r} \left| \gamma_i^{\star\tau} \right| \gtrsim \lambda_{\min}^\star$ and the condition (75) hold. Then conditional on this event $\mathcal{A}$, with probability exceeding $1 - O\left(d^{-11}\right)$ one has*

$$\left\| \overline{\boldsymbol{u}}^\tau - \overline{\boldsymbol{u}}_1^\star \right\|_2 \lesssim \underbrace{\frac{\mu r \log d}{d\sqrt{p}} + \frac{\sigma}{\lambda_{\min}^\star} \sqrt{\frac{rd \log^2 d}{p}} + \sqrt{\frac{\mu r \log d}{d}}}_{=:\,\mathcal{E}_{\mathsf{proj}}}. \tag{79}$$

Proof. See Appendix B.2. ∎

Thus far, we have focused on the ℓ_2 estimation errors. In order to further quantify the ℓ_∞ estimation errors, we need to resort to the leave-one-out estimates $\overline{\boldsymbol{u}}^{\tau,(m)}$ $(1 \le m \le d)$. Specifically, we shall justify in the following two lemmas that: (1) the m-th leave-one-out estimate $\overline{\boldsymbol{u}}^{\tau,(m)}$ is close to the truth at least in the m-th coordinate, and (2) the vector $\overline{\boldsymbol{u}}^\tau$ is extremely close to each of the leave-one-out estimates $\overline{\boldsymbol{u}}^{\tau,(m)}$ $(1 \le m \le d)$. These two observations taken collectively translate to the desired entrywise error control of $\overline{\boldsymbol{u}}^\tau$. Here, we recall that the global sign of $\overline{\boldsymbol{u}}^{\tau,(m)}$ (cf. (78)) and the global sign of $\overline{\boldsymbol{u}}^\tau$ (cf. (76)) are defined in a similar fashion, both using the ground truth information.

Lemma 5.14. *Instate the assumptions of Theorem 5.9. Define $\mathcal{A}$ to be the event such that $\gamma_1^{\star\tau} - \max_{1<i\le r}\left|\gamma_i^{\star\tau}\right| \gtrsim \lambda_{\min}^\star$ and the condition (75) hold. Then conditional on this event $\mathcal{A}$, one has, with probability exceeding $1 - O\left(d^{-10}\right)$, that*

$$\left|\left[\overline{\boldsymbol{u}}^{\tau,(m)} - \overline{\boldsymbol{u}}_1^\star\right]_m\right| \lesssim \mathcal{E}_{\mathsf{op}}\sqrt{\frac{\mu r \log d}{d}}, \tag{80}$$

holds for all $m \in [d]$, where $\mathcal{E}_{\mathsf{op}}$ is defined as follows:

$$\mathcal{E}_{\mathsf{op}} := \frac{\sqrt{\mu r}\log^3 d}{d^{3/2}p} + \frac{\mu r \log^{5/2} d}{d\sqrt{p}} + \frac{\sigma}{\lambda_{\min}^\star}\frac{\log^{7/2} d}{p} + \frac{\sigma}{\lambda_{\min}^\star}\sqrt{\frac{rd\log^5 d}{p}} + \sqrt{\frac{\mu r \log d}{d}}. \tag{81}$$

Proof. See Appendix B.4. ∎

Lemma 5.15. *Instate the assumptions of Theorem 5.9. Define $\mathcal{A}$ to be the event such that $\gamma_1^{\star\tau} - \max_{1<i\le r}\left|\gamma_i^{\star\tau}\right| \gtrsim \lambda_{\min}^\star$ and the condition (75) hold. Then conditional on this event $\mathcal{A}$, one has, with probability at least $1 - O\left(d^{-10}\right)$, for all $m \in [d]$:*

$$\left\|\overline{\boldsymbol{u}}^\tau - \overline{\boldsymbol{u}}^{\tau,(m)}\right\|_2 \lesssim \mathcal{E}_{\mathsf{loo}}\sqrt{\frac{\mu r \log d}{d}}, \tag{82}$$

$$\left\|\overline{\boldsymbol{u}}^\tau - \overline{\boldsymbol{u}}_1^\star\right\|_\infty \lesssim \left(\mathcal{E}_{\mathsf{op}} + \mathcal{E}_{\mathsf{loo}}\right)\sqrt{\frac{\mu r \log d}{d}}, \tag{83}$$

$$\left|\lambda_\tau - \lambda_\tau^{(m)}\right| \lesssim \mathcal{E}_{\mathsf{loo}}\sqrt{\frac{\mu r \log d}{d}}\lambda_{\max}^\star, \tag{84}$$

where $\mathcal{E}_{\mathsf{op}}$ and $\mathcal{E}_{\mathsf{loo}}$ are defined in (81) and (65), respectively.

Proof. See Appendix B.6. ∎

Next, we turn to the estimation accuracy regarding the size of the tensor factors and show that λ_τ (produced in Algorithm 3) is close to the truth as well. As it turns out, a byproduct of this step reveals that $\boldsymbol{\nu}^\tau = \overline{\boldsymbol{u}}^\tau$ and $\boldsymbol{\nu}^{\tau,(m)} = \overline{\boldsymbol{u}}^{\tau,(m)}$, where $\overline{\boldsymbol{u}}^\tau$ and $\overline{\boldsymbol{u}}^{\tau,(m)}$ are an auxiliary vectors defined in (76) and (78), respectively.

Lemma 5.16. *Instate the assumptions of Theorem 5.9. Assume that the results in Lemma 5.13, Lemma 5.14 and Lemma 5.15 hold. Then with probability at least $1 - O\left(d^{-10}\right)$, one has*

$$\left|\lambda_\tau - \lambda_1^\star\right| \lesssim \mathcal{E}_{\mathsf{proj}}\lambda_1^\star. \tag{85}$$

In particular, one has

$$\boldsymbol{\nu}^\tau = \overline{\boldsymbol{u}}^\tau \qquad and \qquad \boldsymbol{\nu}^{\tau,(m)} = \overline{\boldsymbol{u}}^{\tau,(m)}, \quad 1 \le m \le d. \tag{86}$$

Proof. See Appendix B.8. ∎

Thus far, we have only proved that one can find a reliable estimate for each tensor factor within L random trials, provided that L is sufficiently large. To finish up, it remains to show that the pruning procedure PRUNE() is capable of returning a rough estimate for each tensor factor without duplication. This is accomplished in the following lemma.

Lemma 5.17. *Instate the assumptions of Theorem 5.9. On the event that the results in Lemma 5.13, Lemma 5.14, Lemma 5.15 and Lemma 5.16 hold for all $1 \leq i \leq r$, there exists a permutation $\pi(\cdot) : [d] \mapsto [d]$ such that: for each $1 \leq i \leq r$, $\left(\lambda_i, \boldsymbol{w}^i\right)$ and $\left(\lambda_{\pi(i)}^\star, \overline{\boldsymbol{u}}_{\pi(i)}^\star\right)$ satisfy (68a), (68b) and (68c); $\left(\lambda_i, \boldsymbol{w}^i\right)$ and $\left(\lambda_i^{(m)}, \boldsymbol{w}^{i,(m)}\right)_{i=1}^r$ obey (69a), (69b) and (69c) for all $1 \leq m \leq d$, where $\left\{\lambda_i, \boldsymbol{w}^i\right\}_{i=1}^r$ and $\left\{\lambda_i^{(m)}, \boldsymbol{w}^{i,(m)}\right\}_{i=1}^r$ are outputs of Algorithm 3 and Algorithm 6, respectively.*

Proof. See Appendix B.9.   ∎

6 Discussion

The current paper uncovers the possibility of efficiently and stably completing a low-CP-rank tensor from partial and noisy entries. Perhaps somewhat unexpectedly, despite the high degree of nonconvexity, this problem can be solved to optimal statistical accuracy within nearly linear time, provided that the tensor of interest is well-conditioned, incoherent, and of constant rank. To the best of our knowledge, this intriguing message has not been shown in the prior literature.

Moving forward, one pressing issue is to understand how to improve the algorithmic and theoretical dependency upon the tensor rank r of the proposed method. Ideally one would desire a fast algorithm whose sample complexity scales as $rd^{1.5}$, an order that is provably achievable by the sum-of-squares hierarchy. Additionally, in contrast to the matrix counterpart where the rank is upper bounded by the matrix dimension, the tensor CP rank is allowed to rise above d, which is commonly referred to as the over-complete case. Unfortunately, our current initialization scheme (i.e. the spectral method) fails to work unless $r < d$, and our local analysis for GD falls of accommodating the scenario with $r > d$. It would be of great interest to develop more powerful algorithms — in addition to more refined analysis — to tackle such an important over-complete regime.

Another tantalizing research direction is the exploration of landscape design for tensor completion. As our heuristic discussions as well as other prior work (e.g. [RM14]) suggest, randomly initialized gradient descent tailored to (4) seems unlikely to work, unless the sample size is significantly larger than the computational limit. This might mean either that there exist spurious local minima in the natural nonconvex least squares formulation (4), or that the optimization landscape of (4) is too flat around some saddle points and hence not amenable to fast computation. It would be interesting to investigate what families of loss functions allow us to rule out bad local minima and eliminate the need of careful initialization, which might be better suited for tensor recovery problems.

Finally, in statistical inference and decision making, one might not be simply satisfied with obtaining a reliable estimate for each missing entry, but would also like to report a short confidence interval which is likely to contain the true entry. This boils down to the fundamental task of uncertainty quantification for tensor completion, which we leave to future investigation.

Acknowledgements

A Proofs for local convergence of GD

In this section, we establish the key lemmas concerning the convergence properties of GD. As one can easily see, treating $\{E_{i,j,k}\}_{1 \leq i,j,k \leq d}$ (resp. $\{\chi_{i,j,k}\}_{1 \leq i,j,k \leq d}$) as independent random variables — which leads to

asymmetric versions of $\boldsymbol{E}$ and Ω — does not affect the order of our results at all. In light of this, we shall adopt such an independent assumption whenever it simplifies our presentation.

A.1 Proof of Lemma 5.1

For notational convenience, for any matrix $\boldsymbol{M} = [\boldsymbol{m}_1, \ldots, \boldsymbol{m}_r] \in \mathbb{R}^{d \times r}$, let

$$\widetilde{\boldsymbol{M}} := [\boldsymbol{m}_1 \otimes \boldsymbol{m}_1, \ldots, \boldsymbol{m}_r \otimes \boldsymbol{m}_r] \in \mathbb{R}^{d^2 \times r}, \tag{87}$$

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

From the Hessian expression (41), one can decompose

$$\mathsf{vec}(\boldsymbol{V})^\top \nabla^2 f(\boldsymbol{U}) \, \mathsf{vec}(\boldsymbol{V})$$

$$= \underbrace{\frac{1}{3p} \left\| \mathcal{P}_\Omega \Big(\sum_{s \in [r]} \left(\boldsymbol{v}_s \otimes \boldsymbol{u}_s^{\otimes 2} + \boldsymbol{u}_s \otimes \boldsymbol{v}_s \otimes \boldsymbol{u}_s + \boldsymbol{u}_s^{\otimes 2} \otimes \boldsymbol{v}_s \right) \Big) \right\|_{\mathrm{F}}^2 - \frac{1}{3p} \left\| \mathcal{P}_\Omega \Big(\sum_{s \in [r]} \left(\boldsymbol{v}_s \otimes \boldsymbol{u}_s^{\star \otimes 2} + \boldsymbol{u}_s^\star \otimes \boldsymbol{v}_s \otimes \boldsymbol{u}_s^\star + \boldsymbol{u}_s^{\star \otimes 2} \otimes \boldsymbol{v}_s \right) \Big) \right\|_{\mathrm{F}}^2}_{=: \alpha_1}$$

$$+ \underbrace{\frac{1}{3p} \left\| \mathcal{P}_\Omega \Big(\sum_{s \in [r]} \left(\boldsymbol{v}_s \otimes \boldsymbol{u}_s^{\star \otimes 2} + \boldsymbol{u}_s^\star \otimes \boldsymbol{v}_s \otimes \boldsymbol{u}_s^\star + \boldsymbol{u}_s^{\star \otimes 2} \otimes \boldsymbol{v}_s \right) \Big) \right\|_{\mathrm{F}}^2 - \frac{1}{3} \left\| \sum_{s \in [r]} \left(\boldsymbol{v}_s \otimes \boldsymbol{u}_s^{\star \otimes 2} + \boldsymbol{u}_s^\star \otimes \boldsymbol{v}_s \otimes \boldsymbol{u}_s^\star + \boldsymbol{u}_s^{\star \otimes 2} \otimes \boldsymbol{v}_s \right) \right\|_{\mathrm{F}}^2}_{=: \alpha_2}$$

$$+ 2 \underbrace{\Big\langle \frac{1}{p} \mathcal{P}_\Omega \Big(\sum_{s \in [r]} \boldsymbol{u}_s^{\otimes 3} - \boldsymbol{T}^\star \Big), \sum_{s \in [r]} \boldsymbol{v}_s^{\otimes 2} \otimes \boldsymbol{u}_s \Big\rangle}_{=: \alpha_3} + \underbrace{\frac{1}{3} \left\| \sum_{s \in [r]} \left(\boldsymbol{v}_s \otimes \boldsymbol{u}_s^{\star \otimes 2} + \boldsymbol{u}_s^\star \otimes \boldsymbol{v}_s \otimes \boldsymbol{u}_s^\star + \boldsymbol{u}_s^{\star \otimes 2} \otimes \boldsymbol{v}_s \right) \right\|_{\mathrm{F}}^2}_{=: \alpha_4}.$$

In what follows, we shall bound each of the above terms separately.

A.1.1 Bounding α_4

With regards to α_4, by symmetry we have

$$\alpha_4 = \left\| \sum_{s \in [r]} \boldsymbol{v}_s \otimes \boldsymbol{u}_s^{\star \otimes 2} \right\|_{\mathrm{F}}^2 + 2 \Big\langle \sum_{s \in [r]} \boldsymbol{v}_s \otimes \boldsymbol{u}_s^{\star \otimes 2}, \sum_{s \in [r]} \boldsymbol{u}_s^{\star \otimes 2} \otimes \boldsymbol{v}_s \Big\rangle. \tag{88}$$

In order to control (88), we first see that

$$\left\| \sum_{s \in [r]} \boldsymbol{v}_s \otimes \boldsymbol{u}_s^{\star \otimes 2} \right\|_{\mathrm{F}} = \left\| \boldsymbol{V} \widetilde{\boldsymbol{U}}^{\star \top} \right\|_{\mathrm{F}}, \tag{89}$$

where $\widetilde{\boldsymbol{U}}^\star$ is as defined in (87). Similar to the proof of Lemma D.1, we can use the fact that $\langle \boldsymbol{u}_i^{\star \otimes 2}, \boldsymbol{u}_j^{\star \otimes 2} \rangle = \langle \boldsymbol{u}_i^\star, \boldsymbol{u}_j^\star \rangle^2$ and (8c) to deduce that

$$\sigma_{\min}\big(\widetilde{\boldsymbol{U}}^\star\big) = \lambda_{\min}^{\star 2/3} \left(1 + o(1) \right) \qquad \text{and} \qquad \sigma_{\max}\big(\widetilde{\boldsymbol{U}}^\star\big) = \lambda_{\max}^{\star 2/3} \left(1 + o(1) \right), \tag{90}$$

provided that $r \ll d/\mu$. This implies that

$$\frac{19}{20} \lambda_{\min}^{\star 2/3} \|\boldsymbol{V}\|_{\mathrm{F}} \leq \sigma_{\min}\big(\widetilde{\boldsymbol{U}}^\star\big) \|\boldsymbol{V}\|_{\mathrm{F}} \leq \left\| \boldsymbol{V} \widetilde{\boldsymbol{U}}^{\star \top} \right\|_{\mathrm{F}} \leq \sigma_{\max}\big(\widetilde{\boldsymbol{U}}^\star\big) \|\boldsymbol{V}\|_{\mathrm{F}} \leq \frac{11}{10} \lambda_{\max}^{\star 2/3} \|\boldsymbol{V}\|_{\mathrm{F}}. \tag{91}$$

(1) Speaking of an upper bound on α_4, we can invoke the Cauchy-Schwarz inequality followed by (91) to reach

$$\alpha_4 \leq 3 \left\| \sum_{s \in [r]} \boldsymbol{v}_s \otimes \boldsymbol{u}_s^{\star \otimes 2} \right\|_{\mathrm{F}}^2 = 3 \left\| \boldsymbol{V} \widetilde{\boldsymbol{U}}^{\star \top} \right\|_{\mathrm{F}}^2 \leq \frac{7}{2} \lambda_{\max}^{\star 4/3} \|\boldsymbol{V}\|_{\mathrm{F}}^2. \tag{92}$$

(2) When it comes to lower bounding α_4, the main step boils down to controlling the inner product term in (88). Applying the Cauchy-Schwartz inequality gives that

$$\Big\langle \sum_{s\in[r]} \boldsymbol{v}_s \otimes \boldsymbol{u}_s^{\star\otimes 2}, \sum_{s\in[r]} \boldsymbol{u}_s^{\star\otimes 2} \otimes \boldsymbol{v}_s \Big\rangle = \sum_{s\in[r]} \langle \boldsymbol{v}_s, \boldsymbol{u}_s^\star \rangle^2 \|\boldsymbol{u}_s^\star\|_2^2 + \sum_{s_1\neq s_2} \langle \boldsymbol{v}_{s_1}, \boldsymbol{u}_{s_2}^\star \rangle \langle \boldsymbol{u}_{s_1}^\star, \boldsymbol{v}_{s_2} \rangle \langle \boldsymbol{u}_{s_1}^\star, \boldsymbol{u}_{s_2}^\star \rangle$$

$$\geq -\max_{s_1\neq s_2} |\langle \boldsymbol{u}_{s_1}^\star, \boldsymbol{u}_{s_2}^\star \rangle| \sum_{s_1\neq s_2} \|\boldsymbol{v}_{s_1}\|_2 \|\boldsymbol{u}_{s_1}^\star\|_2 \|\boldsymbol{v}_{s_2}\|_2 \|\boldsymbol{u}_{s_2}^\star\|_2$$

$$\geq -\max_{s_1\neq s_2} |\langle \boldsymbol{u}_{s_1}^\star, \boldsymbol{u}_{s_2}^\star \rangle| \Big(\sum_{s\in[r]} \|\boldsymbol{v}_s\|_2 \|\boldsymbol{u}_s^\star\|_2 \Big)^2$$

$$\overset{(i)}{\geq} -\max_{s_1\neq s_2} |\langle \boldsymbol{u}_{s_1}^\star, \boldsymbol{u}_{s_2}^\star \rangle| \|\boldsymbol{U}^\star\|_F^2 \|\boldsymbol{V}\|_F^2$$

$$\geq -r\sqrt{\frac{\mu}{d}} \lambda_{\max}^{\star 4/3} \|\boldsymbol{V}\|_F^2 \geq -\frac{1}{40} \lambda_{\min}^{\star 4/3} \|\boldsymbol{V}\|_F^2,$$

where (i) comes from Cauchy-Schwarz, and the last line follows from (8c), (8d) as well as the condition that $r \ll \sqrt{d/\mu}$ and $\kappa \asymp 1$. Therefore, we can lower bound α_4 by (with the assistance of (91))

$$\alpha_4 = \Big\| \sum_{s\in[r]} \boldsymbol{v}_s \otimes \boldsymbol{u}_s^{\star\otimes 2} \Big\|_F^2 + 2 \Big\langle \sum_{s\in[r]} \boldsymbol{v}_s \otimes \boldsymbol{u}_s^{\star\otimes 2}, \sum_{s\in[r]} \boldsymbol{u}_s^{\star\otimes 2} \otimes \boldsymbol{v}_s \Big\rangle$$

$$\geq \frac{19}{20} \lambda_{\min}^{\star 4/3} \|\boldsymbol{V}\|_F^2 - \frac{1}{20} \lambda_{\min}^{\star 4/3} \|\boldsymbol{V}\|_F^2 \geq \frac{9}{10} \lambda_{\min}^{\star 4/3} \|\boldsymbol{V}\|_F^2. \tag{93}$$

A.1.2 Bounding α_1

When it comes to α_1, we can expand

$$\Big\| \mathcal{P}_\Omega \Big(\sum_{s\in[r]} \boldsymbol{v}_s \otimes \boldsymbol{u}_s^{\otimes 2} \Big) + \mathcal{P}_\Omega \Big(\sum_{s\in[r]} \boldsymbol{u}_s \otimes \boldsymbol{v}_s \otimes \boldsymbol{u}_s \Big) + \mathcal{P}_\Omega \Big(\sum_{s\in[r]} \boldsymbol{u}_s^{\otimes 2} \otimes \boldsymbol{v}_s \Big) \Big\|_F^2$$

$$= 3 \Big\| \mathcal{P}_\Omega \Big(\sum_{s\in[r]} \boldsymbol{v}_s \otimes \boldsymbol{u}_s^{\otimes 2} \Big) \Big\|_F^2 + 6 \Big\langle \mathcal{P}_\Omega \Big(\sum_{s\in[r]} \boldsymbol{v}_s \otimes \boldsymbol{u}_s^{\otimes 2} \Big), \mathcal{P}_\Omega \Big(\sum_{s\in[r]} \boldsymbol{u}_s^{\otimes 2} \otimes \boldsymbol{v}_s \Big) \Big\rangle;$$

we can decompose $\Big\| \mathcal{P}_\Omega \big(\sum_{s\in[r]} \boldsymbol{v}_s \otimes \boldsymbol{u}_s^{\star\otimes 2} \big) + \mathcal{P}_\Omega \big(\sum_{s\in[r]} \boldsymbol{u}_s^\star \otimes \boldsymbol{v}_s \otimes \boldsymbol{u}_s^\star \big) + \mathcal{P}_\Omega \big(\sum_{s\in[r]} \boldsymbol{u}_s^{\star\otimes 2} \otimes \boldsymbol{v}_s \big) \Big\|_F^2$ in a similar way. As a consequence,

$$\alpha_1 = \underbrace{\frac{2}{p} \Big(\Big\langle \mathcal{P}_\Omega \Big(\sum_{s\in[r]} \boldsymbol{v}_s \otimes \boldsymbol{u}_s^{\otimes 2} \Big), \mathcal{P}_\Omega \Big(\sum_{s\in[r]} \boldsymbol{u}_s^{\otimes 2} \otimes \boldsymbol{v}_s \Big) \Big\rangle - \Big\langle \mathcal{P}_\Omega \Big(\sum_{s\in[r]} \boldsymbol{v}_s \otimes \boldsymbol{u}_s^{\star\otimes 2} \Big), \mathcal{P}_\Omega \Big(\sum_{s\in[r]} \boldsymbol{u}_s^{\star\otimes 2} \otimes \boldsymbol{v}_s \Big) \Big\rangle \Big)}_{=:\beta_1}$$

$$+ \underbrace{\frac{1}{p} \Big(\Big\| \mathcal{P}_\Omega \Big(\sum_{s\in[r]} \boldsymbol{v}_s \otimes \boldsymbol{u}_s^{\otimes 2} \Big) \Big\|_F^2 - \Big\| \mathcal{P}_\Omega \Big(\sum_{s\in[r]} \boldsymbol{v}_s \otimes \boldsymbol{u}_s^{\star\otimes 2} \Big) \Big\|_F^2 \Big)}_{=:\beta_2}.$$

We will derive an upper bound on β_1 in the sequel; the same method immediately applies to β_2.

For notational convenience, let us define

$$\boldsymbol{\Delta} := \boldsymbol{U} - \boldsymbol{U}^\star, \quad \boldsymbol{\Delta}_s := \boldsymbol{u}_s - \boldsymbol{u}_s^\star, \quad \widetilde{\boldsymbol{\Delta}} := [\boldsymbol{\Delta}_1 \otimes \boldsymbol{\Delta}_1, \cdots, \boldsymbol{\Delta}_r \otimes \boldsymbol{\Delta}_r] \in \mathbb{R}^{d^2 \times r}. \tag{94}$$

Then one can write

$$\frac{1}{2}\beta_1 = \frac{1}{p} \Big\langle \mathcal{P}_\Omega \Big(\sum_{s\in[r]} \boldsymbol{v}_s \otimes (\boldsymbol{\Delta}_s + \boldsymbol{u}_s^\star)^{\otimes 2} \Big), \mathcal{P}_\Omega \Big(\sum_{s\in[r]} (\boldsymbol{\Delta}_s + \boldsymbol{u}_s^\star)^{\otimes 2} \otimes \boldsymbol{v}_s \Big) \Big\rangle - \frac{1}{p} \Big\langle \mathcal{P}_\Omega \Big(\sum_{s\in[r]} \boldsymbol{v}_s \otimes \boldsymbol{u}_s^{\star\otimes 2} \Big), \mathcal{P}_\Omega \Big(\sum_{s\in[r]} \boldsymbol{u}_s^{\star\otimes 2} \otimes \boldsymbol{v}_s \Big) \Big\rangle$$

$$= \frac{1}{p} \Big\langle \mathcal{P}_\Omega\Big(\sum_{s\in[r]} \boldsymbol{v}_s \otimes \boldsymbol{u}_s^{\star\otimes 2}\Big), \mathcal{P}_\Omega\Big(\sum_{s\in[r]} \boldsymbol{u}_s^\star \otimes \boldsymbol{\Delta}_s \otimes \boldsymbol{v}_s\Big) + \mathcal{P}_\Omega\Big(\sum_{s\in[r]} \boldsymbol{\Delta}_s \otimes \boldsymbol{u}_s^\star \otimes \boldsymbol{v}_s\Big) + \mathcal{P}_\Omega\Big(\sum_{s\in[r]} \boldsymbol{\Delta}_s \otimes \boldsymbol{\Delta}_s \otimes \boldsymbol{v}_s\Big) \Big\rangle$$

$$+ \frac{1}{p} \Big\langle \mathcal{P}_\Omega\Big(\sum_{s\in[r]} \boldsymbol{v}_s \otimes \boldsymbol{u}_s^\star \otimes \boldsymbol{\Delta}_s\Big) + \mathcal{P}_\Omega\Big(\sum_{s\in[r]} \boldsymbol{v}_s \otimes \boldsymbol{\Delta}_s \otimes \boldsymbol{u}_s^\star\Big) + \mathcal{P}_\Omega\Big(\sum_{s\in[r]} \boldsymbol{v}_s \otimes \boldsymbol{\Delta}_s^{\otimes 2}\Big),$$

$$\mathcal{P}_\Omega\Big(\sum_{s\in[r]} \boldsymbol{u}_s^{\star\otimes 2} \otimes \boldsymbol{v}_s\Big) + \mathcal{P}_\Omega\Big(\sum_{s\in[r]} \boldsymbol{u}_s^\star \otimes \boldsymbol{\Delta}_s \otimes \boldsymbol{v}_s\Big) + \mathcal{P}_\Omega\Big(\sum_{s\in[r]} \boldsymbol{\Delta}_s \otimes \boldsymbol{u}_s^\star \otimes \boldsymbol{v}_s\Big) + \mathcal{P}_\Omega\Big(\sum_{s\in[r]} \boldsymbol{\Delta}_s \otimes \boldsymbol{\Delta}_s \otimes \boldsymbol{v}_s\Big) \Big\rangle.$$

Apply the Cauchy-Schwartz inequality to yield that

$$|\beta_1| \lesssim \frac{1}{p} \Big\| \mathcal{P}_\Omega\Big(\sum_{s\in[r]} \boldsymbol{v}_s \otimes \boldsymbol{u}_s^{\star\otimes 2}\Big)\Big\|_{\mathrm{F}} \Big(2\Big\| \mathcal{P}_\Omega\Big(\sum_{s\in[r]} \boldsymbol{u}_s^\star \otimes \boldsymbol{\Delta}_s \otimes \boldsymbol{v}_s\Big)\Big\|_{\mathrm{F}} + \Big\| \mathcal{P}_\Omega\Big(\sum_{s\in[r]} \boldsymbol{\Delta}_s \otimes \boldsymbol{\Delta}_s \otimes \boldsymbol{v}_s\Big)\Big\|_{\mathrm{F}}\Big)$$

$$+ \frac{1}{p} \Big(\Big\| \mathcal{P}_\Omega\Big(\sum_{s\in[r]} \boldsymbol{u}_s^{\star\otimes 2} \otimes \boldsymbol{v}_s\Big)\Big\|_{\mathrm{F}} + 2\Big\| \mathcal{P}_\Omega\Big(\sum_{s\in[r]} \boldsymbol{u}_s^\star \otimes \boldsymbol{\Delta}_s \otimes \boldsymbol{v}_s\Big)\Big\|_{\mathrm{F}} + \Big\| \mathcal{P}_\Omega\Big(\sum_{s\in[r]} \boldsymbol{\Delta}_s \otimes \boldsymbol{\Delta}_s \otimes \boldsymbol{v}_s\Big)\Big\|_{\mathrm{F}}\Big)$$

$$\cdot \Big(2\Big\| \mathcal{P}_\Omega\Big(\sum_{s\in[r]} \boldsymbol{v}_s \otimes \boldsymbol{u}_s^\star \otimes \boldsymbol{\Delta}_s\Big)\Big\|_{\mathrm{F}} + \Big\| \mathcal{P}_\Omega\Big(\sum_{s\in[r]} \boldsymbol{v}_s \otimes \boldsymbol{\Delta}_s^{\otimes 2}\Big)\Big\|_{\mathrm{F}}\Big). \tag{95}$$

Before we bound the above quantities, we pause to make the following observations. In view of the assumptions of this lemma that $\delta \ll 1/\sqrt{r} \leq 1$, the following holds for all $i \in [r]$:

$$\|\boldsymbol{\Delta}_i\|_2 \leq \|\boldsymbol{U} - \boldsymbol{U}^\star\|_{\mathrm{F}} \leq \delta \|\boldsymbol{U}^\star\|_{\mathrm{F}} \leq \delta\sqrt{r} \lambda_{\max}^{\star 1/3} \ll \lambda_{\max}^{\star 1/3}, \tag{96a}$$

$$\|\boldsymbol{\Delta}_i\|_\infty \leq \|\boldsymbol{U} - \boldsymbol{U}^\star\|_{2,\infty} \leq \delta \|\boldsymbol{U}^\star\|_{2,\infty} \leq \delta\sqrt{\frac{\mu r}{d}} \lambda_{\max}^{\star 1/3} \ll \sqrt{\frac{\mu}{d}}\lambda_{\max}^{\star 1/3}, \tag{96b}$$

$$\|\boldsymbol{u}_i\|_2 \leq \|\boldsymbol{u}_i^\star\|_2 + \|\boldsymbol{\Delta}_i\|_2 \leq 2\lambda_{\max}^{\star 1/3}, \tag{96c}$$

$$\|\boldsymbol{u}_i\|_\infty \leq \|\boldsymbol{u}_i^\star\|_\infty + \|\boldsymbol{\Delta}_i\|_\infty \leq 2\sqrt{\frac{\mu}{d}}\lambda_{\max}^{\star 1/3}, \tag{96d}$$

Consequently, we also know that

$$\big\| \widetilde{\boldsymbol{\Delta}}\big\|_{2,\infty} \leq \max_{1\leq i\leq r} \|\boldsymbol{\Delta}_i\|_\infty \|\boldsymbol{\Delta}\|_{2,\infty} \leq \delta^2 \|\boldsymbol{U}^\star\|_{2,\infty}^2. \tag{97}$$

Now, we proceed to prove the claim. Let us define $\mathcal{S}_i := \{\underline{j} \in [d]^2 \mid \chi_{ij_1 j_2} = 1\}$ for each $i \in [d]$. Applying the Chernoff bound and the union bound yields that: with probability at least $1 - O(d^{-10})$ one has

$$\max_{i\in[d]} |\mathcal{S}_i| \lesssim d^2 p. \tag{98}$$

provided $p \gg d^{-2}\log d$. It then follows from the Cauchy-Schwarz inequality that

$$\frac{1}{p} \Big\| \mathcal{P}_\Omega\Big(\sum_{s\in[r]} \boldsymbol{v}_s \otimes \boldsymbol{\Delta}_s^{\otimes 2}\Big)\Big\|_{\mathrm{F}}^2 = \frac{1}{p} \sum_{i\in[d], \underline{j}\in[d]^2} \chi_{ij_1 j_2} \big\langle \boldsymbol{V}_{i,:}, \widetilde{\boldsymbol{\Delta}}_{\underline{j},:}\big\rangle^2$$

$$\leq \frac{1}{p} \sum_{i\in[d]} \|\boldsymbol{V}_{i,:}\|_2^2 \sum_{\underline{j}\in\mathcal{S}_i} \|\widetilde{\boldsymbol{\Delta}}_{\underline{j},:}\|_2^2$$

$$\leq \frac{1}{p} \max_{i\in[d]} |\mathcal{S}_i| \big\|\widetilde{\boldsymbol{\Delta}}\big\|_{2,\infty}^2 \|\boldsymbol{V}\|_{\mathrm{F}}^2$$

$$\overset{(\mathrm{i})}{\lesssim} d^2\delta^4 \|\boldsymbol{U}^\star\|_{2,\infty}^4 \|\boldsymbol{V}\|_{\mathrm{F}}^2 \lesssim \delta^4\mu^2 r^2 \lambda_{\max}^{\star 4/3} \|\boldsymbol{V}\|_{\mathrm{F}}^2,$$

where (i) arises from (97) and (98). In a similar manner, we can derive

$$\frac{1}{p} \Big\| \mathcal{P}_\Omega\Big(\sum_{s\in[r]} \boldsymbol{u}_s^\star \otimes \boldsymbol{v}_s \otimes \boldsymbol{\Delta}_s\Big)\Big\|_{\mathrm{F}}^2 \lesssim \frac{1}{p} \max_{1\leq i\leq d} |\mathcal{S}_i| \max_{1\leq s\leq r} \|\boldsymbol{u}_s^\star\|_\infty^2 \|\boldsymbol{\Delta}\|_{2,\infty}^2 \|\boldsymbol{V}\|_{\mathrm{F}}^2$$

$$\lesssim d^2 \delta^2 \max_{1 \le i \le r} \|\boldsymbol{u}_s^\star\|_\infty^2 \|\boldsymbol{U}^\star\|_{2,\infty}^2 \|\boldsymbol{V}\|_{\mathrm{F}}^2$$

$$\lesssim \delta^2 \mu^2 r \lambda_{\max}^{\star 4/3} \|\boldsymbol{V}\|_{\mathrm{F}}^2 .$$

Regarding $p^{-1} \big\| \mathcal{P}_\Omega \big(\sum_{s \in [r]} \boldsymbol{v}_s \otimes \boldsymbol{u}_s^{\star \otimes 2} \big) \big\|_{\mathrm{F}}^2$, we apply [YZ16, Lemma 5] (with slight modification, which we omit here for brevity) to show that: with probability exceeding $1 - O(d^{-}10)$

$$\frac{1}{\sqrt{p}} \Big\| \mathcal{P}_\Omega \Big(\sum_{s \in [r]} \boldsymbol{v}_s \otimes \boldsymbol{u}_s^{\star \otimes 2} \Big) \Big\|_{\mathrm{F}} \le \frac{3}{2} \Big\| \sum_{s \in [r]} \boldsymbol{v}_s \otimes \boldsymbol{u}_s^{\star \otimes 2} \Big\|_{\mathrm{F}} = \frac{3}{2} \|\boldsymbol{V} \widetilde{\boldsymbol{U}}^{\star \top}\|_{\mathrm{F}} \le \frac{3}{2} \|\widetilde{\boldsymbol{U}}^\star\| \|\boldsymbol{V}\|_{\mathrm{F}} \lesssim \lambda_{\max}^{\star 2/3} \|\boldsymbol{V}\|_{\mathrm{F}} .$$

under the sample size assumptin that $p \gg \mu^2 r^2 d^{-2} \log d$. Here the last inequality makes use of (90). It is self-evident that the above bounds also hold for quantities that appear in (95). Since $0 < \delta \ll 1/\sqrt{r} < 1$, we obtain

$$|\beta_1| \lesssim \delta \mu \sqrt{r} \lambda_{\max}^{\star 4/3} \|\boldsymbol{V}\|_{\mathrm{F}}^2 .$$

The same upper bound holds for any other β_2. Therefore, as long as $0 < \delta \ll 1/(\mu\sqrt{r})$ and $\kappa \asymp 1$, we have

$$|\alpha_1| \le \frac{1}{10} \lambda_{\min}^{\star 4/3} \|\boldsymbol{V}\|_{\mathrm{F}}^2 . \tag{99}$$

A.1.3 Bounding α_2

Regarding α_2, applying [YZ16, Lemma 5] (with slight modification, which we omit here for brevity) implies that: if $p \gg \mu^2 r^2 d^{-2} \log d$, then with probability exceeding $1 - O(d^{-10})$,

$$\left| \frac{1}{2p} \Big\| \mathcal{P}_\Omega \Big(\sum_{s \in [r]} (\boldsymbol{v}_s \otimes \boldsymbol{u}_s^{\star \otimes 2} + \boldsymbol{u}_s^\star \otimes \boldsymbol{v}_s \otimes \boldsymbol{u}_s^\star + \boldsymbol{u}_s^{\star \otimes 2} \otimes \boldsymbol{v}_s) \Big) \Big\|_{\mathrm{F}}^2 \right.$$

$$\left. - \frac{1}{2} \Big\| \sum_{s \in [r]} (\boldsymbol{v}_s \otimes \boldsymbol{u}_s^{\star \otimes 2} + \boldsymbol{u}_s^\star \otimes \boldsymbol{v}_s \otimes \boldsymbol{u}_s^\star + \boldsymbol{u}_s^{\star \otimes 2} \otimes \boldsymbol{v}_s) \Big\|_{\mathrm{F}}^2 \right|$$

$$\le \frac{1}{100} \Big\| \sum_{s \in [r]} (\boldsymbol{v}_s \otimes \boldsymbol{u}_s^{\star \otimes 2} + \boldsymbol{u}_s^\star \otimes \boldsymbol{v}_s \otimes \boldsymbol{u}_s^\star + \boldsymbol{u}_s^{\star \otimes 2} \otimes \boldsymbol{v}_s) \Big\|_{\mathrm{F}}^2 = \frac{1}{100} \alpha_4 \le \frac{1}{10} \lambda_{\min}^{\star 4/3} \|\boldsymbol{V}\|_{\mathrm{F}}^2 .$$

Here the last inequality arises from (92).

A.1.4 Bounding α_3

We now move on to bounding α_3. The triangle inequality gives

$$\left| \Big\langle p^{-1} \mathcal{P}_\Omega \Big(\sum_{s \in [r]} \boldsymbol{u}_s^{\otimes 3} - \boldsymbol{T}^\star \Big), \sum_{s \in [r]} \boldsymbol{v}_s^{\otimes 2} \otimes \boldsymbol{u}_s \Big\rangle \right| \le \sum_{s \in [r]} \left| \Big\langle p^{-1} \mathcal{P}_\Omega \Big(\sum_{i \in [r]} \boldsymbol{u}_i^{\otimes 3} - \boldsymbol{T}^\star \Big) \times_3 \boldsymbol{u}_s, \boldsymbol{v}_s \boldsymbol{v}_s^\top \Big\rangle \right|$$

$$\le \max_{s \in [r]} \Big\| p^{-1} \mathcal{P}_\Omega \Big(\sum_{i \in [r]} \boldsymbol{u}_i^{\otimes 3} - \boldsymbol{T}^\star \Big) \times_3 \boldsymbol{u}_s \Big\| \sum_{s \in [r]} \|\boldsymbol{v}_s\|_2^2$$

$$\le \max_{s \in [r]} \Big\| p^{-1} \mathcal{P}_\Omega \Big(\sum_{i \in [r]} \boldsymbol{u}_i^{\otimes 3} - \boldsymbol{T}^\star \Big) \times_3 \boldsymbol{u}_s \Big\| \|\boldsymbol{V}\|_{\mathrm{F}}^2 .$$

Recall the definitions of $\boldsymbol{\Delta}$ and $\boldsymbol{\Delta}_i$ in (94). Fix an arbitrary $s \in [r]$. From the definition of the operator norm and the triangle inequality, we can derive

$$\Big\| p^{-1} \mathcal{P}_\Omega \Big(\sum_{i \in [r]} \boldsymbol{u}_i^{\otimes 3} - \boldsymbol{T}^\star \Big) \times_3 \boldsymbol{u}_s \Big\| \le \|\boldsymbol{u}_s\|_2 \Big\| p^{-1} \mathcal{P}_\Omega \Big(\sum_{i \in [r]} \boldsymbol{u}_i^{\otimes 3} - \boldsymbol{T}^\star \Big) \Big\| . \tag{100}$$

In order to upper bound $\big\| p^{-1} \mathcal{P}_\Omega \big(\sum_{i \in [r]} \boldsymbol{u}_i^{\otimes 3} - \boldsymbol{T}^\star \big) \big\|$ as required in (100), we invoke the following simple fact, which follows immediately from the definition of the operator norm. Here and throughout, for any tensor $\boldsymbol{A} \in \mathbb{R}^{d \times d \times d}$ we denote

$$|\boldsymbol{A}| := \big[|A_{i,j,k}| \big]_{1 \le i,j,k \le d} \in \mathbb{R}^{d \times d \times d}.$$

Lemma A.1. *Consider any tensor $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{d \times d \times d}$ obeying $|B_{i,j,k}| \geq |A_{i,j,k}|$ for all $1 \leq i, j, k \leq d$. One has*

$$\|\boldsymbol{A}\| \leq \||\boldsymbol{A}|\| \leq \||\boldsymbol{B}|\|. \tag{101}$$

With this lemma in mind, we are ready to derive that

$$\left\| p^{-1} \mathcal{P}_\Omega \Big(\sum_{i \in [r]} \boldsymbol{u}_i^{\otimes 3} - \boldsymbol{T}^\star \Big) \right\| \leq \left\| p^{-1} \mathcal{P}_\Omega \Big(\Big| \sum_{i \in [r]} \boldsymbol{u}_i^{\otimes 3} - \boldsymbol{T}^\star \Big| \Big) \right\|$$

$$\leq \left\| \sum_{i \in [r]} \boldsymbol{u}_i^{\otimes 3} - \boldsymbol{T}^\star \right\|_\infty \left\| p^{-1} \mathcal{P}_\Omega (\boldsymbol{1}^{\otimes 3}) \right\|.$$

Here, $\boldsymbol{1}$ stands for the all-one vector in $\mathbb{R}^d$. This suggests that we shall upper bound $\left\| p^{-1} \mathcal{P}_\Omega (\boldsymbol{1}^{\otimes 3}) \right\|$ and $\left\| \sum_{i \in [r]} \boldsymbol{u}_i^{\otimes 3} - \boldsymbol{T}^\star \right\|_\infty$.

Given that $p \gtrsim d^{-3/2} \log^3 d$, applying Lemma D.2 indicates that with probability at least $1 - O(d^{-10})$,

$$\left\| p^{-1} \mathcal{P}_\Omega (\boldsymbol{1}^{\otimes 3}) - \boldsymbol{1}^{\otimes 3} \right\| \lesssim \frac{\log^3 d}{p} + \sqrt{\frac{d \log^5 d}{p}} \lesssim d^{3/2}. \tag{102}$$

Moreover, it is straightforward to see that $\left\| \boldsymbol{1}^{\otimes 3} \right\| = \|\boldsymbol{1}\|_2^3 = d^{3/2}$. Therefore, one has

$$\left\| p^{-1} \mathcal{P}_\Omega (\boldsymbol{1}^{\otimes 3}) \right\| \lesssim d^{3/2}. \tag{103}$$

Next, we turn to $\|\cdot\|_\infty$. We first expand

$$\sum_{i \in [r]} \boldsymbol{u}_i^{\otimes 3} - \boldsymbol{T}^\star = \sum_{i \in [r]} \big((\boldsymbol{\Delta}_i + \boldsymbol{u}_i^\star)^{\otimes 3} - \boldsymbol{u}_i^{\star \otimes 3} \big) = \sum_{i \in [r]} \boldsymbol{\Delta}_i \otimes \boldsymbol{u}_i^{\star \otimes 2} + \sum_{i \in [r]} \boldsymbol{u}_i^\star \otimes \boldsymbol{\Delta}_i \otimes \boldsymbol{u}_i^\star + \sum_{i \in [r]} \boldsymbol{u}_i^{\star \otimes 2} \otimes \boldsymbol{\Delta}_i$$

$$+ \sum_{i \in [r]} \boldsymbol{\Delta}_i^{\otimes 2} \otimes \boldsymbol{u}_i^\star + \sum_{i \in [r]} \boldsymbol{\Delta}_i \otimes \boldsymbol{u}_i^\star \otimes \boldsymbol{\Delta}_i + \sum_{i \in [r]} \boldsymbol{u}_i^\star \otimes \boldsymbol{\Delta}_i^{\otimes 2} + \sum_{i \in [r]} \boldsymbol{\Delta}_i^{\otimes 3}.$$

By symmetry, it suffices to control $\left\| \sum_{i \in [r]} \boldsymbol{u}_i^{\star \otimes 2} \otimes \boldsymbol{\Delta}_i \right\|$, $\left\| \sum_{i \in [r]} \boldsymbol{\Delta}_i^{\otimes 2} \otimes \boldsymbol{u}_i^\star \right\|$ and $\left\| \sum_{i \in [r]} \boldsymbol{\Delta}_i^{\otimes 3} \right\|$. Let us look at the first term. Towards this, for each $(i, j, k) \in [d]^3$, we can use the Cauchy-Schwartz inequality to control

$$\left| \Big(\sum_{1 \leq s \leq r} \boldsymbol{u}_s^{\star \otimes 2} \otimes \boldsymbol{\Delta}_s \Big)_{i,j,k} \right| = \left| \sum_{1 \leq s \leq r} (\boldsymbol{u}_s^\star)_i (\boldsymbol{u}_s^\star)_j (\boldsymbol{\Delta}_s)_k \right|$$

$$\leq \Big(\sum_{1 \leq s \leq r} \big[(\boldsymbol{u}_s^\star)_i \big]^2 \Big)^{1/2} \Big(\sum_{1 \leq s \leq r} \big[(\boldsymbol{u}_s^\star)_j \big]^2 \big[(\boldsymbol{\Delta}_s)_k \big]^2 \Big)^{1/2}$$

$$\leq \|\boldsymbol{U}^\star\|_{2,\infty} \|\boldsymbol{\Delta}\|_{2,\infty} \max_{1 \leq s \leq r} \|\boldsymbol{u}_s^\star\|_\infty$$

$$\leq \delta \|\boldsymbol{U}^\star\|_{2,\infty}^2 \max_{1 \leq s \leq r} \|\boldsymbol{u}_s^\star\|_\infty,$$

which implies that

$$\left\| \sum_{1 \leq s \leq r} \boldsymbol{u}_s^{\star \otimes 2} \otimes \boldsymbol{\Delta}_s \right\|_\infty \leq \delta \|\boldsymbol{U}^\star\|_{2,\infty}^2 \max_{1 \leq s \leq r} \|\boldsymbol{u}_s^\star\|_\infty \lesssim \frac{\delta \mu^{3/2} r \lambda_{\max}^\star}{d^{3/2}}. \tag{104}$$

In a similar manner, we can control the remaining two terms by

$$\left\| \sum_{i \in [r]} \boldsymbol{\Delta}_i^{\otimes 2} \otimes \boldsymbol{u}_i^\star \right\| \leq \|\boldsymbol{\Delta}\|_{2,\infty}^2 \max_{1 \leq s \leq r} \|\boldsymbol{u}_s^\star\|_\infty \leq \delta^2 \|\boldsymbol{U}^\star\|_{2,\infty}^2 \max_{1 \leq s \leq r} \|\boldsymbol{u}_s^\star\|_\infty \leq \frac{\delta^2 \mu^{3/2} r \lambda_{\max}^\star}{d^{3/2}}; \tag{105}$$

$$\left\| \sum_{i \in [r]} \boldsymbol{\Delta}_i^{\otimes 3} \right\| \leq \|\boldsymbol{\Delta}\|_{2,\infty}^2 \max_{1 \leq s \leq r} \|\boldsymbol{\Delta}_s\|_\infty \leq \delta^3 \|\boldsymbol{U}^\star\|_{2,\infty}^3 \leq \frac{\delta^3 \mu^{3/2} r^{3/2} \lambda_{\max}^\star}{d^{3/2}}. \tag{106}$$

Recall that $0 < \delta \ll 1/r \leq 1$. Putting these together reveals that

$$\left\| \sum_{i \in [r]} \boldsymbol{u}_i^{\otimes 3} - \boldsymbol{T}^\star \right\| \lesssim \frac{\delta \mu^{3/2} r \lambda_{\max}^\star}{d^{3/2}}. \tag{107}$$

32

This combined with (103) yields

$$\left\| p^{-1}\mathcal{P}_\Omega\Big(\sum_{i\in[r]} \boldsymbol{u}_i^{\otimes 3} - \boldsymbol{T}^\star\Big)\right\| \lesssim \delta\mu^{3/2} r\lambda_{\max}^\star,$$

thus indicating that

$$\left\| p^{-1}\mathcal{P}_\Omega\Big(\sum_{i\in[r]} \boldsymbol{u}_i^{\otimes 3} - \boldsymbol{T}^\star\Big) \times_3 \boldsymbol{u}_s\right\| \le \|\boldsymbol{u}_s\|_2 \left\| p^{-1}\mathcal{P}_\Omega\Big(\sum_{i\in[r]} \boldsymbol{u}_i^{\otimes 3} - \boldsymbol{T}^\star\Big)\right\| \lesssim \delta\mu^{3/2} r\lambda_{\max}^{\star 4/3}, \qquad (108)$$

where we use (96c) in the last step. In view of the condition that $\delta \ll 1/(\mu^{3/2}r)$ and the assumption $\kappa \asymp 1$, one has with probability greater than $1 - O(d^{-10})$,

$$|\alpha_3| \le \frac{1}{10}\lambda_{\min}^{\star 4/3} \|\boldsymbol{V}\|_{\mathrm{F}}^2. \qquad (109)$$

A.1.5 Putting all this together

Note that the above bounds hold uniformly for all $\boldsymbol{V}$. Therefore, combining upper bounds for α_i and the union bound, we conclude that with probability exceeding $1 - O(d^{-10})$,

$$\mathsf{vec}(\boldsymbol{V})^\top \nabla^2 f(\boldsymbol{U})\mathsf{vec}(\boldsymbol{V}) \ge \alpha_4 - |\alpha_1| - |\alpha_2| - |\alpha_3| \ge \frac{1}{2}\lambda_{\min}^{\star 4/3} \|\boldsymbol{V}\|_{\mathrm{F}}^2 \qquad (110)$$

$$\mathsf{vec}(\boldsymbol{V})^\top \nabla^2 f(\boldsymbol{U})\mathsf{vec}(\boldsymbol{V}) \le \alpha_4 + |\alpha_1| + |\alpha_2| + |\alpha_3| \le 4\lambda_{\max}^{\star 4/3} \|\boldsymbol{V}\|_{\mathrm{F}}^2 \qquad (111)$$

as claimed.

A.2 Proof of Lemma 5.2

From (47a) and (47c), we use the triangle inequality to obtain

$$\begin{aligned}
\left\|\boldsymbol{U}^{t,(m)} - \boldsymbol{U}^\star\right\|_{\mathrm{F}} &\le \left\|\boldsymbol{U}^{t,(m)} - \boldsymbol{U}^t\right\|_{\mathrm{F}} + \left\|\boldsymbol{U}^t - \boldsymbol{U}^\star\right\|_{\mathrm{F}} \\
&\le \left(2C_1\rho^t\mathcal{E}_{\mathsf{local}} + 2C_2\frac{\sigma}{\lambda_{\min}^\star}\sqrt{\frac{d\log d}{p}}\right)\|\boldsymbol{U}^\star\|_{\mathrm{F}}
\end{aligned}$$

Similarly, we can combine (47b) and (47c) to obtain that

$$\begin{aligned}
\left\|\boldsymbol{U}^{t,(m)} - \boldsymbol{U}^\star\right\|_{2,\infty} &\le \left\|\boldsymbol{U}^{t,(m)} - \boldsymbol{U}^t\right\|_{2,\infty} + \left\|\boldsymbol{U}^t - \boldsymbol{U}^\star\right\|_{\mathrm{F}} \le \left\|\boldsymbol{U}^{t,(m)} - \boldsymbol{U}^t\right\|_{\mathrm{F}} + \left\|\boldsymbol{U}^t - \boldsymbol{U}^\star\right\|_{\mathrm{F}} \\
&\le \left(C_5\rho^t\mathcal{E}_{\mathsf{local}} + C_6\frac{\sigma}{\lambda_{\min}^\star}\sqrt{\frac{d\log d}{p}}\right)\|\boldsymbol{U}^\star\|_{2,\infty} + \left(C_3\rho^t\mathcal{E}_{\mathsf{local}} + C_4\frac{\sigma}{\lambda_{\min}^\star}\sqrt{\frac{d\log d}{p}}\right)\|\boldsymbol{U}^\star\|_{2,\infty} \\
&\le \left((C_3 + C_5)\rho^t\mathcal{E}_{\mathsf{local}} + (C_4 + C_6)\frac{\sigma}{\lambda_{\min}^\star}\sqrt{\frac{d\log d}{p}}\right)\|\boldsymbol{U}^\star\|_{2,\infty}.
\end{aligned}$$

A.3 Proof of Lemma 5.3

In view of the relation (40), one has

$$\begin{aligned}
\left\|\boldsymbol{U}^{t+1} - \boldsymbol{U}^\star\right\|_{\mathrm{F}} &= \left\|\boldsymbol{U}^t - \eta\left(\nabla f_{\mathsf{clean}}\left(\boldsymbol{U}^t\right) - p^{-1}\mathcal{P}_\Omega\left(\boldsymbol{E}\right)\times_1^{\mathsf{seq}} \boldsymbol{U}^t \times_2^{\mathsf{seq}} \boldsymbol{U}^t\right) - \boldsymbol{U}^\star\right\|_{\mathrm{F}} \\
&\le \underbrace{\left\|\boldsymbol{U}^t - \eta\nabla f_{\mathsf{clean}}\left(\boldsymbol{U}^t\right) - \boldsymbol{U}^\star\right\|_{\mathrm{F}}}_{=:\,\alpha_1} + \eta\underbrace{\left\|p^{-1}\mathcal{P}_\Omega\left(\boldsymbol{E}\right)\times_1^{\mathsf{seq}} \boldsymbol{U}^t \times_2^{\mathsf{seq}} \boldsymbol{U}^t\right\|_{\mathrm{F}}}_{=:\,\alpha_2},
\end{aligned}$$

which motivates us to bound α_1 and α_2 separately.

(1) We start with α_1, towards which we find it helpful to define

$$\boldsymbol{U}^t(\tau) := \tau \boldsymbol{U}^t + (1 - \tau) \boldsymbol{U}^\star. \tag{112}$$

Given that $\nabla f_{\mathsf{clean}}(\boldsymbol{U}^\star) = \boldsymbol{0}$ (since $\boldsymbol{U}^\star$ is a global optimizer of f_{clean}), we can use the fundamental theorem of calculus to obtain

$$\mathsf{vec}\left(\boldsymbol{U}^t - \eta \nabla f_{\mathsf{clean}}\left(\boldsymbol{U}^t\right) - \boldsymbol{U}^\star\right) = \mathsf{vec}\left(\boldsymbol{U}^t - \eta \nabla f_{\mathsf{clean}}\left(\boldsymbol{U}^t\right) - \left(\boldsymbol{U}^\star - \eta \nabla f_{\mathsf{clean}}\left(\boldsymbol{U}^\star\right)\right)\right) \tag{113}$$

$$= \mathsf{vec}\left(\boldsymbol{U}^t - \boldsymbol{U}^\star\right) - \eta\, \mathsf{vec}\left(\nabla f_{\mathsf{clean}}\left(\boldsymbol{U}^t\right) - \nabla f_{\mathsf{clean}}\left(\boldsymbol{U}^\star\right)\right) \tag{114}$$

$$= \left(\boldsymbol{I}_{dr} - \eta \underbrace{\int_0^1 \nabla^2 f_{\mathsf{clean}}\left(\boldsymbol{U}^t\left(\tau\right)\right)\,\mathrm{d}\tau}_{=:\,\boldsymbol{\Gamma}}\right) \mathsf{vec}\left(\boldsymbol{U}^t - \boldsymbol{U}^\star\right). \tag{115}$$

It then follows that

$$\left\|\boldsymbol{U}^t - \eta \nabla f_{\mathsf{clean}}\left(\boldsymbol{U}^t\right) - \boldsymbol{U}^\star\right\|_{\mathrm{F}}^2 = \mathsf{vec}\left(\boldsymbol{U}^t - \boldsymbol{U}^\star\right)^\top \left(\boldsymbol{I}_{dr} - \eta\boldsymbol{\Gamma}\right)^2 \mathsf{vec}\left(\boldsymbol{U}^t - \boldsymbol{U}^\star\right)$$

$$\le \left\|\boldsymbol{U}^t - \boldsymbol{U}^\star\right\|_{\mathrm{F}}^2 - 2\eta\, \mathsf{vec}\left(\boldsymbol{U}^t - \boldsymbol{U}^\star\right)^\top \boldsymbol{\Gamma}\, \mathsf{vec}\left(\boldsymbol{U}^t - \boldsymbol{U}^\star\right) + \eta^2 \left\|\boldsymbol{\Gamma}\right\|^2 \left\|\boldsymbol{U}^t - \boldsymbol{U}^\star\right\|_{\mathrm{F}}^2. \tag{116}$$

From the hypothesis (47b) as well as our conditions that $\frac{\sigma}{\lambda_{\min}^\star}\sqrt{\frac{d \log d}{p}} + \mathcal{E}_{\mathsf{local}} \ll \frac{1}{\mu^{3/2} r}$, we know that $\boldsymbol{U}^t(\tau)$ $(0 \le \tau \le 1)$ satisfies the conditions required in Lemma 5.1. Therefore, applying Lemma 5.1 gives that

$$\mathsf{vec}\left(\boldsymbol{U}^t - \boldsymbol{U}^\star\right)^\top \boldsymbol{\Gamma}\, \mathsf{vec}\left(\boldsymbol{U}^t - \boldsymbol{U}^\star\right) \ge \frac{1}{2}\lambda_{\min}^{\star 4/3}\left\|\boldsymbol{U}^t - \boldsymbol{U}^\star\right\|_{\mathrm{F}}^2,$$

$$\left\|\boldsymbol{\Gamma}\right\| \le 4\lambda_{\max}^{\star 4/3}.$$

Substitution into (116) indicates that: if $0 < \eta \le \lambda_{\min}^{\star 4/3}/\left(32\lambda_{\max}^{\star 8/3}\right)$, then

$$\left\|\boldsymbol{U}^t - \eta \nabla f_{\mathsf{clean}}\left(\boldsymbol{U}^t\right) - \boldsymbol{U}^\star\right\|_{\mathrm{F}}^2 \le \left(1 - \lambda_{\min}^{\star 4/3}\eta + 16\lambda_{\max}^{\star 8/3}\eta^2\right)\left\|\boldsymbol{U}^t - \boldsymbol{U}^\star\right\|_{\mathrm{F}}^2 \le \left(1 - \frac{1}{2}\lambda_{\min}^{\star 4/3}\eta\right)\left\|\boldsymbol{U}^t - \boldsymbol{U}^\star\right\|_{\mathrm{F}}^2,$$

which implies that (since $1 - a/2 \ge \sqrt{1-a}$ for $0 < a < 1$)

$$\left\|\boldsymbol{U}^t - \eta \nabla f_{\mathsf{clean}}\left(\boldsymbol{U}^t\right) - \boldsymbol{U}^\star\right\|_{\mathrm{F}} \le \left(1 - \frac{1}{4}\lambda_{\min}^{\star 4/3}\eta\right)\left\|\boldsymbol{U}^t - \boldsymbol{U}^\star\right\|_{\mathrm{F}}. \tag{117}$$

(2) We now turn to α_2. To simplify presentation, we shall assume that $\{E_{i,j,k}\}_{i,j,k\in[d]}$ (resp. $\{\chi_{i,j,k}\}_{i,j,k\in[d]}$) are independent random variables. Fix an arbitrary $s \in [r]$ and $m \in [d]$. The m-th entry of $\mathcal{P}_\Omega\left(\boldsymbol{E}\right) \times_1 \boldsymbol{u}_s^t \times_2 \boldsymbol{u}_s^t$ can be expanded as follows:

$$\left|\left(\mathcal{P}_\Omega\left(\boldsymbol{E}\right) \times_1 \boldsymbol{u}_s^t \times_2 \boldsymbol{u}_s^t\right)_m\right| = \left|\boldsymbol{u}_s^{t\top}\left(\mathcal{P}_\Omega\left(\boldsymbol{E}\right)\right)_{:,:,m}\boldsymbol{u}_s^t\right|$$

$$\le \underbrace{\left|\boldsymbol{u}_s^{t,(m)\top}\left(\mathcal{P}_\Omega\left(\boldsymbol{E}\right)\right)_{:,:,m}\boldsymbol{u}_s^{t,(m)}\right|}_{=:\,\beta_1} + \underbrace{\left|\left(\boldsymbol{u}_s^t - \boldsymbol{u}_s^{t,(m)}\right)^\top\left(\mathcal{P}_\Omega\left(\boldsymbol{E}\right)\right)_{:,:,m}\boldsymbol{u}_s^t\right|}_{=:\,\beta_2}$$

$$+ \underbrace{\left|\boldsymbol{u}_s^{t\top}\left(\mathcal{P}_\Omega\left(\boldsymbol{E}\right)\right)_{:,:,m}\left(\boldsymbol{u}_s^t - \boldsymbol{u}_s^{t,(m)}\right)\right|}_{=:\,\beta_3} + \underbrace{\left|\left(\boldsymbol{u}_s^t - \boldsymbol{u}_s^{t,(m)}\right)^\top\left(\mathcal{P}_\Omega\left(\boldsymbol{E}\right)\right)_{:,:,m}\left(\boldsymbol{u}_s^t - \boldsymbol{u}_s^{t,(m)}\right)\right|}_{=:\,\beta_4}.$$

$$\tag{118}$$

Before continuing, we make the following observations: from the hypotheses (47a), (47b), (47c), as well as our assumption that $\frac{\sigma}{\lambda_{\min}^\star}\sqrt{\frac{d \log d}{p}} + \mathcal{E}_{\mathsf{local}} \ll \frac{1}{\mu^{3/2} r}$, we see that the following holds for all $s \in [r]$:

$$\left\|\boldsymbol{u}_s^t - \boldsymbol{u}_s^\star\right\|_2 \le \left\|\boldsymbol{U}^t - \boldsymbol{U}^\star\right\|_{\mathrm{F}} \ll \frac{1}{r}\left\|\boldsymbol{U}^\star\right\|_{\mathrm{F}} \lesssim \frac{1}{\sqrt{r}}\lambda_{\max}^{\star 1/3}, \tag{119a}$$

$$\left\|\boldsymbol{u}_s^t - \boldsymbol{u}_s^\star\right\|_\infty \leq \left\|\boldsymbol{U}^t - \boldsymbol{U}^\star\right\|_{2,\infty} \ll \frac{1}{r}\left\|\boldsymbol{U}^\star\right\|_{2,\infty} \lesssim \sqrt{\frac{\mu}{rd}}\lambda_{\max}^{\star 1/3}, \tag{119b}$$

$$\left\|\boldsymbol{u}_s^{t,(m)} - \boldsymbol{u}_s^t\right\|_2 \leq \left\|\boldsymbol{U}^{t,(m)} - \boldsymbol{U}^t\right\|_{\mathrm{F}} \ll \frac{1}{r}\left\|\boldsymbol{U}^\star\right\|_{2,\infty} \lesssim \sqrt{\frac{\mu}{rd}}\lambda_{\max}^{\star 1/3}, \tag{119c}$$

$$\left|\left(\boldsymbol{u}_s^{t,(m)} - \boldsymbol{u}_s^\star\right)_m\right| \leq \left\|\left(\boldsymbol{U}^{t,(m)} - \boldsymbol{U}^\star\right)_{m,:}\right\|_2 \ll \frac{1}{r}\left\|\boldsymbol{U}^\star\right\|_{2,\infty} \lesssim \sqrt{\frac{\mu}{rd}}\lambda_{\max}^\star, \tag{119d}$$

$$\left\|\boldsymbol{u}_s^{t,(m)} - \boldsymbol{u}_s^\star\right\|_2 \leq \left\|\boldsymbol{U}^{t,(m)} - \boldsymbol{U}^\star\right\|_{\mathrm{F}} \ll \frac{1}{r}\left\|\boldsymbol{U}^\star\right\|_{\mathrm{F}} \lesssim \frac{1}{\sqrt{r}}\lambda_{\max}^{\star 1/3}, \tag{119e}$$

$$\left\|\boldsymbol{u}_s^t\right\|_2 \leq \left\|\boldsymbol{u}_s^t - \boldsymbol{u}_s^\star\right\|_2 + \left\|\boldsymbol{u}_s^\star\right\|_2 \lesssim \lambda_{\max}^{1/3}, \tag{119f}$$

$$\left\|\boldsymbol{u}_s^t\right\|_\infty \leq \left\|\boldsymbol{u}_s^t - \boldsymbol{u}_s^\star\right\|_\infty + \left\|\boldsymbol{u}_s^\star\right\|_\infty \lesssim \sqrt{\frac{\mu}{d}}\lambda_{\max}^{1/3}, \tag{119g}$$

$$\left\|\boldsymbol{u}_s^{t,(m)}\right\|_2 \leq \left\|\boldsymbol{u}_s^{t,(m)} - \boldsymbol{u}_s^t\right\|_2 + \left\|\boldsymbol{u}_s^t\right\|_2 \lesssim \lambda_{\max}^{\star 1/3}, \tag{119h}$$

$$\left\|\boldsymbol{u}_s^{t,(m)}\right\|_\infty \leq \left\|\boldsymbol{u}_s^{t,(m)} - \boldsymbol{u}_s^t\right\|_2 + \left\|\boldsymbol{u}_s^t\right\|_\infty \lesssim \sqrt{\frac{\mu}{d}}\lambda_{\max}^{1/3}, \tag{119i}$$

$$\left\|\boldsymbol{U}^t\right\|_{\mathrm{F}} \leq \left\|\boldsymbol{U}^t - \boldsymbol{U}^\star\right\|_{\mathrm{F}} + \left\|\boldsymbol{U}^\star\right\|_{\mathrm{F}} \lesssim \left\|\boldsymbol{U}^\star\right\|_{\mathrm{F}}, \tag{119j}$$

$$\left\|\boldsymbol{U}^t\right\|_{2,\infty} \leq \left\|\boldsymbol{U}^t - \boldsymbol{U}^\star\right\|_{2,\infty} + \left\|\boldsymbol{U}^\star\right\|_{2,\infty} \lesssim \left\|\boldsymbol{U}^\star\right\|_{2,\infty}, \tag{119k}$$

$$\left\|\boldsymbol{U}^{t,(m)}\right\|_{\mathrm{F}} \leq \left\|\boldsymbol{U}^{t,(m)} - \boldsymbol{U}^t\right\|_{\mathrm{F}} + \left\|\boldsymbol{U}^t\right\|_{\mathrm{F}} \lesssim \left\|\boldsymbol{U}^\star\right\|_{\mathrm{F}}, \tag{119l}$$

$$\left\|\boldsymbol{U}^{t,(m)}\right\|_{2,\infty} \leq \left\|\boldsymbol{U}^{t,(m)} - \boldsymbol{U}^t\right\|_{\mathrm{F}} + \left\|\boldsymbol{U}^t\right\|_{2,\infty} \lesssim \left\|\boldsymbol{U}^\star\right\|_{2,\infty}. \tag{119m}$$

With these estimates in place, we can upper bound the above four terms in (118) separately.

- For β_1, we note that, by construction, $\boldsymbol{u}^{t,(m)}$ is independent of the m-th mode-3 slice of $\mathcal{P}_\Omega(\boldsymbol{E})$. This tells us that

$$\boldsymbol{u}_s^{t,(m)\top}\left(\mathcal{P}_\Omega(\boldsymbol{E})\right)_{:,:,m}\boldsymbol{u}_s^{t,(m)} = \sum_{i,j\in[d]}\left(\boldsymbol{u}_s^{t,(m)}\right)_i\left(\boldsymbol{u}_s^{t,(m)}\right)_j E_{i,j,m}\chi_{i,j,m}$$

can be viewed as a sum of independent zero-mean random variables (conditional on $\mathcal{P}_{\Omega_{-m}}(\boldsymbol{E})$). It is straightforward to compute that

$$\max_{i,j\in[d]}\left\|\left(\boldsymbol{u}_s^{t,(m)}\right)_i\left(\boldsymbol{u}_s^{t,(m)}\right)_j E_{i,j,m}\chi_{i,j,m}\right\|_{\psi_1} \lesssim \sigma\left\|\boldsymbol{u}_s^{t,(m)}\right\|_\infty^2 =: L,$$

$$\sum_{i,j\in[d]}\mathbb{E}\left[\left(\boldsymbol{u}_s^{t,(m)}\right)_i^2\left(\boldsymbol{u}_s^{t,(m)}\right)_j^2 E_{i,j,m}^2\chi_{i,j,m}^2\right] \lesssim \sigma^2 p\left\|\boldsymbol{u}_s^{t,(m)}\right\|_2^4 \lesssim \sigma^2\lambda_{\max}^{\star 2/3}p\left\|\boldsymbol{u}_s^{t,(m)}\right\|_2^2 =: V,$$

where $\|\cdot\|_{\psi_1}$ denotes the sub-exponential norm and we use (119) in the last inequality. Applying the matrix Bernstein inequality [Kol11, Corollary 2.1] yields that: with probability $1 - O(d^{-20})$,

$$\left|\boldsymbol{u}_s^{t,(m)\top}\left(\mathcal{P}_\Omega(\boldsymbol{E})\right)_{:,:,m}\boldsymbol{u}_s^{t,(m)}\right| \lesssim L\log^2 d + \sqrt{V\log d}$$

$$\lesssim \sigma\log^2 d\left\|\boldsymbol{u}_s^{t,(m)}\right\|_\infty^2 + \sigma\lambda_{\max}^{\star 1/3}\sqrt{p\log d}\left\|\boldsymbol{u}_s^{t,(m)}\right\|_2, \tag{120}$$

- For β_2, we first invoke [CW15, lemma 11] to demonstrate that: with probability at least $1 - O(d^{-11})$,

$$\max_{m\in[d]}\left\|\left(\mathcal{P}_\Omega(\boldsymbol{E})\right)_{:,:,m}\right\| \lesssim \sigma\left(\sqrt{dp} + \log d\right). \tag{121}$$

Next, it is seen that

$$\left|\left(\boldsymbol{u}_s^t - \boldsymbol{u}_s^{t,(m)}\right)^\top\left(\mathcal{P}_\Omega(\boldsymbol{E})\right)_{:,:,m}\boldsymbol{u}_s^t\right| \leq \left\|\left(\mathcal{P}_\Omega(\boldsymbol{E})\right)_{:,:,m}\right\|\left\|\boldsymbol{u}_s^t - \boldsymbol{u}_s^{t,(m)}\right\|_2\left\|\boldsymbol{u}_s^t\right\|_2$$

$$\lesssim \sigma\left(\sqrt{dp} + \log d\right)\left\|\boldsymbol{U}^t - \boldsymbol{U}^{t,(m)}\right\|_{\mathrm{F}}\left\|\boldsymbol{u}_s^t\right\|_2$$

$$\lesssim \sigma\big(\sqrt{dp}+\log d\big)\left(C_5\rho^t\mathcal{E}_{\mathsf{local}}+C_6\frac{\sigma}{\lambda_{\min}^{\star}}\sqrt{\frac{d\log d}{p}}\right)\left\|\boldsymbol{U}^{\star}\right\|_{2,\infty}\left\|\boldsymbol{u}_s^t\right\|_2$$

$$\ll \frac{\lambda_{\min}^{\star 4/3}}{\sqrt{d}}\left(C_5\rho^t\mathcal{E}_{\mathsf{local}}+C_6\frac{\sigma}{\lambda_{\min}^{\star}}\sqrt{\frac{d\log d}{p}}\right)\left\|\boldsymbol{u}_s^t\right\|_2,$$

where the last inequality follows from the noise condition. Clearly, the above bound holds for β_3 as well.

- Regarding β_4, it is easily seen that $\beta_4 \ll \beta_2$, given that $\|\boldsymbol{U}^t - \boldsymbol{U}^{t,(m)}\|_{\mathrm{F}} \ll \|\boldsymbol{U}^{\star}\|_{2,\infty}/r \le \lambda_{\min}^{\star}$ holds according to (119) and $\kappa \asymp 1$.

- Taking together the above bounds and substituting them into (118), we obtain

$$\left|\left(\mathcal{P}_\Omega\left(\boldsymbol{E}\right)\times_1\boldsymbol{u}_s^t\times_2\boldsymbol{u}_s^t\right)_m\right| \lesssim \sigma\log^2 d\left\|\boldsymbol{u}_s^{t,(m)}\right\|_\infty^2 + \sigma\lambda_{\max}^{\star 1/3}\sqrt{p\log d}\left\|\boldsymbol{u}_s^{t,(m)}\right\|_2$$

$$+ o\left(1\right)\frac{\lambda_{\min}^{\star 4/3}}{\sqrt{d}}\left(C_5\rho^t\mathcal{E}_{\mathsf{local}}+C_6\frac{\sigma}{\lambda_{\min}^{\star}}\sqrt{\frac{d\log d}{p}}\right)\left\|\boldsymbol{u}_s^t\right\|_2.$$

Recognizing that this holds for any $m \in [d]$ and $s \in [r]$, one can sum over m and s to deduce that

$$\left\|\mathcal{P}_\Omega\left(\boldsymbol{E}\right)\times_1^{\mathsf{seq}}\boldsymbol{U}^t\times_2^{\mathsf{seq}}\boldsymbol{U}^t\right\|_{\mathrm{F}} \overset{\text{(i)}}{\lesssim} \frac{\sigma\lambda_{\max}^{\star 2/3}\mu\sqrt{r}\log^2 d}{\sqrt{d}}+\sigma\lambda_{\max}^{\star 1/3}\sqrt{dp\log d}\left\|\boldsymbol{U}^{t,(m)}\right\|_{\mathrm{F}}$$

$$+ o\left(1\right)\lambda_{\min}^{\star 4/3}\left(C_5\rho^t\mathcal{E}_{\mathsf{local}}+C_6\frac{\sigma}{\lambda_{\min}^{\star}}\sqrt{\frac{d\log d}{p}}\right)\left\|\boldsymbol{U}^t\right\|_{\mathrm{F}}$$

$$\le C\sigma\lambda_{\min}^{\star 1/3}\sqrt{dp\log d}\left\|\boldsymbol{U}^{\star}\right\|_{\mathrm{F}}$$

$$+ o\left(1\right)\lambda_{\min}^{\star 4/3}\left(C_5\rho^t\mathcal{E}_{\mathsf{local}}+C_6\frac{\sigma}{\lambda_{\min}^{\star}}\sqrt{\frac{d\log d}{p}}\right)\left\|\boldsymbol{U}^{\star}\right\|_{\mathrm{F}}, \qquad (122)$$

for some absolute constant $C > 0$, where (i) is true due to (119); the last inequality follows from the fact that $\|\boldsymbol{U}^{\star}\|_{\mathrm{F}} \ge \lambda_{\min}^{\star 1/3}\sqrt{r}$ as well as the assumptions that $p \gg \mu d^{-2}\log^3 d$ and $\kappa \asymp 1$.

(3) Combining (117) and (122) yields that: with probability at least $1 - O(d^{-10})$,

$$\left\|\boldsymbol{U}^{t+1}-\boldsymbol{U}^{\star}\right\|_{\mathrm{F}} \le \left(1-\frac{1}{4}\lambda_{\min}^{\star 4/3}\eta\right)\left\|\boldsymbol{U}^t-\boldsymbol{U}^{\star}\right\|_{\mathrm{F}}+C\eta\sigma\lambda_{\min}^{\star 1/3}\sqrt{\frac{d\log d}{p}}\left\|\boldsymbol{U}^{\star}\right\|_{\mathrm{F}}$$

$$+ o\left(1\right)\lambda_{\min}^{\star 4/3}\eta\left(C_5\rho^t\mathcal{E}_{\mathsf{local}}+C_6\frac{\sigma}{\lambda_{\min}^{\star}}\sqrt{\frac{d\log d}{p}}\right)\left\|\boldsymbol{U}^{\star}\right\|_{\mathrm{F}}$$

$$\le \left(1-\frac{1}{4}\lambda_{\min}^{\star 4/3}\eta\right)\left(C_1\rho^t\mathcal{E}_{\mathsf{local}}+C_2\frac{\sigma}{\lambda_{\min}^{\star}}\sqrt{\frac{d\log d}{p}}\right)\left\|\boldsymbol{U}^{\star}\right\|_{\mathrm{F}}+C\eta\sigma\lambda_{\min}^{\star 1/3}\sqrt{\frac{d\log d}{p}}\left\|\boldsymbol{U}^{\star}\right\|_{\mathrm{F}}$$

$$+ o\left(1\right)\lambda_{\min}^{\star 4/3}\eta\left(C_5\rho^t\mathcal{E}_{\mathsf{local}}+C_6\frac{\sigma}{\lambda_{\min}^{\star}}\sqrt{\frac{d\log d}{p}}\right)\left\|\boldsymbol{U}^{\star}\right\|_{\mathrm{F}}$$

$$\le \left(1-\frac{1}{5}\lambda_{\min}^{\star 4/3}\eta\right)C_1\rho^t\mathcal{E}_{\mathsf{local}}\left\|\boldsymbol{U}^{\star}\right\|_{\mathrm{F}}+\left(\left(1-\frac{1}{5}\lambda_{\min}^{\star 4/3}\eta\right)C_2+C\eta\lambda_{\min}^{\star 3/4}\right)\frac{\sigma}{\lambda_{\min}^{\star}}\sqrt{\frac{d\log d}{p}}\left\|\boldsymbol{U}^{\star}\right\|_{\mathrm{F}}$$

$$\le C_1\rho^{t+1}\mathcal{E}_{\mathsf{local}}\left\|\boldsymbol{U}^{\star}\right\|_{\mathrm{F}}+C_2\frac{\sigma}{\lambda_{\min}^{\star}}\sqrt{\frac{d\log d}{p}}\left\|\boldsymbol{U}^{\star}\right\|_{\mathrm{F}},$$

with the proviso that $0 < \eta \le \lambda_{\min}^{\star 4/3}/\big(32\lambda_{\max}^{\star 8/3}\big)$, $1-\big(\lambda_{\min}^{\star 4/3}/5\big)\eta \le \rho < 1$ and that C_2 is sufficiently large.

A.4 Proof of Lemma 5.4

Fix an arbitrary $m \in [r]$. From the definition of $f^{(m)}$ in (45) and (46), we can show that

$$
\boldsymbol{U}^{t+1} - \boldsymbol{U}^{t+1,(m)} = \boldsymbol{U}^t - \eta \nabla f(\boldsymbol{U}^t) - \left(\boldsymbol{U}^{t,(m)} - \eta \nabla f^{(m)}(\boldsymbol{U}^{t,(m)}) \right)
$$

$$
= \boldsymbol{U}^t - \boldsymbol{U}^{t,(m)} - \eta \left(\nabla f(\boldsymbol{U}^t) - \nabla f(\boldsymbol{U}^{t,(m)}) \right) - \eta \left(\nabla f(\boldsymbol{U}^{t,(m)}) - \nabla f^{(m)}(\boldsymbol{U}^{t,(m)}) \right)
$$

$$
= \underbrace{\boldsymbol{U}^t - \boldsymbol{U}^{t,(m)} - \eta \left(\nabla f_{\mathsf{clean}}(\boldsymbol{U}^t) - \nabla f_{\mathsf{clean}}(\boldsymbol{U}^{t,(m)}) \right)}_{=: \alpha_1}
$$

$$
\underbrace{- \eta \left(\left(p^{-1} \mathcal{P}_{\Omega_m} - \mathcal{P}_m \right) \left(\sum_{s \in [r]} \left(\boldsymbol{u}_s^{t,(m)} \right)^{\otimes 3} - \boldsymbol{T}^\star \right) \right) \times_1^{\mathsf{seq}} \boldsymbol{U}^{t,(m)} \times_2^{\mathsf{seq}} \boldsymbol{U}^{t,(m)}}_{=: \alpha_2}
$$

$$
+ \underbrace{\eta \, p^{-1} \mathcal{P}_{\Omega_m} (\boldsymbol{E}) \times_1^{\mathsf{seq}} \boldsymbol{U}^{t,(m)} \times_2^{\mathsf{seq}} \boldsymbol{U}^{t,(m)}}_{=: \alpha_3} + \underbrace{\eta p^{-1} \left\{ \mathcal{P}_{\Omega} (\boldsymbol{E}) \times_1^{\mathsf{seq}} \boldsymbol{U}^t \times_2^{\mathsf{seq}} \boldsymbol{U}^t - \mathcal{P}_{\Omega} (\boldsymbol{E}) \times_1^{\mathsf{seq}} \boldsymbol{U}^{t,(m)} \times_2^{\mathsf{seq}} \boldsymbol{U}^{t,(m)} \right\}}_{=: \alpha_4}.
$$

$$(123)$$

Before proceeding to bounding these terms, we pause to define

$$
\boldsymbol{\Delta}_{\boldsymbol{T}}^{t,(m)} := \sum_{s \in [r]} \left(\boldsymbol{u}_s^{t,(m)} \right)^{\otimes 3} - \boldsymbol{T}^\star. \tag{124}
$$

From (119) and hypothesis (47d), one can applying a similar argument as in (107) to find that

$$
\left\| \boldsymbol{\Delta}_{\boldsymbol{T}}^{t,(m)} \right\|_\infty \lesssim \max_{s \in [r]} \left\| \boldsymbol{u}_s^\star \right\|_\infty \left\| \boldsymbol{U}^\star \right\|_{2,\infty} \left\| \boldsymbol{U}^{t,(m)} - \boldsymbol{U}^\star \right\|_{2,\infty} \lesssim \frac{\mu \sqrt{r} \, \lambda_{\max}^{\star 2/3}}{d} \left\| \boldsymbol{U}^{t,(m)} - \boldsymbol{U}^\star \right\|_{2,\infty}. \tag{125}
$$

Now we begin to bound the terms in (123) separately.

(1) We start with α_1. For any $0 \le \tau \le 1$, define

$$
\boldsymbol{U}^{t,(m)}(\tau) := \tau \boldsymbol{U}^t + (1 - \tau) \boldsymbol{U}^{t,(m)}.
$$

The fundamental theorem of calculus yields

$$
\boldsymbol{U}^t - \boldsymbol{U}^{t,(m)} - \eta \left(\nabla f_{\mathsf{clean}}(\boldsymbol{U}^t) - \nabla f_{\mathsf{clean}}(\boldsymbol{U}^{t,(m)}) \right) = \underbrace{\left(\boldsymbol{I}_{dr} - \eta \int_0^1 \nabla^2 f_{\mathsf{clean}} \left(\boldsymbol{U}^{t,(m)}(\tau) \right) \mathrm{d}\tau \right)}_{=: \boldsymbol{\Gamma}} \left(\boldsymbol{U}^t - \boldsymbol{U}^{t,(m)} \right).
$$

By Lemma 5.2 and our assumptions on the noise, we know that $\boldsymbol{U}^{t,(m)}(\tau)$ $(0 \le \tau \le 1)$ satisfies the conditions in Lemma 5.1 for any $\tau \in [0,1]$. Applying the same argument as the one used to bound $\| \boldsymbol{U}^t - \eta \nabla f_{\mathsf{clean}}(\boldsymbol{U}^t) - \boldsymbol{U}^\star \|_{\mathrm{F}}$ in Lemma 5.3, we show that

$$
\left\| \boldsymbol{U}^t - \boldsymbol{U}^{t,(m)} - \eta \left(\nabla f(\boldsymbol{U}^t) - \nabla f(\boldsymbol{U}^{t,(m)}) \right) \right\|_2 \le \left(1 - \frac{1}{4} \lambda_{\min}^{\star 4/3} \eta \right) \left\| \boldsymbol{U}^t - \boldsymbol{U}^{t,(m)} \right\|_{\mathrm{F}}, \tag{126}
$$

provided that $0 < \eta \le \lambda_{\min}^{\star 4/3} / \left(32 \lambda_{\max}^{\star 8/3} \right)$.

In what follows, we shall assume that $\{ E_{i,j,k} \}_{i,j,k \in [d]}$ (resp. $\{ \chi_{i,j,k} \}_{i,j,k \in [d]}$) are independent random variables in order to simplify presentation.

(2) The next step is to bound α_2. For notational simplicity, define

$$
\boldsymbol{V}^{t,(m)} := \left(\left(p^{-1} \mathcal{P}_{\Omega_m} - \mathcal{P}_m \right) \left(\boldsymbol{\Delta}_{\boldsymbol{T}}^{t,(m)} \right) \right) \times_1^{\mathsf{seq}} \boldsymbol{U}^{t,(m)} \times_2^{\mathsf{seq}} \boldsymbol{U}^{t,(m)};
$$

$$
\boldsymbol{v}_s^{t,(m)} := \left(\left(p^{-1} \mathcal{P}_{\Omega_m} - \mathcal{P}_m \right) \left(\boldsymbol{\Delta}_{\boldsymbol{T}}^{t,(m)} \right) \right) \times_1 \boldsymbol{u}_s^{t,(m)} \times_2 \boldsymbol{u}_s^{t,(m)}, \qquad s \in [r].
$$

In order to control the Frobenius norm of $\boldsymbol{V}^{t,(m)}$, we shall start by considering the m-th row of $\boldsymbol{V}^{t,(m)}$. In view of the definitions of $\mathcal{P}_{\Omega_m}$ and $\mathcal{P}_m$ (cf. Appendix 5.1.3), we can expand

$$\boldsymbol{V}_{m,:}^{t,(m)} = \sum_{i,j\in[d]}\sum_{s\in[r]} \big(p^{-1}\chi_{i,j,m} - 1\big) \big(\boldsymbol{\Delta}_{\boldsymbol{T}}^{t,(m)}\big)_{i,j,m} \big(\boldsymbol{u}_s^{t,(m)}\big)_i \big(\boldsymbol{u}_s^{t,(m)}\big)_j \boldsymbol{e}_s^\top .$$

We recognize that $\big\{\boldsymbol{u}_s^{t,(m)}\big\}_{s=1}^r$ is independent of Ω_m, making it convenient for us to upper bound $\big\|\boldsymbol{V}_{m,:}^{t,(m)}\big\|_2$. Specifically, for any $i, j \in [d]$, from (119) and (125) we have

$$\Big\| \sum_{s\in[r]} \big(\boldsymbol{\Delta}_{\boldsymbol{T}}^{t,(m)}\big)_{i,j,m} \big(\boldsymbol{u}_s^{t,(m)}\big)_i \big(\boldsymbol{u}_s^{t,(m)}\big)_j \big(p^{-1}\chi_{i,j,m} - 1\big) \boldsymbol{e}_s^\top \Big\|_2 \leq \frac{1}{p}\big\|\boldsymbol{\Delta}_{\boldsymbol{T}}^{t,(m)}\big\|_\infty \max_{s\in[r]}\big\|\boldsymbol{u}_s^{t,(m)}\big\|_\infty \big\|\boldsymbol{U}^{t,(m)}\big\|_{2,\infty}$$

$$\lesssim \frac{\mu^2 r \lambda_{\max}^{\star 4/3}}{d^2 p} \big\|\boldsymbol{U}^{t,(m)} - \boldsymbol{U}^\star\big\|_{2,\infty} =: L_1 .$$

In addition, it is easy to verify that $\mathbb{E}\big[\boldsymbol{V}_{m,:}^{t,(m)}\big] = \boldsymbol{0}$ and

$$\sum_{i,j\in[d]} \mathbb{E}\left[\Big\| \sum_{s\in[r]} \big(\boldsymbol{\Delta}_{\boldsymbol{T}}^{t,(m)}\big)_{i,j,m} \big(\boldsymbol{u}_s^{t,(m)}\big)_i \big(\boldsymbol{u}_s^{t,(m)}\big)_j (p^{-1}\chi_{i,j,m} - 1) \boldsymbol{e}_s^\top \Big\|_2^2 \right]$$

$$= \sum_{i,j\in[d]}\sum_{s\in[r]} \big(\boldsymbol{\Delta}_{\boldsymbol{T}}^{t,(m)}\big)_{i,j,m}^2 \big(\boldsymbol{u}_s^{t,(m)}\big)_i^2 \big(\boldsymbol{u}_s^{t,(m)}\big)_j^2 \mathbb{E}\big[(p^{-1}\chi_{i,j,m} - 1)^2\big]$$

$$\leq \frac{1}{p}\big\|\boldsymbol{\Delta}_{\boldsymbol{T}}^{t,(m)}\big\|_\infty^2 \max_{s\in[r]}\big\|\boldsymbol{u}_s^{t,(m)}\big\|_2^2 \big\|\boldsymbol{U}^{t,(m)}\big\|_{\mathrm{F}}^2 \lesssim \frac{\mu^2 r^2 \lambda_{\max}^{\star 8/3}}{d^2 p}\big\|\boldsymbol{U}^{t,(m)} - \boldsymbol{U}^\star\big\|_{2,\infty}^2 =: V_1 ,$$

where the last inequality holds due to (119) and (125). We then apply the matrix Bernstein inequality to yield that: with probability exceeding $1 - O(d^{-20})$,

$$\big\|\boldsymbol{V}_{m,:}^{t,(m)}\big\|_2 \lesssim L_1 \log d + \sqrt{V_1 \log d} \lesssim \left\{ \frac{\mu^2 r \lambda_{\max}^{\star 4/3}\log d}{d^2 p} + \frac{\mu r \lambda_{\max}^{\star 4/3}\sqrt{\log d}}{d\sqrt{p}} \right\} \big\|\boldsymbol{U}^{t,(m)} - \boldsymbol{U}^\star\big\|_\infty$$

$$\asymp \frac{\mu r \lambda_{\max}^{\star 4/3}\sqrt{\log d}}{d\sqrt{p}}\big\|\boldsymbol{U}^{t,(m)} - \boldsymbol{U}^\star\big\|_\infty, \tag{127}$$

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

Next, we turn to the k-th row of $\boldsymbol{V}^{t,(m)}$ for any $k \neq m$. For each $s \in [r]$, we have

$$\big(\boldsymbol{v}_s^{t,(m)}\big)_k = \big(\boldsymbol{u}_s^{t,(m)}\big)_m \sum_{j\in[d]} \big(\boldsymbol{\Delta}_{\boldsymbol{T}}^{t,(m)}\big)_{m,j,k} \big(\boldsymbol{u}_s^{t,(m)}\big)_j (p^{-1}\chi_{m,j,k} - 1)$$

$$+ \big(\boldsymbol{u}_s^{t,(m)}\big)_m \sum_{i:i\neq m} \big(\boldsymbol{\Delta}_{\boldsymbol{T}}^{t,(m)}\big)_{i,m,k} \big(\boldsymbol{u}_s^{t,(m)}\big)_i (p^{-1}\chi_{i,m,k} - 1).$$

Similar to the proof of Lemma D.9, we can show that with probability at least $1 - O(d^{-20})$,

$$\sum_{s\in[r]}\sum_{k:k\neq m} \left(\big(\boldsymbol{u}_s^{t,(m)}\big)_m \sum_{j\in[d]} \big(\boldsymbol{\Delta}_{\boldsymbol{T}}^{t,(m)}\big)_{m,j,k} \big(\boldsymbol{u}_s^{t,(m)}\big)_j (p^{-1}\chi_{m,j,k} - 1) \right)^2$$

$$\lesssim \sum_{s\in[r]}\sum_{k:k\neq m} \mathbb{E}\left[\left(\big(\boldsymbol{u}_s^{t,(m)}\big)_m \sum_{j\in[d]} \big(\boldsymbol{\Delta}_{\boldsymbol{T}}^{t,(m)}\big)_{m,j,k} \big(\boldsymbol{u}_s^{t,(m)}\big)_j (p^{-1}\chi_{m,j,k} - 1) \right)^2 \right]$$

$$\lesssim \frac{1}{p} \sum_{s\in[r]} \big(\boldsymbol{u}_s^{t,(m)}\big)_m^2 \sum_{j,k\in[d]} \big(\boldsymbol{\Delta}_{\boldsymbol{T}}^{t,(m)}\big)_{m,j,k}^2 \big(\boldsymbol{u}_s^{t,(m)}\big)_j^2$$

$$\leq \frac{d}{p} \max_{s\in[r]} \big\|\boldsymbol{u}_s^{t,(m)}\big\|_\infty^2 \big\|\boldsymbol{\Delta}_{\boldsymbol{T}}^{t,(m)}\big\|_\infty^2 \big\|\boldsymbol{U}^{t,(m)}\big\|_{\mathrm{F}}^2 \lesssim \frac{\mu^3 r^2 \lambda_{\max}^{\star 8/3}}{d^2 p}\big\|\boldsymbol{U}^{t,(m)} - \boldsymbol{U}^\star\big\|_{2,\infty}^2 ,$$

38

where the last inequality follows from (119), and (125). It is easily seen that the bound also holds for the summation over $i \neq m$. We can then use Cauchy-Schwartz to arrive at

$$\sum_{k:k\neq m} \left\| \boldsymbol{V}_{k,:}^{t,(m)} \right\|_2^2 = \sum_{s\in[r]} \sum_{k:k\neq m} \left(\boldsymbol{v}_s^{t,(m)} \right)_k^2 \lesssim \frac{\mu^3 r^2 \lambda_{\max}^{\star 8/3}}{d^2 p} \left\| \boldsymbol{U}^{t,(m)} - \boldsymbol{U}^\star \right\|_{2,\infty}^2. \tag{128}$$

Combining (128) with (127) and invoking the union bound, we conclude that: with probability exceeding $1 - O(d^{-20})$,

$$\left\| \left(\left(p^{-1} \mathcal{P}_{\Omega_m} - \mathcal{P}_m \right) \left(\boldsymbol{\Delta}_{\boldsymbol{T}}^{t,(m)} \right) \right) \times_1^{\mathsf{seq}} \boldsymbol{U}^{t,(m)} \times_2^{\mathsf{seq}} \boldsymbol{U}^{t,(m)} \right\|_{\mathrm{F}} = \left\| \boldsymbol{V}^{t,(m)} \right\|_{\mathrm{F}} \leq C \frac{\mu^{3/2} r \lambda_{\min}^{\star 4/3} \sqrt{\log d}}{d\sqrt{p}} \left\| \boldsymbol{U}^{t,(m)} - \boldsymbol{U}^\star \right\|_{2,\infty}. \tag{129}$$

for some absolute constant $C > 0$, where we use the assumption that $\kappa \asymp 1$.

(3) For α_3, following a similar argument for α_2, we define

$$\boldsymbol{W}^{t,(m)} := \mathcal{P}_{\Omega_m}\left(\boldsymbol{E} \right) \times_1^{\mathsf{seq}} \boldsymbol{U}_s^{t,(m)} \times_2^{\mathsf{seq}} \boldsymbol{U}_s^{t,(m)}$$
$$\boldsymbol{w}_s^{t,(m)} := \mathcal{P}_{\Omega_m}\left(\boldsymbol{E} \right) \times_1 \boldsymbol{u}_s^{t,(m)} \times_2 \boldsymbol{u}_s^{t,(m)}$$

for each $s \in [r]$. The m-th row of $\boldsymbol{W}^{t,(m)}$ is a sum of independent zero-mean random vectors:

$$\boldsymbol{W}_{m,:}^{t,(m)} = \sum_{i,j\in[d]} \sum_{s\in[r]} \left(\boldsymbol{u}_s^{t,(m)} \right)_i \left(\boldsymbol{u}_s^{t,(m)} \right)_j E_{i,j,m} \chi_{i,j,m} \boldsymbol{e}_s^\top.$$

With (119) in place, it is easy to verify that

$$\max_{i,j\in[d]} \left\| \sum_{s\in[r]} \left(\boldsymbol{u}_s^{t,(m)} \right)_i \left(\boldsymbol{u}_s^{t,(m)} \right)_j E_{i,j,m} \chi_{i,j,m} \boldsymbol{e}_s^\top \right\|_{\psi_1} \leq \sigma \max_{s\in[r]} \left\| \boldsymbol{u}_s^{t,(m)} \right\|_\infty \left\| \boldsymbol{U}^{t,(m)} \right\|_{2,\infty}$$
$$\lesssim \sigma \sqrt{\frac{\mu}{d}} \lambda_{\max}^{\star 1/3} \left\| \boldsymbol{U}^\star \right\|_{2,\infty} =: L_2$$

and

$$\sum_{i,j\in[d]} \mathbb{E}\left[\left\| \sum_{s\in[r]} \left(\boldsymbol{u}_s^{t,(m)} \right)_i \left(\boldsymbol{u}_s^{t,(m)} \right)_j E_{i,j,m} \chi_{i,j,m} \boldsymbol{e}_s^\top \right\|_2^2 \right] = \sum_{i,j\in[d]} \sum_{s\in[r]} \left(\boldsymbol{u}_s^{t,(m)} \right)_i^2 \left(\boldsymbol{u}_s^{t,(m)} \right)_j^2 \mathbb{E}\left[\left(E_{i,j,m} \chi_{i,j,m} \right)^2 \right]$$
$$\leq \sigma^2 p \max_{s\in[r]} \left\| \boldsymbol{u}_s^{t,(m)} \right\|_2^2 \left\| \boldsymbol{U}^{t,(m)} \right\|_{\mathrm{F}}^2$$
$$\lesssim \sigma^2 dp \lambda_{\max}^{\star 2/3} \left\| \boldsymbol{U}^\star \right\|_{2,\infty}^2 =: V_2,$$

where we have used (119) and the fact that $\left\| \boldsymbol{U}^\star \right\|_{\mathrm{F}} \leq \sqrt{d} \left\| \boldsymbol{U}^\star \right\|_{2,\infty}$. Apply the matrix Bernstein inequality to reveal that: with probability at least $1 - O(d^{-20})$,

$$\left\| \boldsymbol{W}_{m,:}^{t,(m)} \right\|_2 = \left\| \sum_{i,j\in[d]} \sum_{s\in[r]} \left(\boldsymbol{u}_s^{t,(m)} \right)_i \left(\boldsymbol{u}_s^{t,(m)} \right)_j E_{i,j,m} \chi_{i,j,m} \boldsymbol{e}_s^\top \right\|_2$$
$$\lesssim L_2 \log^2 d + \sqrt{V_2 \log d} \asymp \sigma \lambda_{\max}^{\star 1/3} \sqrt{dp \log d} \left\| \boldsymbol{U}^\star \right\|_{2,\infty}, \tag{130}$$

where the last inequality holds as long as $p \gg \mu d^{-2} \log^3 d$.

As for the other rows, we have

$$\left(\boldsymbol{w}_s^{t,(m)} \right)_k = \left(\boldsymbol{u}_s^{t,(m)} \right)_m \sum_{j\in[d]} \left(\boldsymbol{u}_s^{t,(m)} \right)_j E_{m,j,k} \chi_{m,j,k} + \left(\boldsymbol{u}_s^{t,(m)} \right)_m \sum_{i:i\neq m} \left(\boldsymbol{u}_s^{t,(m)} \right)_i E_{i,m,k} \chi_{i,m,k}$$

for each $s \in [r]$. Arguing similarly as in the proof of Lemma D.10, we have with probability at least $1 - O(d^{-20})$,

$$\sum_{s \in [r]} \sum_{k:k \neq m} \left(\left(\boldsymbol{u}_s^{t,(m)} \right)_m \sum_{j \in [d]} \left(\boldsymbol{u}_s^{t,(m)} \right)_j E_{m,j,k} \chi_{m,j,k} \right)^2$$

$$\lesssim \sum_{s \in [r]} \sum_{k:k \neq m} \mathbb{E} \left[\left(\left(\boldsymbol{u}_s^{t,(m)} \right)_m \sum_{j \in [d]} \left(\boldsymbol{u}_s^{t,(m)} \right)_j \left(E_{m,j,k} \chi_{m,j,k} \right) \right)^2 \right]$$

$$\leq \sigma^2 dp \sum_{s \in [r]} \left(\boldsymbol{u}_s^{t,(m)} \right)_m^2 \sum_{j \in [d]} \left(\boldsymbol{u}_s^{t,(m)} \right)_j^2$$

$$\leq \sigma^2 dp \max_{s \in [r]} \left\| \boldsymbol{u}_s^{t,(m)} \right\|_2^2 \left\| \boldsymbol{U}^{t,(m)} \right\|_{2,\infty}^2 \lesssim \sigma^2 \lambda_{\max}^{\star 2/3} dp \left\| \boldsymbol{U}^\star \right\|_{2,\infty}^2,$$

where the last inequality follows from (119). Additionally, the summation over $\{i : i \neq m\}$ can be controlled using the same argument. Therefore, we use Cauchy-Schwarz to find that

$$\sum_{k:k \neq m} \left\| \boldsymbol{W}_{k,:}^{t,(m)} \right\|_2^2 \lesssim \sigma^2 \lambda_{\max}^{\star 2/3} dp \left\| \boldsymbol{U}^\star \right\|_{2,\infty}^2. \tag{131}$$

Combined with (130) and the assumption that $\kappa \asymp 1$, we obtain that

$$\left\| p^{-1} \mathcal{P}_{\Omega_m} \left(\boldsymbol{E} \right) \times_1^{\mathsf{seq}} \boldsymbol{U}^{t,(m)} \times_2^{\mathsf{seq}} \boldsymbol{U}^{t,(m)} \right\|_{\mathrm{F}} = \left\| \boldsymbol{W}^{t,(m)} \right\|_{\mathrm{F}} \leq \widetilde{C} \sigma \lambda_{\min}^{\star 1/3} \sqrt{dp \log d} \left\| \boldsymbol{U}^\star \right\|_{2,\infty}, \tag{132}$$

for some absolute constant $\widetilde{C} > 0$.

(4) Regarding α_4, we use the triangle inequality to show that for each $s \in [r]$,

$$\left\| \mathcal{P}_\Omega \left(\boldsymbol{E} \right) \times_1 \boldsymbol{u}_s^t \times_2 \boldsymbol{u}_s^t - \mathcal{P}_\Omega \left(\boldsymbol{E} \right) \times_1 \boldsymbol{u}_s^{t,(m)} \times_2 \boldsymbol{u}_s^{t,(m)} \right\|_2$$

$$\leq \left\| \mathcal{P}_\Omega \left(\boldsymbol{E} \right) \times_1 \left(\boldsymbol{u}_s^t - \boldsymbol{u}_s^{t,(m)} \right) \times_2 \boldsymbol{u}_s^{t,(m)} \right\|_2 + \left\| \mathcal{P}_\Omega \left(\boldsymbol{E} \right) \times_1 \boldsymbol{u}_s^{t,(m)} \times_2 \left(\boldsymbol{u}_s^t - \boldsymbol{u}_s^{t,(m)} \right) \right\|_2$$

$$\quad + \left\| \mathcal{P}_\Omega \left(\boldsymbol{E} \right) \times_1 \left(\boldsymbol{u}_s^t - \boldsymbol{u}_s^{t,(m)} \right) \times_2 \left(\boldsymbol{u}_s^t - \boldsymbol{u}_s^{t,(m)} \right) \right\|_2$$

$$\leq 2 \left\| \mathcal{P}_\Omega \left(\boldsymbol{E} \right) \right\| \left\| \boldsymbol{u}_s^t - \boldsymbol{u}_s^{t,(m)} \right\|_2 \left\| \boldsymbol{u}_s^{t,(m)} \right\|_2 + \left\| \mathcal{P}_\Omega \left(\boldsymbol{E} \right) \right\| \left\| \boldsymbol{u}_s^t - \boldsymbol{u}_s^{t,(m)} \right\|_2^2$$

$$\lesssim \left\| \mathcal{P}_\Omega \left(\boldsymbol{E} \right) \right\| \left\| \boldsymbol{u}_s^t - \boldsymbol{u}_s^{t,(m)} \right\|_2 \left\| \boldsymbol{u}_s^{t,(m)} \right\|_2,$$

where the last line follows from (119). From Corollary D.3, we can further upper bound

$$\left\| \mathcal{P}_\Omega \left(\boldsymbol{E} \right) \times_1 \boldsymbol{u}_s^t \times_2 \boldsymbol{u}_s^t - \mathcal{P}_\Omega \left(\boldsymbol{E} \right) \times_1 \boldsymbol{u}_s^{t,(m)} \times_2 \boldsymbol{u}_s^{t,(m)} \right\|_2 \leq \sigma \left(\sqrt{dp} + \log d \right) \log^{5/2} d \left\| \boldsymbol{u}_s^t - \boldsymbol{u}_s^{t,(m)} \right\|_2 \left\| \boldsymbol{u}_s^{t,(m)} \right\|_2.$$

As a result, we sum over $s \in [r]$ and use the Cauchy-Schwartz inequality to derive

$$\frac{1}{p} \left\| \mathcal{P}_\Omega \left(\boldsymbol{E} \right) \times_1^{\mathsf{seq}} \boldsymbol{U}^t \times_2^{\mathsf{seq}} \boldsymbol{U}^t - \mathcal{P}_\Omega \left(\boldsymbol{E} \right) \times_1^{\mathsf{seq}} \boldsymbol{U}^{t,(m)} \times_2^{\mathsf{seq}} \boldsymbol{U}^{t,(m)} \right\|_{\mathrm{F}}$$

$$\lesssim \frac{\sigma}{p} \left(\sqrt{dp} + \log d \right) \log^{5/2} d \left\| \boldsymbol{U}^t - \boldsymbol{U}^{t,(m)} \right\|_{\mathrm{F}} \left\| \boldsymbol{U}^{t,(m)} \right\|_{\mathrm{F}}$$

$$\lesssim \frac{\sigma}{p} \left(\sqrt{dp} + \log d \right) \log^{5/2} d \left(C_5 \rho^t \mathcal{E}_{\mathsf{local}} + C_6 \frac{\sigma}{\lambda_{\min}^\star} \sqrt{\frac{d \log d}{p}} \right) \left\| \boldsymbol{U}^\star \right\|_{2,\infty}^2$$

$$\ll \lambda_{\min}^{\star 4/3} \left(C_5 \rho^t \mathcal{E}_{\mathsf{local}} + C_6 \frac{\sigma}{\lambda_{\min}^\star} \sqrt{\frac{d \log d}{p}} \right) \left\| \boldsymbol{U}^\star \right\|_{2,\infty}, \tag{133}$$

where the last step arises from conditions that $\frac{\sigma}{\lambda_{\min}^\star} \sqrt{\frac{d \log d}{p}} \ll \frac{1}{\mu^{3/2} r}$ and $\kappa \asymp 1$.

(5) Taking (126), (129), (132) and (133) together, we can invoke the sample size assumption that $p \gg \mu^3 r^2 d^{-2} \log^3 d$ and the union bound to show that: with probability greater than $1 - O(d^{-10})$ one has

$$\left\| \boldsymbol{U}^{t+1} - \boldsymbol{U}^{t+1,(m)} \right\|_{\mathrm{F}}$$

$$
\leq \left(1 - \frac{1}{4}\lambda_{\min}^{\star 4/3}\eta\right)\left\|\boldsymbol{U}^t - \boldsymbol{U}^{t,(m)}\right\|_{\mathrm{F}} + C\frac{\mu^{3/2}r\lambda_{\min}^{\star 4/3}\sqrt{\log d}}{d\sqrt{p}}\left\|\boldsymbol{U}^{t,(m)} - \boldsymbol{U}^{\star}\right\|_{2,\infty}
$$

$$
+ \widetilde{C}\eta\sigma\lambda_{\min}^{\star 1/3}\sqrt{\frac{\mu r d\log d}{p}}\left\|\boldsymbol{U}^{\star}\right\|_{2,\infty} + o\left(1\right)\lambda_{\min}^{\star 4/3}\eta\left(C_5\rho^t\mathcal{E}_{\mathsf{local}} + C_6\frac{\sigma}{\lambda_{\min}^{\star}}\sqrt{\frac{d\log d}{p}}\right)\left\|\boldsymbol{U}^{\star}\right\|_{2,\infty}
$$

$$
\leq \left(1 - \frac{1}{4}\lambda_{\min}^{\star 4/3}\eta\right)\left(C_5\rho^t\mathcal{E}_{\mathsf{local}} + C_6\frac{\sigma}{\lambda_{\min}^{\star}}\sqrt{\frac{d\log d}{p}}\right)\left\|\boldsymbol{U}^{\star}\right\|_{2,\infty} + \widetilde{C}\eta\sigma\lambda_{\min}^{\star 1/3}\sqrt{\frac{d\log d}{p}}\left\|\boldsymbol{U}^{\star}\right\|_{2,\infty}
$$

$$
+ C\eta\lambda_{\max}^{\star 4/3}\frac{\mu^{3/2}r\sqrt{\log d}}{d\sqrt{p}}\left(2C_3\rho^t\mathcal{E}_{\mathsf{local}} + 2C_4\frac{\sigma}{\lambda_{\min}^{\star}}\sqrt{\frac{d\log d}{p}}\right)\left\|\boldsymbol{U}^{\star}\right\|_{2,\infty}
$$

$$
+ o\left(1\right)\lambda_{\min}^{\star 4/3}\eta\left(C_5\rho^t\mathcal{E}_{\mathsf{local}} + C_6\frac{\sigma}{\lambda_{\min}^{\star}}\sqrt{\frac{\mu r d\log d}{p}}\right)\left\|\boldsymbol{U}^{\star}\right\|_{2,\infty}
$$

$$
\leq \left(1 - \frac{1}{5}\lambda_{\min}^{\star 4/3}\eta\right)C_5\rho^t\mathcal{E}_{\mathsf{local}}\left\|\boldsymbol{U}^{\star}\right\|_{2,\infty} + \left(\left(1 - \frac{1}{5}\lambda_{\min}^{\star 4/3}\eta\right)C_6 + \widetilde{C}\lambda_{\min}^{\star 4/3}\eta\right)\frac{\sigma}{\lambda_{\min}^{\star}}\sqrt{\frac{d\log d}{p}}\left\|\boldsymbol{U}^{\star}\right\|_{2,\infty}
$$

$$
\leq \left(1 - \frac{1}{5}\lambda_{\min}^{\star 4/3}\eta\right)C_5\rho^t\mathcal{E}_{\mathsf{local}}\left\|\boldsymbol{U}^{\star}\right\|_{2,\infty} + C_6\frac{\sigma}{\lambda_{\min}^{\star}}\sqrt{\frac{d\log d}{p}}\left\|\boldsymbol{U}^{\star}\right\|_{2,\infty}
$$

$$
\leq \left(C_5\rho^{t+1}\mathcal{E}_{\mathsf{local}} + C_6\frac{\sigma}{\lambda_{\min}^{\star}}\sqrt{\frac{d\log d}{p}}\right)\left\|\boldsymbol{U}^{\star}\right\|_{2,\infty},
$$

provided $0 < \eta \leq \lambda_{\min}^{\star 4/3}/(32\lambda_{\max}^{\star 8/3})$, $1 - (\lambda_{\min}^{\star 4/3}/5)\eta \leq \rho < 1$ and C_6 is sufficiently large.

A.5   Proof of Lemma 5.5

Fix an arbitrary $m \in [d]$. Recall our notation of $\boldsymbol{\Delta}_{\boldsymbol{T}}^{t,(m)}$ in (124). To simplify presentation, we further define

$$
\widehat{\boldsymbol{U}}^{t+1,(m)} := \boldsymbol{U}^{t,(m)} - \eta\left(p^{-1}\mathcal{P}_{\Omega_{-m}}\left(\boldsymbol{\Delta}_{\boldsymbol{T}}^{t,(m)} - \boldsymbol{E}\right)\times_1^{\mathsf{seq}}\boldsymbol{U}^{\star}\times_2^{\mathsf{seq}}\boldsymbol{U}^{\star} + \mathcal{P}_m\left(\boldsymbol{\Delta}_{\boldsymbol{T}}^{t,(m)}\right)\times_1^{\mathsf{seq}}\boldsymbol{U}^{\star}\times_2^{\mathsf{seq}}\boldsymbol{U}^{\star}\right), \quad (134)
$$

$$
\boldsymbol{\Delta}_s^{t,(m)} := \boldsymbol{u}_s^{t,(m)} - \boldsymbol{u}_s^{\star}, \tag{135}
$$

for each $s \in [r]$.

Apply the triangle inequality to yield

$$
\left\|\left(\boldsymbol{U}^{t+1,(m)} - \boldsymbol{U}^{\star}\right)_{m,:}\right\|_2 \leq \underbrace{\left\|\left(\widehat{\boldsymbol{U}}^{t+1,(m)} - \boldsymbol{U}^{\star}\right)_{m,:}\right\|_2}_{=:\alpha_1} + \underbrace{\left\|\left(\boldsymbol{U}^{t+1,(m)} - \widehat{\boldsymbol{U}}^{t+1,(m)}\right)_{m,:}\right\|_2}_{=:\alpha_2},
$$

leaving us with two terms to deal with. As it turns out, we will show that α_1 is the dominant term and α_2 is negligible. To simplify presentation, we shall assume that $\{E_{i,j,k}\}_{i,j,k\in[d]}$ (resp. $\{\chi_{i,j,k}\}_{i,j,k\in[d]}$) are independent random variables.

- Regarding α_1, the definition of $\mathcal{P}_{\Omega_{-m}}$ allows us to derive

$$
\left(\widehat{\boldsymbol{U}}^{t+1,(m)} - \boldsymbol{U}^{\star}\right)_{m,:} = \left(\boldsymbol{U}^{t,(m)} - \boldsymbol{U}^{\star} - \eta\boldsymbol{\Delta}_{\boldsymbol{T}}^{t,(m)}\times_1^{\mathsf{seq}}\boldsymbol{U}^{\star}\times_2^{\mathsf{seq}}\boldsymbol{U}^{\star}\right)_{m,:}
$$

$$
= \left(\boldsymbol{U}^{t,(m)} - \boldsymbol{U}^{\star}\right)_{m,:} - \eta\left(\boldsymbol{\Delta}_{\boldsymbol{T}}^{t,(m)}\times_1^{\mathsf{seq}}\boldsymbol{U}^{\star}\times_2^{\mathsf{seq}}\boldsymbol{U}^{\star}\right)_{m,:}.
$$

We can express $\boldsymbol{\Delta}_{\boldsymbol{T}}^{t,(m)} = \sum_{s\in[r]}\left(\boldsymbol{\Delta}_s^{t,(m)} + \boldsymbol{u}_s^{\star}\right)^{\otimes 3} - \boldsymbol{u}_s^{\star\otimes 3}$ and compute that

$$
\left(\boldsymbol{\Delta}_{\boldsymbol{T}}^{t,(m)}\times_1\boldsymbol{u}_s^{\star}\times_2\boldsymbol{u}_s^{\star}\right)_m
$$

$$
= \left(\left\|\boldsymbol{u}_s^{\star}\right\|_2^2 + \left\langle\boldsymbol{\Delta}_s^{t,(m)}, \boldsymbol{u}_s^{\star}\right\rangle\right)^2\left(\boldsymbol{\Delta}_s^{t,(m)}\right)_m + \left(2\left\|\boldsymbol{u}_s^{\star}\right\|_2^2\left\langle\boldsymbol{\Delta}_s^{t,(m)}, \boldsymbol{u}_s^{\star}\right\rangle + \left\langle\boldsymbol{\Delta}_s^{t,(m)}, \boldsymbol{u}_s^{\star}\right\rangle^2\right)\left(\boldsymbol{u}_s^{\star}\right)_m
$$

$$+ \sum_{i:i\neq s} \left(\langle \boldsymbol{u}_i^\star, \boldsymbol{u}_s^\star \rangle + \langle \boldsymbol{\Delta}_i^{t,(m)}, \boldsymbol{u}_s^\star \rangle \right)^2 (\boldsymbol{\Delta}_i^{t,(m)})_m$$

$$+ \sum_{i:i\neq s} \left(2 \langle \boldsymbol{u}_i^\star, \boldsymbol{u}_s^\star \rangle \langle \boldsymbol{\Delta}_i^{t,(m)}, \boldsymbol{u}_s^\star \rangle + \langle \boldsymbol{\Delta}_i^{t,(m)}, \boldsymbol{u}_s^\star \rangle^2 \right) (\boldsymbol{u}_i^\star)_m \qquad (136)$$

for each $s \in [r]$. This further indicates that

$$\left\| (\widehat{\boldsymbol{U}}^{t+1,(m)} - \boldsymbol{U}^\star)_{m,:} \right\|_2 \leq \underbrace{\left\| \sum_{s\in[r]} \left(1 - \eta \left(\|\boldsymbol{u}_s^\star\|_2^2 + \langle \boldsymbol{\Delta}_s^{t,(m)}, \boldsymbol{u}_s^\star \rangle \right)^2 \right) (\boldsymbol{\Delta}_s^{t,(m)})_m \boldsymbol{e}_s^\top \right\|_2}_{=:\,\beta_1}$$

$$+ \eta \underbrace{\left\| \sum_{s\in[r]} \left(2 \|\boldsymbol{u}_s^\star\|_2^2 \langle \boldsymbol{\Delta}_s^{t,(m)}, \boldsymbol{u}_s^\star \rangle + \langle \boldsymbol{\Delta}_s^{t,(m)}, \boldsymbol{u}_s^\star \rangle^2 \right) (\boldsymbol{u}_s^\star)_m \boldsymbol{e}_s^\top \right\|_2}_{=:\,\beta_2}$$

$$+ \eta \underbrace{\left\| \sum_{s\in[r]} \sum_{i:i\neq s} \left(\langle \boldsymbol{u}_i^\star, \boldsymbol{u}_s^\star \rangle + \langle \boldsymbol{\Delta}_i^{t,(m)}, \boldsymbol{u}_s^\star \rangle \right)^2 (\boldsymbol{\Delta}_i^{t,(m)})_m \boldsymbol{e}_s^\top \right\|_2}_{=:\,\beta_3}$$

$$+ \eta \underbrace{\left\| \sum_{s\in[r]} \sum_{i:i\neq s} \left(2 \langle \boldsymbol{u}_i^\star, \boldsymbol{u}_s^\star \rangle \langle \boldsymbol{\Delta}_i^{t,(m)}, \boldsymbol{u}_s^\star \rangle + \langle \boldsymbol{\Delta}_i^{t,(m)}, \boldsymbol{u}_s^\star \rangle^2 \right) (\boldsymbol{u}_i^\star)_m \boldsymbol{e}_s^\top \right\|_2}_{=:\,\beta_4}.$$

In what follows, we will control the four terms separately.

– For β_1, by (119), we use Cauchy-Schwarz to show that

$$\left(\|\boldsymbol{u}_s^\star\|_2^2 + \langle \boldsymbol{\Delta}_s^{t,(m)}, \boldsymbol{u}_s^\star \rangle \right)^2 \geq \left(\|\boldsymbol{u}_s^\star\|_2^2 - \|\boldsymbol{\Delta}_s^{t,(m)}\|_2 \|\boldsymbol{u}_s^\star\|_2 \right)^2 \geq \frac{2}{3} \|\boldsymbol{u}_s^\star\|_2^4 \geq \frac{2}{3} \lambda_{\min}^{\star 4/3}$$

for each $s \in [r]$. It follows that

$$\beta_1 \leq \left(1 - \frac{2}{3} \lambda_{\min}^{\star 4/3} \eta \right) \left\| (\boldsymbol{U}^{t,(m)} - \boldsymbol{U}^\star)_{m,:} \right\|_2. \qquad (137)$$

– Regarding β_2, by (119), we apply the Cauchy-Schwarz inequality again to get that: for each $s \in [r]$,

$$\left| \left(2 \|\boldsymbol{u}_s^\star\|_2^2 \langle \boldsymbol{\Delta}_s^{t,(m)}, \boldsymbol{u}_s^\star \rangle + \langle \boldsymbol{\Delta}_s^{t,(m)}, \boldsymbol{u}_s^\star \rangle^2 \right) (\boldsymbol{u}_s^\star)_m \right| \leq 3 \|\boldsymbol{u}_s^\star\|_2^3 \|\boldsymbol{u}_s^\star\|_\infty \|\boldsymbol{\Delta}_s^{t,(m)}\|_2 \leq 3 \sqrt{\frac{\mu}{d}} \lambda_{\max}^{\star 4/3} \|\boldsymbol{\Delta}_s^{t,(m)}\|_2,$$

Together with the assumption that $\kappa \asymp 1$, this implies that

$$\beta_2 \leq 3\eta \lambda_{\min}^{\star 4/3} \sqrt{\frac{\mu}{d}} \left\| \boldsymbol{U}^{t,(m)} - \boldsymbol{U}^\star \right\|_{\mathrm{F}}. \qquad (138)$$

– With regards to β_3, we show that for each $s \in [r]$:

$$\left| \sum_{i:i\neq s} \left(\langle \boldsymbol{u}_i^\star, \boldsymbol{u}_s^\star \rangle + \langle \boldsymbol{\Delta}_i^{t,(m)}, \boldsymbol{u}_s^\star \rangle \right)^2 (\boldsymbol{\Delta}_i^{t,(m)})_m \right|$$

$$\lesssim \max_{i:i\neq s} \left| \langle \boldsymbol{u}_i^\star, \boldsymbol{u}_s^\star \rangle \right|^2 \sum_{i:i\neq s} \left| (\boldsymbol{\Delta}_i^{t,(m)})_m \right| + \|\boldsymbol{u}_s^\star\|_2^2 \sum_{i:i\neq s} \|\boldsymbol{\Delta}_i^{t,(m)}\|_2^2 \left| (\boldsymbol{\Delta}_i^{t,(m)})_m \right|$$

$$\leq \lambda_{\max}^{\star 4/3} \frac{\mu\sqrt{r}}{d} \left\| (\boldsymbol{U}^{t,(m)} - \boldsymbol{U}^\star)_{m,:} \right\|_2 + \lambda_{\max}^{\star 2/3} \max_{i:i\neq s} \|\boldsymbol{\Delta}_i^{t,(m)}\|_2 \|\boldsymbol{U}^{t,(m)} - \boldsymbol{U}^\star\|_{\mathrm{F}} \left\| (\boldsymbol{U}^{t,(m)} - \boldsymbol{U}^\star)_{m,:} \right\|_2$$

$$\leq \lambda_{\max}^{\star 4/3} \left(\frac{\mu\sqrt{r}}{d} + o\left(1/r \right) \right) \left\| (\boldsymbol{U}^{t,(m)} - \boldsymbol{U}^\star)_{m,:} \right\|_2 \ll \frac{\lambda_{\max}^{\star 4/3}}{\sqrt{r}} \left\| (\boldsymbol{U}^{t,(m)} - \boldsymbol{U}^\star)_{m,:} \right\|_2,$$

where the last line follows from (119) that $\max_{i:i\neq s} \|\boldsymbol{\Delta}_i^{t,(m)}\|_2 \leq \|\boldsymbol{U}^{t,(m)} - \boldsymbol{U}^\star\|_{\mathrm{F}} \ll \lambda_{\max}^{\star 1/3}/\sqrt{r}$ and the low rank condition $r \ll \sqrt{d/\mu}$. Summing over $s \in [r]$, we get

$$\beta_3 \ll \eta \lambda_{\min}^{\star 4/3} \left\| (\boldsymbol{U}^{t,(m)} - \boldsymbol{U}^\star)_{m,:} \right\|_2 \qquad (139)$$

under the condition $\kappa \asymp 1$.

42

– Turning attention to β_4, we observe that for each $s \in [r]$,

$$\left| \sum_{i:i\neq s} \left(2 \langle \boldsymbol{u}_i^\star, \boldsymbol{u}_s^\star \rangle \langle \boldsymbol{\Delta}_i^{t,(m)}, \boldsymbol{u}_s^\star \rangle + \langle \boldsymbol{\Delta}_i^{t,(m)}, \boldsymbol{u}_s^\star \rangle^2 \right) (\boldsymbol{u}_i^\star)_m \right|$$

$$\lesssim \max_{i:i\neq s} |\langle \boldsymbol{u}_i^\star, \boldsymbol{u}_s^\star \rangle| \|\boldsymbol{u}_s^\star\|_2 \sum_{i:i\neq s} \|\boldsymbol{\Delta}_i^{t,(m)}\|_2 |(\boldsymbol{u}_i^\star)_m| + \|\boldsymbol{u}_s^\star\|_2^2 \sum_{i:i\neq s} \|\boldsymbol{\Delta}_i^{t,(m)}\|_2^2 |(\boldsymbol{u}_i^\star)_m|$$

$$\leq \lambda_{\max}^\star \sqrt{\frac{\mu}{d}} \|\boldsymbol{U}^\star\|_{2,\infty} \|\boldsymbol{U}^{t,(m)} - \boldsymbol{U}^\star\|_{\mathrm{F}} + \lambda_{\max}^{\star 2/3} \max_{i:i\neq s} \|\boldsymbol{\Delta}_i^{t,(m)}\|_2 \|\boldsymbol{U}^\star\|_{2,\infty} \|\boldsymbol{U}^{t,(m)} - \boldsymbol{U}^\star\|_{\mathrm{F}}$$

$$\lesssim \lambda_{\max}^\star \sqrt{\frac{\mu r}{d}} \left(\sqrt{\frac{\mu}{d}} \lambda_{\max}^{\star 1/3} + o(1/\sqrt{r}) \right) \|\boldsymbol{U}^{t,(m)} - \boldsymbol{U}^\star\|_{\mathrm{F}} \ll \lambda_{\max}^{\star 4/3} \sqrt{\frac{\mu}{d}} \|\boldsymbol{U}^{t,(m)} - \boldsymbol{U}^\star\|_{\mathrm{F}},$$

where the last line follows from (119) and the rank assumption $r \ll \sqrt{d/\mu}$. As a consequence,

$$\beta_4 \ll \eta \lambda_{\max}^{\star 4/3} \sqrt{\frac{\mu r}{d}} \|\boldsymbol{U}^{t,(m)} - \boldsymbol{U}^\star\|_{\mathrm{F}}. \tag{140}$$

– Therefore, we have

$$\left\| \left(\widehat{\boldsymbol{U}}^{t+1,(m)} - \boldsymbol{U}^\star \right)_{m,:} \right\|_2 \leq \left(1 - \frac{1}{3} \lambda_{\min}^{\star 4/3} \eta \right) \left\| \left(\boldsymbol{U}^{t,(m)} - \boldsymbol{U}^\star \right)_{m,:} \right\|_2 + 4 \eta \lambda_{\max}^{\star 4/3} \sqrt{\frac{\mu r}{d}} \|\boldsymbol{U}^{t,(m)} - \boldsymbol{U}^\star\|_{\mathrm{F}}. \tag{141}$$

• With regards to α_2, it follows from the definition (46) and (134) that

$$\boldsymbol{U}^{t+1,(m)} - \widehat{\boldsymbol{U}}^{t+1,(m)}$$

$$= -\eta \left(\left(p^{-1} \mathcal{P}_{\Omega_{-m}} + \mathcal{P}_m \right) \left(\boldsymbol{\Delta}_{\boldsymbol{T}}^{t,(m)} \right) \times_1^{\mathsf{seq}} \boldsymbol{U}^{t,(m)} \times_2^{\mathsf{seq}} \boldsymbol{U}^{t,(m)} - \left(p^{-1} \mathcal{P}_{\Omega_{-m}} + \mathcal{P}_m \right) \left(\boldsymbol{\Delta}_{\boldsymbol{T}}^{t,(m)} \right) \times_1^{\mathsf{seq}} \boldsymbol{U}^\star \times_2^{\mathsf{seq}} \boldsymbol{U}^\star \right)$$

$$+ \frac{\eta}{p} \left(\mathcal{P}_{\Omega_{-m}} (\boldsymbol{E}) \times_1^{\mathsf{seq}} \boldsymbol{U}^{t,(m)} \times_2^{\mathsf{seq}} \boldsymbol{U}^{t,(m)} - \mathcal{P}_{\Omega_{-m}} (\boldsymbol{E}) \times_1^{\mathsf{seq}} \boldsymbol{U}^\star \times_2^{\mathsf{seq}} \boldsymbol{U}^\star \right). \tag{142}$$

Recall the definition of $\mathcal{P}_{\Omega_{-m}}$ and $\mathcal{P}_m$. For the m-th row, we have

$$\left(\boldsymbol{U}^{t+1,(m)} - \widehat{\boldsymbol{U}}^{t+1,(m)} \right)_{m,:} = -\eta \left(\boldsymbol{\Delta}_{\boldsymbol{T}}^{t,(m)} \times_1^{\mathsf{seq}} \boldsymbol{U}^{t,(m)} \times_2^{\mathsf{seq}} \boldsymbol{U}^{t,(m)} - \boldsymbol{\Delta}_{\boldsymbol{T}}^{t,(m)} \times_1^{\mathsf{seq}} \boldsymbol{U}^\star \times_2^{\mathsf{seq}} \boldsymbol{U}^\star \right)_{m,:}$$

From the triangle inequality, we can further decompose

$$\left\| \left(\boldsymbol{\Delta}_{\boldsymbol{T}}^{t,(m)} \times_1^{\mathsf{seq}} \boldsymbol{U}^{t,(m)} \times_2^{\mathsf{seq}} \boldsymbol{U}^{t,(m)} - \boldsymbol{\Delta}_{\boldsymbol{T}}^{t,(m)} \times_1^{\mathsf{seq}} \boldsymbol{U}^\star \times_2^{\mathsf{seq}} \boldsymbol{U}^\star \right)_{m,:} \right\|_2$$

$$\leq \underbrace{\left\| \left(\boldsymbol{\Delta}_{\boldsymbol{T}}^{t,(m)} \times_1^{\mathsf{seq}} \left(\boldsymbol{U}^{t,(m)} - \boldsymbol{U}^\star \right) \times_2^{\mathsf{seq}} \boldsymbol{U}^\star \right)_{m,:} \right\|_2}_{=:\gamma_1} + \underbrace{\left\| \left(\boldsymbol{\Delta}_{\boldsymbol{T}}^{t,(m)} \times_1^{\mathsf{seq}} \boldsymbol{U}^\star \times_2^{\mathsf{seq}} \left(\boldsymbol{U}^{t,(m)} - \boldsymbol{U}^\star \right) \right)_{m,:} \right\|_2}_{=:\gamma_2}$$

$$+ \underbrace{\left\| \left(\boldsymbol{\Delta}_{\boldsymbol{T}}^{t,(m)} \times_1^{\mathsf{seq}} \left(\boldsymbol{U}^{t,(m)} - \boldsymbol{U}^\star \right) \times_2^{\mathsf{seq}} \left(\boldsymbol{U}^{t,(m)} - \boldsymbol{U}^\star \right) \right)_{m,:} \right\|_2}_{=:\gamma_3}.$$

Let us consider γ_1 first. It is straightforward to calculate that

$$\left(\boldsymbol{\Delta}_{\boldsymbol{T}}^{t,(m)} \times_1 \boldsymbol{\Delta}_s^{t,(m)} \times_2 \boldsymbol{u}_s^\star \right)_m = \sum_{i \in [r]} \langle \boldsymbol{u}_i^{t,(m)}, \boldsymbol{\Delta}_s^{t,(m)} \rangle \langle \boldsymbol{u}_i^{t,(m)}, \boldsymbol{u}_s^\star \rangle \left(\boldsymbol{\Delta}_i^{t,(m)} \right)_m$$

$$+ \sum_{i \in [r]} \left(\langle \boldsymbol{\Delta}_i^{t,(m)}, \boldsymbol{\Delta}_s^{t,(m)} \rangle \langle \boldsymbol{u}_i^{t,(m)}, \boldsymbol{u}_s^{t,(m)} \rangle + \langle \boldsymbol{u}_i^\star, \boldsymbol{\Delta}_s^{t,(m)} \rangle \langle \boldsymbol{\Delta}_i^{t,(m)}, \boldsymbol{u}_s^\star \rangle + \langle \boldsymbol{\Delta}_i^{t,(m)}, \boldsymbol{\Delta}_s^{t,(m)} \rangle \langle \boldsymbol{\Delta}_i^{t,(m)}, \boldsymbol{u}_s^{t,(m)} \rangle \right) (\boldsymbol{u}_i^\star)_m.$$

for each $s \in [r]$. From (119), we use the triangle inequality and the Cauchy-Schwarz inequality to upper bound

$$\left| \left(\boldsymbol{\Delta}_{\boldsymbol{T}}^{t,(m)} \times_1 \boldsymbol{\Delta}_s^{t,(m)} \times_2 \boldsymbol{u}_s^\star \right)_m \right| \lesssim \|\boldsymbol{\Delta}_s^{t,(m)}\|_2 \|\boldsymbol{u}_s^\star\|_2 \sum_{i \in [r]} \|\boldsymbol{u}_i^{t,(m)}\|_2^2 \left| \left(\boldsymbol{\Delta}_i^{t,(m)} \right)_m \right|$$

43

$$
+ \left\| \boldsymbol{\Delta}_s^{t,(m)} \right\|_2 \big(\left\| \boldsymbol{u}_s^\star \right\|_2 + \left\| \boldsymbol{u}_s^{t,(m)} \right\|_2 \big) \sum_{i\in[r]} \left\| \boldsymbol{\Delta}_i^{t,(m)} \right\|_2 \big(\left\| \boldsymbol{u}_i^{t,(m)} \right\|_2 + \left\| \boldsymbol{u}_i^\star \right\|_2 + \left\| \boldsymbol{\Delta}_i^{t,(m)} \right\|_2 \big) \left| (\boldsymbol{u}_i^\star)_m \right|
$$

$$
\lesssim \lambda_{\max}^{\star 2/3} \left\| \boldsymbol{\Delta}_s^{t,(m)} \right\|_2 \sum_{i\in[r]} \left\| \boldsymbol{u}_i^{t,(m)} \right\|_2 \left| (\boldsymbol{\Delta}_i^{t,(m)})_m \right| + \lambda_{\max}^{\star 2/3} \sqrt{\frac{\mu}{d}} \left\| \boldsymbol{\Delta}_s^{t,(m)} \right\|_2 \sum_{i\in[r]} \left\| \boldsymbol{\Delta}_i^{t,(m)} \right\|_2 \left\| \boldsymbol{u}_i^\star \right\|_2
$$

$$
\lesssim \lambda_{\max}^{\star 2/3} \left\| \boldsymbol{\Delta}_s^{t,(m)} \right\|_2 \left\| \boldsymbol{U}^{t,(m)} \right\|_{\mathrm{F}} \left\| (\boldsymbol{U}^{t,(m)} - \boldsymbol{U}^\star)_{m,:} \right\|_2 + \lambda_{\max}^{\star 2/3} \sqrt{\frac{\mu}{d}} \left\| \boldsymbol{\Delta}_s^{t,(m)} \right\|_2 \left\| \boldsymbol{U}^\star \right\|_{\mathrm{F}} \left\| \boldsymbol{U}^{t,(m)} - \boldsymbol{U}^\star \right\|_{\mathrm{F}}
$$

$$
\lesssim \lambda_{\max}^\star \sqrt{r} \left\| \boldsymbol{\Delta}_s^{t,(m)} \right\|_2 \left\| (\boldsymbol{U}^{t,(m)} - \boldsymbol{U}^\star)_{m,:} \right\|_2 + \lambda_{\max}^\star \sqrt{\frac{\mu r}{d}} \left\| \boldsymbol{\Delta}_s^{t,(m)} \right\|_2 \left\| \boldsymbol{U}^{t,(m)} - \boldsymbol{U}^\star \right\|_{\mathrm{F}}
$$

We then sum over $s \in [r]$ to find that

$$
\left\| \left(\boldsymbol{\Delta}_{\boldsymbol{T}}^{t,(m)} \times_1^{\mathsf{seq}} (\boldsymbol{U}^{t,(m)} - \boldsymbol{U}^\star) \times_2^{\mathsf{seq}} \boldsymbol{U}^\star \right)_{m,:} \right\|_2 \lesssim \lambda_{\max}^\star \sqrt{r} \left\| \boldsymbol{U}^{t,(m)} - \boldsymbol{U}^\star \right\|_{\mathrm{F}} \left\| (\boldsymbol{U}^{t,(m)} - \boldsymbol{U}^\star)_{m,:} \right\|_2
$$
$$
+ \lambda_{\max}^\star \sqrt{\frac{\mu r}{d}} \left\| \boldsymbol{U}^{t,(m)} - \boldsymbol{U}^\star \right\|_{\mathrm{F}}^2, \tag{143}
$$

Moreover, it is easy to see that the upper bound also holds for γ_2. As for γ_3, we can express

$$
\left(\boldsymbol{\Delta}_{\boldsymbol{T}}^{t,(m)} \times_1 \boldsymbol{\Delta}_s^{t,(m)} \times_2 \boldsymbol{\Delta}_s^{t,(m)} \right)_m = \sum_{i\in[r]} \left(2 \left\langle \boldsymbol{u}_i^\star, \boldsymbol{\Delta}_s^{t,(m)} \right\rangle \left\langle \boldsymbol{\Delta}_i^{t,(m)}, \boldsymbol{\Delta}_s^{t,(m)} \right\rangle + \left\langle \boldsymbol{\Delta}_i^{t,(m)}, \boldsymbol{\Delta}_s^{t,(m)} \right\rangle^2 \right) (\boldsymbol{u}_i^\star)_m
$$
$$
+ \sum_{i\in[r]} \left\langle \boldsymbol{u}_i^{t,(m)}, \boldsymbol{\Delta}_s^{t,(m)} \right\rangle^2 (\boldsymbol{\Delta}_i^{t,(m)})_m. \tag{144}
$$

Similarly, we combine (119) wtih the triangle inequality and the Cauchy-Schwarz inequality to bound

$$
\left| \left(\boldsymbol{\Delta}_{\boldsymbol{T}}^{t,(m)} \times_1 \boldsymbol{\Delta}_s^{t,(m)} \times_2 \boldsymbol{\Delta}_s^{t,(m)} \right)_m \right|
$$
$$
\lesssim \left\| \boldsymbol{\Delta}_s^{t,(m)} \right\|_2^2 \sum_{i\in[r]} \left\| \boldsymbol{\Delta}_i^{t,(m)} \right\|_2 \left\| \boldsymbol{u}_i^\star \right\|_2 \left| (\boldsymbol{u}_i^\star)_m \right| + \left\| \boldsymbol{\Delta}_s^{t,(m)} \right\|_2^2 \sum_{i\in[r]} \left\| \boldsymbol{u}_i^{t,(m)} \right\|_2^2 \left| (\boldsymbol{\Delta}_i^{t,(m)})_m \right|
$$
$$
\lesssim \lambda_{\max}^{\star 1/3} \left\| \boldsymbol{\Delta}_s^{t,(m)} \right\|_2^2 \sum_{i\in[r]} \left\| \boldsymbol{\Delta}_i^{t,(m)} \right\|_2 \left| (\boldsymbol{u}_i^\star)_m \right| + \lambda_{\max}^{\star 1/3} \left\| \boldsymbol{\Delta}_s^{t,(m)} \right\|_2^2 \sum_{i\in[r]} \left\| \boldsymbol{u}_i^{t,(m)} \right\|_2 \left| (\boldsymbol{\Delta}_i^{t,(m)})_m \right|
$$
$$
\leq \lambda_{\max}^{\star 1/3} \left\| \boldsymbol{\Delta}_s^{t,(m)} \right\|_2^2 \left\| \boldsymbol{U}^\star \right\|_{2,\infty} \left\| \boldsymbol{U}^{t,(m)} - \boldsymbol{U}^\star \right\|_{\mathrm{F}} + \lambda_{\max}^{\star 1/3} \left\| \boldsymbol{\Delta}_s^{t,(m)} \right\|_2^2 \left\| \boldsymbol{U}^{t,(m)} \right\|_{\mathrm{F}} \left\| (\boldsymbol{U}^{t,(m)} - \boldsymbol{U}^\star)_{m,:} \right\|_2
$$
$$
\leq \lambda_{\max}^{\star 1/3} \left\| \boldsymbol{U}^\star \right\|_{2,\infty} \left\| \boldsymbol{U}^{t,(m)} - \boldsymbol{U}^\star \right\|_{\mathrm{F}}^2 \left\| \boldsymbol{\Delta}_s^{t,(m)} \right\|_2
$$
$$
+ \lambda_{\max}^{\star 1/3} \left\| \boldsymbol{U}^\star \right\|_{\mathrm{F}} \left\| (\boldsymbol{U}^{t,(m)} - \boldsymbol{U}^\star)_{m,:} \right\|_2 \left\| \boldsymbol{U}^{t,(m)} - \boldsymbol{U}^\star \right\|_{\mathrm{F}} \left\| \boldsymbol{\Delta}_s^{t,(m)} \right\|_2.
$$

Sum over $s \in [r]$ to obtain

$$
\left\| \left(\boldsymbol{\Delta}_{\boldsymbol{T}}^{t,(m)} \times_1^{\mathsf{seq}} (\boldsymbol{U}^{t,(m)} - \boldsymbol{U}^\star) \times_2^{\mathsf{seq}} (\boldsymbol{U}^{t,(m)} - \boldsymbol{U}^\star) \right)_{m,:} \right\|_2
$$
$$
\leq \lambda_{\max}^{\star 1/3} \sqrt{r} \left\| \boldsymbol{U}^{t,(m)} - \boldsymbol{U}^\star \right\|_{\mathrm{F}}^2 \left\| (\boldsymbol{U}^{t,(m)} - \boldsymbol{U}^\star)_{m,:} \right\|_2 + \lambda_{\max}^{\star 1/3} \sqrt{\frac{\mu r}{d}} \left\| \boldsymbol{U}^{t,(m)} - \boldsymbol{U}^\star \right\|_{\mathrm{F}}^3.
$$

Since $\left\| \boldsymbol{U}^{t,(m)} - \boldsymbol{U}^\star \right\|_{\mathrm{F}} \ll \lambda_{\max}^{\star 1/3}/\sqrt{r} \leq 1$ by (119), combined with (143), we find that

$$
\left\| \left(\boldsymbol{\Delta}_{\boldsymbol{T}}^{t,(m)} \times_1^{\mathsf{seq}} \boldsymbol{U}^{t,(m)} \times_2^{\mathsf{seq}} \boldsymbol{U}^{t,(m)} - \boldsymbol{\Delta}_{\boldsymbol{T}}^{t,(m)} \times_1^{\mathsf{seq}} \boldsymbol{U}^\star \times_2^{\mathsf{seq}} \boldsymbol{U}^\star \right)_{m,:} \right\|_2
$$
$$
\lesssim \lambda_{\max}^{\star 1/3} \sqrt{r} \left\| \boldsymbol{U}^{t,(m)} - \boldsymbol{U}^\star \right\|_{\mathrm{F}} \left\| (\boldsymbol{U}^{t,(m)} - \boldsymbol{U}^\star)_{m,:} \right\|_2 + \lambda_{\max}^{\star 1/3} \sqrt{\frac{\mu r}{d}} \left\| \boldsymbol{U}^{t,(m)} - \boldsymbol{U}^\star \right\|_{\mathrm{F}}^2
$$
$$
\ll \lambda_{\min}^{\star 4/3} \left\| (\boldsymbol{U}^{t,(m)} - \boldsymbol{U}^\star)_{m,:} \right\|_2 + \lambda_{\min}^{\star 4/3} \sqrt{\frac{\mu r}{d}} \left\| \boldsymbol{U}^{t,(m)} - \boldsymbol{U}^\star \right\|_{\mathrm{F}}. \tag{145}
$$

where the last inequality follows from the assumption that $\kappa \asymp 1$.

44

- Putting (141) and (145) together, we reach the conclusion from (48) and the condition $r \ll \sqrt{d/\mu}$ that,

$$
\begin{aligned}
\left\| \left(\boldsymbol{U}^{t+1,(m)} - \boldsymbol{U}^\star \right)_{m,:} \right\|_2 &\leq \left(1 - \frac{1}{4}\lambda_{\min}^{\star 4/3}\eta\right) \left\| \left(\boldsymbol{U}^{t+1} - \boldsymbol{U}^\star\right)_{m,:} \right\|_2 + 5\eta\lambda_{\min}^{\star 4/3}\sqrt{\frac{\mu r}{d}} \left\| \boldsymbol{U}^{t,(m)} - \boldsymbol{U}^\star \right\|_{\mathrm{F}} \\
&\leq \left(1 - \frac{1}{4}\lambda_{\min}^{\star 4/3}\eta\right) \left(C_7 \rho^t \mathcal{E}_{\mathsf{local}} + C_8 \frac{\sigma}{\lambda_{\min}^\star}\sqrt{\frac{d\log d}{p}} \right) \left\| \boldsymbol{U}^\star \right\|_{2,\infty} \\
&\quad + 5\eta\lambda_{\min}^{\star 4/3} r \sqrt{\frac{\mu}{d}} \left(2C_1 \rho^t \mathcal{E}_{\mathsf{local}} + 2C_2 \frac{\sigma}{\lambda_{\min}^\star}\sqrt{\frac{d\log d}{p}} \right) \left\| \boldsymbol{U}^\star \right\|_{2,\infty} \\
&\leq \left(\left(1 - \frac{1}{4}\lambda_{\min}^{\star 4/3}\eta\right)C_7 + o\left(1\right)C_1\lambda_{\min}^{\star 4/3}\eta \right) \rho^t \mathcal{E}_{\mathsf{local}} \left\| \boldsymbol{U}^\star \right\|_{2,\infty} \\
&\quad + \left(\left(1 - \lambda_{\min}^{\star 4/3}\eta\right)C_8 + o\left(1\right)C_2\lambda_{\min}^{\star 4/3}\eta \right) \frac{\sigma}{\lambda_{\min}^\star}\sqrt{\frac{d\log d}{p}} \left\| \boldsymbol{U}^\star \right\|_{2,\infty} \\
&\leq \left(C_7 \rho^{t+1} \mathcal{E}_{\mathsf{local}} + C_8 \frac{\sigma}{\lambda_{\min}^\star}\sqrt{\frac{d\log d}{p}} \right) \left\| \boldsymbol{U}^\star \right\|_{2,\infty},
\end{aligned}
$$

provided that $0 < \eta \leq \lambda_{\min}^{\star 4/3}/(32\lambda_{\max}^{\star 8/3})$, $1 - \lambda_{\min}^{\star 4/3}\eta/5 \leq \rho < 1$, C_7, C_8 are sufficiently large.

Recognizing that the above bound holds for any $1 \leq m \leq d$, we conclude the proof.

A.6   Proof of Lemma 5.6

It is easy to see that

$$
\left\| \left(\boldsymbol{U}^{t+1} - \boldsymbol{U}^\star\right)_{m,:} \right\|_2 \leq \left\| \boldsymbol{U}^{t+1} - \boldsymbol{U}^{t+1,(m)} \right\|_{\mathrm{F}} + \left\| \left(\boldsymbol{U}^{t+1,(m)} - \boldsymbol{U}^\star\right)_{m,:} \right\|_2. \tag{146}
$$

Combining Lemma 5.4 and Lemma 5.5, we conclude that with probability at least $1 - O(d^{-10})$,

$$
\begin{aligned}
\left\| \left(\boldsymbol{U}^{t+1} - \boldsymbol{U}^\star\right)_{m,:} \right\|_2 &\leq \left(C_5 \rho^t \mathcal{E}_{\mathsf{local}} + C_6 \frac{\sigma}{\lambda_{\min}^\star}\sqrt{\frac{d\log d}{p}} \right) \left\| \boldsymbol{U}^\star \right\|_{2,\infty} \\
&\quad + \left(C_7 \rho^t \mathcal{E}_{\mathsf{local}} + C_8 \frac{\sigma}{\lambda_{\min}^\star}\sqrt{\frac{d\log d}{p}} \right) \left\| \boldsymbol{U}^\star \right\|_{2,\infty} \\
&\leq \left(C_3 \rho^t \mathcal{E}_{\mathsf{local}} + C_4 \frac{\sigma}{\lambda_{\min}^\star}\sqrt{\frac{d\log d}{p}} \right) \left\| \boldsymbol{U}^\star \right\|_{2,\infty},
\end{aligned}
$$

with the proviso that $C_3/(C_5 + C_7)$ and $C_4/(C_6 + C_8)$ are both sufficiently large.

B   Proofs for retrieving tensor components

B.1   Proof of Lemma 5.12

We shall often operate upon the event where the claims in Lemma 5.7 hold, which happens with very high probability (i.e. at least $1 - O\left(d^{-10}\right)$). Recall the definition of $\boldsymbol{\gamma}^{\star\tau}$ in (72). Since $\boldsymbol{\theta}^\tau = \boldsymbol{U}\boldsymbol{U}^\top \boldsymbol{g}^\tau$, this allows us to write that: for each $1 \leq i \leq r$,

$$
\gamma_i^{\star\tau} = \lambda_i^\star \left\langle \boldsymbol{U}\boldsymbol{U}^\top \overline{\boldsymbol{u}}_i^\star, \boldsymbol{g}^\tau \right\rangle,
$$

where we recall that $\lambda_i^\star = \|\boldsymbol{u}_i^\star\|_2^3$ and $\overline{\boldsymbol{u}}_i^\star = \boldsymbol{u}_i^\star/\|\boldsymbol{u}_i^\star\|_2$. Given that $\boldsymbol{g}^\tau$ is a Gaussian vector independent of $\boldsymbol{U}$, we observe that $\boldsymbol{\gamma}^{\star\tau}$ is zero-mean Gaussian conditional on Ω and $\boldsymbol{E}$. In order to understand the order statistics associated with this vector, we first look at its covariance matrix.

Denote by $\boldsymbol{\Sigma}^\tau$ the covariance matrix of $\boldsymbol{\gamma}^{\star\tau}$ (conditional on $\boldsymbol{U}$). Then we have

$$\Sigma_{i,i}^\tau = \lambda_i^{\star 2} \left\| \boldsymbol{U}\boldsymbol{U}^\top \overline{\boldsymbol{u}}_i^\star \right\|_2^2 = \lambda_i^{\star 2} \left\| \mathcal{P}_{\boldsymbol{U}}\left(\overline{\boldsymbol{u}}_i^\star\right) \right\|_2^2, \tag{147a}$$

$$\Sigma_{i,j}^\tau = \lambda_i^\star \lambda_j^\star \left\langle \boldsymbol{U}\boldsymbol{U}^\top \overline{\boldsymbol{u}}_i^\star, \boldsymbol{U}\boldsymbol{U}^\top \overline{\boldsymbol{u}}_j^\star \right\rangle = \lambda_i^\star \lambda_j^\star \left\langle \overline{\boldsymbol{u}}_i^\star, \mathcal{P}_{\boldsymbol{U}}\left(\overline{\boldsymbol{u}}_j^\star\right) \right\rangle$$

$$= \lambda_i^\star \lambda_j^\star \left(\left\langle \overline{\boldsymbol{u}}_i^\star, \overline{\boldsymbol{u}}_j^\star \right\rangle - \left\langle \overline{\boldsymbol{u}}_i^\star, \mathcal{P}_{\boldsymbol{U}^\perp}\left(\overline{\boldsymbol{u}}_j^\star\right) \right\rangle \right), \tag{147b}$$

where we denote by $\mathcal{P}_{\boldsymbol{U}}(\boldsymbol{z}) = \boldsymbol{U}\boldsymbol{U}^\top \boldsymbol{z}$ and $\mathcal{P}_{\boldsymbol{U}^\perp}(\boldsymbol{z}) = \left(\boldsymbol{I} - \boldsymbol{U}\boldsymbol{U}^\top\right)\boldsymbol{z}$. In addition, since the unit vector $\overline{\boldsymbol{u}}_i^\star$ lies in the span of the columns of $\boldsymbol{U}_{\mathsf{orth}}^\star$ (cf. (54)), it follows from Lemma D.6 that

$$\left\| \mathcal{P}_{\boldsymbol{U}}\left(\overline{\boldsymbol{u}}_i^\star\right) \right\|_2 = \left\| \boldsymbol{U}\boldsymbol{U}^\top \overline{\boldsymbol{u}}_i^\star \right\|_2 = \left\| \boldsymbol{U}\boldsymbol{R}\left(\boldsymbol{U}\boldsymbol{R}\right)^\top \overline{\boldsymbol{u}}_i^\star \right\|_2 \geq \sqrt{1 - \left\| \boldsymbol{U}\boldsymbol{R} - \boldsymbol{U}_{\mathsf{orth}}^\star \right\|^2}, \tag{148a}$$

$$\left\| \mathcal{P}_{\boldsymbol{U}^\perp}\left(\overline{\boldsymbol{u}}_i^\star\right) \right\|_2 = \left\| \left(\boldsymbol{I} - \boldsymbol{U}\boldsymbol{U}^\top\right)\overline{\boldsymbol{u}}_i^\star \right\|_2 = \left\| \left(\boldsymbol{I} - \boldsymbol{U}\boldsymbol{R}\left(\boldsymbol{U}\boldsymbol{R}\right)^\top\right)\overline{\boldsymbol{u}}_i^\star \right\|_2 \leq \left\| \boldsymbol{U}\boldsymbol{R} - \boldsymbol{U}_{\mathsf{orth}}^\star \right\|, \tag{148b}$$

where $\boldsymbol{R}$ is a rotation matrix defined in (55). This together with (147a) gives

$$\lambda_i^{\star 2}\left(1 - \left\| \boldsymbol{U}\boldsymbol{R} - \boldsymbol{U}_{\mathsf{orth}}^\star \right\|^2\right) \leq \Sigma_{i,i}^\tau \leq \lambda_i^{\star 2}, \tag{149}$$

where we have also used the fact that $\left\| \mathcal{P}_{\boldsymbol{U}}\left(\overline{\boldsymbol{u}}_i^\star\right) \right\|_2 \leq \left\| \overline{\boldsymbol{u}}_i^\star \right\|_2 = 1$. Moreover, taking together (147a), (148b) and the incoherence condition, we see that for any $1 \leq i \neq j \leq r$,

$$\left| \Sigma_{i,j}^\tau \right| \leq \lambda_i^\star \lambda_j^\star \left\{ \left| \left\langle \overline{\boldsymbol{u}}_i^\star, \overline{\boldsymbol{u}}_j^\star \right\rangle \right| + \left\| \overline{\boldsymbol{u}}_i^\star \right\|_2 \left\| \mathcal{P}_{\boldsymbol{U}^\perp}\left(\overline{\boldsymbol{u}}_j^\star\right) \right\|_2 \right\} \leq \lambda_i^\star \lambda_j^\star \underbrace{\left(\sqrt{\mu/d} + \left\| \boldsymbol{U}\boldsymbol{R} - \boldsymbol{U}_{\mathsf{orth}}^\star \right\|\right)}_{=:\delta_1}, \tag{150}$$

which is expected to be small if δ_1 is small.

From our assumptions on the sample size, the rank and the condition number, we can invoke Lemma 5.7 to see that $\kappa r \delta_1 \ll 1$, where δ_1 is defined in (150) and $\kappa = \lambda_{\max}^\star / \lambda_{\min}^\star$. Thus, we can decompose $\boldsymbol{\Sigma}^\tau$ into two components as follows

$$\boldsymbol{\Sigma}^\tau = \underbrace{\left(1 - \kappa r \delta_1\right) \boldsymbol{D}^{\star 2}}_{=:\widehat{\boldsymbol{\Sigma}}^\tau} + \underbrace{\boldsymbol{\Sigma}^\tau - \left(1 - \kappa r \delta_1\right) \boldsymbol{D}^{\star 2}}_{=:\widecheck{\boldsymbol{\Sigma}}^\tau},$$

where

$$\boldsymbol{D}^\star := \mathsf{diag}\left(\lambda_1^\star, \cdots, \lambda_r^\star\right) \in \mathbb{R}^{r \times r}.$$

As it turns out, both $\widehat{\boldsymbol{\Sigma}}^\tau$ and $\widecheck{\boldsymbol{\Sigma}}^\tau$ are positive definite. Indeed, we first learn from (149) and (150) that: the i-th digaonal entry of $\widecheck{\boldsymbol{\Sigma}}^\tau$ obeys

$$\widecheck{\Sigma}_{i,i}^\tau \geq \lambda_i^{\star 2}\left(1 - \left\| \boldsymbol{U}\boldsymbol{R} - \boldsymbol{U}_{\mathsf{orth}}^\star \right\|^2\right) - \left(1 - \kappa r \delta_1\right) \lambda_i^{\star 2}$$

$$= \lambda_i^{\star 2}\left(\kappa r \delta_1 - \left\| \boldsymbol{U}\boldsymbol{R} - \boldsymbol{U}_{\mathsf{orth}}^\star \right\|^2\right) \geq \lambda_i^{\star 2}\left(\kappa r \delta_1 - \delta_1^2\right)$$

$$\overset{\text{(i)}}{>} \kappa \lambda_i^{\star 2}\left(r - 1\right) \delta_1 \overset{\text{(ii)}}{\geq} \lambda_i^\star \lambda_{\max}^\star \left(r - 1\right) \delta_1$$

$$\geq \sum_{j:j \neq i} \left| \widecheck{\Sigma}_{i,j}^\tau \right|,$$

where (i) holds since $\delta_1 < 1$ under our assumptions, (ii) follows since $\kappa \lambda_i^\star \geq \kappa \lambda_{\min}^\star = \lambda_{\max}^\star$, and the last line makes use of (150). This implies that $\widecheck{\boldsymbol{\Sigma}}^\tau$ is diagonally dominant, and hence $\widecheck{\boldsymbol{\Sigma}}^\tau \succeq \boldsymbol{0}$. In conclusion, both $\widehat{\boldsymbol{\Sigma}}^\tau$ and $\widecheck{\boldsymbol{\Sigma}}^\tau$ are positive definite.

Let $\widehat{\boldsymbol{\gamma}}^{\star\tau}$ and $\widecheck{\boldsymbol{\gamma}}^{\star\tau}$ be independent zero-mean Gaussian random vectors with covariance matrices $\widehat{\boldsymbol{\Sigma}}^\tau$ and $\widecheck{\boldsymbol{\Sigma}}^\tau$, respectively. Clearly, the distribution of $\boldsymbol{\gamma}^{\star\tau}$ is identical to that of $\widehat{\boldsymbol{\gamma}}^{\star\tau} + \widecheck{\boldsymbol{\gamma}}^{\star\tau}$. Consequently, it allows us to look at the distributions of these two random vectors separately.

- In view of (149) and the fact $\kappa r \delta_1 < 1$, one has

$$\widecheck{\Sigma}_{i,i}^\tau \leq \lambda_i^{\star 2} - \left(1 - \kappa r \delta_1\right) \lambda_i^{\star 2} = \kappa r \delta_1 \lambda_i^{\star 2}. \tag{151}$$

46

Thus, with probability at least $1 - O(d^{-10})$, we have

$$\|\breve{\boldsymbol{\gamma}}^{\star\tau}\|_\infty \lesssim \mathbb{E}\left[\max_{1\leq i\leq r}|\breve{\gamma}_i^{\star\tau}|\right] + \max_{1\leq i\leq r}\sqrt{\mathsf{Var}(\breve{\gamma}_i^{\star\tau})\log d} \lesssim \lambda_{\max}^\star\sqrt{\kappa r\delta_1\log d} \ll \lambda_{\min}^\star/(r\sqrt{\log d}), \qquad (152)$$

where the last step arises from $\kappa^3 r^3 \delta_1 \log^2 d \ll 1$ under our sample size, noise and rank conditions. By the condition that $L \asymp r^{2\kappa^2}\log^{3/2} r$, $r \ll d$ and $\kappa \asymp 1$, we take a union bound over $\tau \in [L]$ to find that each entry of $\breve{\boldsymbol{\gamma}}^{\star\tau}$ is fairly small for all $\tau \in [L]$. Another immediate consequence of (152) is that: for any fixed vector $\boldsymbol{v} \in \mathbb{R}^r$, with probability at least $1 - O(d^{-10})$, for all $\tau \in [L]$,

$$\|\breve{\boldsymbol{\gamma}}^{\star\tau}\|_2 \leq \sqrt{r}\|\breve{\boldsymbol{\gamma}}^{\star\tau}\|_\infty \ll \lambda_{\min}^\star, \tag{153a}$$

$$|\langle\boldsymbol{v},\breve{\boldsymbol{\gamma}}^{\star\tau}\rangle| \leq \|\boldsymbol{v}\|_2\|\breve{\boldsymbol{\gamma}}^{\star\tau}\|_2 \ll \|\boldsymbol{v}\|_2\lambda_{\min}^\star. \tag{153b}$$

- We then turn attention to $\widehat{\boldsymbol{\gamma}}^{\star\tau}$, which is composed of independent Gaussian random variables. Let us define $\widehat{\Delta}^\tau := \widehat{\gamma}_1^{\star\tau} - \max_{1<i\leq r}|\widehat{\gamma}_i^{\star\tau}|$ for each $1 \leq \tau \leq L$ and let $\widehat{\Delta}_1^{(1)} \geq \widehat{\Delta}_1^{(2)} \geq \cdots \geq \widehat{\Delta}_1^{(L)}$ denote the order statistics of $\{\widehat{\Delta}_1^\tau\}_{\tau=1}^L$ in descending order. Fix any small constant $\delta > 0$. Invoke Lemma D.5 to demonstrate that: with probability greater than $1 - \delta/r$,

$$\widehat{\Delta}_1^{(1)} \gtrsim \lambda_{\min}^\star, \tag{154a}$$

$$\widehat{\Delta}_1^{(1)} - \widehat{\Delta}_1^{(2)} \gtrsim \frac{\lambda_{\min}^\star}{r\sqrt{\log d}}, \tag{154b}$$

where we use the conditions that $L \asymp r^{2\kappa^2}\log^{3/2} r$, $r \ll d$ and $\kappa \asymp 1$. In addition, let $\widehat{\boldsymbol{\gamma}}_{\searrow 1}^{\star\tau} := \left[\widehat{\gamma}_2^{\star\tau}, \cdots, \widehat{\gamma}_r^{\star\tau}\right]^\top \in \mathbb{R}^{r-1}$. We know from standard Gaussian concentration inequalities and union bounds that for any fixed vector $\boldsymbol{v} \in \mathbb{R}^r$, with probability $1 - O(d^{-20})$, for all $\tau \in [L]$,

$$\widehat{\gamma}_1^{\star\tau} \lesssim \left(\sqrt{\log L} + \sqrt{\log d}\right)\lambda_{\max}^\star \asymp \sqrt{\log d}\,\lambda_{\max}^\star, \tag{155a}$$

$$\|\widehat{\boldsymbol{\gamma}}^{\star\tau}\|_2 \leq \widehat{\gamma}_1^{\star\tau} + \|\widehat{\boldsymbol{\gamma}}_{\searrow 1}^{\star\tau}\|_2 \lesssim \left(\sqrt{\log d} + \sqrt{r\log d}\right)\lambda_{\max}^\star \lesssim \sqrt{r\log d}\,\lambda_{\max}^\star, \tag{155b}$$

$$|\langle\boldsymbol{v},\widehat{\boldsymbol{\gamma}}^{\star\tau}\rangle| \leq \widehat{\gamma}_1^{\star\tau}\|\boldsymbol{v}\|_2 + |\langle\boldsymbol{v}_{\searrow 1},\widehat{\boldsymbol{\gamma}}_{\searrow 1}^{\star\tau}\rangle| \lesssim \|\boldsymbol{v}\|_2\sqrt{\log d}\,\lambda_{\max}^\star, \tag{155c}$$

where $\boldsymbol{v}_{\searrow 1} := [v_2, \cdots, \cdots, v_r] \in \mathbb{R}^{r-1}$.

- Putting (152) and (154) together and invoking the triangle inequality immediately establish (74a) and (74b). On the other hand, combining (152) with (155a) proves (75a); (153a) and (155b) taken collectively establish (75b), whereas (153b) and (155c) prove (75c).

B.2 Proof of Lemma 5.13

Recall that the vector of interest $\overline{\boldsymbol{u}}^\tau$ is the leading singular vector of $\boldsymbol{M}^\tau$ (as constructed in (70c)), where $\boldsymbol{M}^\tau$ satisfies

$$\begin{aligned}
\boldsymbol{M}^\tau &= p^{-1}\boldsymbol{T} \times_3 \boldsymbol{\theta}^\tau = \boldsymbol{T}^\star \times_3 \boldsymbol{\theta}^\tau + \left(p^{-1}\boldsymbol{T} - \boldsymbol{T}^\star\right) \times_3 \boldsymbol{\theta}^\tau \\
&= \gamma_1^{\star\tau}\overline{\boldsymbol{u}}_1^\star\overline{\boldsymbol{u}}_1^{\star\top} + \sum_{s:s\neq 1}\gamma_s^{\star\tau}\overline{\boldsymbol{u}}_s^\star\overline{\boldsymbol{u}}_s^{\star\top} + \left(p^{-1}\boldsymbol{T} - \boldsymbol{T}^\star\right) \times_3 \boldsymbol{\theta}^\tau \\
&= \underbrace{\gamma_1^{\star\tau}\overline{\boldsymbol{u}}_1^\star\overline{\boldsymbol{u}}_1^{\star\top} + \sum_{s:s\neq 1}\gamma_s^{\star\tau}\left(\boldsymbol{I} - \overline{\boldsymbol{u}}_1^\star\overline{\boldsymbol{u}}_1^{\star\top}\right)\overline{\boldsymbol{u}}_s^\star\overline{\boldsymbol{u}}_s^{\star\top}\left(\boldsymbol{I} - \overline{\boldsymbol{u}}_1^\star\overline{\boldsymbol{u}}_1^{\star\top}\right)}_{=:\boldsymbol{M}^{\star\tau}} \\
&\quad + \underbrace{\sum_{s:s\neq 1}\gamma_s^{\star\tau}\langle\overline{\boldsymbol{u}}_s^\star,\overline{\boldsymbol{u}}_1^\star\rangle\left(\overline{\boldsymbol{u}}_1^\star\overline{\boldsymbol{u}}_s^{\star\top} + \overline{\boldsymbol{u}}_s^\star\overline{\boldsymbol{u}}_1^{\star\top}\right) - \sum_{s:s\neq 1}\gamma_s^{\star\tau}\langle\overline{\boldsymbol{u}}_s^\star,\overline{\boldsymbol{u}}_1^\star\rangle^2\overline{\boldsymbol{u}}_s^\star\overline{\boldsymbol{u}}_s^{\star\top}}_{=:\boldsymbol{C}^\tau} + \underbrace{\left(p^{-1}\boldsymbol{T} - \boldsymbol{T}^\star\right) \times_3 \boldsymbol{\theta}^\tau}_{=:\boldsymbol{F}^\tau}, \quad (156)
\end{aligned}$$

and $\gamma_i^{\star\tau}$ $(1 \leq i \leq r)$ is defined in (72).

In what follows, we shall view $\boldsymbol{C}^\tau$ and $\boldsymbol{F}^\tau$ as perturbation terms superimposed on $\boldsymbol{M}^{\star\tau}$. Lemma B.1 below proves that their operator norms are all small under our sample size, noise and rank conditions, which enables to apply Wedin's theorem to justify the ℓ_2 proximity between $\overline{\boldsymbol{u}}^\tau$ and $\overline{\boldsymbol{u}}_1^\star$.

Lemma B.1. *Instate the assumptions of Lemma 5.13. With probability at least $1 - O\left(d^{-10}\right)$, one has*

$$\left\|\boldsymbol{F}^\tau\right\| \lesssim \frac{\sqrt{\mu r}\,\lambda_{\max}^\star \log^3 d}{d^{3/2}p} + \frac{\mu r \lambda_{\max}^\star \log^{5/2} d}{d\sqrt{p}} + \frac{\sigma \log^{7/2} d}{p} + \sigma\sqrt{\frac{rd\log^5 d}{p}}, \tag{157}$$

$$\left\|\boldsymbol{C}^\tau\right\| \lesssim \sqrt{\frac{\mu r \log d}{d}}\,\lambda_{\max}^\star, \tag{158}$$

$$\left\|\boldsymbol{F}^\tau \overline{\boldsymbol{u}}_1^\star\right\|_2 \lesssim \frac{\mu r \lambda_{\max}^\star \log d}{d\sqrt{p}} + \sigma\sqrt{\frac{rd\log^2 d}{p}}, \tag{159}$$

$$\left\|\boldsymbol{C}^\tau \overline{\boldsymbol{u}}_1^\star\right\|_2 \lesssim \sqrt{\frac{\mu r \log d}{d}}\,\lambda_{\max}^\star. \tag{160}$$

Proof. See Appendix B.3. ∎

As a consequence, recalling the definition of $\mathcal{E}_{\mathsf{op}}$ and $\mathcal{E}_{\mathsf{proj}}$ in (81) and (79) respectively, one has

$$\left\|\boldsymbol{M}^\tau - \boldsymbol{M}^{\star\tau}\right\| \leq \left\|\boldsymbol{F}^\tau\right\| + \left\|\boldsymbol{C}^\tau\right\|$$

$$\lesssim \underbrace{\frac{\sqrt{\mu r}\,\lambda_{\max}^\star \log^3 d}{d^{3/2}p} + \frac{\mu r \lambda_{\max}^\star \log^{5/2} d}{d\sqrt{p}} + \frac{\sigma \log^{7/2} d}{p} + \sigma\sqrt{\frac{rd\log^5 d}{p}} + \sqrt{\frac{\mu r \log d}{d}}\,\lambda_{\max}^\star}_{= \mathcal{E}_{\mathsf{op}} \cdot \lambda_{\min}^\star}, \tag{161}$$

and

$$\left\|(\boldsymbol{M}^\tau - \boldsymbol{M}^{\star\tau})\overline{\boldsymbol{u}}_1^\star\right\|_2 \leq \left\|\boldsymbol{F}^\tau \overline{\boldsymbol{u}}_1^\star\right\|_2 + \left\|\boldsymbol{C}^\tau \overline{\boldsymbol{u}}_1^\star\right\|_2 \lesssim \underbrace{\frac{\mu r \lambda_{\max}^\star \log d}{d\sqrt{p}} + \sigma\sqrt{\frac{rd\log^2 d}{p}} + \sqrt{\frac{\mu r \log d}{d}}\,\lambda_{\max}^\star}_{= \mathcal{E}_{\mathsf{proj}} \cdot \lambda_{\min}^\star}. \tag{162}$$

It then follows from Weyl's inequality that

$$\max_{i\in[d]}\left|\sigma_i\left(\boldsymbol{M}^\tau\right) - \sigma_i\left(\boldsymbol{M}^{\star\tau}\right)\right| \leq \left\|\boldsymbol{M}^\tau - \boldsymbol{M}^{\star\tau}\right\| \lesssim \mathcal{E}_{\mathsf{op}} \cdot \lambda_{\min}^\star \ll \lambda_{\min}^\star \tag{163}$$

where $\sigma_i(\boldsymbol{Z})$ denotes the i-th largest singular value of a matrix $\boldsymbol{Z}$ and we use the condition that $\mathcal{E}_{\mathsf{op}} \ll 1$. All in all, these arguments justify that the spectrum of $\boldsymbol{M}^\tau$ is fairly close to that of $\boldsymbol{M}^{\star\tau}$.

Next, we look at the gap between the two leading singular values of $\boldsymbol{M}^{\star\tau}$. To begin with, it is self-evident from the definition of $\boldsymbol{M}^{\star\tau}$ that: $\overline{\boldsymbol{u}}_1^\star$ is the singular vector of $\boldsymbol{M}^{\star\tau}$. In fact, we claim one further result, that is, $\overline{\boldsymbol{u}}_1^\star$ is indeed the leading singular vector of $\boldsymbol{M}^{\star\tau}$ whose singular value is given by $\sigma_1\left(\boldsymbol{M}^{\star\tau}\right) = \gamma_1^{\star\tau}$. Towards this end, let us define

$$\boldsymbol{U}_{\smallsetminus 1}^\tau := \left(\boldsymbol{I} - \overline{\boldsymbol{u}}_1^\star \overline{\boldsymbol{u}}_1^{\star\top}\right)\overline{\boldsymbol{U}}^\star \in \mathbb{R}^{d\times(r-1)} \qquad \text{and} \qquad \boldsymbol{\gamma}_{\smallsetminus 1}^{\star\tau} := \left[\gamma_2^{\star\tau}, \cdots, \gamma_r^{\star\tau}\right]^\top \in \mathbb{R}^{r-1},$$

allowing us to write

$$\sum_{s:s\neq 1} \gamma_s^{\star\tau}\left(\boldsymbol{I} - \overline{\boldsymbol{u}}_1^\star \overline{\boldsymbol{u}}_1^{\star\top}\right)\overline{\boldsymbol{u}}_s^\star \overline{\boldsymbol{u}}_s^{\star\top}\left(\boldsymbol{I} - \overline{\boldsymbol{u}}_1^\star \overline{\boldsymbol{u}}_1^{\star\top}\right) = \boldsymbol{U}_{\smallsetminus 1}^\tau \mathsf{diag}\left(\boldsymbol{\gamma}_{\smallsetminus 1}^{\star\tau}\right)\boldsymbol{U}_{\smallsetminus 1}^{\tau\top} =: \boldsymbol{M}_{\smallsetminus 1}^{\star\tau}.$$

We note that from Lemma D.1, one has

$$\left\|\boldsymbol{U}_{\smallsetminus 1}^\tau\right\| = \left\|\left(\boldsymbol{I} - \overline{\boldsymbol{u}}_1^\star \overline{\boldsymbol{u}}_1^{\star\top}\right)\overline{\boldsymbol{U}}^\star\right\| \leq \left\|\overline{\boldsymbol{U}}^\star\right\| \leq \sqrt{1 + r\sqrt{\mu/d}}.$$

Let $\left|\gamma^{\star\tau}\right|_{(1)} \geq \cdots \geq \left|\gamma^{\star\tau}\right|_{(r)}$ denote the absolute values of $\{\gamma_i^{\star\tau}\}_{i=1}^r$ in descending order. This together with Lemma 5.12 implies that

$$\left\|\boldsymbol{M}_{\smallsetminus 1}^{\star\tau}\right\| \leq \left|\gamma^{\star\tau}\right|_{(2)}\left\|\boldsymbol{U}_{\smallsetminus 1}^\tau\right\|^2 \leq \left|\gamma^{\star\tau}\right|_{(2)}\left(1 + r\sqrt{\mu/d}\right)$$

$$\leq \gamma_1^{\star\tau} - \left(\gamma_1^{\star\tau} - \left|\gamma^{\star\tau}\right|_{(2)}\right) + r\sqrt{\mu/d}\,\gamma_1^{\star\tau} < \gamma_1^{\star\tau},$$

as long as $\kappa r\sqrt{(\mu\log d)/d} \ll 1$. Given that $\overline{\boldsymbol{u}}_1^{\star}$ is the singular vector of $\boldsymbol{M}^{\star\tau}$ with singular value $\gamma_1^{\star\tau}$, we can conclude that $\sigma_1(\boldsymbol{M}^{\star\tau}) = \gamma_1^{\star\tau}$. This also allows us to lower bound the gap between the two largest singular values $\boldsymbol{M}^{\star\tau}$ as follows

$$\sigma_1(\boldsymbol{M}^{\star\tau}) - \sigma_2(\boldsymbol{M}^{\star\tau}) \geq \gamma_1^{\star\tau} - \left\|\boldsymbol{M}_{\searrow 1}^{\star\tau}\right\| \geq \gamma_1^{\star\tau} - \left|\gamma^{\star\tau}\right|_{(2)}\left(1 + r\sqrt{\mu/d}\right)$$
$$\gtrsim \gamma_1^{\star\tau} - \left|\gamma^{\star\tau}\right|_{(2)} \gtrsim \lambda_{\min}^{\star}, \tag{164}$$

provided that $\kappa r\sqrt{(\mu\log d)/d} \ll 1$. We also know from (163) and (164) that

$$\sigma_1(\boldsymbol{M}^{\tau}) - \sigma_2(\boldsymbol{M}^{\tau}) \geq \sigma_1(\boldsymbol{M}^{\star\tau}) - \sigma_2(\boldsymbol{M}^{\star\tau}) - 2\left\|\boldsymbol{M}^{\tau} - \boldsymbol{M}^{\star\tau}\right\| \gtrsim \lambda_{\min}^{\star}.$$

Combined with (162) and Wedin's theorem, we conclude that

$$\left\|\overline{\boldsymbol{u}}^{\tau} - \overline{\boldsymbol{u}}_1^{\star}\right\|_2 \leq \frac{\left\|(\boldsymbol{M}^{\tau} - \boldsymbol{M}^{\star\tau})\overline{\boldsymbol{u}}_1^{\star}\right\|_2}{\sigma_1(\boldsymbol{M}^{\tau}) - \sigma_2(\boldsymbol{M}^{\tau}) - \left\|\boldsymbol{M}^{\tau} - \boldsymbol{M}^{\star\tau}\right\|} \lesssim \frac{\mu r\log d}{d\sqrt{p}} + \frac{\sigma}{\lambda_{\min}^{\star}}\sqrt{\frac{rd\log^2 d}{p}} + \sqrt{\frac{\mu r\log d}{d}}. \tag{165}$$

Here, we have made use of the fact that $\overline{\boldsymbol{u}}^{\tau}$ is the leading singular vector of $\boldsymbol{M}^{\tau}$ obeying $\langle\overline{\boldsymbol{u}}^{\tau}, \overline{\boldsymbol{u}}_1^{\star}\rangle \geq 0$.

B.3 Proof of Lemma B.1

B.3.1 Controlling $\boldsymbol{F}^{\tau}$

- We first consider the spectral norm of $\boldsymbol{F}^{\tau}$. Recall the definition that $\boldsymbol{\theta}^{\tau} = \boldsymbol{U}\boldsymbol{U}^{\top}\boldsymbol{g}^{\tau}$. Let us define

$$\boldsymbol{\theta}^{\star\tau} := \boldsymbol{U}_{\mathsf{orth}}^{\star}\boldsymbol{U}_{\mathsf{orth}}^{\star\top}\boldsymbol{g}^{\tau}$$

and decompose

$$\boldsymbol{F}^{\tau} = \left(p^{-1}\boldsymbol{T} - \boldsymbol{T}^{\star}\right) \times_3 \boldsymbol{\theta}^{\tau} = \underbrace{\left(p^{-1}\boldsymbol{T} - \boldsymbol{T}^{\star}\right) \times_3 \boldsymbol{\theta}^{\star\tau}}_{=:\,\boldsymbol{X}} + \underbrace{\left(p^{-1}\boldsymbol{T} - \boldsymbol{T}^{\star}\right) \times_3 \left(\boldsymbol{\theta}^{\tau} - \boldsymbol{\theta}^{\star\tau}\right)}_{=:\,\boldsymbol{Y}}.$$

In the sequel, we shall control these two terms separately.

- To bound $\|\boldsymbol{X}\|$, observe that $\boldsymbol{\theta}^{\star\tau}$ is independent of $p^{-1}\boldsymbol{T} - \boldsymbol{T}^{\star}$. By Lemma D.4, one has

$$\left\|\left(p^{-1}\boldsymbol{T} - \boldsymbol{T}^{\star}\right) \times_3 \boldsymbol{\theta}^{\star\tau}\right\| \lesssim \left\|\boldsymbol{\theta}^{\star\tau}\right\|_{\infty}\sqrt{\frac{\mu r\log d}{dp}}\lambda_{\max}^{\star} + \left\|\boldsymbol{\theta}^{\star\tau}\right\|_{\infty}\frac{\sigma\log^{5/2}d}{p} + \left\|\boldsymbol{\theta}^{\star\tau}\right\|_2\sigma\sqrt{\frac{d\log d}{p}}. \tag{166}$$

This suggests that we need to control the ℓ_{∞} and ℓ_2 norms of $\boldsymbol{\theta}^{\star\tau}$. Using standard results on Gaussian random vectors and Lemma D.1, we know that with probability at least $1 - O(d^{-20})$,

$$\left\|\boldsymbol{\theta}^{\star\tau}\right\|_{\infty} = \left\|\boldsymbol{U}_{\mathsf{orth}}^{\star}\boldsymbol{U}_{\mathsf{orth}}^{\star\top}\boldsymbol{g}^{\tau}\right\|_{\infty} \lesssim \left\|\boldsymbol{U}_{\mathsf{orth}}^{\star}\right\|_{2,\infty}\sqrt{\log d} \leq \sqrt{\frac{\mu r\log d}{d}}, \tag{167}$$

$$\left\|\boldsymbol{\theta}^{\star\tau}\right\|_2 = \left\|\boldsymbol{U}_{\mathsf{orth}}^{\star}\boldsymbol{U}_{\mathsf{orth}}^{\star\top}\boldsymbol{g}^{\tau}\right\|_2 \lesssim \left\|\boldsymbol{U}_{\mathsf{orth}}^{\star}\right\|_{\mathsf{F}}\sqrt{\log d} = \sqrt{r\log d}. \tag{168}$$

Combining (166) with (167) and (168) reveals that with probability exceeding $1 - O(d^{-20})$,

$$\left\|\left(p^{-1}\boldsymbol{T} - \boldsymbol{T}^{\star}\right) \times_3 \boldsymbol{\theta}^{\star\tau}\right\| \lesssim \frac{\mu r\lambda_{\max}^{\star}\log d}{d\sqrt{p}} + \frac{\sigma}{p}\sqrt{\frac{\mu r\log^6 d}{d}} + \sigma\sqrt{\frac{rd\log^2 d}{p}}$$

$$\asymp \frac{\mu r\lambda_{\max}^{\star}\log d}{d\sqrt{p}} + \sigma\sqrt{\frac{rd\log^2 d}{p}}, \tag{169}$$

where the last inequality holds as long as $p \gtrsim \mu d^{-2}\log^4 d$.

– Turning to $\boldsymbol{Y}$, we can simply upper bound

$$\left\| \left(p^{-1}\boldsymbol{T} - \boldsymbol{T}^\star \right) \times_3 \left(\boldsymbol{\theta}^\tau - \boldsymbol{\theta}^{\star\tau} \right) \right\| \leq \left\| p^{-1}\boldsymbol{T} - \boldsymbol{T}^\star \right\| \left\| \boldsymbol{\theta}^\tau - \boldsymbol{\theta}^{\star\tau} \right\|_2.$$

Since $\mathsf{rank}\left(\boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{U}^\star_{\mathsf{orth}}\boldsymbol{U}^{\star\top}_{\mathsf{orth}} \right) \leq 2r$, Lemma 5.7 and the standard result of Gaussian random vectors yields that: with probability at least $1 - O\left(d^{-12} \right)$,

$$\begin{aligned}
\left\| \boldsymbol{\theta}^\tau - \boldsymbol{\theta}^{\star\tau} \right\|_2 &= \left\| \left(\boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{U}^\star_{\mathsf{orth}}\boldsymbol{U}^{\star\top}_{\mathsf{orth}} \right) \boldsymbol{g}^\tau \right\|_2 \lesssim \left\| \boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{U}^\star_{\mathsf{orth}}\boldsymbol{U}^{\star\top}_{\mathsf{orth}} \right\|_{\mathrm{F}} \sqrt{\log d} \\
&\leq \left\| \boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{U}^\star_{\mathsf{orth}}\boldsymbol{U}^{\star\top}_{\mathsf{orth}} \right\| \sqrt{2r\log d} \lesssim \mathcal{E}_{\mathsf{se}} \sqrt{r\log d} \ll 1,
\end{aligned} \tag{170}$$

where we recall the definition of $\mathcal{E}_{\mathsf{se}}$ in (57) and that $\mathcal{E}_{\mathsf{se}} \ll 1/\sqrt{r\log d}$ by our conditions. Moreover, by Lemma D.2, we know that with probability exceeding $1 - O\left(d^{-10} \right)$,

$$\begin{aligned}
\left\| p^{-1}\boldsymbol{T} - \boldsymbol{T}^\star \right\| &\leq \left\| p^{-1}\mathcal{P}_\Omega\left(\boldsymbol{T}^\star \right) - \boldsymbol{T}^\star \right\| + \left\| p^{-1}\mathcal{P}_\Omega\left(\boldsymbol{E} \right) \right\| \\
&\lesssim \frac{\sqrt{\mu r}\,\lambda^\star_{\max}\log^3 d}{d^{3/2}p} + \frac{\mu\sqrt{r}\,\lambda^\star_{\max}\log^{5/2} d}{d\sqrt{p}} + \frac{\sigma\log^{7/2} d}{p} + \sigma\sqrt{\frac{d\log^5 d}{p}}.
\end{aligned} \tag{171}$$

Combining (170) and (171), we find that

$$\left\| \left(p^{-1}\boldsymbol{T} - \boldsymbol{T}^\star \right) \times_3 \left(\boldsymbol{\theta}^\tau - \boldsymbol{\theta}^{\star\tau} \right) \right\| \lesssim \frac{\sqrt{\mu r}\,\lambda^\star_{\max}\log^3 d}{d^{3/2}p} + \frac{\mu\sqrt{r}\,\lambda^\star_{\max}\log^{5/2} d}{d\sqrt{p}} + \frac{\sigma\log^{7/2} d}{p} + \sigma\sqrt{\frac{d\log^5 d}{p}}. \tag{172}$$

Putting (169) and (172) together shows that

$$\left\| \boldsymbol{F}^\tau \right\| \lesssim \frac{\sqrt{\mu r}\,\lambda^\star_{\max}\log^3 d}{d^{3/2}p} + \frac{\mu r\lambda^\star_{\max}\log^{5/2} d}{d\sqrt{p}} + \frac{\sigma\log^{7/2} d}{p} + \sigma\sqrt{\frac{rd\log^5 d}{p}}.$$

• Next, we turn to $\left\| \boldsymbol{F}^\tau \overline{\boldsymbol{u}}^\star_1 \right\|_2$. By the definition of the operator norm, we know that

$$\left\| \boldsymbol{F}^\tau \overline{\boldsymbol{u}}^\star_1 \right\|_2 = \left\| \left(p^{-1}\boldsymbol{T} - \boldsymbol{T}^\star \right) \times_2 \overline{\boldsymbol{u}}^\star_1 \times_3 \boldsymbol{\theta}^\tau \right\|_2 \leq \left\| \left(p^{-1}\boldsymbol{T} - \boldsymbol{T}^\star \right) \times_2 \overline{\boldsymbol{u}}^\star_1 \right\| \left\| \boldsymbol{\theta}^\tau \right\|_2.$$

Applying Lemma D.4 again reveals that with probability at least $1 - O\left(d^{-12} \right)$,

$$\begin{aligned}
\left\| \left(p^{-1}\boldsymbol{T} - \boldsymbol{T}^\star \right) \times_2 \overline{\boldsymbol{u}}^\star_1 \right\| &\lesssim \left\| \overline{\boldsymbol{u}}^\star_1 \right\|_\infty \sqrt{\frac{\mu r\log d}{dp}}\,\lambda^\star_{\max} + \left\| \overline{\boldsymbol{u}}^\star_1 \right\|_\infty \frac{\sigma\log^{5/2} d}{p} + \left\| \overline{\boldsymbol{u}}^\star_1 \right\|_2 \sigma\sqrt{\frac{d\log d}{p}} \\
&\lesssim \frac{\mu\sqrt{r}\,\lambda^\star_{\max}\sqrt{\log d}}{d\sqrt{p}} + \frac{\sigma\sqrt{\mu\log^5 d}}{\sqrt{d}\,p} + \sigma\sqrt{\frac{d\log d}{p}} \\
&\asymp \frac{\mu\sqrt{r}\,\lambda^\star_{\max}\sqrt{\log d}}{d\sqrt{p}} + \sigma\sqrt{\frac{d\log d}{p}},
\end{aligned} \tag{173}$$

where the last step arises from the condition that $p \gtrsim \mu d^{-2}\log^4 d$. In addition, realizing that $\boldsymbol{U}$ consists of eigenvectors, standard Gaussian random vectors results give that with probability at least $1 - O\left(d^{-12} \right)$,

$$\left\| \boldsymbol{\theta}^\tau \right\|_2 = \left\| \boldsymbol{U}\boldsymbol{U}^\top\boldsymbol{g}^\tau \right\|_2 \lesssim \left\| \boldsymbol{U} \right\|_{\mathrm{F}} \sqrt{\log d} = \sqrt{r\log d}. \tag{174}$$

Combining (173) and (174) shows that with probability exceeding $1 - O\left(d^{-10} \right)$,

$$\left\| \boldsymbol{F}^\tau \overline{\boldsymbol{u}}^\star_1 \right\|_2 \lesssim \frac{\mu r\lambda^\star_{\max}\log d}{d\sqrt{p}} + \sigma\sqrt{\frac{rd\log^2 d}{p}}.$$

50

B.3.2 Controlling C^τ

Recall the definition of C^τ in (156). We first consider the spectral norm of C^τ. It is straightforward to compute that

$$
\begin{aligned}
\big\|C^\tau\big\| &\leq 2\Big\|\sum_{s:s\neq 1}\gamma_s^{\star\tau}\,\langle\overline{u}_s^\star,\overline{u}_1^\star\rangle\,\overline{u}_s^\star\Big\|_2 + \Big|\sum_{s:s\neq 1}\gamma_s^{\star\tau}\,\langle\overline{u}_s^\star,\overline{u}_1^\star\rangle^2\,\Big| \\
&\lesssim \max_{s:s\neq 1}|\langle\overline{u}_s^\star,\overline{u}_1^\star\rangle|\,\big\|\overline{U}^\star\big\|\big\|\gamma^{\star\tau}\big\|_2 + \max_{s:s\neq 1}\langle\overline{u}_s^\star,\overline{u}_1^\star\rangle^2\,\sqrt{r}\big\|\gamma^{\star\tau}\big\|_2 \\
&\lesssim \Big(\sqrt{\tfrac{\mu}{d}}+\tfrac{\mu\sqrt{r}}{d}\Big)\sqrt{r\log d}\,\lambda_{\max}^\star \asymp \sqrt{\tfrac{\mu r\log d}{d}}\,\lambda_{\max}^\star
\end{aligned}
\tag{175}
$$

if $\mu r/d\lesssim 1$, where we recall that $\overline{U}^\star = [\overline{u}_1^\star,\cdots,\overline{u}_r^\star]$. Here, the last line holds owing to (8c), Lemma D.1 (which justifies that $\big\|\overline{U}^\star\big\|\lesssim 1$ if $r\sqrt{\mu/d}\leq 1$) and Lemma 5.12.

The claim (160) arises from the definition of the spectral norm that $\big\|C^\tau\overline{u}_1^\star\big\|_2\leq\big\|C^\tau\big\|$.

B.4 Proof of Lemma 5.14

Let us fix an arbitrary $m\in[d]$. We remind the readers of several definitions: (1) $\gamma_1^{\star\tau}$: see (72); (2) $M^{\star\tau}$: see (156); and (3) $M^{\tau,(m)}$: see (70d).

Before continuing, we state two immediate facts. First, it has already been observed in Appendix B.2 that $\overline{u}_1^\star$ is a singular vector of $M^{\star\tau}$ with singular value $\gamma_1^{\star\tau}$, and hence

$$
\big(\overline{u}_1^\star\big)_m = \big(\gamma_1^{\star\tau}\big)^{-1}M_{m,:}^{\star\tau}\overline{u}_1^\star.
\tag{176}
$$

Here and throughout, $A_{m,:}$ denotes the m-th row of a matrix A. Second, $\overline{u}^{\tau,(m)}$ is the top singular vector of $M^{\tau,(m)}$ such that $\langle\overline{u}^{\tau,(m)},\overline{u}_1^\star\rangle\geq 0$, and we denote by $\gamma_\tau^{(m)}$ the associated singular value. Recall our definition of $\nu^{\tau,(m)}$ in Algorithm 6. Similar to the case of ν^τ, we will show shortly in Lemma 5.16 that the global signs of $\nu^{\tau,(m)}$ and $\overline{u}^{\tau,(m)}$ coincide, and hence

$$
\nu^{\tau,(m)} = \overline{u}^{\tau,(m)}.
\tag{177}
$$

As a result, the proof of this lemma boils down to showing that $\overline{u}^{\tau,(m)}$ (and hence $\nu^{\tau,(m)}$) is sufficiently close to $\overline{u}_1^\star$ in the m-th entry. Towards this end, observe that

$$
\big(\overline{u}^{\tau,(m)}\big)_m = \big(\gamma_\tau^{(m)}\big)^{-1}M_{m,:}^{\tau,(m)}\overline{u}^{\tau,(m)}.
\tag{178}
$$

The above two facts (176) and (178) together with the triangle inequality lead to

$$
\begin{aligned}
\Big|\big(\overline{u}^{\tau,(m)}-\overline{u}_1^\star\big)_m\Big| &\leq \Big|\big\{\big(\gamma_\tau^{(m)}\big)^{-1}-\big(\gamma_1^{\star\tau}\big)^{-1}\big\}M_{m,:}^{\tau,(m)}\overline{u}^{\tau,(m)}\Big| \\
&\quad + \big(\gamma_1^{\star\tau}\big)^{-1}\Big|\big(M^{\tau,(m)}-M^{\star\tau}\big)_{m,:}\overline{u}^{\tau,(m)}\Big| \\
&\quad + \big(\gamma_1^{\star\tau}\big)^{-1}\Big|M_{m,:}^{\star\tau}\big(\overline{u}^{\tau,(m)}-\overline{u}_1^\star\big)\Big| \\
&\leq \underbrace{\Big|\big(\gamma_\tau^{(m)}\big)^{-1}-\big(\gamma_1^{\star\tau}\big)^{-1}\Big|\big\|M_{m,:}^{\tau,(m)}\big\|_2\big\|\overline{u}^{\tau,(m)}\big\|_2}_{=:\,\alpha_1} \\
&\quad + \underbrace{\big(\gamma_1^{\star\tau}\big)^{-1}\big\|\big(M^{\tau,(m)}-M^{\star\tau}\big)_{m,:}\big\|_2\big\|\overline{u}^{\tau,(m)}\big\|_2}_{=:\,\alpha_2} \\
&\quad + \underbrace{\big(\gamma_1^{\star\tau}\big)^{-1}\big\|M_{m,:}^{\star\tau}\big\|_2\big\|\overline{u}^{\tau,(m)}-\overline{u}_1^\star\big\|_2}_{=:\,\alpha_3}.
\end{aligned}
\tag{179}
$$

Therefore, it suffices to upper bound the above three quantities separately.

B.4.1 Controlling α_3

The first step to bound α_3 (cf. (179)) is to control $\left\|\boldsymbol{M}_{m,:}^{\star\tau}\right\|_2$. Towards this end, we first observe from the incoherence conditions that

$$\max_{s:s\neq 1}\left|\left(\overline{\boldsymbol{u}}_s^\star - \langle\overline{\boldsymbol{u}}_s^\star,\overline{\boldsymbol{u}}_1^\star\rangle\overline{\boldsymbol{u}}_1^\star\right)_m\right| \leq \max_{s:s\neq 1}\|\overline{\boldsymbol{u}}_s^\star\|_\infty + \max_{s:s\neq 1}\left|\langle\overline{\boldsymbol{u}}_s^\star,\overline{\boldsymbol{u}}_1^\star\rangle\right|\|\overline{\boldsymbol{u}}_1^\star\|_\infty$$

$$\leq \max_{s:s\neq 1}\|\overline{\boldsymbol{u}}_s^\star\|_\infty + \max_{s:s\neq 1}\|\overline{\boldsymbol{u}}_s^\star\|_2\|\overline{\boldsymbol{u}}_1^\star\|_2\|\overline{\boldsymbol{u}}_1^\star\|_\infty \lesssim \sqrt{\frac{\mu}{d}}. \tag{180}$$

When combined with the definition (156), this gives

$$\left\|\boldsymbol{M}_{m,:}^{\star\tau}\right\|_2 = \left\|\gamma_1^{\star\tau}(\overline{\boldsymbol{u}}_1^\star)_m\overline{\boldsymbol{u}}_1^{\star\top} + \sum_{s:s\neq 1}\gamma_s^{\star\tau}\left(\overline{\boldsymbol{u}}_s^\star - \langle\overline{\boldsymbol{u}}_s^\star,\overline{\boldsymbol{u}}_1^\star\rangle\overline{\boldsymbol{u}}_1^\star\right)_m\overline{\boldsymbol{u}}_s^{\star\top}\left(\boldsymbol{I} - \overline{\boldsymbol{u}}_1^\star\overline{\boldsymbol{u}}_1^{\star\top}\right)\right\|_2$$

$$\leq \gamma_1^{\star\tau}\|\overline{\boldsymbol{u}}_1^\star\|_\infty\|\overline{\boldsymbol{u}}_1^\star\|_2 + \left\|\sum_{s:s\neq 1}\gamma_s^{\star\tau}\left(\overline{\boldsymbol{u}}_s^\star - \langle\overline{\boldsymbol{u}}_s^\star,\overline{\boldsymbol{u}}_1^\star\rangle\overline{\boldsymbol{u}}_1^\star\right)_m\overline{\boldsymbol{u}}_s^{\star\top}\right\|_2$$

$$\leq \gamma_1^{\star\tau}\|\overline{\boldsymbol{u}}_1^\star\|_\infty + \max_{s:s\neq 1}\left|\left(\overline{\boldsymbol{u}}_s^\star - \langle\overline{\boldsymbol{u}}_s^\star,\overline{\boldsymbol{u}}_1^\star\rangle\overline{\boldsymbol{u}}_1^\star\right)_m\right|\|\boldsymbol{\gamma}^{\star\tau}\|_2\|\overline{\boldsymbol{U}}^\star\|$$

$$\overset{\text{(i)}}{\lesssim} \gamma_1^{\star\tau}\sqrt{\frac{\mu}{d}} + \|\boldsymbol{\gamma}^{\star\tau}\|_2\sqrt{\frac{\mu}{d}} \asymp \|\boldsymbol{\gamma}^{\star\tau}\|_2\sqrt{\frac{\mu}{d}} \lesssim \sqrt{\frac{\mu r\log d}{d}}\lambda_{\max}^\star, \tag{181}$$

where (i) arises from (180) and $\|\overline{\boldsymbol{U}}^\star\| \lesssim 1$ if $r\sqrt{\mu/d} \ll 1$, and the last step comes from Lemma 5.12.

The second step is to upper bound $\left\|\overline{\boldsymbol{u}}^{\tau,(m)} - \overline{\boldsymbol{u}}_1^\star\right\|_2$. Towards this, we resort to Wedin's theorem as follows

$$\left\|\overline{\boldsymbol{u}}^{\tau,(m)} - \overline{\boldsymbol{u}}_1^\star\right\|_2 \leq \frac{\left\|\left(\boldsymbol{M}^{\tau,(m)} - \boldsymbol{M}^{\star\tau}\right)\overline{\boldsymbol{u}}_1^\star\right\|_2}{\sigma_1\left(\boldsymbol{M}^{\star\tau}\right) - \sigma_2\left(\boldsymbol{M}^{\star\tau}\right) - \left\|\boldsymbol{M}^{\tau,(m)} - \boldsymbol{M}^{\star\tau}\right\|}, \tag{182}$$

where we rely on the fact that $\langle\overline{\boldsymbol{u}}^{\tau,(m)},\overline{\boldsymbol{u}}_1^\star\rangle \geq 0$. To complete this bound, we need to control $\boldsymbol{M}^{\tau,(m)} - \boldsymbol{M}^{\star\tau}$. Before we move on, we find it helpful to introduce

$$\widehat{\boldsymbol{M}}^{\tau,(m)} := p^{-1}\boldsymbol{T} \times_3 \boldsymbol{\theta}^{\tau,(m)}. \tag{183}$$

Let $\widehat{\boldsymbol{u}}^{\tau,(m)}$ denote the top left singular vector of $\widehat{\boldsymbol{M}}^{\tau,(m)}$ such that

$$\langle\widehat{\boldsymbol{u}}^{\tau,(m)},\overline{\boldsymbol{u}}_1^\star\rangle \geq 0. \tag{184}$$

Since we have already bounded $\left\|\boldsymbol{M}^\tau - \boldsymbol{M}^{\star\tau}\right\|$ in Lemma B.1, we can decompose

$$\boldsymbol{M}^{\tau,(m)} - \boldsymbol{M}^{\star\tau} = \boldsymbol{M}^{\tau,(m)} - \boldsymbol{M}^\tau + \boldsymbol{M}^\tau - \boldsymbol{M}^{\star\tau} = \boldsymbol{M}^{\tau,(m)} - \widehat{\boldsymbol{M}}^{\tau,(m)} + \widehat{\boldsymbol{M}}^{\tau,(m)} - \boldsymbol{M}^\tau + \boldsymbol{M}^\tau - \boldsymbol{M}^{\star\tau}.$$

With these definitions in place, Lemma B.2 below provides the desired bounds.

Lemma B.2. *Instate the assumptions of Lemma 5.14. With probability at least $1 - O\left(d^{-10}\right)$, the following holds simultaneously for all $1 \leq m \leq d$:*

$$\left\|\widehat{\boldsymbol{M}}^{\tau,(m)} - \boldsymbol{M}^{\tau,(m)}\right\| \lesssim \frac{\mu r\lambda_{\max}^\star\sqrt{\log d}}{d\sqrt{p}} + \sigma\sqrt{\frac{rd\log d}{p}}, \tag{185}$$

$$\left\|\widehat{\boldsymbol{M}}^{\tau,(m)} - \boldsymbol{M}^\tau\right\| \lesssim \mathcal{E}_{\mathsf{loo}}\sqrt{\frac{\mu r\log d}{d}}\lambda_{\max}^\star, \tag{186}$$

where $\mathcal{E}_{\mathsf{loo}}$ is defined in (65). As a result, one has

$$\left\|\left(\boldsymbol{M}^{\tau,(m)} - \boldsymbol{M}^{\star\tau}\right)\overline{\boldsymbol{u}}_1^\star\right\|_2 \leq \left\|\boldsymbol{M}^\tau - \boldsymbol{M}^{\tau,(m)}\right\| \lesssim \frac{\mu r\lambda_{\max}^\star\sqrt{\log d}}{d\sqrt{p}} + \sigma\sqrt{\frac{rd\log d}{p}}. \tag{187}$$

Proof. See Appendix B.5. $\blacksquare$

We then can further combine (161) and (162) to deduce that

$$\left\| \boldsymbol{M}^{\tau,(m)} - \boldsymbol{M}^{\star\tau} \right\| \leq \left\| \boldsymbol{M}^{\tau} - \boldsymbol{M}^{\tau,(m)} \right\| + \left\| \boldsymbol{M}^{\tau} - \boldsymbol{M}^{\star\tau} \right\|$$

$$\lesssim \underbrace{\frac{\sqrt{\mu r}\,\lambda_{\max}^{\star}\log^3 d}{d^{3/2}p} + \frac{\mu r \lambda_{\max}^{\star}\log^{5/2} d}{d\sqrt{p}} + \frac{\sigma \log^{7/2} d}{p} + \sigma\sqrt{\frac{rd\log^5 d}{p}} + \sqrt{\frac{\mu r \log d}{d}}\,\lambda_{\max}^{\star}}_{=\,\mathcal{E}_{\mathsf{op}}\cdot\lambda_{\min}^{\star}},$$

(188)

and

$$\left\| \left(\boldsymbol{M}^{\tau,(m)} - \boldsymbol{M}^{\star\tau} \right)\overline{\boldsymbol{u}}_1^{\star} \right\|_2 \lesssim \underbrace{\frac{\mu r \lambda_{\max}^{\star}\log d}{d\sqrt{p}} + \sigma\sqrt{\frac{rd\log^2 d}{p}} + \sqrt{\frac{\mu r \log d}{d}}\,\lambda_{\max}^{\star}}_{=\,\mathcal{E}_{\mathsf{proj}}\cdot\lambda_{\min}^{\star}}.$$

(189)

In particular, we have $\left\| \boldsymbol{M}^{\tau,(m)} - \boldsymbol{M}^{\star\tau} \right\| \ll \lambda_{\min}^{\star}$ under our conditions, and it follows from (164) that

$$\sigma_1\big(\boldsymbol{M}^{\star\tau}\big) - \sigma_2\big(\boldsymbol{M}^{\star\tau}\big) - \left\| \boldsymbol{M}^{\tau,(m)} - \boldsymbol{M}^{\star\tau} \right\| \gtrsim \lambda_{\min}^{\star}.$$

Invoke the bound (182) to obtain

$$\left\| \overline{\boldsymbol{u}}^{\tau,(m)} - \overline{\boldsymbol{u}}_1^{\star} \right\|_2 \lesssim \frac{1}{\lambda_{\min}^{\star}}\left\| \left(\boldsymbol{M}^{\tau,(m)} - \boldsymbol{M}^{\star\tau} \right)\overline{\boldsymbol{u}}_1^{\star} \right\|_2 \lesssim \underbrace{\frac{\mu r \log d}{d\sqrt{p}} + \frac{\sigma}{\lambda_{\min}^{\star}}\sqrt{\frac{rd\log^2 d}{p}} + \sqrt{\frac{\mu r \log d}{d}}}_{=\,\mathcal{E}_{\mathsf{proj}}}.$$

(190)

To finish up, combine this with (181) and the spectral condition to arrive at

$$\alpha_3 \lesssim \mathcal{E}_{\mathsf{proj}}\sqrt{\frac{\mu r \log d}{d}} \lesssim \mathcal{E}_{\mathsf{op}}\sqrt{\frac{\mu r \log d}{d}},$$

(191)

which results from the fact that $\mathcal{E}_{\mathsf{proj}} \leq \mathcal{E}_{\mathsf{op}}$ (cf. (81)).

B.4.2   Controlling α_2

We then turn to α_2 (cf. (179)). Recall the definition of $\boldsymbol{M}^{\tau,(m)}$ in (70c). It is straightforward to verify that

$$\big(\boldsymbol{M}^{\tau,(m)} - \boldsymbol{M}^{\star\tau}\big)_{m,:} = \big(p^{-1}\boldsymbol{T}^{\tau,(m)} \times_3 \boldsymbol{\theta}^{\tau,(m)} - \boldsymbol{T}^{\star} \times_3 \boldsymbol{\theta}^{\tau}\big)_{m,:} + \big(\boldsymbol{T}^{\star} \times_3 \boldsymbol{\theta}^{\tau} - \boldsymbol{M}^{\star\tau}\big)_{m,:}$$

$$= \big(p^{-1}\boldsymbol{T}^{\tau,(m)} \times_3 \boldsymbol{\theta}^{\tau,(m)} - \boldsymbol{T}^{\star} \times_3 \boldsymbol{\theta}^{\tau}\big)_{m,:} + \boldsymbol{C}_{m,:}^{\tau}$$

$$= \boldsymbol{T}_{m,:,:}^{\star} \times_3 \big(\boldsymbol{\theta}^{\tau,(m)} - \boldsymbol{\theta}^{\tau}\big) + \boldsymbol{C}_{m,:}^{\tau},$$

where $\boldsymbol{C}^{\tau}$ is defined in (156).

From the incoherence conditions, we can upper bound

$$\left\| \boldsymbol{C}_{m,:}^{\tau} \right\|_2 = \left\| \sum_{s:s\neq 1} \gamma_s^{\star\tau}\langle \overline{\boldsymbol{u}}_s^{\star}, \overline{\boldsymbol{u}}_1^{\star}\rangle\big((\overline{\boldsymbol{u}}_1^{\star})_m\,\overline{\boldsymbol{u}}_s^{\star\top} + (\overline{\boldsymbol{u}}_s^{\star})_m\,\overline{\boldsymbol{u}}_1^{\star\top}\big) - \sum_{s:s\neq 1}\gamma_s^{\star\tau}\langle \overline{\boldsymbol{u}}_s^{\star}, \overline{\boldsymbol{u}}_1^{\star}\rangle^2(\overline{\boldsymbol{u}}_1^{\star})_m\,\overline{\boldsymbol{u}}_1^{\star\top} \right\|_2$$

$$\leq \left\| \sum_{s:s\neq 1}\gamma_s^{\star\tau}\langle \overline{\boldsymbol{u}}_s^{\star}, \overline{\boldsymbol{u}}_1^{\star}\rangle(\overline{\boldsymbol{u}}_1^{\star})_m\,\overline{\boldsymbol{u}}_s^{\star\top} \right\|_2 + \left| \sum_{s:s\neq 1}\gamma_s^{\star\tau}\langle \overline{\boldsymbol{u}}_s^{\star}, \overline{\boldsymbol{u}}_1^{\star}\rangle(\overline{\boldsymbol{u}}_s^{\star})_m \right| + \left| \sum_{s:s\neq 1}\gamma_s^{\star\tau}\langle \overline{\boldsymbol{u}}_s^{\star}, \overline{\boldsymbol{u}}_1^{\star}\rangle^2(\overline{\boldsymbol{u}}_1^{\star})_m \right|$$

$$\lesssim \left\| \boldsymbol{\gamma}^{\star\tau} \right\|_2\Big(\max_{s:s\neq 1}|\langle \overline{\boldsymbol{u}}_s^{\star}, \overline{\boldsymbol{u}}_1^{\star}\rangle|\,\|\overline{\boldsymbol{u}}_1^{\star}\|_{\infty}\,\|\overline{\boldsymbol{U}}^{\star}\| + \max_{s:s\neq 1}|\langle \overline{\boldsymbol{u}}_s^{\star}, \overline{\boldsymbol{u}}_1^{\star}\rangle|\,\|\overline{\boldsymbol{U}}^{\star}\|_{2,\infty} + \max_{s:s\neq 1}|\langle \overline{\boldsymbol{u}}_s^{\star}, \overline{\boldsymbol{u}}_1^{\star}\rangle|^2\,\|\overline{\boldsymbol{u}}_1^{\star}\|_{\infty}\sqrt{r}\Big)$$

$$\lesssim \left\| \boldsymbol{\gamma}^{\star\tau} \right\|_2\left(\sqrt{\frac{\mu}{d}}\sqrt{\frac{\mu}{d}} + \sqrt{\frac{\mu}{d}}\sqrt{\frac{\mu r}{d}} + \sqrt{\frac{\mu}{d}}\sqrt{\frac{\mu r}{d}}\right) \lesssim \sqrt{\frac{\mu r}{d}}\sqrt{\frac{\mu r \log d}{d}}\,\lambda_{\max}^{\star},$$

(192)

where we use the fact that $\left\| \boldsymbol{\gamma}^{\star\tau} \right\|_2 \lesssim \sqrt{r\log d}\,\lambda_{\max}^{\star}$ from Lemma 5.12 and $\left\| \overline{\boldsymbol{U}}^{\star} \right\| \lesssim 1$ from Lemma D.1.

In addition, we can express

$$\boldsymbol{T}^\star_{m,:,:} \times_3 \left(\boldsymbol{\theta}^{\tau,(m)} - \boldsymbol{\theta}^\tau\right) = \sum_{s \in [r]} \lambda^\star_s \left(\overline{\boldsymbol{u}}^\star_s\right)_m \langle \overline{\boldsymbol{u}}^\star_s, \boldsymbol{\theta}^{\tau,(m)} - \boldsymbol{\theta}^\tau \rangle \overline{\boldsymbol{u}}^\star_s.$$

By construction, we know that

$$\boldsymbol{\theta}^\tau - \boldsymbol{\theta}^{\tau,(m)} = \left(\boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{U}^{(m)}\boldsymbol{U}^{(m)\top}\right)\boldsymbol{g}^\tau$$

is a zero-mean Gaussian random vector conditional on $\mathcal{P}_\Omega\left(\boldsymbol{E}\right)$. Using standard results on Gaussian random vectors, one has: with probability at least $1 - O\left(d^{-11}\right)$, for each $s \in [r]$ and $m \in [d]$,

$$\begin{aligned}
\left|\langle \overline{\boldsymbol{u}}^\star_s, \boldsymbol{\theta}^\tau - \boldsymbol{\theta}^{\tau,(m)}\rangle\right| &= \left|\langle \boldsymbol{g}^\tau, \left(\boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{U}^{(m)}\boldsymbol{U}^{(m)\top}\right)\overline{\boldsymbol{u}}^\star_s \rangle\right| \lesssim \left\|\left(\boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{U}^{(m)}\boldsymbol{U}^{(m)\top}\right)\overline{\boldsymbol{u}}^\star_s\right\|_2 \sqrt{\log d} \\
&\leq \left\|\boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{U}^{(m)}\boldsymbol{U}^{(m)\top}\right\| \sqrt{\log d}
\end{aligned} \tag{193}$$

and

$$\left\|\boldsymbol{\theta}^\tau - \boldsymbol{\theta}^{\tau,(m)}\right\|_2 \lesssim \left\|\boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{U}^{(m)}\boldsymbol{U}^{(m)\top}\right\|_{\mathrm{F}} \sqrt{\log d}. \tag{194}$$

Therefore, we have

$$\begin{aligned}
\left\|\boldsymbol{T}^\star_{m,:,:} \times_3 \left(\boldsymbol{\theta}^{\tau,(m)} - \boldsymbol{\theta}^\tau\right)\right\|_2 &\leq \max_{s \in [r]} \left|\lambda^\star_s \langle \overline{\boldsymbol{u}}^\star_s, \boldsymbol{\theta}^{\tau,(m)} - \boldsymbol{\theta}^\tau \rangle\right| \left\|\overline{\boldsymbol{U}}^\star\right\|_{2,\infty} \left\|\overline{\boldsymbol{U}}^\star\right\| \\
&\lesssim \lambda^\star_{\max} \left\|\overline{\boldsymbol{U}}^\star\right\|_{2,\infty} \left\|\boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{U}^{(m)}\boldsymbol{U}^{(m)\top}\right\| \sqrt{\log d} \\
&\lesssim \mathcal{E}_{\mathsf{loo}} \sqrt{\frac{\mu r}{d}} \sqrt{\frac{\mu r \log d}{d}} \lambda^\star_{\max} \ll \sqrt{\frac{\mu r}{d}} \sqrt{\frac{\mu r \log d}{d}} \lambda^\star_{\max}
\end{aligned}$$

where we have used Lemma 5.8 as well as the conditions that $\mathcal{E}_{\mathsf{loo}} \ll 1$, $\left\|\overline{\boldsymbol{U}}^\star\right\| \lesssim 1$ and $\left\|\overline{\boldsymbol{U}}^\star\right\|_{2,\infty} \lesssim \sqrt{\mu r/d}$.

Putting the above bounds together, we arrive at

$$\left\|\left(\boldsymbol{M}^{\tau,(m)} - \boldsymbol{M}^{\star\tau}\right)_{m,:}\right\|_2 \leq \left\|\boldsymbol{T}^\star_{m,:,:} \times_3 \left(\boldsymbol{\theta}^{\tau,(m)} - \boldsymbol{\theta}^\tau\right)\right\|_2 + \left\|\boldsymbol{C}^\tau_{m,:}\right\|_2 \lesssim \sqrt{\frac{\mu r}{d}} \sqrt{\frac{\mu r \log d}{d}} \lambda^\star_{\max}. \tag{195}$$

We therefore conclude that

$$\alpha_2 \lesssim \sqrt{\frac{\mu r}{d}} \sqrt{\frac{\mu r \log d}{d}} \lambda^\star_{\max} \lesssim \mathcal{E}_{\mathsf{op}} \sqrt{\frac{\mu r \log d}{d}} \lambda^\star_{\max}$$

where we remind the reader of the definition of $\mathcal{E}_{\mathsf{op}}$ in (81).

B.4.3 Controlling α_1

The remaining quantity to control is α_1 (see (179)). Invoke Weyl's inequality to show that

$$\left|\gamma^{\star\tau}_1 - \gamma^{(m)}_\tau\right| \leq \left\|\boldsymbol{M}^{\tau,(m)} - \boldsymbol{M}^{\star\tau}\right\| \lesssim \mathcal{E}_{\mathsf{op}} \lambda^\star_{\min} \ll \gamma^{\star\tau}_1,$$

where the last inequality arises from (188) and Lemma 5.12. Under our sample size, rank and noise conditions, we have

$$\frac{1}{2}\gamma^{\star\tau}_1 \leq \gamma^{\star\tau}_1 - \left|\gamma^{\star\tau}_1 - \gamma^{(m)}_\tau\right| \leq \gamma^{(m)}_\tau \leq \left|\gamma^{\star\tau}_1 - \gamma^{(m)}_\tau\right| + \gamma^{\star\tau}_1 \leq 2\gamma^{\star\tau}_1.$$

This indicates that

$$\frac{\left|\gamma^{\star\tau}_1 - \gamma^{(m)}_\tau\right|}{\gamma^{\star\tau}_1 \gamma^{(m)}_\tau} \lesssim \frac{1}{\left(\gamma^{\star\tau}_1\right)^2} \left\|\boldsymbol{M}^{\tau,(m)} - \boldsymbol{M}^{\star\tau}\right\| \lesssim \frac{1}{\lambda^\star_{\min}} \mathcal{E}_{\mathsf{op}}.$$

Moreover, we learn from (195) and (181) that

$$\left\|\boldsymbol{M}^{\tau,(m)}_{m,:}\right\|_2 \leq \left\|\left(\boldsymbol{M}^{\tau,(m)} - \boldsymbol{M}^{\star\tau}\right)_{m,:}\right\|_2 + \left\|\boldsymbol{M}^{\star\tau}_{m,:}\right\|_2$$

54

$$\lesssim \frac{\mu r \sqrt{\log d}}{d} \lambda^\star_{\max} + \sqrt{\frac{\mu r \log d}{d}} \lambda^\star_{\max} \asymp \sqrt{\frac{\mu r \log d}{d}} \lambda^\star_{\min},$$

where the last step follows from the fact that $\mu r \leq d$ and $\kappa \asymp 1$. Hence, we reach the conclusion that

$$\alpha_1 \lesssim \frac{\left|\gamma_1^{\star\tau} - \gamma_\tau^{(m)}\right|}{\gamma_1^{\star\tau} \gamma_\tau^{(m)}} \left\|\boldsymbol{M}_{m,:}^{\tau,(m)}\right\|_2 \lesssim \mathcal{E}_{\mathsf{op}} \sqrt{\frac{\mu r \log d}{d}}.$$

B.4.4 Combining α_1, α_2 and α_3

Putting together all of the preceding bounds on α_1, α_2 and α_3 immediately establishes the lemma.

B.5 Proof of Lemma B.2

First of all, if the claims (185) and (186) can be established, then putting them together yields

$$\left\|\boldsymbol{M}^\tau - \boldsymbol{M}^{\tau,(m)}\right\| \leq \left\|\boldsymbol{M}^\tau - \widehat{\boldsymbol{M}}^{\tau,(m)}\right\| + \left\|\widehat{\boldsymbol{M}}^{\tau,(m)} - \boldsymbol{M}^{\tau,(m)}\right\| \lesssim \frac{\mu r \lambda^\star_{\max} \sqrt{\log d}}{d\sqrt{p}} + \sigma\sqrt{\frac{r d \log d}{p}}, \qquad (196)$$

where we recall the definition of $\mathcal{E}_{\mathsf{loo}}$ in (65) and use the sample size, noise and rank conditions. The rest of the proof is thus dedicated to establishing (185) and (186). In what follows, we shall assume $\{E_{i,j,k}\}_{i,j,k\in[d]}$ (resp. $\{\chi_{i,j,k}\}_{i,j,k\in[d]}$) are independent random variables to simplify presentation.

B.5.1 Proximity of $\boldsymbol{M}^{\tau,(m)}$ and $\widehat{\boldsymbol{M}}^{\tau,(m)}$

Recall the definition of $\boldsymbol{M}^{\tau,(m)} = p^{-1}\boldsymbol{T}^{(m)} \times_3 \boldsymbol{\theta}^{\tau,(m)}$ in (70c). Comparing this with the definition of $\widehat{\boldsymbol{M}}^{\tau,(m)}$ in (183), we see that

$$\left(\widehat{\boldsymbol{M}}^{\tau,(m)} - \boldsymbol{M}^{\tau,(m)}\right)_{i,j} = \theta_m^{\tau,(m)}\left(T^\star_{i,j,m}(p^{-1}\chi_{i,j,m} - 1) + p^{-1}E_{i,j,m}\chi_{i,j,m}\right), \; i \neq m, j \neq m, \qquad (197\text{a})$$

$$\left(\widehat{\boldsymbol{M}}^{\tau,(m)} - \boldsymbol{M}^{\tau,(m)}\right)_{i,m} = \theta_k^{\tau,(m)} \sum_{k\in[d]}\left(T^\star_{i,m,k}(p^{-1}\chi_{i,m,k} - 1) + p^{-1}E_{i,m,k}\chi_{i,m,k}\right), \; i \neq m, \qquad (197\text{b})$$

$$\left(\widehat{\boldsymbol{M}}^{\tau,(m)} - \boldsymbol{M}^{\tau,(m)}\right)_{m,j} = \theta_k^{\tau,(m)} \sum_{k\in[d]}\left(T^\star_{m,j,k}(p^{-1}\chi_{m,j,k} - 1) + p^{-1}E_{m,j,k}\chi_{m,j,k}\right). \qquad (197\text{c})$$

Note that $\boldsymbol{\theta}^{\tau,(m)} \sim \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{U}^{(m)}\boldsymbol{U}^{(m)\top}\right)$ conditional on $\mathcal{P}_\Omega(\boldsymbol{E})$. Standard Gaussian concentration inequalities reveal that with probability exceeding $1 - O(d^{-10})$,

$$\left\|\boldsymbol{\theta}^{\tau,(m)}\right\|_2 \lesssim \sqrt{r \log d}. \qquad (198)$$

From Lemmas 5.7-5.8 and the fact that $\max\{\mathcal{E}_{\mathsf{se}}, \mathcal{E}_{\mathsf{loo}}\} \ll 1$, we have

$$\max_{i\in[d]}\mathsf{Var}\left(\theta_i^\tau\right) = \|\boldsymbol{U}\|_{2,\infty}^2 \lesssim \frac{\mu r}{d} + \|\boldsymbol{U}^\star_{\mathsf{orth}}\|_{2,\infty}^2 \asymp \frac{\mu r}{d},$$

$$\max_{i\in[d]}\mathsf{Var}\left(\theta_i^{\tau,(m)}\right) = \left\|\boldsymbol{U}^{(m)}\right\|_{2,\infty}^2 \lesssim \|\boldsymbol{U}\|_{2,\infty}^2 \lesssim \frac{\mu r}{d}, \quad 1 \leq m \leq d.$$

As a consequence, standard concentration results assert that with probability $1 - O(d^{-10})$,

$$\left\|\boldsymbol{\theta}^\tau\right\|_\infty \leq \sqrt{\max_{i\in[d]}\mathsf{Var}\left(\theta_i^\tau\right)\log d} \lesssim \sqrt{\frac{\mu r \log d}{d}}; \qquad (199)$$

$$\left\|\boldsymbol{\theta}^{\tau,(m)}\right\|_\infty \leq \sqrt{\max_{i\in[d]}\mathsf{Var}\left(\theta_i^{\tau,(m)}\right)\log d} \lesssim \sqrt{\frac{\mu r \log d}{d}}, \quad 1 \leq m \leq d. \qquad (200)$$

- Regarding the m-th row of $\widehat{\boldsymbol{M}}^{\tau,(m)} - \boldsymbol{M}^{\tau,(m)}$, apply Lemma D.9 to show that with probability $1 - O\left(d^{-11}\right)$,

$$\sum_{j\in[d]} \left(\sum_{k\in[d]} T^{\star}_{m,j,k} \theta^{\tau,(m)}_k (p^{-1}\chi_{m,j,k} - 1)\right)^2 \lesssim \frac{\mu r \lambda^{\star 2}_{\max}}{dp} \left\|\boldsymbol{\theta}^{\tau,(m)}\right\|^2_{\infty} \lesssim \frac{\mu^2 r^2 \lambda^{\star 2}_{\max} \log d}{d^2 p},$$

where the last inequality comes from (200). In addition, Lemma D.10 indicates that with probability exceeding $1 - O\left(d^{-11}\right)$,

$$\sum_{j\in[d]} \left(\sum_{k\in[d]} \theta^{\tau,(m)}_k E_{m,j,k} \chi_{m,j,k}\right)^2 \lesssim \sigma^2 dp \left\|\boldsymbol{\theta}^{\tau,(m)}\right\|^2_2 + \sigma^2 \left\|\boldsymbol{\theta}^{\tau,(m)}\right\|^2_{\infty} \log^5 d$$

$$\lesssim \sigma^2 rdp \log d + \frac{\sigma^2 \mu r \log^6 d}{d} \asymp \sigma^2 rdp \log d,$$

where the second line comes from (200) and (198), and the last inequality holds as long as $p \gg \mu d^{-2} \log^5 d$. These together with (197c) allow us to obtain

$$\left\|\left(\widehat{\boldsymbol{M}}^{\tau,(m)} - \boldsymbol{M}^{\tau,(m)}\right)_{m,:}\right\|^2_2 \lesssim \frac{\mu^2 r^2 \lambda^{\star 2}_{\max} \log d}{d^2 p} + \frac{\sigma^2 rd \log d}{p}. \tag{201}$$

Clearly, this bound is also valid for $\sum_{i:i\neq m}\left\{\left(\widehat{\boldsymbol{M}}^{\tau,(m)} - \boldsymbol{M}^{\tau,(m)}\right)_{i,m}\right\}^2$, namely,

$$\sum_{i:i\neq m} \left\{\left(\widehat{\boldsymbol{M}}^{\tau,(m)} - \boldsymbol{M}^{\tau,(m)}\right)_{i,m}\right\}^2 \lesssim \frac{\mu^2 r^2 \lambda^{\star 2}_{\max} \log d}{d^2 p} + \frac{\sigma^2 rd \log d}{p}. \tag{202}$$

- When it comes to the remaining entries of $\widehat{\boldsymbol{M}}^{\tau,(m)} - \boldsymbol{M}^{\tau,(m)}$, by the fact that the spectral norm of a submatrix is always less than or equal to that of the whole matrix, applying the matrix Bernstein inequality gives that with probability $1 - O\left(d^{-11}\right)$,

$$\left\|\left[\left(\widehat{\boldsymbol{M}}^{\tau,(m)} - \boldsymbol{M}^{\tau,(m)}\right)_{i,j}\right]_{i,j\neq m}\right\| \lesssim \left\|\boldsymbol{\theta}^{\tau,(m)}\right\|_{\infty} \left(\frac{\log d}{p} \left\|\boldsymbol{A}^{\star}\right\|_{\infty} + \sqrt{\frac{\log d}{p}} \left\|\boldsymbol{A}^{\star\top}\right\|_{2,\infty} + \frac{\sigma \log^2 d}{p} + \sigma\sqrt{\frac{d \log d}{p}}\right)$$

$$\lesssim \sqrt{\frac{\mu r \log d}{d}} \left(\frac{\sqrt{\mu r}\, \lambda^{\star}_{\max} \log d}{d^{3/2} p} + \frac{\mu\sqrt{r}\, \lambda^{\star}_{\max} \sqrt{\log d}}{d\sqrt{p}} + \frac{\sigma \log^2 d}{p} + \sigma\sqrt{\frac{d \log d}{p}}\right)$$

$$\lesssim \frac{\mu r \lambda^{\star}_{\max} \sqrt{\log d}}{d\sqrt{p}} + \sigma\sqrt{\frac{rd \log d}{p}},$$

as long as our sample size and rank condition holds.

- Putting the preceding bounds together yields

$$\left\|\widehat{\boldsymbol{M}}^{\tau,(m)} - \boldsymbol{M}^{\tau,(m)}\right\| \lesssim \frac{\mu r \lambda^{\star}_{\max} \sqrt{\log d}}{d\sqrt{p}} + \sigma\sqrt{\frac{rd \log d}{p}}. \tag{203}$$

B.5.2 Proximity of $\boldsymbol{M}^{\tau}$ and $\widehat{\boldsymbol{M}}^{\tau,(m)}$

Recall the definitions of $\boldsymbol{M}^{\tau}$ and $\widehat{\boldsymbol{M}}^{\tau,(m)}$ in (279b) and (183), respectively. From the definition of the operator norm and the triangle inequality, we have

$$\left\|\boldsymbol{M}^{\tau} - \widehat{\boldsymbol{M}}^{\tau,(m)}\right\| \leq \underbrace{\left\|\boldsymbol{T}^{\star} \times_3 \left(\boldsymbol{\theta}^{\tau} - \boldsymbol{\theta}^{\tau,(m)}\right)\right\|}_{=:\,\alpha_1} + \underbrace{\left\|\left(p^{-1}\boldsymbol{T} - \boldsymbol{T}^{\star}\right) \times_3 \left(\boldsymbol{\theta}^{\tau} - \boldsymbol{\theta}^{\tau,(m)}\right)\right\|}_{=:\,\alpha_2}. \tag{204}$$

- To control α_1, we can express

$$\boldsymbol{T}^{\star} \times_3 \left(\boldsymbol{\theta}^{\tau} - \boldsymbol{\theta}^{\tau,(m)}\right) = \sum_{s\in[r]} \lambda^{\star}_s \left\langle \overline{\boldsymbol{u}}^{\star}_s, \boldsymbol{\theta}^{\tau} - \boldsymbol{\theta}^{\tau,(m)}\right\rangle \overline{\boldsymbol{u}}^{\star}_s \overline{\boldsymbol{u}}^{\star\top}_s. \tag{205}$$

As shown in (193), with probability at least $1 - O\left(d^{-12}\right)$,

$$\left|\langle \overline{\boldsymbol{u}}_s^\star, \boldsymbol{\theta}^{\tau,(m)} - \boldsymbol{\theta}^\tau \rangle\right| \lesssim \left\|\boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{U}^{(m)}\boldsymbol{U}^{(m)\top}\right\| \sqrt{\log d}.$$

Consequently, we know from Lemma 5.8 that with probability at least $1 - O\left(d^{-10}\right)$,

$$\left\|\boldsymbol{T}^\star \times_3 \left(\boldsymbol{\theta}^\tau - \boldsymbol{\theta}^{\tau,(m)}\right)\right\| \leq \max_{s \in [r]} \left|\lambda_s^\star \langle \overline{\boldsymbol{u}}_s^\star, \boldsymbol{\theta}^\tau - \boldsymbol{\theta}^{\tau,(m)} \rangle\right| \left\|\overline{\boldsymbol{U}}^\star\right\|^2 \lesssim \left\|\boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{U}^{(m)}\boldsymbol{U}^{(m)\top}\right\| \sqrt{\log d} \, \lambda_{\max}^\star$$

$$\lesssim \mathcal{E}_{\mathsf{loo}} \sqrt{\frac{\mu r \log d}{d}} \, \lambda_{\max}^\star, \tag{206}$$

where we use the fact that $\left\|\overline{\boldsymbol{U}}^\star\right\| \lesssim 1$ if $r\sqrt{\mu/d} \ll 1$.

- When it comes to α_2, combining (171) and (194) with our sample size, noise and rank conditions, one has

$$\left\|\left(p^{-1}\boldsymbol{T} - \boldsymbol{T}^\star\right) \times_3 \left(\boldsymbol{\theta}^\tau - \boldsymbol{\theta}^{\tau,(m)}\right)\right\| \leq \left\|p^{-1}\boldsymbol{T} - \boldsymbol{T}^\star\right\| \left\|\boldsymbol{\theta}^\tau - \boldsymbol{\theta}^{\tau,(m)}\right\|_2 \ll \lambda_{\max}^\star \left\|\boldsymbol{\theta}^\tau - \boldsymbol{\theta}^{\tau,(m)}\right\|_2$$

$$\lesssim \left\|\boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{U}^{(m)}\boldsymbol{U}^{(m)\top}\right\|_{\mathrm{F}} \sqrt{\log d} \, \lambda_{\max}^\star$$

$$\lesssim \mathcal{E}_{\mathsf{loo}} \sqrt{\frac{\mu r \log d}{d}} \, \lambda_{\max}^\star. \tag{207}$$

- Combining (204), (206) and (207), we conclude that

$$\left\|\boldsymbol{M}^\tau - \widehat{\boldsymbol{M}}^{\tau,(m)}\right\| \lesssim \mathcal{E}_{\mathsf{loo}} \sqrt{\frac{\mu r \log d}{d}} \, \lambda_{\max}^\star.$$

B.6 Proof of Lemma 5.15

- We start with the first claim regarding $\|\overline{\boldsymbol{u}}^\tau - \boldsymbol{\nu}^{\tau,(m)}\|_2$, or equivalently, $\|\overline{\boldsymbol{u}}^\tau - \overline{\boldsymbol{u}}^{\tau,(m)}\|_2$ (as argued in the proof of Lemma 5.14). By the triangle inequality, we can upper bound the following two terms separately:

$$\left\|\overline{\boldsymbol{u}}^\tau - \overline{\boldsymbol{u}}^{\tau,(m)}\right\|_2 \leq \underbrace{\left\|\overline{\boldsymbol{u}}^\tau - \widehat{\boldsymbol{u}}^{\tau,(m)}\right\|_2}_{=:\beta_1} + \underbrace{\left\|\widehat{\boldsymbol{u}}^{\tau,(m)} - \overline{\boldsymbol{u}}^{\tau,(m)}\right\|_2}_{=:\beta_2}. \tag{208}$$

Here, we remind the reader that $\widehat{\boldsymbol{u}}^{\tau,(m)}$ is the top left singular vector of $\widehat{\boldsymbol{M}}^{\tau,(m)}$ (see (183)) obeying $\langle \widehat{\boldsymbol{u}}^{\tau,(m)}, \overline{\boldsymbol{u}}_1^\star \rangle \geq 0$.

  – The first term β_1 shall be bounded via Wedin's theorem. From (163) and (164), we have

$$\sigma_1\left(\boldsymbol{M}^\tau\right) - \sigma_2(\boldsymbol{M}^\tau) \geq \sigma_1\left(\boldsymbol{M}^{\star\tau}\right) - \sigma_2(\boldsymbol{M}^{\star\tau}) - 2\left\|\boldsymbol{M}^\tau - \boldsymbol{M}^{\star\tau}\right\| \gtrsim \lambda_{\min}^\star. \tag{209}$$

Combined with Lemma B.2, one has

$$\sigma_1\left(\boldsymbol{M}^\tau\right) - \sigma_2(\boldsymbol{M}^\tau) - \left\|\boldsymbol{M}^\tau - \widehat{\boldsymbol{M}}^{\tau,(m)}\right\| \gtrsim \lambda_{\min}^\star.$$

Note that we have already shown in the proof of Lemma B.2 and Lemma B.4 that $\left\|\overline{\boldsymbol{u}}^\tau - \overline{\boldsymbol{u}}_1^\star\right\|_2 = o(1)$ and $\left\|\widehat{\boldsymbol{u}}^{\tau,(m)} - \overline{\boldsymbol{u}}_1^\star\right\|_2 = o(1)$, which implies that $\boldsymbol{u}^\tau$ and $\widehat{\boldsymbol{u}}^{\tau,(m)}$ are positively correlated. Thus, one can invoke Wedin's theorem and use the bound (186) to reach

$$\left\|\overline{\boldsymbol{u}}^\tau - \widehat{\boldsymbol{u}}^{\tau,(m)}\right\|_2 \leq \frac{\left\|\boldsymbol{M}^\tau - \widehat{\boldsymbol{M}}^{\tau,(m)}\right\|}{\sigma_1\left(\boldsymbol{M}^\tau\right) - \sigma_2(\boldsymbol{M}^\tau) - \left\|\boldsymbol{M}^\tau - \widehat{\boldsymbol{M}}^{\tau,(m)}\right\|} \lesssim \frac{1}{\lambda_{\min}^\star} \left\|\boldsymbol{M}^\tau - \widehat{\boldsymbol{M}}^{\tau,(m)}\right\|$$

$$\lesssim \mathcal{E}_{\mathsf{loo}} \sqrt{\frac{\mu r \log d}{d}} \leq \mathcal{E}_{\mathsf{loo}} \sqrt{r \log d} \, \max\left\{\sqrt{\mu/d}, \left\|\overline{\boldsymbol{u}}^\tau\right\|_\infty\right\}. \tag{210}$$

In addition to this bound on β_1, we also make note of the following simple bound

$$\left\|\widehat{\boldsymbol{u}}^{\tau,(m)}\right\|_\infty \leq \left\|\overline{\boldsymbol{u}}^\tau - \widehat{\boldsymbol{u}}^{\tau,(m)}\right\|_\infty + \left\|\overline{\boldsymbol{u}}^\tau\right\|_\infty \leq \left\|\overline{\boldsymbol{u}}^\tau - \widehat{\boldsymbol{u}}^{\tau,(m)}\right\|_2 + \left\|\overline{\boldsymbol{u}}^\tau\right\|_\infty \lesssim \max\left\{\sqrt{\mu/d}, \left\|\overline{\boldsymbol{u}}^\tau\right\|_\infty\right\}, \tag{211}$$

where the last inequality follows from our sample size, noise and rank condition that $\mathcal{E}_{\mathsf{loo}} \sqrt{r \log d} \ll 1$.

– The second term β_2 is also controlled via Wedin's theorem:

$$\left\|\widehat{\boldsymbol{u}}^{\tau,(m)} - \overline{\boldsymbol{u}}^{\tau,(m)}\right\|_2 \leq \frac{\left\|\left(\widehat{\boldsymbol{M}}^{\tau,(m)} - \boldsymbol{M}^{\tau,(m)}\right)\overline{\boldsymbol{u}}^{\tau,(m)}\right\|_2}{\sigma_1\left(\boldsymbol{M}^{\tau,(m)}\right) - \sigma_2\left(\boldsymbol{M}^{\tau,(m)}\right) - \left\|\widehat{\boldsymbol{M}}^{\tau,(m)} - \boldsymbol{M}^{\tau,(m)}\right\|}. \tag{212}$$

The denominator term is easy to handle. With (196) and (209) in mind, we can apply Weyl's inequality to obtain

$$\sigma_1\left(\boldsymbol{M}^{\tau,(m)}\right) - \sigma_2\left(\boldsymbol{M}^{\tau,(m)}\right) \geq \sigma_1\left(\boldsymbol{M}^{\tau}\right) - \sigma_2\left(\boldsymbol{M}^{\tau}\right) - 2\left\|\boldsymbol{M}^{\tau} - \boldsymbol{M}^{\tau,(m)}\right\| \gtrsim \lambda_{\min}^{\star}. \tag{213}$$

From Lemma B.2, one has $\left\|\boldsymbol{M}^{\tau} - \boldsymbol{M}^{\tau,(m)}\right\| \ll \lambda_{\min}^{\star}$. Therefore, we know that

$$\sigma_1\left(\boldsymbol{M}^{\tau,(m)}\right) - \sigma_2\left(\boldsymbol{M}^{\tau,(m)}\right) - \left\|\widehat{\boldsymbol{M}}^{\tau,(m)} - \boldsymbol{M}^{\tau,(m)}\right\| \gtrsim \lambda_{\min}^{\star}.$$

In addition, Lemma B.3 below develops an upper bound on the numerator term:

Lemma B.3. *Instate the assumptions of Lemma 5.15. With probability at least $1 - O\left(d^{-10}\right)$, one has*

$$\left\|\left(\widehat{\boldsymbol{M}}^{\tau,(m)} - \boldsymbol{M}^{\tau,(m)}\right)\overline{\boldsymbol{u}}^{\tau,(m)}\right\|_2 \lesssim \left\{\frac{\mu r \lambda_{\max}^{\star} \log d}{d\sqrt{p}} + \sigma\sqrt{\frac{rd\log^2 d}{p}}\right\}\left\|\overline{\boldsymbol{u}}^{\tau,(m)}\right\|_{\infty}. \tag{214}$$

Proof. See Appendix B.7. ∎

– Substitution of the above bounds into (212) yields

$$\left\|\widehat{\boldsymbol{u}}^{\tau,(m)} - \overline{\boldsymbol{u}}^{\tau,(m)}\right\|_2 \lesssim \left\{\frac{\mu r \log d}{d\sqrt{p}} + \frac{\sigma}{\lambda_{\min}^{\star}}\sqrt{\frac{rd\log^2 d}{p}}\right\}\left\|\overline{\boldsymbol{u}}^{\tau,(m)}\right\|_{\infty} \ll \left\|\overline{\boldsymbol{u}}^{\tau,(m)}\right\|_{\infty},$$

where the last step holds as long as $p \gg \mu^2 r^2 d^{-2} \log^2 d$ and $\sigma/\lambda_{\min}^{\star} \ll \sqrt{p/(rd\log^2 d)}$. In addition, from (211), we observe that

$$\left\|\overline{\boldsymbol{u}}^{\tau,(m)}\right\|_{\infty} \leq \left\|\widehat{\boldsymbol{u}}^{\tau,(m)} - \overline{\boldsymbol{u}}^{\tau,(m)}\right\|_{\infty} + \left\|\widehat{\boldsymbol{u}}^{\tau,(m)}\right\|_{\infty} \leq \left\|\widehat{\boldsymbol{u}}^{\tau,(m)} - \overline{\boldsymbol{u}}^{\tau,(m)}\right\|_2 + \left\|\widehat{\boldsymbol{u}}^{\tau,(m)}\right\|_{\infty}$$
$$\lesssim o\left(1\right)\left\|\overline{\boldsymbol{u}}^{\tau,(m)}\right\|_{\infty} + \left\|\widehat{\boldsymbol{u}}^{\tau,(m)}\right\|_{\infty} \leq o\left(1\right)\left\|\overline{\boldsymbol{u}}^{\tau,(m)}\right\|_{\infty} + \max\left\{\sqrt{\mu/d},\, \left\|\overline{\boldsymbol{u}}^{\tau}\right\|_{\infty}\right\},$$

from which we can deduce that

$$\left\|\overline{\boldsymbol{u}}^{\tau,(m)}\right\|_{\infty} \lesssim \max\left\{\sqrt{\mu/d},\, \left\|\overline{\boldsymbol{u}}^{\tau}\right\|_{\infty}\right\}.$$

As a consequence, one immediately obtains

$$\left\|\widehat{\boldsymbol{u}}^{\tau,(m)} - \overline{\boldsymbol{u}}^{\tau,(m)}\right\|_2 \lesssim \left\{\frac{\mu r \log d}{d\sqrt{p}} + \frac{\sigma}{\lambda_{\min}^{\star}}\sqrt{\frac{rd\log^2 d}{p}}\right\}\max\left\{\sqrt{\mu/d},\, \left\|\overline{\boldsymbol{u}}^{\tau}\right\|_{\infty}\right\}. \tag{215}$$

– Combining (208), (210) and (215) and the definition of $\mathcal{E}_{\mathsf{loo}}$, we arrive at

$$\left\|\overline{\boldsymbol{u}}^{\tau} - \overline{\boldsymbol{u}}^{\tau,(m)}\right\|_2 \lesssim \mathcal{E}_{\mathsf{loo}}\sqrt{r\log d}\max\left\{\sqrt{\mu/d},\, \left\|\overline{\boldsymbol{u}}^{\tau}\right\|_{\infty}\right\}. \tag{216}$$

Comparing this bound with the first claim of the lemma, we see that the claim can be established as long as we can show that

$$\left\|\overline{\boldsymbol{u}}^{\tau}\right\|_{\infty} \lesssim \sqrt{\mu/d}. \tag{217}$$

To justify this bound (217), we make use of Lemma 5.14 to derive that

$$\left|\left(\overline{\boldsymbol{u}}^{\tau} - \overline{\boldsymbol{u}}_1^{\star}\right)_m\right| \leq \left|\left(\overline{\boldsymbol{u}}^{\tau} - \overline{\boldsymbol{u}}^{\tau,(m)}\right)_m\right| + \left|\left(\overline{\boldsymbol{u}}^{\tau,(m)} - \overline{\boldsymbol{u}}_1^{\star}\right)_m\right|$$

58

$$\leq \left\| \overline{\boldsymbol{u}}^\tau - \overline{\boldsymbol{u}}^{\tau,(m)} \right\|_2 + \left| \left(\overline{\boldsymbol{u}}^{\tau,(m)} - \overline{\boldsymbol{u}}_1^\star \right)_m \right|$$

$$\lesssim \left(\mathcal{E}_{\mathsf{loo}} + \mathcal{E}_{\mathsf{op}} \right) \sqrt{r \log d} \, \max \left\{ \sqrt{\mu/d}, \left\| \overline{\boldsymbol{u}}^\tau \right\|_\infty \right\}$$

for each $m \in [d]$. Maximizing over $m \in [d]$ gives that

$$\left\| \overline{\boldsymbol{u}}^\tau - \overline{\boldsymbol{u}}_1^\star \right\|_\infty \lesssim \left(\mathcal{E}_{\mathsf{loo}} + \mathcal{E}_{\mathsf{op}} \right) \sqrt{r \log d} \, \max \left\{ \sqrt{\mu/d}, \left\| \overline{\boldsymbol{u}}^\tau \right\|_\infty \right\} \tag{218}$$

$$\ll \max \left\{ \sqrt{\mu/d}, \left\| \overline{\boldsymbol{u}}^\tau \right\|_\infty \right\}, \tag{219}$$

where we use the condition that $\left(\mathcal{E}_{\mathsf{loo}} + \mathcal{E}_{\mathsf{op}} \right) \sqrt{r \log d} \ll 1$. Apply the triangle inequality to yield

$$\left\| \overline{\boldsymbol{u}}^\tau \right\|_\infty \leq \left\| \overline{\boldsymbol{u}}^\tau - \overline{\boldsymbol{u}}_1^\star \right\|_\infty + \left\| \overline{\boldsymbol{u}}_1^\star \right\|_\infty \leq o\left(1\right) \left\| \overline{\boldsymbol{u}}^\tau \right\|_\infty + \sqrt{\mu/d}.$$

These allow us to establish the claim (217), which in turn finishes the proof for the first claim of this lemma.

- The second claim (83) of this lemma follows immediately from (217) and (218).

- It remains to prove the last claim (84). Recall the definition of λ_τ and $\lambda_\tau^{(m)}$ in (71). We can decompose

$$\left\langle p^{-1}\boldsymbol{T}^{(m)}, (\overline{\boldsymbol{u}}^{\tau,(m)})^{\otimes 3} \right\rangle - \left\langle p^{-1}\boldsymbol{T}, (\overline{\boldsymbol{u}}^\tau)^{\otimes 3} \right\rangle = \underbrace{\left\langle p^{-1}\boldsymbol{T}^{(m)} - p^{-1}\boldsymbol{T}, (\overline{\boldsymbol{u}}^{\tau,(m)})^{\otimes 3} \right\rangle}_{=:\,\beta_1}$$

$$+ \underbrace{\left\langle p^{-1}\boldsymbol{T}, (\overline{\boldsymbol{u}}^{\tau,(m)})^{\otimes 3} - (\overline{\boldsymbol{u}}^\tau)^{\otimes 3} \right\rangle}_{=:\,\beta_2} \tag{220}$$

In what follows, we will control β_1 and β_2 seperately.

For β_1, we note that all non-zero entries of $\boldsymbol{T}^{(m)} - \boldsymbol{T}$ are located in the mth slices, and are independent of $\overline{\boldsymbol{u}}^{\tau,(m)}$. This type of quantities have appeared many times and we omit the detailed proof for conciseness. By the Bernstein inequality, one can show that with probability at least $1 - O\left(d^{-10}\right)$,

$$\left| \left\langle p^{-1}\boldsymbol{T}^{(m)} - p^{-1}\boldsymbol{T}, (\overline{\boldsymbol{u}}^{\tau,(m)})^{\otimes 3} \right\rangle \right| \lesssim \sqrt{\frac{\mu r \log d}{d^2 p}} \sqrt{\frac{\mu}{d}} \lambda_{\max}^\star. \tag{221}$$

Next, we turn to β_2. From our sample size and noise condition, Lemma D.1 and Corollary D.3 demonstrates that with probability at least $1 - O\left(d^{-10}\right)$,

$$\left\| p^{-1}\boldsymbol{T} \right\| \leq \left\| p^{-1}\boldsymbol{T} - \boldsymbol{T}^\star \right\| + \left\| \boldsymbol{T}^\star \right\| \leq \left\| p^{-1}\boldsymbol{T} - \boldsymbol{T}^\star \right\| + \left\| \boldsymbol{A}^\star \right\| \lesssim \lambda_{\max}^\star.$$

where we use the fact that the tensor spectral norm is always less than or equal to that of its matricization. By the definition of the operator norm, one has

$$|\beta_2| \leq 3 \left| \left\langle p^{-1}\boldsymbol{T}, (\overline{\boldsymbol{u}}^{\tau,(m)})^{\otimes 2} \otimes \left(\overline{\boldsymbol{u}}^{\tau,(m)} - \overline{\boldsymbol{u}}^\tau \right) \right\rangle \right| + 3 \left| \left\langle p^{-1}\boldsymbol{T}, (\overline{\boldsymbol{u}}^{\tau,(m)}) \otimes \left(\overline{\boldsymbol{u}}^{\tau,(m)} - \overline{\boldsymbol{u}}^\tau \right)^{\otimes 2} \right\rangle \right|$$

$$+ \left| \left\langle p^{-1}\boldsymbol{T}, \left(\overline{\boldsymbol{u}}^{\tau,(m)} - \overline{\boldsymbol{u}}^\tau \right)^{\otimes 3} \right\rangle \right|$$

$$\lesssim \left\| p^{-1}\boldsymbol{T} \right\| \left(\left\| \overline{\boldsymbol{u}}^{\tau,(m)} - \overline{\boldsymbol{u}}^\tau \right\|_2 + \left\| \overline{\boldsymbol{u}}^{\tau,(m)} - \overline{\boldsymbol{u}}^\tau \right\|_2^2 + \left\| \overline{\boldsymbol{u}}^{\tau,(m)} - \overline{\boldsymbol{u}}^\tau \right\|_2^3 \right)$$

$$\lesssim \mathcal{E}_{\mathsf{loo}} \sqrt{\frac{\mu r \log d}{d}} \lambda_{\max}^\star \tag{222}$$

where we use (216) and (217) in the last step and the fact that $\mathcal{E}_{\mathsf{loo}} \sqrt{\mu r \log d / d} \ll 1$.

Combining (221) and (222) immediately establishes (84).

B.7 Proof of Lemma B.3

Recalling the definitions of $\widehat{\boldsymbol{M}}^{\tau,(m)}$ and $\boldsymbol{M}^{\tau,(m)}$ in (183) and (70c), respectively, we observe that $\widehat{\boldsymbol{M}}^{\tau,(m)} - \boldsymbol{M}^{\tau,(m)}$ is independent of $\overline{\boldsymbol{u}}^{\tau,(m)}$ conditional on $\mathcal{P}_{\Omega_{-m}}(\boldsymbol{E})$ and $\boldsymbol{g}$.

- The m-th entry of $\big(\widehat{\boldsymbol{M}}^{\tau,(m)} - \boldsymbol{M}^{\tau,(m)}\big)\overline{\boldsymbol{u}}^{\tau,(m)}$ can be written as

$$
\begin{aligned}
\big(\widehat{\boldsymbol{M}}^{\tau,(m)} - \boldsymbol{M}^{\tau,(m)}\big)_{m,:}\overline{\boldsymbol{u}}^{\tau,(m)} &= \big(p^{-1}\big(\boldsymbol{T} - \boldsymbol{T}^{(m)}\big) \times_3 \boldsymbol{\theta}^{\tau,(m)}\big)_{m,:}\overline{\boldsymbol{u}}^{\tau,(m)} \\
&= \underbrace{\sum_{j,k\in[d]} T^\star_{m,j,k}\overline{u}^{\tau,(m)}_j \theta^{\tau,(m)}_k (p^{-1}\chi_{m,j,k} - 1)}_{=:\alpha_1} + \underbrace{\sum_{j,k\in[d]} \overline{u}^{\tau,(m)}_j \theta^{\tau,(m)}_k p^{-1} E_{m,j,k}\chi_{m,j,k}}_{=:\alpha_2} .
\end{aligned} \tag{223}
$$

  - For the first term α_1, it is easily seen from (200) and incoherence conditions that

$$
\begin{aligned}
L_1 := \max_{j,k\in[d]} \Big| T^\star_{m,j,k}\overline{u}^{\tau,(m)}_j \theta^{\tau,(m)}_k (p^{-1}\chi_{m,j,k} - 1)\Big| &\le p^{-1}\|\boldsymbol{T}^\star\|_\infty \big\|\overline{\boldsymbol{u}}^{\tau,(m)}\big\|_\infty \big\|\boldsymbol{\theta}^{\tau,(m)}\big\|_\infty \\
&\lesssim \frac{\mu r\lambda^\star_{\max}\sqrt{\log d}}{d^2 p}\big\|\overline{\boldsymbol{u}}^{\tau,(m)}\big\|_\infty ;
\end{aligned}
$$

  and

$$
\begin{aligned}
V_1 := \sum_{j,k\in[d]} T^{\star 2}_{m,j,k}\big(\overline{u}^{\tau,(m)}_j\big)^2\big(\theta^{\tau,(m)}_k\big)^2 \mathbb{E}\big[(p^{-1}\chi_{m,j,k} - 1)^2\big] &\le p^{-1}\|\boldsymbol{A}^\star\|^2_{2,\infty}\big\|\overline{\boldsymbol{u}}^{\tau,(m)}\big\|^2_\infty\big\|\boldsymbol{\theta}^{\tau,(m)}\big\|^2_\infty \\
&\lesssim \frac{\mu^2 r^2 \lambda^{\star 2}_{\max}\log d}{d^2 p}\big\|\overline{\boldsymbol{u}}^{\tau,(m)}\big\|^2_\infty .
\end{aligned}
$$

  Apply the Bernstein inequality to yield that with probability at least $1 - O\big(d^{-11}\big)$,

$$
\sum_{j,k\in[d]} T^\star_{m,j,k}\overline{u}^{\tau,(m)}_j \theta^{\tau,(m)}_k (p^{-1}\chi_{m,j,k} - 1) \lesssim \sqrt{V_1\log d} + L_1\log d \lesssim \frac{\mu r\lambda^\star_{\max}\log d}{d\sqrt{p}}\big\|\overline{\boldsymbol{u}}^{\tau,(m)}\big\|_\infty,
$$

  where the last inequality holds as long as $p \gg d^{-2}\log^2 d$.

  - Regarding α_2 (cf. (223)), it is straightforward to compute that

$$
\begin{aligned}
L_2 := \big\|p^{-1}\overline{u}^{\tau,(m)}_j \theta^{\tau,(m)}_k E_{m,j,k}\chi_{m,j,k}\big\|_{\psi_1} &\le \frac{\sigma}{p}\big\|\overline{\boldsymbol{u}}^{\tau,(m)}\big\|_\infty\big\|\boldsymbol{\theta}^{\tau,(m)}\big\|_\infty \\
&\lesssim \frac{\sigma}{p}\sqrt{\frac{\mu r\log d}{d}}\big\|\overline{\boldsymbol{u}}^{\tau,(m)}\big\|_\infty,
\end{aligned}
$$

  with $\|\cdot\|_{\psi_1}$ denoting the sub-exponential norm, and

$$
\begin{aligned}
V_2 := \mathbb{E}\Big[\big(\sum_{j,k\in[d]}\overline{u}^{\tau,(m)}_j \theta^{\tau,(m)}_k p^{-1}E_{m,j,k}\chi_{m,j,k}\big)^2\Big] &= \sum_{j,k\in[d]}\big(\overline{u}^{\tau,(m)}_j\big)^2\big(\theta^{\tau,(m)}_k\big)^2 \mathbb{E}\big[p^{-2}E^2_{m,j,k}\chi_{m,j,k}\big] \\
&\le \frac{\sigma^2}{p}\big\|\overline{\boldsymbol{u}}^{\tau,(m)}\big\|^2_2\big\|\boldsymbol{\theta}^{\tau,(m)}\big\|^2_2 \lesssim \frac{\sigma^2 r d\log d}{p}\big\|\overline{\boldsymbol{u}}^{\tau,(m)}\big\|^2_\infty .
\end{aligned}
$$

  Then the Bernstein inequality reveals that with probability at least $1 - O\big(d^{-11}\big)$,

$$
\Big|p^{-1}\sum_{j,k\in[d]}\overline{u}^{\tau,(m)}_j \theta^{\tau,(m)}_k E_{m,j,k}\chi_{m,j,k}\Big| \lesssim L_2\log^2 d + \sqrt{V_2\log d} \le \sigma\sqrt{\frac{r d\log^2 d}{p}}\big\|\overline{\boldsymbol{u}}^{\tau,(m)}\big\|_\infty,
$$

  where the last inequality follows from our sample size condition.

– Substituting these into (223), we arrive at

$$\left| \left(\widehat{\boldsymbol{M}}^{\tau,(m)} - \boldsymbol{M}^{\tau,(m)} \right)_{m,:} \overline{\boldsymbol{u}}^{\tau,(m)} \right| \lesssim \left\{ \frac{\mu r \lambda_{\max}^{\star} \log d}{d\sqrt{p}} + \sigma \sqrt{\frac{r d \log^2 d}{p}} \right\} \left\| \overline{\boldsymbol{u}}^{\tau,(m)} \right\|_{\infty}. \tag{224}$$

- For the remaining entries of $\left(\widehat{\boldsymbol{M}}^{\tau,(m)} - \boldsymbol{M}^{\tau,(m)} \right) \overline{\boldsymbol{u}}^{\tau,(m)}$, we have

$$\left(\widehat{\boldsymbol{M}}^{\tau,(m)} - \boldsymbol{M}^{\tau,(m)} \right)_{i,:} \overline{\boldsymbol{u}}^{\tau,(m)}$$
$$= \theta_m^{\tau,(m)} \Big(\sum_{j:j\neq m} \overline{u}_j^{\tau,(m)} \big(T_{i,j,m}^{\star}(p^{-1}\chi_{i,j,m} - 1) + p^{-1}E_{i,j,m}\chi_{i,j,m} \big) \Big)$$
$$+ \overline{v}_m^{\tau,(m)} \Big(\sum_{k \in [d]} \theta_k^{\tau,(m)} \big(T_{i,m,k}^{\star}(p^{-1}\chi_{i,m,k} - 1) + p^{-1}E_{i,m,k}\chi_{i,m,k} \big) \Big)$$

for any $i \neq m$. From Lemma D.9 and (200), we have, with probability at least $1 - O(d^{-11})$, that

$$\big(\theta_m^{\tau,(m)}\big)^2 \sum_{i:i\neq m} \Big(\sum_{j:j\neq m} T_{i,j,m}^{\star} \overline{u}_j^{\tau,(m)}(p^{-1}\chi_{i,j,m} - 1) \Big)^2 + \big(\overline{v}_m^{\tau,(m)}\big)^2 \sum_{i:i\neq m} \Big(\sum_{k \in [d]} T_{i,m,k}^{\star} \theta_k^{\tau,(m)}(p^{-1}\chi_{i,m,k} - 1) \Big)^2$$
$$\lesssim \frac{\mu r \lambda_{\max}^{\star 2}}{dp} \big\| \boldsymbol{\theta}^{\tau,(m)} \big\|_{\infty}^2 \big\| \overline{\boldsymbol{u}}^{\tau,(m)} \big\|_{\infty}^2$$
$$\lesssim \frac{\mu^2 r^2 \lambda_{\max}^{\star 2} \log d}{d^2 p} \big\| \overline{\boldsymbol{u}}^{\tau,(m)} \big\|_{\infty}^2.$$

Combined with (198) and (200), Lemma D.10 reveals that with probability at least $1 - O(d^{-11})$,

$$\big(\theta_m^{\tau,(m)}\big)^2 \sum_{i:i\neq m} \Big(p^{-1} \sum_{j:j\neq m} \overline{u}_j^{\tau,(m)} E_{i,j,m}\chi_{i,j,m} \Big)^2 + \big(\overline{v}_m^{\tau,(m)}\big)^2 \sum_{i:i\neq m} \Big(p^{-1} \sum_{k \in [d]} \theta_k^{\tau,(m)} E_{i,m,k}\chi_{i,m,k} \Big)^2$$
$$\lesssim \frac{\sigma^2 d}{p} \big\| \overline{\boldsymbol{u}}^{\tau,(m)} \big\|_{2}^2 \big\| \boldsymbol{\theta}^{\tau,(m)} \big\|_{\infty}^2 + \frac{\sigma^2 d}{p} \big\| \boldsymbol{\theta}^{\tau,(m)} \big\|_{2}^2 \big\| \overline{\boldsymbol{u}}^{\tau,(m)} \big\|_{\infty}^2 + \frac{\sigma^2 \log^5 d}{p^2} \big\| \overline{\boldsymbol{u}}^{\tau,(m)} \big\|_{\infty}^2 \big\| \boldsymbol{\theta}^{\tau,(m)} \big\|_{\infty}^2$$
$$\lesssim \frac{\sigma^2 r d \log d}{p} \big\| \overline{\boldsymbol{u}}^{\tau,(m)} \big\|_{\infty}^2,$$

which implies that

$$\sum_{i:i\neq m} \Big(\big(\widehat{\boldsymbol{M}}^{\tau,(m)} - \boldsymbol{M}^{\tau,(m)} \big)_{i,:} \overline{\boldsymbol{u}}^{\tau,(m)} \Big)^2 \lesssim \left\{ \frac{\mu^2 r^2 \lambda_{\max}^{\star 2} \log d}{d^2 p} + \frac{\sigma^2 r d \log d}{p} \right\} \big\| \overline{\boldsymbol{u}}^{\tau,(m)} \big\|_{\infty}^2. \tag{225}$$

- Therefore, combine (224) and (225) to obtain that

$$\big\| \big(\widehat{\boldsymbol{M}}^{\tau,(m)} - \boldsymbol{M}^{\tau,(m)} \big) \overline{\boldsymbol{u}}^{\tau,(m)} \big\|_{2} \lesssim \left\{ \frac{\mu r \lambda_{\max}^{\star} \log d}{d\sqrt{p}} + \sigma \sqrt{\frac{r d \log^2 d}{p}} \right\} \big\| \overline{\boldsymbol{u}}^{\tau,(m)} \big\|_{\infty}.$$

B.8 Proof of Lemma 5.16

By definition, the only possible difference between $\overline{\boldsymbol{u}}^{\tau}$ and $\boldsymbol{\nu}^{\tau}$ lies in how their global signs are chosen. To show that $\overline{\boldsymbol{u}}^{\tau} = \boldsymbol{\nu}^{\tau}$, we first claim for the moment that

$$\big| \big\langle p^{-1}\boldsymbol{T}, (\overline{\boldsymbol{u}}^{\tau})^{\otimes 3} \big\rangle - \lambda_1^{\star} \big| \lesssim \mathcal{E}_{\mathsf{proj}} \cdot \lambda_1^{\star}, \tag{226}$$

where $\mathcal{E}_{\mathsf{proj}}$ is defined in (79). Given that $\mathcal{E}_{\mathsf{proj}} \ll 1$ under our sample size, noise and rank condition, this immediately implies that $\big\langle p^{-1}\boldsymbol{T}, (\overline{\boldsymbol{u}}^{\tau})^{\otimes 3} \big\rangle > 0$. Consequently, by construction, the global signs of $\overline{\boldsymbol{u}}^{\tau}$ and $\boldsymbol{\nu}^{\tau}$ coincide. Moreover, from (84) and the condition that $\mathcal{E}_{\mathsf{loo}}\sqrt{\mu r \log d / d} \ll 1$, one also knows that $\big\langle p^{-1}\boldsymbol{T}^{(m)}, (\overline{\boldsymbol{u}}^{\tau,(m)})^{\otimes 3} \big\rangle > 0$ and hence the global signs of $\overline{\boldsymbol{u}}^{\tau,(m)}$ and $\boldsymbol{\nu}^{\tau,(m)}$ also coincide.

In addition, recall that $\lambda_\tau = \langle p^{-1}\boldsymbol{T}, (\boldsymbol{\nu}^\tau)^{\otimes 3}\rangle$. One thus has

$$\lambda_\tau = \langle p^{-1}\boldsymbol{T}, (\boldsymbol{\nu}^\tau)^{\otimes 3}\rangle = \langle p^{-1}\boldsymbol{T}, (\overline{\boldsymbol{u}}^\tau)^{\otimes 3}\rangle, \tag{227}$$

which taken collectively with (226) justifies (85).

The rest of the proof then comes down to establishing the claim (226). Towards this, we first decompose

$$\langle p^{-1}\boldsymbol{T}, (\overline{\boldsymbol{u}}^\tau)^{\otimes 3}\rangle - \lambda_1^\star = \underbrace{\langle p^{-1}\boldsymbol{T}, (\overline{\boldsymbol{u}}^\tau)^{\otimes 3}\rangle - \langle \boldsymbol{T}^\star, (\overline{\boldsymbol{u}}^\tau)^{\otimes 3}\rangle}_{=:\,\beta_1}$$
$$+ \underbrace{\langle \boldsymbol{T}^\star, (\overline{\boldsymbol{u}}^\tau)^{\otimes 3}\rangle - \langle \boldsymbol{T}^\star, (\overline{\boldsymbol{u}}_1^\star)^{\otimes 3}\rangle}_{=:\,\beta_2} + \underbrace{\langle \boldsymbol{T}^\star, (\overline{\boldsymbol{u}}_1^\star)^{\otimes 3}\rangle - \lambda_1^\star}_{=:\,\beta_3}. \tag{228}$$

In what follows, we shall upper bound these three terms separately.

B.8.1 Controlling β_1

Let us start with β_1, For simplicity of notation, let us define $\boldsymbol{\Delta}_1 := \overline{\boldsymbol{u}}^\tau - \overline{\boldsymbol{u}}_1^\star$. By construction, $\boldsymbol{T}$ and $\boldsymbol{T}^\star$ are symmetric. We then can expand

$$\langle p^{-1}\boldsymbol{T} - \boldsymbol{T}^\star, (\overline{\boldsymbol{u}}^\tau)^{\otimes 3}\rangle = \langle p^{-1}\boldsymbol{T} - \boldsymbol{T}^\star, (\overline{\boldsymbol{u}}_1^\star + \boldsymbol{\Delta}_1)^{\otimes 3}\rangle$$
$$= \langle p^{-1}\boldsymbol{T} - \boldsymbol{T}^\star, (\overline{\boldsymbol{u}}_1^\star)^{\otimes 3}\rangle + 3\langle p^{-1}\boldsymbol{T} - \boldsymbol{T}^\star, \boldsymbol{\Delta}_1 \otimes (\overline{\boldsymbol{u}}_1^\star)^{\otimes 2}\rangle + 3\langle p^{-1}\boldsymbol{T} - \boldsymbol{T}^\star, \boldsymbol{\Delta}_1^{\otimes 2} \otimes \overline{\boldsymbol{u}}_1^\star\rangle$$
$$+ \langle p^{-1}\boldsymbol{T} - \boldsymbol{T}^\star, \boldsymbol{\Delta}_1^{\otimes 3}\rangle. \tag{229}$$

We first look at the first term of (229) which only consists of $\overline{\boldsymbol{u}}_1^\star$. As shown in (173), with probability at least $1 - O\left(d^{-11}\right)$, one has

$$\left\|\left(p^{-1}\boldsymbol{T} - \boldsymbol{T}^\star\right) \times_3 \overline{\boldsymbol{u}}_1^\star\right\| \lesssim \frac{\mu\sqrt{r}\,\lambda_{\max}^\star\sqrt{\log d}}{d\sqrt{p}} + \sigma\sqrt{\frac{d\log d}{p}} \leq \mathcal{E}_{\mathsf{proj}}\lambda_{\min}^\star.$$

It follows that

$$\left|\langle p^{-1}\boldsymbol{T} - \boldsymbol{T}^\star, (\overline{\boldsymbol{u}}_1^\star)^{\otimes 3}\rangle\right| \leq \left\|\left(p^{-1}\boldsymbol{T} - \boldsymbol{T}^\star\right) \times_3 \overline{\boldsymbol{u}}_1^\star\right\| \|\overline{\boldsymbol{u}}_1^\star\|_2^2 \lesssim \mathcal{E}_{\mathsf{proj}}\lambda_{\min}^\star,$$

where we recall the definition of $\mathcal{E}_{\mathsf{proj}}$ in (79). As for the term linear in $\boldsymbol{\Delta}_1$, by Lemma 5.13, we know that $\|\boldsymbol{\Delta}_1\|_2 \lesssim \mathcal{E}_{\mathsf{proj}}$. As a result, one has

$$\left|\langle p^{-1}\boldsymbol{T} - \boldsymbol{T}^\star, \boldsymbol{\Delta}_1 \otimes (\overline{\boldsymbol{u}}_1^\star)^{\otimes 2}\rangle\right| \leq \left\|\left(p^{-1}\boldsymbol{T} - \boldsymbol{T}^\star\right) \times_3 \overline{\boldsymbol{u}}_1^\star\right\| \|\overline{\boldsymbol{u}}_1^\star\|_2 \|\boldsymbol{\Delta}_1\|_2 \lesssim \mathcal{E}_{\mathsf{proj}}^2\lambda_{\min}^\star.$$

We then turn to the quadratic terms in $\boldsymbol{\Delta}_1$. Similar to the above arguments, one can deduce that

$$\left|\langle p^{-1}\boldsymbol{T} - \boldsymbol{T}^\star, \boldsymbol{\Delta}_1^{\otimes 2} \otimes \overline{\boldsymbol{u}}_1^\star\rangle\right| \leq \left\|\left(p^{-1}\boldsymbol{T} - \boldsymbol{T}^\star\right) \times_3 \overline{\boldsymbol{u}}_1^\star\right\| \|\boldsymbol{\Delta}_1\|_2^2 \lesssim \mathcal{E}_{\mathsf{proj}}^3\lambda_{\min}^\star.$$

Finally, we can simply upper bound the last term in (173) by

$$\left|\langle p^{-1}\boldsymbol{T} - \boldsymbol{T}^\star, \boldsymbol{\Delta}_1^{\otimes 3}\rangle\right| \leq \left\|p^{-1}\boldsymbol{T} - \boldsymbol{T}^\star\right\| \|\boldsymbol{\Delta}_1\|_2^3 \lesssim \mathcal{E}_{\mathsf{op}}\mathcal{E}_{\mathsf{proj}}^3\lambda_{\min}^\star \ll \mathcal{E}_{\mathsf{proj}}^3\lambda_{\min}^\star,$$

where the last step is due to the fact that $\mathcal{E}_{\mathsf{op}} \ll 1$. By our sample size, noise and rank conditions, one has $\mathcal{E}_{\mathsf{proj}} \ll 1$. Putting these bounds together reveals that

$$\left|\langle p^{-1}\boldsymbol{T} - \boldsymbol{T}^\star, (\overline{\boldsymbol{u}}^\tau)^{\otimes 3}\rangle\right| \lesssim \mathcal{E}_{\mathsf{proj}}\lambda_{\min}^\star. \tag{230}$$

B.8.2 Controlling β_2

Recall the definition of β_2 in (228) and $\boldsymbol{\Delta}_1 = \overline{\boldsymbol{u}}^\tau - \overline{\boldsymbol{u}}_1^\star$. We can further decompose

$$\left\langle \boldsymbol{T}^\star, (\overline{\boldsymbol{u}}^\tau)^{\otimes 3} - (\overline{\boldsymbol{u}}_1^\star)^{\otimes 3} \right\rangle = 3\left\langle \boldsymbol{T}^\star, (\overline{\boldsymbol{u}}_1^\star)^{\otimes 2} \otimes \boldsymbol{\Delta}_1 \right\rangle + 3\left\langle \boldsymbol{T}^\star, \boldsymbol{\Delta}_1^{\otimes 2} \otimes \overline{\boldsymbol{u}}_1^\star \right\rangle + \left\langle \boldsymbol{T}^\star, \boldsymbol{\Delta}_1^{\otimes 3} \right\rangle.$$

We first consider the first term which is linear in $\boldsymbol{\Delta}_1$. Since $\boldsymbol{T}^\star$ is a symmetric tensor, we have

$$\boldsymbol{T}^\star \times_1 \overline{\boldsymbol{u}}_1^\star = \boldsymbol{T}^\star \times_2 \overline{\boldsymbol{u}}_1^\star = \boldsymbol{T}^\star \times_3 \overline{\boldsymbol{u}}_1^\star = \sum_{s\in[r]} \lambda_1^\star \langle \overline{\boldsymbol{u}}_s^\star, \overline{\boldsymbol{u}}_1^\star \rangle \, \overline{\boldsymbol{u}}_s^\star \overline{\boldsymbol{u}}_s^{\star\top}.$$

By Lemma D.1, one has

$$\left\| \boldsymbol{T}^\star \times_1 \overline{\boldsymbol{u}}_1^\star \right\| \le \max_{1\le s\le r} \left| \lambda_1^\star \langle \overline{\boldsymbol{u}}_s^\star, \overline{\boldsymbol{u}}_1^\star \rangle \right| \left\| \overline{\boldsymbol{U}}^\star \right\|^2 \le (\lambda_1^\star + \lambda_{\max}^\star \sqrt{\mu/d}) \lesssim \lambda_1^\star,$$

which arises from $\max_{s\ne i} \left| \langle \overline{\boldsymbol{u}}_s^\star, \overline{\boldsymbol{u}}_1^\star \rangle \right| \le \sqrt{\mu/d}$ and $\left\| \overline{\boldsymbol{U}}^\star \right\| \lesssim 1$ as long as $r\sqrt{\mu/d} \ll 1$. As a result, one has

$$\left| \left\langle \boldsymbol{T}^\star, (\overline{\boldsymbol{u}}_1^\star)^{\otimes 2} \otimes \boldsymbol{\Delta}_1 \right\rangle \right| \le \left\| \boldsymbol{T}^\star \times_1 \overline{\boldsymbol{u}}_1^\star \right\| \left\| \overline{\boldsymbol{u}}_1^\star \right\|_2 \left\| \boldsymbol{\Delta}_1 \right\|_2 \lesssim \lambda_1^\star \left\| \boldsymbol{\Delta}_1 \right\|_2 \lesssim \mathcal{E}_{\mathsf{proj}} \lambda_1^\star.$$

In a similar manner, we also know that

$$\left| \left\langle \boldsymbol{T}^\star, \boldsymbol{\Delta}_1^{\otimes 2} \otimes \overline{\boldsymbol{u}}_1^\star \right\rangle \right| \le \left\| \boldsymbol{T}^\star \times_1 \overline{\boldsymbol{u}}_1^\star \right\| \left\| \boldsymbol{\Delta}_1 \right\|_2^2 \lesssim \mathcal{E}_{\mathsf{proj}}^2 \lambda_1^\star.$$

Finally, using the fact that the tensor spectral norm is always less than or equal to that of its matricization, we find that

$$\left| \left\langle \boldsymbol{T}^\star, \boldsymbol{\Delta}_1^{\otimes 3} \right\rangle \right| \le \left\| \boldsymbol{T}^\star \right\| \left\| \boldsymbol{\Delta}_1 \right\|_2^3 \le \left\| \boldsymbol{A}^\star \right\| \left\| \boldsymbol{\Delta}_1 \right\|_2^3 \lesssim \mathcal{E}_{\mathsf{proj}}^3 \lambda_1^\star.$$

Combining this with the fact that $\mathcal{E}_{\mathsf{proj}} \ll 1$, we conclude that

$$\left| \left\langle \boldsymbol{T}^\star, (\overline{\boldsymbol{u}}^\tau)^{\otimes 3} - (\overline{\boldsymbol{u}}_1^\star)^{\otimes 3} \right\rangle \right| \lesssim \mathcal{E}_{\mathsf{proj}} \lambda_1^\star. \tag{231}$$

B.8.3 Controlling β_3

It remains to control β_3. Straightforward calculation reveals that

$$\left\langle \boldsymbol{T}^\star, (\overline{\boldsymbol{u}}_1^\star)^{\otimes 3} \right\rangle = \sum_{s\in[r]} \lambda_s^\star \langle (\overline{\boldsymbol{u}}_s^\star)^{\otimes 3}, (\overline{\boldsymbol{u}}_1^\star)^{\otimes 3} \rangle = \lambda_1^\star \left\| \overline{\boldsymbol{u}}_1^\star \right\|_2^6 + \sum_{s:s\ne 1} \lambda_s^\star \langle \overline{\boldsymbol{u}}_s^\star, \overline{\boldsymbol{u}}_1^\star \rangle^6.$$

By the incoherence conditions, we can upper bound

$$\left| \left\langle \boldsymbol{T}^\star, (\overline{\boldsymbol{u}}_1^\star)^{\otimes 3} \right\rangle - \lambda_1^\star \right| = \sum_{s:s\ne 1} \lambda_1^\star \langle \overline{\boldsymbol{u}}_s^\star, \overline{\boldsymbol{u}}_1^\star \rangle^6 \le r \max_{s:s\ne 1} \left| \langle \overline{\boldsymbol{u}}_s^\star, \overline{\boldsymbol{u}}_1^\star \rangle \right|^6 \lambda_{\max}^\star \le \frac{\mu^3 r \lambda_{\max}^\star}{d^3}. \tag{232}$$

B.8.4 Combining β_1, β_2 and β_3

Putting (230), (231) and (232) together, we find that

$$\left| \left\langle p^{-1} \boldsymbol{T}, (\overline{\boldsymbol{u}}^\tau)^{\otimes 3} \right\rangle - \lambda_1^\star \right| \lesssim \mathcal{E}_{\mathsf{proj}} \lambda_1^\star + \frac{\mu r \lambda_{\max}^\star}{d^3} \le \left(\mathcal{E}_{\mathsf{proj}} + \frac{\mu^3 r}{d^3} \right) \lambda_1^\star \asymp \mathcal{E}_{\mathsf{proj}} \lambda_1^\star,$$

where the last step follows from the condition $\mu \le d$, $r \ll \sqrt{d/\mu}$ and the definition $\mathcal{E}_{\mathsf{proj}} \ge \sqrt{\mu r/d}$ (cf. (79)).

B.9   Proof of Lemma 5.17

We first show that for each $i \in [r]$, $\left(\boldsymbol{w}_i, \lambda_i\right)$ and $\left(\boldsymbol{w}_i^{(m)}, \lambda_i^{(m)}\right)$ (returned by PRUNE() in Algorithm 3 and Algorithm 6 respectively) satisfy (80), (82) and (84); in other words, we want to show that they correspond to the same index $\tau \in [L]$ (and are hence produced using the same Gaussian random vector $\boldsymbol{g}^\tau$). The proof idea is this: given that the proposed algorithms select the pair with the largest spectral gap in each round of PRUNE(), it suffices to ensure that there is sufficient separation between the largest and the second largest spectral gaps (so that both algorithms can identify the same τ).

By Lemma 5.12 and union bounds, we know that with probability at least $1 - \delta$, for each $i \in [r]$,

$$\Delta_i^{(1)} - \Delta_i^{(2)} \gtrsim \lambda_{\min}^\star / \left(r \sqrt{\log d} \right), \tag{233}$$

where we recall that $\Delta_i^\tau := \gamma_i^{\star\tau} - \max_{j:j\neq i}\left|\gamma_j^{\star\tau}\right|$, and $\Delta_i^{(1)} \geq \Delta_i^{(2)} \geq \cdots \geq \Delta_i^{(L)}$ denote the order statistics of $\left\{\Delta_i^\tau\right\}_{\tau=1}^L$ in descending order. As shown in the proof of Lemma 5.13, the spectral gap of $\boldsymbol{M}^\tau$ is well approximated by $\max_i \Delta_i^\tau$, namely,

$$\left| \max_{1 \leq i \leq r} \Delta_i^\tau - \left(\sigma_1\left(\boldsymbol{M}^\tau\right) - \sigma_2\left(\boldsymbol{M}^\tau\right) \right) \right| \lesssim \left\| \boldsymbol{M}^\tau - \boldsymbol{M}^{\tau\star} \right\| \ll \frac{\lambda_{\min}^\star}{r\sqrt{\log d}}$$

under our sample size, noise and rank conditions (67). Moreover, from Lemma B.2, we see that $\boldsymbol{M}^\tau$ and $\boldsymbol{M}^{\tau,(m)}$ are extremely close in terms of the spectral norm, i.e.

$$\left\| \boldsymbol{M}^\tau - \boldsymbol{M}^{\tau,(m)} \right\| \ll \frac{\lambda_{\min}^\star}{r\sqrt{\log d}}.$$

This implies that the perturbation incurred by the leave-out-one procedure is relatively small compared to the difference between the largest and the second largest spectral gaps of $\boldsymbol{M}^\tau$. Consequently, the leave-one-out estimates $\left\{\left(\boldsymbol{w}_i^{(m)}, \lambda_i^{(m)}\right)\right\}_{i=1}^r$ returned by Algorithm 6 and the true estimates $\left\{\left(\boldsymbol{w}_i, \lambda_i\right)\right\}_{i=1}^r$ should correspond to the same trials and should be generated by the same set of Gaussian random vectors. As a result, they obey (69a), (69b) and (69c) for all $1 \leq m \leq d$.

From the discussion above, we also know that as long as $\sigma_1\left(\boldsymbol{M}^\tau\right) - \sigma_2\left(\boldsymbol{M}^\tau\right) \gtrsim \lambda_{\min}^\star$, one has $\left\| \boldsymbol{\nu}^\tau - \overline{\boldsymbol{u}}_i^\star \right\|_2 \lesssim \mathcal{E}_{\mathsf{proj}}$ for some $i \in [r]$. This is an immediate consequence of Lemma 5.13 and the fact that the spectral gap of $\boldsymbol{M}^\tau$ and $\max_i \Delta_i^\tau$ are extremely close.

It remains to show that our pruning procedure can return estimates of tensor factors without duplicates. Suppose that there exist $1 \leq \tau_1 \neq \tau_2 \leq L$ such that $\left\| \boldsymbol{\nu}^{\tau_1} - \overline{\boldsymbol{u}}_i^\star \right\|_2 \lesssim \mathcal{E}_{\mathsf{proj}}$ and $\left\| \boldsymbol{\nu}^{\tau_2} - \overline{\boldsymbol{u}}_i^\star \right\|_2 \lesssim \mathcal{E}_{\mathsf{proj}}$ for some $i \in [r]$. By the triangle inequality, one has

$$\begin{aligned}
\left| \langle \boldsymbol{\nu}^{\tau_1}, \boldsymbol{\nu}^{\tau_2} \rangle \right| &= \left| \left\| \overline{\boldsymbol{u}}_i^\star \right\|_2^2 + \langle \boldsymbol{\nu}^{\tau_1} - \overline{\boldsymbol{u}}_i^\star, \overline{\boldsymbol{u}}_i^\star \rangle + \langle \overline{\boldsymbol{u}}_i^\star, \boldsymbol{\nu}^{\tau_2} - \overline{\boldsymbol{u}}_i^\star \rangle + \langle \boldsymbol{\nu}^{\tau_1} - \overline{\boldsymbol{u}}_i^\star, \boldsymbol{\nu}^{\tau_2} - \overline{\boldsymbol{u}}_i^\star \rangle \right| \\
&\geq 1 - \left\| \boldsymbol{\nu}^{\tau_1} - \overline{\boldsymbol{u}}_i^\star \right\|_2 - \left\| \boldsymbol{\nu}^{\tau_2} - \overline{\boldsymbol{u}}_i^\star \right\|_2 - \left\| \boldsymbol{\nu}^{\tau_1} - \overline{\boldsymbol{u}}_i^\star \right\|_2 \left\| \boldsymbol{\nu}^{\tau_2} - \overline{\boldsymbol{u}}_i^\star \right\|_2 \\
&\geq 1 - 2\mathcal{E}_{\mathsf{proj}} - \mathcal{E}_{\mathsf{proj}}^2 \geq 1 - 3\mathcal{E}_{\mathsf{proj}},
\end{aligned}$$

provided that $\mathcal{E}_{\mathsf{proj}} \ll 1$. In addition, for any $j \neq i, j \in [r]$, we know that exists some $1 \leq \tau_3 \leq L$ such that

$$\left\| \boldsymbol{\nu}^{\tau_3} - \overline{\boldsymbol{u}}_j^\star \right\|_2 \lesssim \mathcal{E}_{\mathsf{proj}}.$$

Recall our incoherence condition in (8c). It is easy to see that

$$\begin{aligned}
\left| \langle \boldsymbol{\nu}^{\tau_1}, \boldsymbol{\nu}^{\tau_3} \rangle \right| &= \left| \langle \overline{\boldsymbol{u}}_i^\star, \overline{\boldsymbol{u}}_j^\star \rangle + \langle \boldsymbol{\nu}^{\tau_1} - \overline{\boldsymbol{u}}_i^\star, \overline{\boldsymbol{u}}_i^\star \rangle + \langle \overline{\boldsymbol{u}}_i^\star, \boldsymbol{\nu}^{\tau_3} - \overline{\boldsymbol{u}}_j^\star \rangle + \langle \boldsymbol{\nu}^{\tau_1} - \overline{\boldsymbol{u}}_i^\star, \boldsymbol{\nu}^{\tau_3} - \overline{\boldsymbol{u}}_j^\star \rangle \right| \\
&\leq \left| \langle \overline{\boldsymbol{u}}_i^\star, \overline{\boldsymbol{u}}_j^\star \rangle \right| + \left\| \boldsymbol{\nu}^{\tau_1} - \overline{\boldsymbol{u}}_i^\star \right\|_2 + \left\| \boldsymbol{\nu}^{\tau_3} - \overline{\boldsymbol{u}}_j^\star \right\|_2 + \left\| \boldsymbol{\nu}^{\tau_1} - \overline{\boldsymbol{u}}_i^\star \right\|_2 \left\| \boldsymbol{\nu}^{\tau_3} - \overline{\boldsymbol{u}}_j^\star \right\|_2 \\
&\leq \sqrt{\mu/d} + 2\mathcal{E}_{\mathsf{proj}} + \mathcal{E}_{\mathsf{proj}}^2 \leq \sqrt{\mu/d} + 3\mathcal{E}_{\mathsf{proj}} \ll 1 - 3\mathcal{E}_{\mathsf{proj}},
\end{aligned}$$

with the proviso that $\mu \ll d$ and $\mathcal{E}_{\mathsf{proj}} \ll 1$.

The above argument reveals a clear separation between $\left| \langle \boldsymbol{\nu}^{\tau_1}, \boldsymbol{\nu}^{\tau_3} \rangle \right|$ and $\left| \langle \boldsymbol{\nu}^{\tau_1}, \boldsymbol{\nu}^{\tau_2} \rangle \right|$. As an immediate consequence, the proposed pruning procedure successfully removes all duplication while securing an estimate for each tensor factor.

B.10 Proof of Corollary 5.11

Fix any arbitrary small constant $\delta > 0$. From Theorems 5.9-5.10 and the assumptions of Theorem 2.8, one knows that with probability exceeding $1 - \delta$, there exists a permutation $\pi(\cdot) : [d] \mapsto [d]$ such that for all $1 \leq i \leq r$,

$$\left\| \boldsymbol{w}^i - \overline{\boldsymbol{u}}^\star_{\pi(i)} \right\|_2 \leq \delta, \qquad \left\| \boldsymbol{w}^i - \overline{\boldsymbol{u}}^\star_{\pi(i)} \right\|_\infty \leq \delta\sqrt{\frac{1}{d}}, \qquad \left| \lambda_i - \lambda^\star_{\pi(i)} \right| \leq \delta\lambda^\star_{\max},$$

$$\left\| \boldsymbol{w}^i - \boldsymbol{w}^{i,(m)} \right\|_2 \leq \delta\sqrt{\frac{1}{d}}, \qquad \left| \lambda_i - \lambda^{(m)}_i \right| \leq \delta\sqrt{\frac{1}{d}}\lambda^\star_{\max}, \qquad \left| \left(\boldsymbol{w}^i - \overline{\boldsymbol{u}}^\star_{\pi(i)} \right)_m \right|_2 \leq \delta\sqrt{\frac{1}{d}}$$

for some $0 < \delta \ll 1/(\mu^{3/2}r) < 1$. To prove the corollary, we shall just combine the above results.

Without loss of generality, assume that $\pi(i) = i$ for each $i \in [r]$. Given that $\delta \ll 1$ and $\kappa \asymp 1$, by the triangle inequality, one has $\lambda_i \asymp \lambda^\star_i$ for all $i \in [r]$, which further implies that

$$\left| \lambda^{1/3}_i - \lambda^{\star 1/3}_i \right| \lesssim \frac{|\lambda_i - \lambda^\star_i|}{\lambda^{\star 2/3}_i} \lesssim \delta\lambda^{\star 1/3}_{\max}.$$

Consequently, we can apply the triangle inequality to demonstrate that: for each $1 \leq i \leq r$,

$$\left\| \lambda^{1/3}_i \boldsymbol{w}^i - \boldsymbol{u}^\star_i \right\|_2 \leq \left| \lambda^{1/3}_i - \lambda^{\star 1/3}_i \right| \left\| \overline{\boldsymbol{u}}^\star_i \right\|_2 + \lambda^{\star 1/3}_i \left\| \boldsymbol{w}^i - \overline{\boldsymbol{u}}^\star_i \right\|_2 \lesssim \delta\lambda^{\star 1/3}_{\max}.$$

Arguing similarly, we also see that

$$\left\| \lambda^{1/3}_i \boldsymbol{w}^i - \boldsymbol{u}^\star_i \right\|_\infty \lesssim \delta\sqrt{\frac{1}{d}}\lambda^{\star 1/3}_{\max},$$

$$\left\| \lambda^{1/3}_i \boldsymbol{w}^i - \left(\lambda^{(m)}_i \right)^{1/3} \boldsymbol{w}^{i,(m)} \right\|_2 \lesssim \delta\sqrt{\frac{1}{d}}\lambda^{\star 1/3}_{\max},$$

$$\left| \left(\left(\lambda^{(m)}_i \right)^{1/3} \boldsymbol{w}^{i,(m)} - \boldsymbol{u}^\star_i \right)_m \right| \lesssim \delta\sqrt{\frac{1}{d}}\lambda^{\star 1/3}_{\max}$$

hold for all $i \in [r]$ and $m \in [d]$. Recall that $\boldsymbol{U}^0 = \left[\lambda^{1/3}_i \boldsymbol{w}^i \right]_{1 \leq i \leq r}$. One can deduce that

$$\left\| \boldsymbol{U}^0 - \boldsymbol{U}^\star \right\|_{\mathrm{F}} \lesssim \delta\sqrt{r}\lambda^{\star 1/3}_{\max} \lesssim \delta \left\| \boldsymbol{U}^\star \right\|_{\mathrm{F}},$$

$$\left\| \boldsymbol{U}^0 - \boldsymbol{U}^\star \right\|_{2,\infty} \lesssim \delta\sqrt{\frac{r}{d}}\lambda^{\star 1/3}_{\max} \lesssim \delta \left\| \boldsymbol{U}^\star \right\|_{2,\infty},$$

$$\left\| \boldsymbol{U}^0 - \boldsymbol{U}^{0,(m)} \right\|_{2,\infty} \lesssim \delta\sqrt{\frac{r}{d}}\lambda^{\star 1/3}_{\max} \lesssim \delta \left\| \boldsymbol{U}^\star \right\|_{2,\infty},$$

$$\left\| \left(\boldsymbol{U}^{0,(m)} - \boldsymbol{U}^\star \right)_{m,:} \right\|_2 \lesssim \delta\sqrt{\frac{r}{d}}\lambda^{\star 1/3}_{\max} \lesssim \delta \left\| \boldsymbol{U}^\star \right\|_{2,\infty},$$

where we have used the condition that $\kappa \asymp 1$ and the fact that $\left\| \boldsymbol{U}^\star \right\|_{\mathrm{F}} \geq \sqrt{r}\lambda^{\star 1/3}_{\min}$ and $\left\| \boldsymbol{U}^\star \right\|_{2,\infty} \geq \left\| \boldsymbol{U}^\star \right\|_{\mathrm{F}}/\sqrt{d}$.

C Proof of Corollary 2.9

This section establishes Corollary 2.9. First of all, it is easy to see that: given the estimation accuracy established in Theorem 2.8, the permutation matrices that best match $\boldsymbol{U}^t$ to $\boldsymbol{U}^\star$ remain unchanged as t increases. Therefore, we assume without loss of generality that $\boldsymbol{I}_r = \arg\min_{\boldsymbol{\Pi} \in \mathsf{perm}_r} \left\| \boldsymbol{U}^t \boldsymbol{\Pi} - \boldsymbol{U}^\star \right\|$ for all $t \geq 0$.

Suppose that $r\sqrt{\mu/d} \ll 1$. We claim for the moment that: if a matrix $\boldsymbol{U} = [\boldsymbol{u}_1, \cdots, \boldsymbol{u}_r] \in \mathbb{R}^{d \times r}$ satisfies

$$\left\| \boldsymbol{U} - \boldsymbol{U}^\star \right\|_{\mathrm{F}} \leq \delta \left\| \boldsymbol{U}^\star \right\|_{\mathrm{F}} \qquad \text{and} \qquad \left\| \boldsymbol{U} - \boldsymbol{U}^\star \right\|_{2,\infty} \leq \delta \left\| \boldsymbol{U}^\star \right\|_{2,\infty}$$

for any $0 \leq \delta \ll 1/(\mu^{3/2}r) \leq 1$, then one has

$$\|\boldsymbol{T} - \boldsymbol{T}^\star\|_{\mathrm{F}} \lesssim \delta \|\boldsymbol{T}^\star\|_{\mathrm{F}} \qquad \text{and} \qquad \|\boldsymbol{T} - \boldsymbol{T}^\star\|_\infty \lesssim \delta\sqrt{\mu^3 r}\, \|\boldsymbol{T}^\star\|_\infty \,, \tag{234}$$

where $\boldsymbol{T} := \sum_{i=1}^r \boldsymbol{u}_i \otimes \boldsymbol{u}_i \otimes \boldsymbol{u}_i$. As already shown in the analysis of Theorem 2.8, one has

$$\|\boldsymbol{U}^t - \boldsymbol{U}^\star\|_{\mathrm{F}} \lesssim \left(C_1 \frac{\rho^{t+1}}{\mu^{3/2}r} + C_2 \frac{\sigma}{\lambda^\star_{\min}}\sqrt{\frac{d\log d}{p}} \right) \|\boldsymbol{U}^\star\|_{\mathrm{F}} \,,$$

$$\|\boldsymbol{U}^t - \boldsymbol{U}^\star\|_{2,\infty} \lesssim \left(C_3\rho^{t+1} \frac{\rho^{t+1}}{\mu^{3/2}r} + C_4 \frac{\sigma}{\lambda^\star_{\min}}\sqrt{\frac{d\log d}{p}} \right) \|\boldsymbol{U}^\star\|_{2,\infty} \,,$$

from which Corollary 2.9 follows immediately.

It remains to prove the claim (234). For notational convenience, let us define $\boldsymbol{\Delta} := \boldsymbol{U} - \boldsymbol{U}^\star$ and $\boldsymbol{\Delta}_s := \boldsymbol{u}_s - \boldsymbol{u}_s^\star$ for each $1 \leq s \leq r$. Then we can expand

$$\boldsymbol{T} - \boldsymbol{T}^\star = \sum_{1 \leq s \leq r} \boldsymbol{u}_s^{\otimes 3} - \boldsymbol{u}_s^{\star \otimes 3} = \sum_{1 \leq s \leq r} \boldsymbol{u}_s^{\star \otimes 2} \otimes \boldsymbol{\Delta}_s + \sum_{1 \leq s \leq r} \boldsymbol{u}_s^\star \otimes \boldsymbol{\Delta}_s \otimes \boldsymbol{u}_s^\star + \sum_{1 \leq s \leq r} \boldsymbol{\Delta}_s \otimes \boldsymbol{u}_s^{\star \otimes 2}$$

$$+ \sum_{1 \leq s \leq r} \boldsymbol{u}_s^\star \otimes \boldsymbol{\Delta}_s^{\otimes 2} + \sum_{1 \leq s \leq r} \boldsymbol{\Delta}_s \otimes \boldsymbol{u}_s^\star \otimes \boldsymbol{\Delta}_s + \sum_{1 \leq s \leq r} \boldsymbol{\Delta}_s^{\otimes 2} \otimes \boldsymbol{u}_s^\star + \sum_{1 \leq s \leq r} \boldsymbol{\Delta}_s^{\otimes 3}. \tag{235}$$

(1) Euclidean loss. We first look at the loss measured by $\|\cdot\|_{\mathrm{F}}$. In view of the symmetric structure of tensors, it suffices to control $\sum_{1 \leq s \leq r} \boldsymbol{u}_s^{\star \otimes 2} \otimes \boldsymbol{\Delta}_s$, $\sum_{1 \leq s \leq r} \boldsymbol{u}_s^\star \otimes \boldsymbol{\Delta}_s^{\otimes 2}$ and $\sum_{1 \leq s \leq r} \boldsymbol{\Delta}_s^{\otimes 3}$.

Let us define $\boldsymbol{W}_1 := [\boldsymbol{u}_s^\star \otimes \boldsymbol{\Delta}_s]_{1 \leq s \leq r} \in \mathbb{R}^{d^2 \times r}$ and $\boldsymbol{W}_2 := \left[\boldsymbol{\Delta}_s^{\otimes 2}\right]_{1 \leq s \leq r} \in \mathbb{R}^{d^2 \times r}$. Recalling the fact that $\|\boldsymbol{U}^\star\| \leq \|\overline{\boldsymbol{U}}^\star\|\lambda_{\max}^{\star 1/3} \lesssim \lambda_{\max}^{\star 1/3}$ (established in Lemma D.1), we have

$$\left\|\sum\nolimits_{1 \leq s \leq r} \boldsymbol{u}_s^{\star \otimes 2} \otimes \boldsymbol{\Delta}_s\right\|_{\mathrm{F}} = \left\|\boldsymbol{U}^\star \boldsymbol{W}_1^\top\right\|_{\mathrm{F}} \leq \|\boldsymbol{U}^\star\| \|\boldsymbol{W}_1\|_{\mathrm{F}} \lesssim \lambda_{\max}^{\star 1/3} \|\boldsymbol{W}_1\|_{\mathrm{F}} \,,$$

$$\left\|\sum\nolimits_{1 \leq s \leq r} \boldsymbol{u}_s^\star \otimes \boldsymbol{\Delta}_s^{\otimes 2}\right\|_{\mathrm{F}} = \left\|\boldsymbol{U}^\star \boldsymbol{W}_2^\top\right\|_{\mathrm{F}} \leq \|\boldsymbol{U}^\star\| \|\boldsymbol{W}_2\|_{\mathrm{F}} \lesssim \lambda_{\max}^{\star 1/3} \|\boldsymbol{W}_2\|_{\mathrm{F}} \,,$$

$$\left\|\sum\nolimits_{1 \leq s \leq r} \boldsymbol{\Delta}_s^{\otimes 3}\right\|_{\mathrm{F}} = \left\|\boldsymbol{\Delta} \boldsymbol{W}_2^\top\right\|_{\mathrm{F}} \leq \|\boldsymbol{\Delta}\| \|\boldsymbol{W}_2\|_{\mathrm{F}} \leq \|\boldsymbol{\Delta}\|_{\mathrm{F}} \|\boldsymbol{W}_2\|_{\mathrm{F}} \,.$$

It then suffices to control $\|\boldsymbol{W}_1\|_{\mathrm{F}}$ and $\|\boldsymbol{W}_2\|_{\mathrm{F}}$. If $\|\boldsymbol{\Delta}\|_{\mathrm{F}} \leq \delta \|\boldsymbol{U}^\star\|_{\mathrm{F}} \leq \delta\sqrt{r}\,\lambda_{\max}^{\star 1/3}$, then it is easy to bound

$$\|\boldsymbol{W}_1\|_{\mathrm{F}}^2 = \sum_{1 \leq s \leq r} \|\boldsymbol{u}_s^\star \otimes \boldsymbol{\Delta}_s\|_2^2 = \sum_{1 \leq s \leq r} \|\boldsymbol{u}_s^\star\|_2^2 \|\boldsymbol{\Delta}_s\|_2^2 \leq \max_{1 \leq s \leq r} \|\boldsymbol{u}_s^\star\|_2^2 \|\boldsymbol{\Delta}\|_{\mathrm{F}}^2 \leq \delta^2 r \lambda_{\max}^{\star 4/3} \,,$$

$$\|\boldsymbol{W}_2\|_{\mathrm{F}}^2 = \sum_{1 \leq s \leq r} \|\boldsymbol{\Delta}_s^{\otimes 2}\|_2^2 = \sum_{1 \leq s \leq r} \|\boldsymbol{\Delta}_s\|_2^4 \leq \max_{1 \leq s \leq r} \|\boldsymbol{\Delta}_s\|_2^2 \|\boldsymbol{\Delta}\|_{\mathrm{F}}^2 \leq \|\boldsymbol{\Delta}\|_{\mathrm{F}}^4 \leq \delta^4 r^2 \lambda_{\max}^{\star 4/3} \,.$$

Therefore, one has

$$\left\|\sum\nolimits_{1 \leq s \leq r} \boldsymbol{u}_s^{\star \otimes 2} \otimes \boldsymbol{\Delta}_s\right\|_{\mathrm{F}} \lesssim \delta\sqrt{r}\lambda_{\max}^\star, \tag{236}$$

$$\left\|\sum\nolimits_{1 \leq s \leq r} \boldsymbol{u}_s^\star \otimes \boldsymbol{\Delta}_s^{\otimes 2}\right\|_{\mathrm{F}} \lesssim \delta^2 r\lambda_{\max}^\star, \tag{237}$$

$$\left\|\sum\nolimits_{1 \leq s \leq r} \boldsymbol{\Delta}_s^{\otimes 3}\right\|_{\mathrm{F}} \lesssim \delta^3 r^{3/2}\lambda_{\max}^\star. \tag{238}$$

Since $0 \leq \delta \ll r^{-1} \leq 1$, combining (236), (237) and (238) with the fact that $\|\boldsymbol{T}^\star\|_{\mathrm{F}} \geq \sqrt{r}\lambda_{\min}^\star/2$ (established in Lemma D.1), we conclude that

$$\|\boldsymbol{T} - \boldsymbol{T}^\star\| \lesssim \delta\sqrt{r}\lambda_{\max}^\star \lesssim \delta \|\boldsymbol{T}^\star\|_{\mathrm{F}} \,. \tag{239}$$

(2) ℓ_∞ loss. Next, we turn to the $\|\cdot\|_\infty$ loss. Again, it suffices to focus on $\sum_{1 \le s \le r} \boldsymbol{u}_s^{\star \otimes 2} \otimes \boldsymbol{\Delta}_s$, $\sum_{1 \le s \le r} \boldsymbol{u}_s^\star \otimes \boldsymbol{\Delta}_s^{\otimes 2}$ and $\sum_{1 \le s \le r} \boldsymbol{\Delta}_s^{\otimes 3}$. From (104), (105) and (106) shown in the proof of Lemma 5.1, one has

$$\Big\| \sum\nolimits_{1 \le s \le r} (\boldsymbol{u}_s^\star)^{\otimes 2} \otimes \boldsymbol{\Delta}_s \Big\|_\infty \le \delta \max_{1 \le s \le r} \|\boldsymbol{u}_s^\star\|_\infty \|\boldsymbol{U}^\star\|_{2,\infty}^2 \le \delta \frac{\mu^{3/2} r \lambda_{\max}^\star}{d^{3/2}}, \tag{240}$$

$$\Big\| \sum\nolimits_{1 \le s \le r} \boldsymbol{u}_s^\star \otimes \boldsymbol{\Delta}_s^{\otimes 2} \Big\|_\infty \le \delta^2 \max_{1 \le s \le r} \|\boldsymbol{u}_s^\star\|_\infty \|\boldsymbol{U}^\star\|_{2,\infty}^2 \le \frac{\delta^2 \mu^{3/2} r \lambda_{\max}^\star}{d^{3/2}}, \tag{241}$$

$$\Big\| \sum\nolimits_{1 \le s \le r} \boldsymbol{\Delta}_s^{\otimes 3} \Big\|_\infty \le \delta^3 \|\boldsymbol{U}^\star\|_{2,\infty}^3 \le \frac{\delta^3 \mu^{3/2} r^{3/2} \lambda_{\max}^\star}{d^{3/2}}. \tag{242}$$

Putting (240), (241) and (242) together with the condition that $0 < \delta \ll r^{-1} \le 1$, we arrive at

$$\|\boldsymbol{T} - \boldsymbol{T}^\star\|_\infty \lesssim \frac{\delta \mu^{3/2} r \lambda_{\max}^\star}{d^{3/2}}.$$

In addition, from the lower bound on $\|\boldsymbol{T}^\star\|_{\mathrm{F}}$, one has

$$\|\boldsymbol{T}^\star\|_\infty \ge \frac{1}{d^{3/2}} \|\boldsymbol{T}^\star\|_{\mathrm{F}} \gtrsim \sqrt{\frac{r}{d^3}} \lambda_{\min}^\star,$$

which allows us to conclude that

$$\|\boldsymbol{T} - \boldsymbol{T}^\star\|_\infty \lesssim \sqrt{\mu^3 r} \, \delta \, \|\boldsymbol{T}^\star\|_\infty.$$

D   Auxiliary lemmas

This section gathers several auxiliary lemmas that prove useful when establishing our main results.

D.1   Statements of auxiliary lemmas

We begin by stating all auxiliary lemmas formally, with the proofs postponed to subsequent subsections. We shall define

$$\overline{\boldsymbol{U}}^\star := [\overline{\boldsymbol{u}}_1^\star, \cdots, \overline{\boldsymbol{u}}_r^\star], \qquad \text{with} \qquad \overline{\boldsymbol{u}}_i^\star := \boldsymbol{u}_i^\star / \|\boldsymbol{u}_i^\star\|_2. \tag{243}$$

Lemma D.1. *Suppose that Assumption 2.1 holds, and assume that $r\sqrt{\mu/d} \le c_1$ for some sufficiently small universal constant $c_2 > 0$. Then for d sufficiently large, the matrices $\boldsymbol{A}^\star$, $\boldsymbol{B}^\star$ and $\boldsymbol{U}_{\mathsf{orth}}^\star$ (defined respectively in (24), (63) and (54)) obey*

$$\frac{1}{2}\lambda_{\min}^\star \le \|\boldsymbol{A}^\star\|_{\mathrm{F}} \le 2\lambda_{\max}^\star, \quad \|\boldsymbol{A}^\star\|_\infty \le \frac{\sqrt{2\mu r}\lambda_{\max}^\star}{d^{3/2}}, \quad \|\boldsymbol{A}^\star\|_{2,\infty} \le \sqrt{\frac{2\mu r}{d}} \lambda_{\max}^\star, \quad \|\boldsymbol{A}^{\star\top}\|_{2,\infty} \le \frac{\mu\sqrt{2r}\,\lambda_{\max}^\star}{d},$$

$$\|\boldsymbol{A}^\star\| = \lambda_{\max}^\star \left(1 + O\left(r\sqrt{\tfrac{\mu}{d}}\right)\right), \qquad \lambda_i(\boldsymbol{B}^\star) = \lambda_{(i)}^{\star 2} \left(1 + O\left(r\sqrt{\tfrac{\mu}{d}}\right)\right), \quad i \in [r],$$

$$\|\boldsymbol{B}^\star\|_{2,\infty} \le 2\sqrt{\frac{\mu r}{d}} \lambda_{\max}^{\star 2}, \qquad \|\boldsymbol{U}_{\mathsf{orth}}^\star\|_{2,\infty} \le \sqrt{\frac{2\mu r}{d}}, \qquad \|\overline{\boldsymbol{U}}^{\star\top}\overline{\boldsymbol{U}}^\star - \boldsymbol{I}\| \le r\sqrt{\frac{\mu}{d}}.$$

Here, $\|\boldsymbol{A}\|_{2,\infty} := \max_i \|\boldsymbol{A}_{i,:}\|_2$, $\lambda_{(i)}^\star$ stands for the i-th largest value in $\{\lambda_i^\star\}_{1 \le i \le r}$ (or equivalently $\{\|\boldsymbol{u}_i^\star\|_2^3\}_{1 \le i \le r}$), and $\lambda_i(\boldsymbol{B}^\star)$ represents the i-th largest eigenvalue of $\boldsymbol{B}^\star$.

Proof. See Appendix D.2. ∎

Lemma D.2. *Let $\boldsymbol{R} \in \mathbb{R}^{d \times d \times d}$ be a random order-3 tensor with independent entries $\{R_{i,j,k}\}_{i,j,k \in [d]}$ obeying*

$$\mathbb{E}\left[R_{i,j,k}\right] = 0, \quad \max_{i,j,k \in [d]} |R_{i,j,k}| \le B.$$

Define

$$\sigma_{\mathsf{mode}}^2 := \max_{j,k \in [d]} \sum_{i \in [d]} \mathbb{E}[R_{i,j,k}^2] + \max_{i,k \in [d]} \sum_{j \in [d]} \mathbb{E}[R_{i,j,k}^2] + \max_{i,j \in [d]} \sum_{k \in [d]} \mathbb{E}[R_{i,j,k}^2]. \tag{244}$$

Then with probability exceeding $1 - O\left(d^{-10}\right)$, *one has*

$$\|\boldsymbol{R}\| \lesssim B \log^3 d + \sigma_{\mathsf{mode}} \log^{5/2} d. \tag{245}$$

Proof. See Appendix D.3. ∎

An immediate consequence of this lemma is the following:

Corollary D.3. *With probability at least* $1 - O\left(d^{-10}\right)$, *one has*

$$\left\|p^{-1}\mathcal{P}_\Omega(\boldsymbol{T}^\star) - \boldsymbol{T}^\star\right\| \lesssim \frac{\sqrt{\mu r}\,\lambda_{\max}^\star \log^3 d}{d^{3/2}p} + \frac{\mu\sqrt{r}\,\lambda_{\max}^\star \log^{5/2} d}{d\sqrt{p}}; \tag{246}$$

$$\left\|\mathcal{P}_\Omega(\boldsymbol{E})\right\| \lesssim \sigma\left(\log^{7/2} d + \sqrt{dp}\,\log^{5/2} d\right). \tag{247}$$

Proof. See Appendix D.3. ∎

Lemma D.4. *Suppose that* $p \gtrsim d^{-2}\log^3 d$ *and that* $\mu \log^2 d \lesssim d$. *Then for any fixed vector* $\boldsymbol{w} \in \mathbb{R}^d$, *with probability* $1 - O\left(d^{-10}\right)$, *one has*

$$\left\|(p^{-1}\boldsymbol{T} - \boldsymbol{T}^\star) \times_3 \boldsymbol{w}\right\| \lesssim \|\boldsymbol{w}\|_\infty \sqrt{\frac{\mu r \log d}{dp}}\,\lambda_{\max}^\star + \|\boldsymbol{w}\|_\infty \frac{\sigma \log^{5/2} d}{p} + \|\boldsymbol{w}\|_2\,\sigma \sqrt{\frac{d \log d}{p}},$$

where $\times_3$ *is defined in Section 2.4. The results also holds if we replace* $\times_3$ *with* $\times_1$ *or* $\times_2$.

Proof. See Appendix D.4. ∎

Lemma D.5. *Let* $\{X_{i,j}\}_{1 \leq i \leq r, 1 \leq j \leq L}$ *be a sequence of i.i.d. standard Gaussian random variables, where* $r \geq 2$ *and* $L \geq 1$. *Consider some quantities* $\kappa \geq 1, \Delta > 0, 0 < \delta < 1/2$. *There exists some universal constant* $C > 0$ *such that if*

$$L \geq Cr^{2\kappa^2}\left(\kappa\sqrt{\log r} + \Delta\right)\exp(\Delta^2)\log \frac{1}{\delta},$$

then with probability at least $1 - \delta$, *there exists some* $1 \leq j_0 \leq L$ *such that*

$$X_{1,j_0} > \kappa \max_{i:1<i\leq r} |X_{i,j_0}| + \Delta.$$

In addition, define $\Delta_j := X_{1,j} - \kappa \max_{i:1<i\leq r} |X_{i,j}|$ *for each* $1 \leq j \leq L$. *Then with probability at least* $1 - 2\delta$,

$$\Delta_{(1)} - \Delta_{(2)} \gtrsim \frac{\delta}{\sqrt{\log L} + \sqrt{\log(1/\delta)}}.$$

where $\Delta_{(1)} \geq \Delta_{(2)} \geq \cdots \geq \Delta_{(L)}$ *denote the order statistics of* $\{\Delta_j\}_{j=1}^L$ *in descending order.*

Proof. See Appendix D.5. ∎

Lemma D.6. *Let* $\boldsymbol{U}$ *(resp.* $\boldsymbol{V}$*) be a* $d \times r$ *matrix with orthonormal columns. Suppose that* $\|\boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{V}\boldsymbol{V}^\top\| \leq \delta$. *Then for any unit vector* $\boldsymbol{u}_0 \in \mathbb{R}^d$ *lying in* $\mathsf{span}(\boldsymbol{U})$, *we have*

$$\left\|\mathcal{P}_{\boldsymbol{V}}(\boldsymbol{u}_0)\right\|_2 \geq \sqrt{1-\delta^2} \qquad and \qquad \left\|\mathcal{P}_{\boldsymbol{V}^\perp}(\boldsymbol{u}_0)\right\|_2 \leq \delta, \tag{248}$$

where we denote by $\mathcal{P}_{\boldsymbol{V}}(\boldsymbol{u}_0) := \boldsymbol{V}\boldsymbol{V}^\top\boldsymbol{u}_0$ *and* $\mathcal{P}_{\boldsymbol{V}^\perp}(\boldsymbol{u}_0) = \left(\boldsymbol{I}_d - \boldsymbol{V}\boldsymbol{V}^\top\right)\boldsymbol{u}_0$.

Proof. See Appendix D.6. ∎

Additionally, we record several facts concerning the set of Bernoulli random variables $\{\chi_{i,j,k}\}_{1 \leq i,j,k \leq d}$. We recall that

$$\chi_{i,j,k} := \mathbb{1}\{(i,j,k) \in \Omega\}, \tag{249}$$

which is a Bernoulli random variable with mean p.

Lemma D.7. *Suppose that $p \gtrsim d^{-2} \log d$. With probability exceeding $1 - O\left(d^{-10}\right)$, one has*

$$\sum_{i,j \in [d]} T_{i,j,k}^{\star 2} (p^{-1} \chi_{i,j,k} - 1)^2 \lesssim \frac{\mu r \lambda_{\max}^{\star 2}}{dp}, \qquad 1 \leq k \leq d. \tag{250}$$

Proof. See Appendix D.7. ∎

Lemma D.8. *Suppose that $p \gtrsim d^{-2} \log^2 d$. With probability exceeding $1 - O\left(d^{-10}\right)$, one has*

$$\sum_{i,j \in [d]} \left(p^{-1} E_{i,j,k} \chi_{i,j,k}\right)^2 \lesssim \sigma^2 d^2 / p, \qquad 1 \leq k \leq d. \tag{251}$$

Proof. See Appendix D.8. ∎

Lemma D.9. *Suppose $p \gtrsim d^{-2} \log d$ and $\mu \log^2 d \lesssim d$. Consider any fixed vector $\boldsymbol{w} \in \mathbb{R}^d$ and any $1 \leq i \leq d$. With probability exceeding $1 - O(d^{-10})$, one has*

$$\sum_{j \in [d]} \left(\sum_{k \in [d]} T_{i,j,k}^{\star} w_k (p^{-1} \chi_{i,j,k} - 1) \right)^2 \lesssim \frac{\mu r \lambda_{\max}^{\star 2}}{dp} \|\boldsymbol{w}\|_\infty^2. \tag{252}$$

Proof. See Appendix D.9. ∎

Lemma D.10. *Consider any fixed vector $\boldsymbol{w} \in \mathbb{R}^d$. With probability $1 - O(d^{-10})$ one has*

$$\sum_{j \in [d]} \left(\sum_{k \in [d]} w_k E_{i,j,k} \chi_{i,j,k} \right)^2 \lesssim \sigma^2 dp \|\boldsymbol{w}\|_2^2 + \sigma^2 \|\boldsymbol{w}\|_\infty^2 \log^5 d, \qquad 1 \leq i \leq d. \tag{253}$$

Proof. See Appendix D.10. ∎

D.2 Proof of Lemma D.1

1. To begin with, the incoherence condition (8b) gives

$$
\begin{aligned}
\|\boldsymbol{T}^\star\|_{\mathrm{F}}^2 &= \left\langle \sum_{i \in [r]} \boldsymbol{u}_i^{\star \otimes 3}, \sum_{i \in [r]} \boldsymbol{u}_i^{\star \otimes 3} \right\rangle \\
&= \sum_{1 \leq i \leq r} \left\| \boldsymbol{u}_i^{\star \otimes 3} \right\|_{\mathrm{F}}^2 + \sum_{1 \leq i \neq j \leq r} \left\langle \boldsymbol{u}_i^{\star \otimes 3}, \boldsymbol{u}_j^{\star \otimes 3} \right\rangle \\
&\overset{\text{(i)}}{=} \sum_{1 \leq i \leq r} \|\boldsymbol{u}_i^\star\|_2^6 + \sum_{1 \leq i \neq j \leq r} \left\langle \boldsymbol{u}_i^\star, \boldsymbol{u}_j^\star \right\rangle^3 \\
&\leq r \max_{1 \leq i \leq r} \|\boldsymbol{u}_i^\star\|_2^6 + r^2 \max_{1 \leq i \neq j \leq r} \left| \left\langle \boldsymbol{u}_i^\star, \boldsymbol{u}_j^\star \right\rangle^3 \right| \\
&\leq r \lambda_{\max}^{\star 2} + r^2 \left(\frac{\mu}{d} \right)^{3/2} \lambda_{\max}^{\star 2} \\
&\overset{\text{(ii)}}{\leq} 2r \lambda_{\max}^{\star 2},
\end{aligned}
$$

where we use the fact that $\left\langle \boldsymbol{u}^{\otimes 3}, \boldsymbol{v}^{\otimes 3} \right\rangle = \left\langle \boldsymbol{u}, \boldsymbol{v} \right\rangle^3$ in (i), and (ii) arises due to the condition that $3r \leq \sqrt{d/\mu} \leq (d/\mu)^{3/2}$. Using a similar argument, we also know that $\|\boldsymbol{T}^\star\|_{\mathrm{F}}^2 \geq r \lambda_{\min}^{\star 2}/2$. This combined with the incoherence condition in (8a) yields

$$\|\boldsymbol{A}^\star\|_\infty = \|\boldsymbol{T}^\star\|_\infty \leq \sqrt{\frac{\mu}{d^3}} \|\boldsymbol{T}^\star\|_{\mathrm{F}} \leq \frac{\sqrt{2\mu r}}{d^{3/2}} \lambda_{\max}^\star.$$

69

2. For any $1 \leq i \leq d$, the ℓ_2 norm of the i-th row of $\boldsymbol{A}^\star$ can be bounded by

$$
\begin{aligned}
\left\|\boldsymbol{A}_{i,:}^\star\right\|_2^2 &= \left\|\sum_{1 \leq s \leq r} \left(\boldsymbol{u}_s^\star\right)_i \left(\boldsymbol{u}_s^\star \otimes \boldsymbol{u}_s^\star\right)^\top\right\|_2^2 \\
&= \sum_{1 \leq s \leq r} \left(\boldsymbol{u}_s^\star\right)_i^2 \|\boldsymbol{u}_s^\star \otimes \boldsymbol{u}_s^\star\|_2^2 + \sum_{1 \leq s_1 \neq s_2 \leq r} \left(\boldsymbol{u}_{s_1}^\star\right)_i \left(\boldsymbol{u}_{s_2}^\star\right)_i \langle \boldsymbol{u}_{s_1}^\star \otimes \boldsymbol{u}_{s_1}^\star, \boldsymbol{u}_{s_2}^\star \otimes \boldsymbol{u}_{s_2}^\star \rangle \\
&\leq r \max_{s \in [r]} \|\boldsymbol{u}_s^\star\|_\infty^2 \max_{s \in [r]} \|\boldsymbol{u}_s^\star\|_2^4 + r^2 \max_{1 \leq s \leq r} \|\boldsymbol{u}_s^\star\|_\infty^2 \max_{s_1 \neq s_2} \langle \boldsymbol{u}_{s_1}^\star, \boldsymbol{u}_{s_2}^\star \rangle^2 \\
&\leq \lambda_{\max}^{\star 2} \left(\frac{\mu r}{d} + \frac{\mu^2 r^2}{d^2}\right) \leq \frac{2\mu r \lambda_{\max}^{\star 2}}{d}, \tag{254}
\end{aligned}
$$

where the first inequality follows from the fact that $\langle \boldsymbol{u} \otimes \boldsymbol{u}, \boldsymbol{v} \otimes \boldsymbol{v} \rangle = \langle \boldsymbol{u}, \boldsymbol{v} \rangle^2$, the second inequality holds true due to (8b) and (8c), and the last inequality holds as long as $r \leq d/\mu$. This immediately yields the advertised bound on $\|\boldsymbol{A}^\star\|_{2,\infty}$.

3. For any $\underline{j} \in [d]^2$ (which corresponds to (j_1, j_2)), the ℓ_2 norm of the $\underline{j}$-th column of $\boldsymbol{A}^\star$ can be upper bounded similarly by

$$
\begin{aligned}
\left\|\boldsymbol{A}_{:,\underline{j}}^\star\right\|_2^2 &= \left\|\sum_{1 \leq s \leq r} \left(\boldsymbol{u}_s^\star \otimes \boldsymbol{u}_s^\star\right)_{\underline{j}} \boldsymbol{u}_s^\star\right\|_2^2 \\
&= \sum_{1 \leq s \leq r} \left(\boldsymbol{u}_s^\star\right)_{j_1}^2 \left(\boldsymbol{u}_s^\star\right)_{j_2}^2 \|\boldsymbol{u}_s^\star\|_2^2 + \sum_{s_1 \neq s_2} \left(\boldsymbol{u}_{s_1}^\star\right)_{j_1} \left(\boldsymbol{u}_{s_1}^\star\right)_{j_2} \left(\boldsymbol{u}_{s_2}^\star\right)_{j_1} \left(\boldsymbol{u}_{s_2}^\star\right)_{j_2} \langle \boldsymbol{u}_{s_1}^\star, \boldsymbol{u}_{s_2}^\star \rangle \\
&\leq r \max_{1 \leq s \leq r} \|\boldsymbol{u}_s^\star\|_\infty^4 \max_{s \in [r]} \|\boldsymbol{u}_s^\star\|_2^2 + r^2 \max_{1 \leq s \leq r} \|\boldsymbol{u}_s^\star\|_\infty^4 \max_{1 \leq s_1 \neq s_2 \leq r} \left|\langle \boldsymbol{u}_{s_1}^\star, \boldsymbol{u}_{s_2}^\star \rangle\right| \\
&\leq \lambda_{\max}^{\star 2} \left(\frac{\mu^2 r}{d^2} + \frac{\mu^{5/2} r^2}{d^{5/2}}\right) \leq \frac{2\mu^2 r \lambda_{\max}^{\star 2}}{d^2}, \tag{255}
\end{aligned}
$$

where the second inequality is valid due to (8b) and (8c), and the last inequality holds as long as $r \leq \sqrt{d/\mu}$. This yields the claimed bound regarding $\|\boldsymbol{A}^{\star\top}\|_{2,\infty}$.

4. Regarding the spectrum of $\boldsymbol{A}^\star$, $\boldsymbol{B}^\star$, $\boldsymbol{U}_{\text{orth}}^\star$ and $\overline{\boldsymbol{U}}^\star$, we refer the reader to the proof of [CLC$^+$20, Corollary 1].

5. We now move on to $\|\boldsymbol{B}^\star\|_{2,\infty}$. For any $i \in [d]$, it is seen that

$$
\begin{aligned}
\left\|\boldsymbol{B}_{i,:}^\star\right\|_2^2 &= \sum_{j \in [d]} \left(\boldsymbol{A}_{i,:}^\star \boldsymbol{A}_{j,:}^{\star\top}\right)^2 = \boldsymbol{A}_{i,:}^\star \left(\sum_{j \in [d]} \boldsymbol{A}_{j,:}^{\star\top} \boldsymbol{A}_{j,:}^\star\right) \boldsymbol{A}_{i,:}^{\star\top} = \boldsymbol{A}_{i,:}^\star \left(\boldsymbol{A}^{\star\top} \boldsymbol{A}^\star\right) \boldsymbol{A}_{i,:}^{\star\top} \\
&\leq \left\|\boldsymbol{A}^{\star\top} \boldsymbol{A}^\star\right\| \left\|\boldsymbol{A}_{i,:}^\star\right\|_2^2 \leq 2\lambda_{\max}^{\star 2} \left\|\boldsymbol{A}_{i,:}^\star\right\|_2^2.
\end{aligned}
$$

Here, the last line makes use of the bound $\|\boldsymbol{A}^\star\| \leq \lambda_{\max}^\star(1 + O(r\sqrt{\mu/d})) \leq 2\lambda_{\max}^\star$, which holds if $r\sqrt{\mu/d} \leq c_1$ for some sufficiently small constant $c_1 > 0$. It then follows from (254) that

$$
\|\boldsymbol{B}^\star\|_{2,\infty} \leq \sqrt{2}\,\lambda_{\max}^\star \|\boldsymbol{A}^\star\|_{2,\infty} \leq 2\lambda_{\max}^{\star 2}\sqrt{\frac{\mu r}{d}}. \tag{256}
$$

D.3 Proof of Lemma D.2 and Corollary D.3

D.3.1 Proof of Lemma D.2

We start by making the following simple observation: the tensor spectral norm is a 1-Lipschitz function (w.r.t. the Frobenius norm) of the entries of the tensor. This follows since $\left|\|\boldsymbol{T}\| - \|\boldsymbol{R}\|\right| \leq \|\boldsymbol{T} - \boldsymbol{R}\| \leq \|\boldsymbol{T} - \boldsymbol{R}\|_{\text{F}}$ holds for any tensor $\boldsymbol{T}, \boldsymbol{R} \in \mathbb{R}^{d \times d \times d}$. This allows us to invoke standard concentration results regarding functions of independent random variables.

We shall first develop an upper bound on the mean $\mathbb{E}\big[\big\|\mathcal{P}_\Omega(\boldsymbol{R})\big\|\big]$. In view of [NDT15, Corollary 4] and Jensen's inequality, one has

$$
\mathbb{E}\big[\,\|\boldsymbol{R}\|\,\big] \le \sqrt{\mathbb{E}\big[\,\|\boldsymbol{R}\|^2\,\big]}
$$

$$
\lesssim \left(\mathbb{E}\left[\max_{j,k\in[d]}\sum_{i\in[d]} R_{i,j,k}^2 + \max_{i,j\in[d]}\sum_{k\in[d]} R_{i,j,k}^2 + \max_{i,k\in[d]}\sum_{j\in[d]} R_{i,j,k}^2\right]\right)^{1/2} \log^{5/2} d. \tag{257}
$$

We then need to bound the quantity presented in (257).

For some $\beta > 0$ to be specified later, one can upper bound

$$
\mathbb{E}\Big[\max_{j,k\in[d]}\sum_{i\in[d]} R_{i,j,k}^2\Big] = \int_0^\infty \mathbb{P}\Big\{\max_{j,k\in[d]}\sum_{i\in[d]} R_{i,j,k}^2 > t\Big\}\,\mathrm{d}t
$$

$$
\le \beta + \int_\beta^\infty \mathbb{P}\Big\{\max_{j,k\in[d]}\sum_{i\in[d]} R_{i,j,k}^2 > t\Big\}\,\mathrm{d}t
$$

$$
\le \beta + d^2 \int_\beta^\infty \mathbb{P}\Big\{\sum_{i\in[d]} R_{i,j,k}^2 > t\Big\}\,\mathrm{d}t. \tag{258}
$$

We shall resort to the Bernstein inequality to bound $\mathbb{P}\big\{\sum_{i\in[d]} R_{i,j,k}^2 > t\big\}$. It is straightforward to compute that

$$
M := \sum_{i\in[d]} \mathbb{E}\big[R_{i,j,k}^2\big] \le \sigma_{\mathsf{mode}}^2,
$$

$$
L := \max_{i\in[d]} \big|R_{i,j,k}^2\big| \le B^2,
$$

$$
S^2 := \sum_{i\in[d]} \mathbb{E}\big[R_{i,j,k}^4\big] \le B^2 \sigma_{\mathsf{mode}}^2.
$$

The Bernstein inequality then tells us that

$$
\mathbb{P}\Big\{\sum_{i\in[d]} R_{i,j,k}^2 - M > t\Big\} \le \exp\left(-\frac{3}{8}\min\left\{\frac{t^2}{S^2}, \frac{t}{L}\right\}\right), \quad t > 0. \tag{259}
$$

In particular, this implies that with probability exceeding $1 - O\big(d^{-20}\big)$,

$$
\sum_{i\in[d]} R_{i,j,k}^2 \lesssim M + L\log d + S\sqrt{\log d} \lesssim \sigma_{\mathsf{mode}}^2 + B^2\log d + \sqrt{B^2\sigma_{\mathsf{mode}}^2 \log d}
$$

$$
\asymp \sigma_{\mathsf{mode}}^2 + B^2\log d,
$$

where we have used the AM-GM inequality in the last step. Therefore, by taking

$$
\beta := C\,\big(\sigma_{\mathsf{mode}}^2 + B^2\log d\big)^{1/2}
$$

for some sufficiently large constant $C > 0$, we arrive at

$$
\beta \ge \frac{C}{3}\Big(M + L\log d + S\sqrt{\log d}\Big) \gg M + L\log d + S\sqrt{\log d}. \tag{260}
$$

Given that $\beta \gg M$, for any $t \ge \beta$ one has the following relations about several events

$$
\Big\{\sum_{i\in[d]} R_{i,j,k}^2 > t\Big\} = \Big\{\sum_{i\in[d]} R_{i,j,k}^2 - M > t - M\Big\} \subset \Big\{\sum_{i\in[d]} R_{i,j,k}^2 - M > t - \beta/2\Big\} \tag{261}
$$

In addition, it is easily seen that

$$
\min\big\{t^2/S^2, t/L\big\} \ge \frac{t}{\max\big\{S/\sqrt{\log d}, L\big\}} \tag{262}
$$

71

for any $t \geq \beta$ (with β obeying (260)). As a result, one can bound

$$
\begin{aligned}
\int_\beta^\infty \mathbb{P}\Big\{ \sum_{i\in[d]} R_{i,j,k}^2 > t \Big\} \mathrm{d}t &\overset{\text{(i)}}{\leq} \int_\beta^\infty \mathbb{P}\Big\{ \sum_{i\in[d]} R_{i,j,k}^2 - M > t - \beta/2 \Big\} \mathrm{d}t \\
&= \int_{\beta/2}^\infty \mathbb{P}\Big\{ \sum_{i\in[d]} R_{i,j,k}^2 - M > t \Big\} \mathrm{d}t \\
&\overset{\text{(ii)}}{\leq} \int_{\beta/2}^\infty \exp\left(-\frac{3}{8}\min\Big\{ \frac{t^2}{S^2}, \frac{t}{L} \Big\} \right) \mathrm{d}t \\
&\overset{\text{(iii)}}{\leq} \int_{\beta/2}^\infty \exp\left(-\frac{3}{8}\frac{t}{\max\{S/\sqrt{\log d},\, L\}} \right) \mathrm{d}t \\
&\lesssim \max\Big\{ S/\sqrt{\log d},\, L \Big\} \exp\left(-\frac{3}{16}\frac{\beta}{\max\{S/\sqrt{\log d},\, L\}} \right) \\
&\overset{\text{(iv)}}{\lesssim} \beta \exp\left(-\frac{1}{16}C\log d \right) \ll \beta/d^2,
\end{aligned}
$$

where (i) follows from (261), (ii) comes from (259), (iii) is a consequence of (262), and (iv) holds true when $C > 0$ is sufficiently large. Consequently,

$$
\mathbb{E}\Big[\max_{j,k\in[d]} \sum_{i\in[d]} R_{i,j,k}^2 \Big] \lesssim \beta \lesssim B\sqrt{\log d} + \sigma_{\mathsf{mode}}.
$$

Clearly, the same bound holds for $\mathbb{E}\big[\max_{i,k\in[d]} \sum_{j\in[d]} R_{i,j,k}^2 \big]$ and $\mathbb{E}\big[\max_{i,j\in[d]} \sum_{k\in[d]} R_{i,j,k}^2 \big]$.

Substitution into (257) yields

$$
\mathbb{E}\big[\|\boldsymbol{R}\| \big] \lesssim B\log^3 d + \sigma_{\mathsf{mode}}\log^{5/2} d. \tag{263}
$$

Recognizing that the magnitudes of all entries of $\boldsymbol{R}$ are bounded by B, we can invoke Talagrand's concentration inequality [Ver18, Theorem 5.2.16] for convex Lipschitz functions of independent bounded random variables to show that with probability $1 - O(d^{-10})$,

$$
\big| \|\boldsymbol{R}\| - \mathbb{E}\big[\|\boldsymbol{R}\| \big] \big| \lesssim B\sqrt{\log d}
$$

and, therefore,

$$
\|\boldsymbol{R}\| \lesssim B\log^3 d + \sigma_{\mathsf{mode}}\log^{5/2} d. \tag{264}
$$

D.3.2 Proof of Corollary D.3

Now we apply Lemma D.2 to our concrete setting. We first look at $p^{-1}\mathcal{P}_\Omega\left(\boldsymbol{T}^\star\right) - \boldsymbol{T}^\star$ and treat it as $\boldsymbol{R}$ in Lemma D.2. With the help of Lemma D.1, it is straightforward to compute that

$$
\max_{i,j,k\in[d]} \big| T_{i,j,k}^\star \left(p^{-1}\chi_{i,j,k} - 1 \right) \big| \lesssim \frac{1}{p}\|\boldsymbol{A}^\star\|_\infty \lesssim \frac{\sqrt{\mu r}\,\lambda_{\max}^\star}{d^{3/2}p},
$$

and

$$
\max_{i,j\in[d]} \sum_{k\in[d]} \mathbb{E}\big[T_{i,j,k}^{\star 2}(p^{-1}\chi_{i,j,k} - 1)^2 \big] \lesssim \frac{1}{p}\big\|\boldsymbol{A}^{\star\top}\big\|_{2,\infty}^2 \lesssim \frac{\mu^2 r\lambda_{\max}^{\star 2}}{d^2 p}.
$$

Clearly, $\max_{i,k\in[d]} \sum_{j\in[d]} \mathbb{E}\big[T_{i,j,k}^{\star 2}(p^{-1}\chi_{i,j,k} - 1)^2 \big]$ and $\max_{j,k\in[d]} \sum_{i\in[d]} \mathbb{E}\big[T_{i,j,k}^{\star 2}(p^{-1}\chi_{i,j,k} - 1)^2 \big]$ can be controlled in the same way. Substitution into (264) proves the claim (246).

We then turn to $\mathcal{P}_\Omega\left(\boldsymbol{E}\right)$. Recognizing that the entries of $\boldsymbol{E}$ might be unbounded, we invoke the following truncation trick to cope with this unboundedness issue. Specifically, define $\widetilde{\boldsymbol{E}} = [\widetilde{E}_{i,j,k}]_{1\leq i,j,k\leq d}$ where

$$
\widetilde{E}_{i,j,k} := E_{i,j,k}\mathbb{1}\big\{ |E_{i,j,k}| \leq c_1\sigma\sqrt{\log d}\big\}, \qquad 1 \leq i,j,k \leq d \tag{265}
$$

for some some sufficiently large constant $c_1 > 0$. Moreover, $\widetilde{E}_{i,j,k}$ is zero-mean because we assume that the distribution of $E_{i,j,k}$ is symmetric about 0. Standard concentration inequalities reveal that: with probability exceeding $1 - O(d^{-10})$, one has $\boldsymbol{E} = \widetilde{\boldsymbol{E}}$. Hence, it suffices to bound $\|\mathcal{P}_\Omega(\widetilde{\boldsymbol{E}})\|$. Towards this end, simple calculation reveals that

$$B = \max_{i,j,k \in [d]} |\widetilde{E}_{i,j,k} \chi_{i,j,k}| \lesssim \|\widetilde{\boldsymbol{E}}\|_\infty \lesssim \sigma \sqrt{\log d},$$

$$\sigma^2_{\mathsf{mode}} \leq \max_{i,j \in [d]} \sum_{k \in [d]} \mathbb{E}[E^2_{i,j,k} \chi_{i,j,k}] + \max_{i,k \in [d]} \sum_{j \in [d]} \mathbb{E}[E^2_{i,j,k} \chi_{i,j,k}] + \max_{j,k \in [d]} \sum_{i \in [d]} \mathbb{E}[E^2_{i,j,k} \chi_{i,j,k}] \lesssim p\sigma^2 d.$$

This together with (264) as well as the high-probability event $\boldsymbol{E} = \widetilde{\boldsymbol{E}}$ completes the proof.

D.4 Proof of Lemma D.4

For notational simplicity, let us denote

$$\boldsymbol{X} := (p^{-1}\boldsymbol{T} - \boldsymbol{T}^\star) \times_3 \boldsymbol{w}.$$

Observe that $\boldsymbol{X}$ is a zero-mean random matrix in $\mathbb{R}^{d \times d}$ with independent entries

$$X_{i,j} = \sum_{k \in [d]} w_k \left\{ T^\star_{i,j,k} (p^{-1}\chi_{i,j,k} - 1) + p^{-1}E_{i,j,k}\chi_{i,j,k} \right\}, \quad (i,j) \in [d]^2.$$

We shall apply the truncated matrix Bernstein inequality to control the spectral norm of $\boldsymbol{X}$.

- First, it is straightforward to bound

$$V := \max \left\{ \max_{i \in [d]} \sum_{j \in [d]} \mathbb{E}[X^2_{i,j}], \max_{j \in [d]} \sum_{i \in [d]} \mathbb{E}[X^2_{i,j}] \right\}$$

$$= \max \left\{ \max_{i \in [d]} \sum_{j,k \in [d]} p^{-1} w_k^2 (T^{\star 2}_{i,j,k} + \mathbb{E}[E^2_{i,j,k}]), \max_{j \in [d]} \sum_{i,k \in [d]} p^{-1} w_k^2 (T^{\star 2}_{i,j,k} + \mathbb{E}[E^2_{i,j,k}]) \right\}$$

$$\leq \frac{1}{p} \left(\|\boldsymbol{w}\|_\infty^2 \|\boldsymbol{A}^\star\|_{2,\infty}^2 + \|\boldsymbol{w}\|_2^2 \sigma^2 d \right).$$

- Second, using the same truncation argument as in the proof of Lemma D.2 in Appendix D.3, we can assume $|E_{i,j,k}| \lesssim \sigma\sqrt{\log d}$ for all $1 \leq i,j,k \leq d$ (which holds with very high probability). The Bernstein inequality reveals that

$$\mathbb{P}\{|X_{i,j}| > t\} \leq 2\exp\left(-\frac{3}{8}\min\left\{\frac{t^2}{S^2}, \frac{t}{L}\right\}\right), \quad t > 0$$

for each $(i,j) \in [d]^2$, where

$$L := \max_{k \in [d]} \left\{ |w_k| |T^\star_{i,j,k}(p^{-1}\chi_{i,j,k} - 1) + p^{-1}E_{i,j,k}\chi_{i,j,k}| \right\} \lesssim \frac{1}{p}\|\boldsymbol{w}\|_\infty \left(\|\boldsymbol{A}^\star\|_\infty + \sigma\sqrt{\log d}\right);$$

$$S^2 := \mathbb{E}[X^2_{i,j}] \asymp \sum_{k \in [d]} p^{-1} w_k^2 (T^{\star 2}_{i,j,k} + \mathbb{E}[E^2_{i,j,k}]) \leq \frac{1}{p}\left(\|\boldsymbol{w}\|_\infty^2 \|\boldsymbol{A}^{\star\top}\|_{2,\infty}^2 + \|\boldsymbol{w}\|_2^2 \sigma^2\right).$$

This implies that with probability exceeding $1 - O(d^{-20})$,

$$\max_{i,j \in [d]} |X_{i,j}| \lesssim L\log d + S\sqrt{\log d}$$

$$\lesssim \frac{\|\boldsymbol{w}\|_\infty \log d}{p}\left(\|\boldsymbol{A}^\star\|_\infty + \sigma\sqrt{\log d}\right) + \sqrt{\frac{\log d}{p}}\left(\|\boldsymbol{w}\|_\infty \|\boldsymbol{A}^{\star\top}\|_{2,\infty} + \|\boldsymbol{w}\|_2 \sigma\right).$$

Therefore, if we choose

$$\beta := C \left\{ \frac{\|\boldsymbol{w}\|_\infty \log d}{p} \left(\|\boldsymbol{A}^\star\|_\infty + \sigma \sqrt{\log d} \right) + \sqrt{\frac{\log d}{p}} \left(\|\boldsymbol{w}\|_\infty \left\|\boldsymbol{A}^{\star\top}\right\|_{2,\infty} + \|\boldsymbol{w}\|_2 \sigma \right) \right\}$$

for some sufficiently large constant $C > 0$, then one has

$$\beta \geq \frac{C}{2} \left(L \log d + S \sqrt{\log d} \right).$$

- Third, it is easy to bound

$$\mathbb{E}\big[\, |X_{i,j}|\, \mathbb{1}\left\{|X_{i,j}| \geq \beta\right\} \big] \leq \beta \cdot \mathbb{P}\left\{|X_{i,j}| \geq \beta\right\} + \int_\beta^\infty \mathbb{P}\left\{|X_{i,j}| \geq t\right\} \mathrm{d}t$$

$$\leq \beta \cdot O\left(d^{-20}\right) + \int_\beta^\infty \mathbb{P}\left\{|X_{i,j}| \geq t\right\} \mathrm{d}t.$$

In view of our choice of β, we know that $\min\left\{t^2/S^2, t/L\right\} \geq t/\max\left\{S/\sqrt{\log d}, L\right\}$ for any $t \geq \beta$. As a result, for d sufficiently large, we have

$$\int_\beta^\infty \mathbb{P}\left\{|X_{i,j}| \geq t\right\} \mathrm{d}t \leq 2 \int_\beta^\infty \exp\left(-\frac{3}{8} \min\left\{\frac{t^2}{S^2}, \frac{t}{L}\right\}\right) \mathrm{d}t$$

$$\leq 2 \int_\beta^\infty \exp\left(-\frac{3}{8} \frac{t}{\max\left\{S/\sqrt{\log d}, L\right\}}\right) \mathrm{d}t$$

$$\lesssim \max\left\{S/\sqrt{\log d}, L\right\} \exp\left(-\frac{3}{8} \frac{\beta}{\max\left\{S/\sqrt{\log d}, L\right\}}\right)$$

$$\lesssim \max\left\{S/\sqrt{\log d}, L\right\} \exp\left(-\frac{3}{8} C \log d\right) \ll \frac{\beta}{d^2}.$$

Consequently, we have established that

$$q := \sum_{i,j} \mathbb{E}\big[\, |X_{i,j}|\, \mathbb{1}\left\{|X_{i,j}| \geq \beta\right\} \big] \ll \beta.$$

Invoke the matrix Bernstein inequality to demonstrate that with probability $1 - O\left(d^{-10}\right)$,

$$\|\boldsymbol{X}\| \lesssim q + \beta \log d + \sqrt{V \log d} \asymp \beta \log d + \sqrt{V \log d}$$

$$\lesssim \frac{\|\boldsymbol{w}\|_\infty \log^2 d}{p} \left(\|\boldsymbol{A}^\star\|_\infty + \sigma \sqrt{\log d} \right) + \frac{\log^{3/2} d}{\sqrt{p}} \left(\|\boldsymbol{w}\|_\infty \left\|\boldsymbol{A}^{\star\top}\right\|_{2,\infty} + \sigma \|\boldsymbol{w}\|_2 \right)$$

$$+ \sqrt{\frac{\log d}{p}} \left(\|\boldsymbol{w}\|_\infty \|\boldsymbol{A}^\star\|_{2,\infty} + \|\boldsymbol{w}\|_2 \sigma \sqrt{d} \right)$$

$$\asymp \|\boldsymbol{w}\|_\infty \left\{ \frac{\|\boldsymbol{A}^\star\|_\infty \log^2 d}{p} + \frac{\left\|\boldsymbol{A}^{\star\top}\right\|_{2,\infty} \log^{3/2} d}{\sqrt{p}} + \frac{\|\boldsymbol{A}^\star\|_{2,\infty} \sqrt{\log d}}{\sqrt{p}} + \frac{\sigma \log^{5/2} d}{p} \right\} + \|\boldsymbol{w}\|_2 \sigma \sqrt{\frac{d \log d}{p}}$$

$$\overset{\mathrm{(i)}}{\lesssim} \|\boldsymbol{w}\|_\infty \left\{ \frac{\sqrt{\mu r} \log^2 d}{d^{3/2} p} + \sqrt{\frac{\mu^2 r \log^3 d}{d^2 p}} + \sqrt{\frac{\mu r \log d}{dp}} \right\} \lambda_{\max}^\star + \|\boldsymbol{w}\|_\infty \frac{\sigma \log^{5/2} d}{p} + \|\boldsymbol{w}\|_2 \sigma \sqrt{\frac{d \log d}{p}}$$

$$\overset{\mathrm{(ii)}}{\lesssim} \|\boldsymbol{w}\|_\infty \sqrt{\frac{\mu r \log d}{dp}} \lambda_{\max}^\star + \|\boldsymbol{w}\|_\infty \frac{\sigma \log^{5/2} d}{p} + \|\boldsymbol{w}\|_2 \sigma \sqrt{\frac{d \log d}{p}},$$

where (i) is due to Lemma D.1, and (ii) follows as long as $p \gtrsim d^{-2} \log^3 d$ and $\mu \log^2 d \lesssim d$.

74

D.5 Proof of Lemma D.5

Recall that for a standard Gaussian random variable $Z \sim \mathcal{N}(0,1)$, one has

$$\frac{1}{5t\sqrt{2\pi}} \exp\left(-t^2/2\right) \leq \left(\frac{1}{t} - \frac{1}{t^3}\right)\frac{1}{\sqrt{2\pi}} \exp\left(-t^2/2\right) \leq \mathbb{P}\{Z \geq t\} \leq \frac{1}{t}\frac{1}{\sqrt{2\pi}} \exp\left(-t^2/2\right) \tag{266}$$

for all $t > \sqrt{5/4}$. Observing that $\kappa\sqrt{2\log r} + \Delta \geq \sqrt{5/4}$ since $\kappa \geq 1$ and $r \geq 2$, we can invoke the above tail bound to deduce that

$$\mathbb{P}\left\{X_{1,j} \geq \kappa\sqrt{2\log r} + \Delta\right\} > \frac{1}{5\sqrt{2\pi}\left(\kappa\sqrt{2\log r} + \Delta\right)} \exp\left(-\left(\kappa\sqrt{2\log r} + \Delta\right)^2/2\right)$$

$$\geq \frac{1}{5\sqrt{2\pi}\left(\kappa\sqrt{2\log r} + \Delta\right) r^{2\kappa^2}\exp(\Delta^2)}, \tag{267}$$

where we use the elementary inequality $\left(\kappa\sqrt{2\log r} + \Delta\right)^2 \leq 4\kappa^2\log r + 2\Delta^2$. In addition, it follows from the union bound that

$$\mathbb{P}\left\{\max_{1<i\leq r}|X_{i,j}| < \sqrt{2\log r}\right\} \geq 1 - r\,\mathbb{P}\left\{|X_{i,j}| > \sqrt{2\log r}\right\} \geq 1 - r\left\{\frac{1}{2\sqrt{\pi\log r}}\exp(-\log r)\right\}$$

$$\geq 1 - \frac{1}{\sqrt{\pi\log r}} \geq 1 - \frac{1}{\sqrt{\pi\log 2}}. \tag{268}$$

To prove the claim, it is sufficient to choose L such that

$$\mathbb{P}\left\{\forall j : X_{1,j} < \kappa\sqrt{2\log r} + \Delta \text{ or } \max_{1<i\leq r}|X_{i,j}| \geq \sqrt{2\log r}\right\} \leq \delta,$$

or equivalently,

$$\left(1 - \mathbb{P}\{X_{1,j} \geq \kappa\sqrt{2\log r} + \Delta\}\mathbb{P}\{\max_{1<i\leq r}|X_{i,j}| < \sqrt{2\log r}\}\right)^L \leq \delta. \tag{269}$$

Note that $\log(1-x) \leq -1/(2x)$ for $0 < x < 1/4$. In view of (267) and (268), one can verify that the above inequality (269) as long as

$$L \geq C\left(\kappa\sqrt{\log r} + \Delta\right)r^{2\kappa^2}\exp(\Delta^2)\log\frac{1}{\delta}, \tag{270}$$

where $C > 0$ is some universal constant.

To prove the second claim, recall the definitions that

$$\Delta_j := X_{1,j} - \max_{1<i\leq r}\kappa X_{i,j}, \qquad 1 \leq j \leq L.$$

and that $\Delta_{(1)} \geq \Delta_{(2)} \geq \ldots, \geq \Delta_{(L)}$ denote $\{\Delta_j\}_{j=1}^L$ in descending order. For any $\epsilon > 0$, one has

$$\mathbb{P}\left\{\Delta_{(1)} - \Delta_{(2)} < \epsilon\right\} = \sum_{1\leq j\leq L}\mathbb{P}\left\{\Delta_j - \max_{k:k\neq j}\Delta_k < \epsilon \,\big|\, \Delta_j = \Delta_{(1)}\right\}\mathbb{P}\left\{\Delta_j = \Delta_{(1)}\right\}$$

$$= \mathbb{P}\left\{\Delta_1 - \max_{k:k\neq 1}\Delta_k > \epsilon \,\big|\, \Delta_1 = \Delta_{(1)}\right\},$$

where the last line holds because the distribution of $\Delta_j - \max_{k:k\neq j}\Delta_k$ conditional on $\Delta_j = \max_{1\leq k\leq r}\Delta_k$ is identical for all $1 \leq j \leq L$. In addition, it is straightforward to see that

$$\Delta_1 = \Delta_{(1)} \iff \Delta_1 \geq \max_{k:k\neq 1}\Delta_k$$

$$\iff X_{1,1} \geq \max_{1<i\leq r}\kappa X_{i,1} + \max_{k:k\neq 1}\Delta_k =: Y_{1,1}.$$

Hence, we have

$$\mathbb{P}\left\{\Delta_{(1)} - \Delta_{(2)} < \epsilon\right\} = \mathbb{P}\left\{\Delta_1 - \max_{k:k\neq 1}\Delta_k < \epsilon \,\middle|\, \Delta_1 = \Delta_{(1)}\right\} = \mathbb{P}\left\{X_{1,1} - Y_{1,1} < \epsilon \,\middle|\, X_{1,1} \geq Y_{1,1}\right\}.$$

Next, observe that $X_{1,1}$ is independent of $Y_{1,1}$, and hence we have

$$\mathbb{P}\left\{X_{1,1} - Y_{1,1} < \epsilon \,\middle|\, X_{1,1} \geq Y_{1,1}, Y_{1,1} = x\right\} = \mathbb{P}\left\{X_{1,1} < x + \epsilon \,\middle|\, X_{1,1} \geq y, Y_{1,1} = x\right\}$$
$$= \mathbb{P}\left\{X_{1,1} < x + \epsilon \,\middle|\, X_{1,1} \geq x\right\} = \frac{\mathbb{P}\left\{x \leq X_{1,1} < x + \epsilon\right\}}{\mathbb{P}\left\{X_{1,1} \geq x\right\}}$$

for any $x \geq 0$. In order to study this function, we define $f_\epsilon(x) := \frac{\mathbb{P}\{x \leq Z \leq x + \epsilon\}}{\mathbb{P}\{Z \geq x\}}$ with $Z \sim \mathcal{N}(0,1)$. Taking the derivative of $f_\epsilon(\cdot)$ w.r.t. x gives: for any $\epsilon > 0$,

$$f_\epsilon'(x) = \frac{\left\{\exp(-(x+\epsilon)^2/2) - \exp(-x^2/2)\right\}\mathbb{P}\left\{Z \geq x\right\} + \exp\left(-x^2/2\right)\mathbb{P}\left\{x \leq Z \leq x + \epsilon\right\}}{\sqrt{2\pi}\left(\mathbb{P}\left\{Z \geq x\right\}\right)^2}$$

$$= \frac{\exp\left(-(x+\epsilon)^2/2\right)\mathbb{P}\left\{Z \geq x\right\} - \exp(-\frac{x^2}{2})\mathbb{P}\left\{Z \geq x + \epsilon\right\}}{\sqrt{2\pi}\left(\mathbb{P}\left\{N \geq x\right\}\right)^2}$$

$$= \frac{\exp\left(x^2/2\right)\mathbb{P}\left\{Z \geq x\right\} - \exp\left((x+\epsilon)^2/2\right)\mathbb{P}\left\{Z \geq x + \epsilon\right\}}{\sqrt{2\pi}\exp\left((x+\epsilon)^2/2\right)\exp\left(x^2/2\right)\left(\mathbb{P}\left\{Z \geq x\right\}\right)^2}$$

$$= \frac{\int_0^\infty\left(\exp\left(-\frac{t^2+2tx}{2}\right) - \exp\left(-\frac{t^2+2t(x+\epsilon)}{2}\right)\right)\,\mathrm{d}t}{2\pi\exp\left((x+\epsilon)^2/2\right)\exp\left(x^2/2\right)\left(\mathbb{P}\left\{Z \geq x\right\}\right)^2} > 0.$$

In other words, $f_\epsilon(x)$ is monotonically increasing in x for any given $\varepsilon > 0$. Therefore, for any $0 \leq x < B$ for some sufficiently large $B > \sqrt{5/4}$ (to be specified later), the above bounds taken together give

$$\mathbb{P}\left\{X_{1,1} - Y_{1,1} < \epsilon \,\middle|\, X_{1,1} \geq Y_{1,1}, Y_{1,1} = x\right\} \overset{(i)}{\leq} \frac{\mathbb{P}\left\{B \leq X_{1,1} < B + \epsilon\right\}}{\mathbb{P}\left\{X_{1,1} \geq B\right\}} \overset{(ii)}{\leq} \frac{\epsilon\exp\left(-B^2/2\right)}{\frac{1}{5B}\exp\left(-B^2/2\right)} = 5\epsilon B,$$

where (i) arises from the monotonicity of $f_\epsilon(\cdot)$, and (ii) relies on (266). By taking $\epsilon = \delta/(5B)$, we obtain

$$\mathbb{P}\left\{X_{1,1} - Y_{1,1} < \epsilon \,\middle|\, X_{1,1} \geq Y_{1,1}, Y_{1,1} = x\right\} \leq \delta$$

for any $0 \leq x \leq B$. Recall that $Y_{1,1} = \max_{1 < i \leq r}\kappa X_{i,1} + \max_{k:k\neq 1}\Delta_k$. By standard Gaussian concentration inequalities, with probability at least $1 - \delta$ one has

$$Y_{1,1} \lesssim \kappa\sqrt{\log r} + \sqrt{\log L} + \sqrt{\log(1/\delta)} \asymp \sqrt{\log L} + \sqrt{\log(1/\delta)}.$$

where the last step arises from the lower bound on L in (270). If we choose $B = C\left(\sqrt{\log L} + \sqrt{\log(1/\delta)}\right)$ for some sufficiently large universal constant $C > 0$, then this immediately implies that

$$\mathbb{P}\left\{\Delta_{(1)} - \Delta_{(2)} < \epsilon\right\} = \mathbb{P}\left\{X_{1,1} - Y_{1,1} < \epsilon \,\middle|\, X_{1,1} \geq Y_{1,1}\right\}$$
$$\leq \mathbb{P}\left\{Y_{1,1} > B\right\} + \mathbb{P}\left\{X_{1,1} - Y_{1,1} < \epsilon \,\middle|\, X_{1,1} \geq Y_{1,1}, Y_{1,1} = B\right\} \leq 2\delta.$$

We have therefore concluded the proof.

D.6   Proof of Lemma D.6

To begin with, it is self-evident that

$$\mathcal{P}_{\boldsymbol{V}}(\boldsymbol{u}_0) = \boldsymbol{V}\boldsymbol{V}^\top\boldsymbol{u}_0 = \boldsymbol{U}\boldsymbol{U}^\top\boldsymbol{u}_0 + \left(\boldsymbol{V}\boldsymbol{V}^\top - \boldsymbol{U}\boldsymbol{U}^\top\right)\boldsymbol{u}_0 = \boldsymbol{u}_0 + \left(\boldsymbol{V}\boldsymbol{V}^\top - \boldsymbol{U}\boldsymbol{U}^\top\right)\boldsymbol{u}_0,$$

where the last identity follows since $\boldsymbol{u}_0$ is assumed to lie within $\mathsf{span}(\boldsymbol{U})$. As a result,

$$\mathcal{P}_{\boldsymbol{V}^\perp}(\boldsymbol{u}_0) = \boldsymbol{u}_0 - \mathcal{P}_{\boldsymbol{V}}(\boldsymbol{u}_0) = -\left(\boldsymbol{V}\boldsymbol{V}^\top - \boldsymbol{U}\boldsymbol{U}^\top\right)\boldsymbol{u}_0$$

$$\implies \quad \left\|\mathcal{P}_{\boldsymbol{V}^\perp}(\boldsymbol{u}_0)\right\|_2 \leq \left\|\boldsymbol{V}\boldsymbol{V}^\top - \boldsymbol{U}\boldsymbol{U}^\top\right\| \cdot \|\boldsymbol{u}_0\|_2 \leq \delta.$$

The Pythagorean theorem then gives $\left\|\mathcal{P}_{\boldsymbol{V}}(\boldsymbol{u}_0)\right\|_2 = \sqrt{\|\boldsymbol{u}_0\|_2 - \left\|\mathcal{P}_{\boldsymbol{V}^\perp}(\boldsymbol{u}_0)\right\|_2^2} \geq \sqrt{1 - \delta^2}.$

D.7 Proof of Lemma D.7

By virtue of Lemma D.1, we can compute

$$\sum_{i,j\in[d]} \mathbb{E}\left[T_{i,j,k}^{\star 2}(p^{-1}\chi_{i,j,k}-1)^2\right] \le \frac{1}{p}\sum_{i,j\in[d]} T_{i,j,k}^{\star 2} \le \frac{1}{p}\|\boldsymbol{A}^\star\|_{2,\infty}^2 \lesssim \frac{\mu r\lambda_{\max}^{\star 2}}{dp} =: M;$$

$$\left|T_{i,j,k}^{\star 2}(p^{-1}\chi_{i,j,k}-1)^2\right| \le \frac{1}{p^2}\|\boldsymbol{T}^\star\|_\infty^2 = \frac{1}{p^2}\|\boldsymbol{A}^\star\|_\infty^2 \lesssim \frac{\mu r\lambda_{\max}^{\star 2}}{d^3p^2} =: L;$$

$$\sum_{i,j\in[d]} \mathsf{Var}\left(T_{i,j,k}^{\star 2}(p^{-1}\chi_{i,j,k}-1)^2\right) \lesssim \frac{1}{p^3}\sum_{i,j\in[d]} T_{i,j,k}^{\star 4} \le \frac{1}{p^3}\|\boldsymbol{A}^\star\|_\infty^2\|\boldsymbol{A}^\star\|_{2,\infty}^2 \lesssim \frac{\mu^2 r^2\lambda_{\max}^{\star 4}}{d^4p^3} =: V.$$

Invoke the Bernstein inequality to show that: with probability exceeding $1 - O(d^{-20})$,

$$\sum_{i,j\in[d]} T_{i,j,k}^{\star 2}(p^{-1}\chi_{i,j,k}-1)^2 \lesssim M + L\log d + \sqrt{V\log d}$$

$$\lesssim \frac{\mu r\lambda_{\max}^{\star 2}}{dp} + \frac{\mu r\lambda_{\max}^{\star 2}\log d}{d^3p^2} + \sqrt{\frac{\mu^2 r^2\lambda_{\max}^{\star 4}\log d}{d^4p^3}}$$

$$\asymp \frac{\mu r\lambda_{\max}^{\star 2}}{dp},$$

where the last line holds with the proviso that $p \gtrsim d^{-2}\log d$.

D.8 Proof of Lemma D.8

Since the $E_{i,j,k}$'s are independent sub-Gaussian random variables with variance at most σ^2, one has

$$\sum_{i,j\in[d]} \mathbb{E}\left[(E_{i,j,k}\chi_{i,j,k})^2\right] \lesssim \sigma^2 d^2 p =: M;$$

$$\left\|(E_{i,j,k}\chi_{i,j,k})^2\right\|_{\psi_1} \lesssim \sigma^2 =: L;$$

$$\sum_{i,j\in[d]} \mathsf{Var}\left[(E_{i,j,k}\chi_{i,j,k})^2\right] \lesssim \sigma^4 d^2 p =: V.$$

Here, $\|\cdot\|_{\psi_1}$ denotes the sub-exponential norm [Ver10]. Taken together with the Bernstein inequality, these yield that with probability exceeding $1 - O(d^{-20})$,

$$\sum_{i,j\in[d]} (E_{i,j,k}\chi_{i,j,k})^2 \lesssim M + \sqrt{V\log d} + L\log^2 d \lesssim \sigma^2 d^2 p + \sqrt{\sigma^4 d^2 p\log d} + \sigma^2\log^2 d \asymp \sigma^2 d^2 p,$$

provided that $p \gtrsim d^{-2}\log^2 d$.

D.9 Proof of Lemma D.9

Fix an arbitrary $1 \le i \le d$. We first define a sequence of independent zero-mean random variables $\{X_j\}_{j\in[d]}$ as follows

$$X_j := \sum_{k\in[d]} T_{i,j,k}^\star w_k(p^{-1}\chi_{i,j,k}-1).$$

One can easily show that

$$\max_{k\in[d]}\left|T_{i,j,k}^\star w_k(p^{-1}\chi_{i,j,k}-1)\right| \le \frac{1}{p}\|\boldsymbol{A}^\star\|_\infty\|\boldsymbol{w}\|_\infty =: L,$$

$$\mathbb{E}[X_j^2] = \sum_{k\in[d]} T_{i,j,k}^{\star 2} w_k^2 \mathbb{E}\left[(p^{-1}\chi_{i,j,k}-1)^2\right] \le \frac{1}{p}\|\boldsymbol{w}\|_\infty^2 \sum_{k\in[d]} T_{i,j,k}^{\star 2} = \frac{1}{p}\|\boldsymbol{w}\|_\infty^2\|\boldsymbol{A}^{\star\top}\|_{2,\infty}^2 =: V.$$

The Bernstein inequality indicates that: with probability at least $1 - O(d^{-20})$,

$$|X_j| \lesssim L \log d + \sqrt{V \log d} \lesssim \frac{1}{p} \|\boldsymbol{A}^\star\|_\infty \|\boldsymbol{w}\|_\infty \log d + \sqrt{\frac{\log d}{p}} \|\boldsymbol{A}^{\star\top}\|_{2,\infty} \|\boldsymbol{w}\|_\infty := L_j. \tag{271}$$

Moreover, we can also bound the variance of X_j as follows

$$\mathsf{Var}(X_j^2) \le \mathbb{E}[X_j^4] \lesssim \frac{1}{p^3} \sum_{k\in[d]} T_{i,j,k}^{\star 4} w_k^4 + \frac{1}{p^2} \sum_{k_1 \ne k_2} T_{i,j,k_1}^{\star 2} T_{i,j,k_2}^{\star 2} w_{k_1}^2 w_{k_2}^2$$

$$\lesssim \frac{1}{p^3} \|\boldsymbol{w}\|_\infty^4 \|\boldsymbol{A}^\star\|_\infty^2 \sum_{k\in[d]} T_{i,j,k}^{\star 2} + \frac{1}{p^2} \|\boldsymbol{w}\|_\infty^4 \|\boldsymbol{A}^{\star\top}\|_{2,\infty}^2 \sum_{k\in[d]} T_{i,j,k}^{\star 2}.$$

Given that X_j might be overly large in some rare case, we introduce a sequence $\{Y_j\}$, where we denote by Y_j the truncated version of X_j as follows

$$Y_j := X_j \, \mathbb{1}\{|X_j| \lesssim L_j\}.$$

We have learn from (271) and the union bound that with probability at least $1 - O(d^{-15})$, one has $Y_j = X_j$ for all $j \in [d]$.

Using the above bounds on the X_j's, one observes that $\{Y_j\}_{j\in[d]}$ is a sequence of independent random variables satisfying

$$\sum_{j\in[d]} \mathbb{E}[Y_j^2] \le \sum_{j\in[d]} \mathbb{E}[X_j^2] \le \frac{1}{p} \|\boldsymbol{A}^\star\|_{2,\infty}^2 \|\boldsymbol{w}\|_\infty^2;$$

$$\max_{j\in[d]} Y_j^2 \le \max_{j\in[d]} L_j^2 \lesssim \frac{1}{p^2} \|\boldsymbol{A}^\star\|_\infty^2 \|\boldsymbol{w}\|_\infty^2 \log d + \frac{\log d}{p} \|\boldsymbol{A}^{\star\top}\|_{2,\infty}^2 \|\boldsymbol{w}\|_\infty^2;$$

$$\sum_{j\in[d]} \mathsf{Var}(Y_j^2) \le \sum_{j\in[d]} \mathbb{E}[X_j^4] \lesssim \frac{1}{p^3} \|\boldsymbol{A}^\star\|_\infty^2 \|\boldsymbol{A}^\star\|_{2,\infty}^2 \|\boldsymbol{w}\|_\infty^4 + \frac{1}{p^2} \|\boldsymbol{A}^{\star\top}\|_{2,\infty}^2 \|\boldsymbol{A}^\star\|_{2,\infty}^2 \|\boldsymbol{w}\|_\infty^4.$$

We can apply the Bernstein inequality to conclude that: with probability greater than $1 - O(d^{-15})$,

$$\sum_{j\in[d]} Y_j^2 \lesssim \sum_{j\in[d]} \mathbb{E}[Y_j^2] + \max_{j\in[d]} Y_j^2 \log d + \sqrt{\sum_{j\in[d]} \mathsf{Var}(Y_j^2) \log d}$$

$$\lesssim \frac{1}{p} \|\boldsymbol{A}^\star\|_{2,\infty}^2 \|\boldsymbol{w}\|_\infty^2 + \frac{\log^2 d}{p^2} \|\boldsymbol{A}^\star\|_\infty^2 \|\boldsymbol{w}\|_\infty^2 + \frac{\log^2 d}{p} \|\boldsymbol{A}^{\star\top}\|_{2,\infty}^2 \|\boldsymbol{w}\|_\infty^2$$

$$+ \sqrt{\frac{\log d}{p^3} \|\boldsymbol{A}^\star\|_\infty^2 \|\boldsymbol{A}^\star\|_{2,\infty}^2 \|\boldsymbol{w}\|_\infty^4 + \frac{\log d}{p^2} \|\boldsymbol{w}\|_\infty^4 \|\boldsymbol{A}^{\star\top}\|_{2,\infty}^2 \|\boldsymbol{A}^\star\|_{2,\infty}^2}$$

$$\overset{(i)}{\asymp} \frac{1}{p} \|\boldsymbol{A}^\star\|_{2,\infty}^2 \|\boldsymbol{w}\|_\infty^2 + \frac{\log^2 d}{p^2} \|\boldsymbol{A}^\star\|_\infty^2 \|\boldsymbol{w}\|_\infty^2 + \frac{\log^2 d}{p} \|\boldsymbol{A}^{\star\top}\|_{2,\infty}^2 \|\boldsymbol{w}\|_\infty^2$$

$$\overset{(ii)}{\lesssim} \left(\frac{\mu r \lambda_{\max}^{\star 2}}{dp} + \frac{\mu r \lambda_{\max}^{\star 2} \log^2 d}{d^3 p^2} + \frac{\mu^2 r \lambda_{\max}^{\star 2} \log^2 d}{d^2 p} \right) \|\boldsymbol{w}\|_\infty^2$$

$$\asymp \frac{\mu r \lambda_{\max}^{\star 2}}{dp} \|\boldsymbol{w}\|_\infty^2,$$

where (i) is due to the AM-GM inequality, (ii) makes use of Lemma D.1, and the last line follows under the conditions $p \gtrsim d^{-2} \log^2 d$ and $\mu \log^2 d \lesssim d$. This together with the high-probability fact $Y_j = X_j$ ($\forall j \in [d]$) concludes the proof.

D.10 Proof of Lemma D.10

Fix any $1 \le i \le d$. To begin with, define

$$Z_j := \sum_{k\in[d]} w_k E_{i,j,k} \chi_{i,j,k},$$

78

which is a zero-mean random variable. In order to bound Z_j, one observes that

$$\|w_k E_{i,j,k} \chi_{i,j,k}\|_{\psi_1} \lesssim \sigma \|\boldsymbol{w}\|_\infty =: L;$$

$$\sum_{k \in [d]} \mathsf{Var}\big(w_k E_{i,j,k} \chi_{i,j,k}\big) \leq \sigma^2 p \sum_{k \in [d]} w_k^2 = \sigma^2 p \|\boldsymbol{w}\|_2^2 =: V,$$

where $\|\cdot\|_{\psi_1}$ denotes the sub-exponential norm. Apply the Bernstein inequality for the sum of sub-exponential random variables to obtain

$$|Z_j| \lesssim \sqrt{V \log d} + L \log^2 d \lesssim \sigma \|\boldsymbol{w}\|_2 \sqrt{p \log d} + \sigma \|\boldsymbol{w}\|_\infty \log^2 d =: L_j \tag{272}$$

with probability exceeding $1 - O(d^{-20})$. Further, given that Z_j is not necessarily bounded, we introduce a sequence of truncated random variables as follows

$$Y_j := Z_j \, \mathbb{1}\{|Z_j| \lesssim L_j\}. \tag{273}$$

According to the above bound, one has $Y_j = Z_j$ $(\forall j)$ with probability at least $1 - O(d^{-19})$.

We then turn attention to bounding $\sum_{j \in [d]} Y_j^2$. To this end, observe that

$$\sum_{j \in [d]} \mathbb{E}[Y_j^2] \leq \sum_{j \in [d]} \mathbb{E}\Big[\Big(\sum_{k \in [d]} w_k E_{i,j,k} \chi_{i,j,k}\Big)^2\Big] \lesssim \sigma^2 p \sum_{j \in [d]} \sum_{k \in [d]} w_k^2$$

$$= \sigma^2 p d \|\boldsymbol{w}\|_2^2 =: M_0.$$

Additionally,

$$\sum_{j \in [d]} \mathbb{E}\big[Y_j^4\big] \leq \sum_{j \in [d]} \mathbb{E}\Big[\Big(\sum_{k \in [d]} w_k E_{i,j,k} \chi_{i,j,k}\Big)^4\Big]$$

$$\leq \sum_{j \in [d]} \sum_{k \in [d]} \mathbb{E}\big[w_k^4 E_{i,j,k}^4 \chi_{i,j,k}^4\big] + \sum_{j \in [d]} \sum_{k_1 \neq k_2} \mathbb{E}\big[w_{k_1}^2 w_{k_2}^2 E_{i,j,k_1}^2 E_{i,j,k_2}^2 \chi_{i,j,k_1}^2 \chi_{i,j,k_2}^2\big]$$

$$\lesssim \sigma^4 p \sum_{j \in [d]} \sum_{k \in [d]} w_k^4 + \sigma^4 p^2 \sum_{j \in [d]} \sum_{1 \leq k_1 \neq k_2 \leq d} w_{k_1}^2 w_{k_2}^2$$

$$\lesssim \sigma^4 p d \|\boldsymbol{w}\|_2^2 \|\boldsymbol{w}\|_\infty^2 + \sigma^4 p^2 d \|\boldsymbol{w}\|_2^4 =: V_0.$$

Invokde the Bernstein inequality to arrive at: with probability at least $1 - O(d^{-20})$,

$$\sum_{j \in [d]} Y_j^2 \lesssim M_0 + \sqrt{V_0 \log d} + \max_{j \in [d]} L_j^2 \log d$$

$$\lesssim \sigma^2 p d \|\boldsymbol{w}\|_2^2 + \sqrt{\sigma^4 p d \|\boldsymbol{w}\|_2^2 \big(p \|\boldsymbol{w}\|_2^2 + \|\boldsymbol{w}\|_\infty^2\big) \log d} + \big(\sigma^2 p \|\boldsymbol{w}\|_2^2 \log^2 d + \sigma^2 \|\boldsymbol{w}\|_\infty^2 \log^5 d\big)$$

$$\asymp \sigma^2 p d \|\boldsymbol{w}\|_2^2 + \sigma^2 \|\boldsymbol{w}\|_\infty^2 \log^5 d.$$

This together with the high-probability fact $Y_j = Z_j$ $(\forall j)$ completes the proof.

E   Extension to asymmetric tensors

Thus far, we have focused on the case where the tensor of interest is symmetric. In this section, we discuss how to generalize our algorithm and analysis to accommodate asymmetric tensors.

E.1   Problem settings

Suppose that the unknown tensor $\boldsymbol{T}^\star \in \mathbb{R}^{d_1 \times d_2 \times d_3}$ is a rank-r tensor with CP decomposition

$$\boldsymbol{T}^\star = \sum_{i=1}^r \boldsymbol{u}_i^\star \otimes \boldsymbol{v}_i^\star \otimes \boldsymbol{w}_i^\star, \tag{274}$$

where $\boldsymbol{u}_i^\star \in \mathbb{R}^{d_1}$, $\boldsymbol{v}_i^\star \in \mathbb{R}^{d_2}$, $\boldsymbol{w}_i^\star \in \mathbb{R}^{d_3}$ represent the tensor factors of interest. Apparently, there is an unavoidable global scaling ambiguity issue (for instance, multiplying $\boldsymbol{u}_i^\star$ by a constant c and multiplying $\boldsymbol{v}_i^\star$ by $1/c$ accordingly result in the same tensor). Without loss of generality, we shall assume throughout that

$$\|\boldsymbol{u}_i^\star\|_2 = \|\boldsymbol{v}_i^\star\|_2 = \|\boldsymbol{w}_i^\star\|_2, \qquad 1 \le i \le r. \tag{275}$$

In addition, we assume that each entry (j, k, l) is included in the sampling set Ω independently with probability p, and that each observed entry $T_{j,k,l}^\star$ is corrupted by an independent zero-mean sub-Guassian noise $E_{j,k,l}$ (cf. Assumption 2.3). Our goal is to (1) estimate $\{\boldsymbol{u}_i^\star, \boldsymbol{v}_i^\star, \boldsymbol{w}_i^\star\}_{i=1}^r$ faithfully, modulo global permutation and global signs, and (2) estimate $\boldsymbol{T}^\star$ in a reliable manner.

E.2 Algorithms

We now move on to present an extension of our nonconvex algorithm to handle the noisy scenario.

First of all, setting

$$\boldsymbol{U} := [\boldsymbol{u}_1, \cdots, \boldsymbol{u}_r] \in \mathbb{R}^{d_1 \times r}, \quad \boldsymbol{V} = [\boldsymbol{v}_1, \cdots, \boldsymbol{v}_r] \in \mathbb{R}^{d_2 \times r} \quad \text{and} \quad \boldsymbol{W} = [\boldsymbol{w}_1, \cdots, \boldsymbol{w}_r] \in \mathbb{R}^{d_3 \times r},$$

we can define the following regularized squared loss function

$$g(\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}) := \frac{1}{6p} \left\| \mathcal{P}_\Omega \Big(\sum_{i=1}^r \boldsymbol{u}_i \otimes \boldsymbol{v}_i \otimes \boldsymbol{w}_i - \boldsymbol{T} \Big) \right\|_{\mathrm{F}}^2 + \mathsf{reg}(\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}), \tag{276}$$

where the regularization term $\mathsf{reg}(\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W})$ is given by

$$\mathsf{reg}(\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}) := \frac{1}{24} \sum_{i=1}^r \alpha_i \left\{ \big(\|\boldsymbol{u}_i\|_2^2 - \|\boldsymbol{v}_i\|_2^2 \big)^2 + \big(\|\boldsymbol{u}_i\|_2^2 - \|\boldsymbol{w}_i\|_2^2 \big)^2 + \big(\|\boldsymbol{v}_i\|_2^2 - \|\boldsymbol{w}_i\|_2^2 \big)^2 \right\} \tag{277}$$

for some positive regularization parameters $\{\alpha_i\}_{i=1}^r$ to be specified momentarily. In contrast to the symmetric case, the addition term $\mathsf{reg}(\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W})$ is included to help ensure that the sizes of $\boldsymbol{U}, \boldsymbol{V}$ and $\boldsymbol{W}$ stay close — an algorithmic trick that has proved useful in other problems like nonconvex rectangular matrix recovery [TBS+16, ZL16, CLL19].

We are now ready to present our nonconvex algorithm that accommodates the case with asymmetric tensors. As before, the proposed algorithm is initialized by a spectral method, followed by gradient descent designed to minimize the regularized loss function (276). The precise procedure is described in Algorithm 7 (which invokes Algorithms 8-9).

Before proceeding, we find it helpful to record closed-form expressions for the gradients, which are a crucial part when implementing the nonconvex gradient descent algorithm. Specifically, the gradients of $g(\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W})$ can be computed as follows

$$\nabla_{\boldsymbol{u}_i} g(\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}) = \frac{1}{3p} \mathcal{P}_\Omega \Big(\sum_{i=1}^r \boldsymbol{u}_i \otimes \boldsymbol{v}_i \otimes \boldsymbol{w}_i - \boldsymbol{T} \Big) \times_2 \boldsymbol{v}_i \times_3 \boldsymbol{w}_i + \frac{1}{6} \alpha_i \big(2 \|\boldsymbol{u}_i\|_2^2 - \|\boldsymbol{v}_i\|_2^2 - \|\boldsymbol{w}_i\|_2^2 \big) \boldsymbol{u}_i, \tag{278a}$$

$$\nabla_{\boldsymbol{v}_i} g(\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}) = \frac{1}{3p} \mathcal{P}_\Omega \Big(\sum_{i=1}^r \boldsymbol{u}_i \otimes \boldsymbol{v}_i \otimes \boldsymbol{w}_i - \boldsymbol{T} \Big) \times_1 \boldsymbol{u}_i \times_3 \boldsymbol{w}_i + \frac{1}{6} \alpha_i \big(2 \|\boldsymbol{v}_i\|_2^2 - \|\boldsymbol{u}_i\|_2^2 - \|\boldsymbol{w}_i\|_2^2 \big) \boldsymbol{v}_i, \tag{278b}$$

$$\nabla_{\boldsymbol{w}_i} g(\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}) = \frac{1}{3p} \mathcal{P}_\Omega \Big(\sum_{i=1}^r \boldsymbol{u}_i \otimes \boldsymbol{v}_i \otimes \boldsymbol{w}_i - \boldsymbol{T} \Big) \times_1 \boldsymbol{u}_i \times_2 \boldsymbol{v}_i + \frac{1}{6} \alpha_i \big(2 \|\boldsymbol{w}_i\|_2^2 - \|\boldsymbol{u}_i\|_2^2 - \|\boldsymbol{v}_i\|_2^2 \big) \boldsymbol{w}_i \tag{278c}$$

for each $1 \le i \le r$, where $\times_1, \times_2$ and $\times_3$ have been defined in Section 2.4.

E.3 Numerical experiments

In order to validate the effectiveness of the proposed algorithm, we conduct a series of numerical experiments.

Algorithm 7 Gradient descent for nonconvex tensor completion (asymmetric case)

1: Generate initial estimates $\boldsymbol{U}^0 \in \mathbb{R}^{d_1 \times r}$, $\boldsymbol{V}^0 \in \mathbb{R}^{d_2 \times r}$, $\boldsymbol{W}^0 \in \mathbb{R}^{d_3 \times r}$ via Algorithm 8.
2: **for** $t = 0, 1, \ldots, t_0 - 1$ **do**

$$\boldsymbol{U}^{t+1} = \boldsymbol{U}^t - \eta_t \nabla_{\boldsymbol{U}} g(\boldsymbol{U}^t, \boldsymbol{V}^t, \boldsymbol{W}^t),$$
$$\boldsymbol{V}^{t+1} = \boldsymbol{V}^t - \eta_t \nabla_{\boldsymbol{V}} g(\boldsymbol{U}^t, \boldsymbol{V}^t, \boldsymbol{W}^t),$$
$$\boldsymbol{W}^{t+1} = \boldsymbol{W}^t - \eta_t \nabla_{\boldsymbol{W}} g(\boldsymbol{U}^t, \boldsymbol{V}^t, \boldsymbol{W}^t),$$

where the gradients are given in (278).

Algorithm 8 Spectral initialization for nonconvex tensor completion (asymmetric case)

1: Let $\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top$ be the rank-r eigen-decomposition of $\mathcal{P}_{\mathsf{off\text{-}diag}}(\boldsymbol{A}\boldsymbol{A}^\top)$ where $\boldsymbol{A} = \mathsf{unfold}(\boldsymbol{T})$ is the mode-1 matricization of $\boldsymbol{T}$, and $\mathcal{P}_{\mathsf{off\text{-}diag}}(\boldsymbol{Z})$ extracts out the off-diagonal entries of $\boldsymbol{Z}$.
2: **Output:** initial estimates $\boldsymbol{U}^0 \in \mathbb{R}^{d_1 \times r}$, $\boldsymbol{V}^0 \in \mathbb{R}^{d_2 \times r}$, $\boldsymbol{W}^0 \in \mathbb{R}^{d_3 \times r}$ on the basis of $\boldsymbol{U} \in \mathbb{R}^{d_1 \times r}$ using Algorithm 9.

To begin with, let us generate the true tensor $\boldsymbol{T}^\star = \sum_{1 \le i \le r} \boldsymbol{u}_i^\star \otimes \boldsymbol{v}_i^\star \otimes \boldsymbol{w}_i^\star$ via the following procedure: (1) generate $\widehat{\boldsymbol{u}}_i^\star \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{d_1})$, $\widehat{\boldsymbol{v}}_i^\star \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{d_2})$ and $\widehat{\boldsymbol{w}}_i^\star \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{d_3})$, and (2) set $\lambda_i^\star := \|\widehat{\boldsymbol{u}}_i^\star\|_2 \|\widehat{\boldsymbol{v}}_i^\star\|_2 \|\widehat{\boldsymbol{w}}_i^\star\|_2$, $\boldsymbol{u}_i^\star = \lambda_i^{\star 1/3} \widehat{\boldsymbol{u}}_i^\star$, $\boldsymbol{v}_i^\star = \lambda_i^{\star 1/3} \widehat{\boldsymbol{v}}_i^\star$ and $\boldsymbol{w}_i^\star = \lambda_i^{\star 1/3} \widehat{\boldsymbol{w}}_i^\star$. Akin to the symmetric case, we choose the algorithmic parameters to be $L = r^2$, $\epsilon_{\mathsf{th}} = 0.4$ and $t_0 = 100$. The noise components are generated as i.i.d. Gaussians, namely, $E_{i,j,k} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$. The signal-to-noise ratio (SNR) is defined as $\mathrm{SNR} = \|\boldsymbol{T}^\star\|_{\mathrm{F}}^2 / (\sigma^2 d_1 d_2 d_3)$, and $\mathrm{SNR} = \infty$ stands for the noiseless case (i.e. the case with $\sigma = 0$). The stepsize is set to be $\eta_t \equiv \eta = 1/(2 \max_i \|\boldsymbol{u}_i^0\|_2^{4/3})$, where $\boldsymbol{U}^0 = [\boldsymbol{u}_1^0, \cdots, \boldsymbol{u}_r^0]$ is the initial estimate generated by Algorithm 8. Figure 2 illustrates the numerical convergence rates of the proposed algorithm, where we set $d_1 = 100$, $d_2 = 150$, $d_3 = 200$, $r = 4$ and $p = 0.05$. Since one can only recover the tensor factors up to global signs and global permutation, the relative estimation errors of $\boldsymbol{U}^t$ are defined to be

$$\min_{\boldsymbol{\Pi} \in \mathsf{perm}_r, \, \boldsymbol{S} = \mathsf{diag}(s_i) \in \mathbb{R}^{r \times r}, s_i = \pm 1} \left\| \boldsymbol{U}^t \boldsymbol{\Pi} \boldsymbol{S} - \boldsymbol{U}^\star \right\|_{\mathrm{F}} / \left\| \boldsymbol{U}^\star \right\|_{\mathrm{F}},$$

$$\min_{\boldsymbol{\Pi} \in \mathsf{perm}_r, \, \boldsymbol{S} = \mathsf{diag}(s_i) \in \mathbb{R}^{r \times r}, s_i = \pm 1} \left\| \boldsymbol{U}^t \boldsymbol{\Pi} \boldsymbol{S} - \boldsymbol{U}^\star \right\|_{2, \infty} / \left\| \boldsymbol{U}^\star \right\|_{2, \infty},$$

where perm_r stands for the set of $r \times r$ permutation matrices. The error metrics for $\boldsymbol{V}^t$ and $\boldsymbol{W}^t$ can be defined analogously. The relative Euclidean and $\ell_{2,\infty}$ estimation errors of $\boldsymbol{T}$ are defined as $\|\boldsymbol{T}^t - \boldsymbol{T}^\star\|_{\mathrm{F}} / \|\boldsymbol{T}^\star\|_{\mathrm{F}}$ and $\|\boldsymbol{T}^t - \boldsymbol{T}^\star\|_{2,\infty} / \|\boldsymbol{T}^\star\|_{2,\infty}$, respectively, where $\boldsymbol{T}^t = \sum_{1 \le i \le r} \boldsymbol{u}_i^t \otimes \boldsymbol{v}_i^t \otimes \boldsymbol{w}_i^t$. As can be seen from the plots, the estimation errors decay geometrically fast in the noiseless case. In the noisy case, the numerical estimation errors of the algorithm also converge geometrically fast until an error floor is hit.

Moving beyond the above synthetic data, we apply our methods to an simulated MRI brain image dataset [CKK+97], which is available online at the McGill University McConnell Brain Imaging Centre and has also been studied in prior work [XYZ17]. The database consists of pre-computed simulated brain data, and a set of parameters can be set to generate the data accordingly. In this series of experiments, we choose the parameters to be T1 modality, 1mm slice thickness, 1% noise and 20% RF. The resulting data is a three-order tensor in $\mathbb{R}^{181 \times 217 \times 181}$, where each slice in any mode corresponds to a brain image. We use the tensor decomposition algorithm in [AGJ15] to decompose the original data, and keep the top-36 components whose energy accounts for 85% of the original tensor. Hence, the resulting low-rank tensor $\boldsymbol{T}^\star$ preserves the major information of the original tensor, and is used as the ground truth in the simulation. We sample each entry of $\boldsymbol{T}^\star$ independently with probability p, and then inject i.i.d. Gaussian noise $\mathcal{N}(0, \sigma^2)$ to each observed entry (so as to simulate the noisy scenario). Based on the incomplete set of noisy samples, we apply the proposed algorithm to reconstruct $\boldsymbol{T}^\star$; the algorithmic parameters are chosen as above, and we denote by $\boldsymbol{T}$ the resulting tensor estimate. Table 2 reports the relative errors $\|\boldsymbol{T} - \boldsymbol{T}^\star\|_{\mathrm{F}} / \|\boldsymbol{T}^\star\|_{\mathrm{F}}$ and $\|\boldsymbol{T} - \boldsymbol{T}^\star\|_{\infty} / \|\boldsymbol{T}^\star\|_{\infty}$. As can be seen in the table, the performance of our algorithm is quite favorable for low-rank tensor reconstruction. In particular, given that these real data tensor examples might not satisfy

---

Algorithm 9 Retrieval of low-rank tensor factors from a given subspace estimate (asymmetric case)

1: **Input:** number of restarts L, pruning threshold ϵ_{th}, subspace estimate $\boldsymbol{U} \in \mathbb{R}^{d_1 \times r}$ given by Algorithm 8.
2: **for** $\tau = 1, \ldots, L$ **do**
3:     Generate an independent Gaussian vector $\boldsymbol{g}^\tau \sim \mathcal{N}(0, \boldsymbol{I}_{d_1})$.
4:

$$\left(\boldsymbol{\nu}^{(1),\tau}, \boldsymbol{\nu}^{(2),\tau}, \boldsymbol{\nu}^{(3),\tau}, \lambda_\tau, \mathsf{gap}_\tau\right) \leftarrow \text{RETRIEVE-ONE-TENSOR-FACTOR-ASYM}(\boldsymbol{T}, p, \boldsymbol{U}, \boldsymbol{g}^\tau).$$

5: Generate tensor factor estimates

$$\left\{(\boldsymbol{u}^1, \boldsymbol{v}^1, \boldsymbol{w}^1, \lambda_1), \ldots, (\boldsymbol{u}^r, \boldsymbol{v}^r, \boldsymbol{w}^r, \lambda_r)\right\} \leftarrow \text{PRUNE-ASYM}\big(\big\{\left(\boldsymbol{\nu}^{(1),\tau}, \boldsymbol{\nu}^{(2),\tau}, \boldsymbol{\nu}^{(3),\tau}, \lambda_\tau, \mathsf{gap}_\tau\right)\big\}_{\tau=1}^L, \epsilon_{\mathsf{th}}\big).$$

6: **Output:** initial estimate

$$\boldsymbol{U}^0 = \big[\lambda_1^{1/3}\boldsymbol{u}^1, \ldots, \lambda_r^{1/3}\boldsymbol{u}^r\big], \qquad \boldsymbol{V}^0 = \big[\lambda_1^{1/3}\boldsymbol{v}^1, \ldots, \lambda_r^{1/3}\boldsymbol{v}^r\big], \qquad \text{and} \qquad \boldsymbol{W}^0 = \big[\lambda_1^{1/3}\boldsymbol{w}^1, \ldots, \lambda_r^{1/3}\boldsymbol{w}^r\big].$$

1: **function** RETRIEVE-ONE-TENSOR-FACTOR-ASYM$(\boldsymbol{T}, p, \boldsymbol{U}, \boldsymbol{g})$
2:     Compute

$$\boldsymbol{\theta} = \boldsymbol{U}\boldsymbol{U}^\top \boldsymbol{g} =: \mathcal{P}_{\boldsymbol{U}}(\boldsymbol{g}), \tag{279a}$$

$$\boldsymbol{M} = p^{-1}\boldsymbol{T} \times_1 \boldsymbol{\theta}, \tag{279b}$$

where $\times_1$ is defined in Section 2.4.
3:     Let $\boldsymbol{\nu}^{(2)}$ (resp. $\boldsymbol{\nu}^{(3)}$) be the leading left (resp. right) singular vector of $\boldsymbol{M}$. Let $\boldsymbol{\nu}^{(1)} = p^{-1}\boldsymbol{T} \times_2 \boldsymbol{\nu}^{(2)} \times_3 \boldsymbol{\nu}^{(3)}$ and set $\lambda = \|\boldsymbol{\nu}^{(1)}\|_2^{1/3}$.
4:     **return** $\left(\boldsymbol{\nu}^{(1)}, \boldsymbol{\nu}^{(2)}, \boldsymbol{\nu}^{(3)}, \lambda, \sigma_1(\boldsymbol{M}) - \sigma_2(\boldsymbol{M})\right)$.

---

the assumptions imposed in our main theorems, these numerical experiments hint at the applicability of our algorithm to a broader set of problems.

E.4   Analysis ideas

Before describing the proof ideas, we define the following incoherence parameters and condition number that, similar to the symmetric case, play a crucial role in our theoretical development.

Definition E.1. *Define the incoherence parameters and the condition number of $\boldsymbol{T}^\star$ as follows*

$$\mu_0 := \frac{d_1 d_2 d_3 \|\boldsymbol{T}^\star\|_\infty^2}{\|\boldsymbol{T}^\star\|_{\mathrm{F}}^2}, \tag{280a}$$

$$\mu_1 := \max\left\{\frac{d_1\|\boldsymbol{u}_i^\star\|_\infty^2}{\|\boldsymbol{u}_i^\star\|_2^2}, \frac{d_2\|\boldsymbol{v}_i^\star\|_\infty^2}{\|\boldsymbol{v}_i^\star\|_2^2}, \frac{d_3\|\boldsymbol{w}_i^\star\|_\infty^2}{\|\boldsymbol{w}_i^\star\|_2^2}\right\}, \tag{280b}$$

$$\mu_2 := \max\left\{\frac{d_1\langle\boldsymbol{u}_i^\star, \boldsymbol{u}_j^\star\rangle^2}{\|\boldsymbol{u}_i^\star\|_2^2\|\boldsymbol{u}_j^\star\|_2^2}, \frac{d_2\langle\boldsymbol{v}_i^\star, \boldsymbol{v}_j^\star\rangle^2}{\|\boldsymbol{v}_i^\star\|_2^2\|\boldsymbol{v}_j^\star\|_2^2}, \frac{d_3\langle\boldsymbol{w}_i^\star, \boldsymbol{w}_j^\star\rangle^2}{\|\boldsymbol{w}_i^\star\|_2^2\|\boldsymbol{w}_j^\star\|_2^2}\right\}, \tag{280c}$$

$$\kappa := \frac{\max_i\left\{\|\boldsymbol{u}_i^\star\|_2\|\boldsymbol{v}_i^\star\|_2\|\boldsymbol{w}_i^\star\|_2\right\}}{\min_i\left\{\|\boldsymbol{u}_i^\star\|_2\|\boldsymbol{v}_i^\star\|_2\|\boldsymbol{w}_i^\star\|_2\right\}}. \tag{280d}$$

For notational convenience, we shall set

$$\mu := \max\left\{\mu_0, \mu_1, \mu_2\right\}, \quad d_{\min} := \min\{d_1, d_2, d_3\} \quad \text{and} \quad d_{\max} := \max\{d_1, d_2, d_3\}. \tag{280e}$$

As has been made clear in the symmetric case, at the heart of our analysis lie two crucial components: (1) the geometric property of a noiseless version of the loss function, and (2) reasonably well initial estimates

1: **function** PRUNE-ASYM$\left(\left\{\left(\boldsymbol{\nu}^{(1),\tau}, \boldsymbol{\nu}^{(2),\tau}, \boldsymbol{\nu}^{(3),\tau}, \lambda_\tau, \mathsf{gap}_\tau\right)\right\}_{\tau=1}^L, \epsilon_{\mathsf{th}}\right)$

2:     Set $\Theta = \left\{\left(\boldsymbol{\nu}^{(1),\tau}, \boldsymbol{\nu}^{(2),\tau}, \boldsymbol{\nu}^{(3),\tau}, \lambda_\tau, \mathsf{gap}_\tau\right)\right\}_{\tau=1}^L$.

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

4:         Choose $\left(\boldsymbol{\nu}^{(1),\tau}, \boldsymbol{\nu}^{(2),\tau}, \boldsymbol{\nu}^{(3),\tau}, \lambda_\tau, \mathsf{gap}_\tau\right)$ from Θ with the largest gap_τ; set $\boldsymbol{u}^i = \boldsymbol{\nu}^{(1),\tau}, \boldsymbol{v}^i = \boldsymbol{\nu}^{(2),\tau}, \boldsymbol{w}^i = \boldsymbol{\nu}^{(3),\tau}$ and $\lambda_i = \lambda_\tau$.

5:         Update $\Theta \leftarrow \Theta \setminus \left\{\left(\boldsymbol{\nu}^{(1),\tau}, \boldsymbol{\nu}^{(2),\tau}, \boldsymbol{\nu}^{(3),\tau}, \lambda_\tau, \mathsf{gap}_\tau\right) \in \Theta : |\langle \boldsymbol{\nu}^{(2),\tau}, \boldsymbol{v}^i \rangle| > 1 - \epsilon_{\mathsf{th}}\right\}$.

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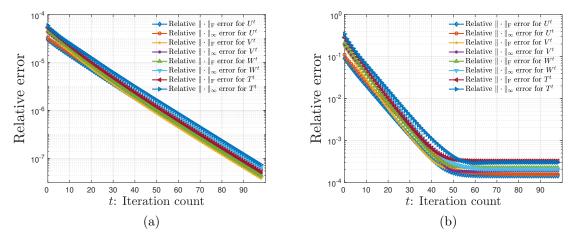

Figure 2: (a) relative errors of the estimates $\boldsymbol{U}^t, \boldsymbol{V}^t, \boldsymbol{W}^t$ and $\boldsymbol{T}^t$ vs. iteration count t for noiseless tensor completion, where $d_1 = 100$, $d_2 = 150$, $d_3 = 200$, $r = 4$, $p = 0.05$; (b) relative errors of the estimates $\boldsymbol{U}^t, \boldsymbol{V}^t, \boldsymbol{W}^t$ and $\boldsymbol{T}^t$ vs. iteration count t for noisy tensor completion, where $d_1 = 100$, $d_2 = 150$, $d_3 = 200$, $r = 4$, $p = 0.05$, SNR = 10.

for tensor factors (in the entrywise sense). Rather than providing a complete analysis for the asymmetric case (which would be very long), we shall only point out the important steps needed to extend these two parts for the asymmetric case.

1. Local optimization landscape. Similar to the symmetric counterpart, the key step of the local convergence analysis lies in establishing the favorable geometric property (i.e. local strong convexity and smoothness) of the following noiseless regularized loss function

$$g_{\mathsf{clean}}(\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}) = \frac{1}{6p}\left\|\mathcal{P}_\Omega\left(\sum_{i=1}^r \boldsymbol{u}_i \otimes \boldsymbol{v}_i \otimes \boldsymbol{w}_i - \boldsymbol{T}^\star\right)\right\|_{\mathrm{F}}^2 + \mathsf{reg}(\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}), \tag{281}$$

where the regularization term $\mathsf{reg}(\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W})$ is defined in (277). In words, this is a simplified version of the original loss function (276) by dropping the influence of the noise. Lemma E.2 below demonstrates that g_{clean} is locally strongly convex and smooth in the neighborhood of the ground truth.

Lemma E.2 (Local strong convexity and smoothness). *Suppose that*

$$p \geq c_0 \max\left\{\frac{\log^3 d_{\max}}{\sqrt{d_1 d_2 d_3}}, \frac{\mu^2 r^2 d_{\max} \log^5 d_{\max}}{d_1 d_2 d_3}\right\}, \qquad r \leq c_1 \sqrt{\frac{d_{\min}}{\mu}} \tag{282}$$

and that the regularization parameter obeys $\left|\alpha_i - \lambda_i^{\star 2/3}\right| \leq c_2 \lambda_{\min}^{\star 2/3}$ for some sufficiently large (resp. small)

83

Table 2: Relative ℓ_2 and $\ell_{2,\infty}$ errors for varying p and SNR in the MRI data experiments.

(p, SNR)	$\|\boldsymbol{T} - \boldsymbol{T}^\star\|_\mathrm{F}/\|\boldsymbol{T}^\star\|_\mathrm{F}$	$\|\boldsymbol{T} - \boldsymbol{T}^\star\|_\infty/\|\boldsymbol{T}^\star\|_\infty$
$(0.05, 10)$	0.2231	0.8774
$(0.05, \infty)$	0.2149	0.7856
$(0.1, 10)$	0.1472	0.5343
$(0.1, \infty)$	0.1398	0.3991
$(0.2, 10)$	0.0841	0.1417
$(0.2, \infty)$	0.0686	0.1193

constant $c_0 > 0$ (resp. $c_1, c_2 > 0$). Then with probability greater than $1 - O\big(d_{\min}^{-10}\big)$,

$$\mathsf{vec}\left(\begin{bmatrix} \boldsymbol{X} \\ \boldsymbol{Y} \\ \boldsymbol{Z} \end{bmatrix}\right)^\top \nabla^2 g_{\mathsf{clean}}(\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W})\, \mathsf{vec}\left(\begin{bmatrix} \boldsymbol{X} \\ \boldsymbol{Y} \\ \boldsymbol{Z} \end{bmatrix}\right) \leq 4\lambda_{\max}^{\star 4/3}\left(\|\boldsymbol{X}\|_\mathrm{F}^2 + \|\boldsymbol{Y}\|_\mathrm{F}^2 + \|\boldsymbol{Z}\|_\mathrm{F}^2\right), \tag{283}$$

$$\mathsf{vec}\left(\begin{bmatrix} \boldsymbol{X} \\ \boldsymbol{Y} \\ \boldsymbol{Z} \end{bmatrix}\right)^\top \nabla^2 g_{\mathsf{clean}}(\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W})\, \mathsf{vec}\left(\begin{bmatrix} \boldsymbol{X} \\ \boldsymbol{Y} \\ \boldsymbol{Z} \end{bmatrix}\right) \geq \frac{1}{2}\lambda_{\min}^{\star 4/3}\left(\|\boldsymbol{X}\|_\mathrm{F}^2 + \|\boldsymbol{Y}\|_\mathrm{F}^2 + \|\boldsymbol{Z}\|_\mathrm{F}^2\right) \tag{284}$$

holds simultaneously for all $\boldsymbol{X} \in \mathbb{R}^{d_1 \times r}, \boldsymbol{Y} \in \mathbb{R}^{d_2 \times r}, \boldsymbol{Z} \in \mathbb{R}^{d_3 \times r}$ and all $\boldsymbol{U} \in \mathbb{R}^{d_1 \times r}, \boldsymbol{V} \in \mathbb{R}^{d_2 \times r}, \boldsymbol{W} \in \mathbb{R}^{d_3 \times r}$ obeying

$$\|\boldsymbol{U} - \boldsymbol{U}^\star \boldsymbol{S}^{(1)}\|_\mathrm{F} \leq \delta\|\boldsymbol{U}^\star\|_\mathrm{F} \quad and \quad \left\|\boldsymbol{U} - \boldsymbol{U}^\star \boldsymbol{S}^{(1)}\right\|_{2,\infty} \leq \delta\|\boldsymbol{U}^\star\|_{2,\infty}; \tag{285a}$$

$$\|\boldsymbol{V} - \boldsymbol{V}^\star \boldsymbol{S}^{(2)}\|_\mathrm{F} \leq \delta\|\boldsymbol{V}^\star\|_\mathrm{F} \quad and \quad \left\|\boldsymbol{V} - \boldsymbol{V}^\star \boldsymbol{S}^{(2)}\right\|_{2,\infty} \leq \delta\|\boldsymbol{V}^\star\|_{2,\infty}; \tag{285b}$$

$$\|\boldsymbol{W} - \boldsymbol{W}^\star \boldsymbol{S}^{(3)}\|_\mathrm{F} \leq \delta\|\boldsymbol{V}^\star\|_\mathrm{F} \quad and \quad \left\|\boldsymbol{W} - \boldsymbol{W}^\star \boldsymbol{S}^{(3)}\right\|_{2,\infty} \leq \delta\|\boldsymbol{W}^\star\|_{2,\infty}. \tag{285c}$$

Here, $\delta \leq c_3/(\mu^{3/2} r)$ for some sufficiently small constant $c_3 > 0$, and $\boldsymbol{S}^{(1)}, \boldsymbol{S}^{(2)}$ and $\boldsymbol{S}^{(3)}$ are some diagonal matrices in $\mathbb{R}^{r \times r}$ such that for each $1 \leq i \leq r$, two of $S_{i,i}^{(1)}, S_{i,i}^{(2)}$ and $S_{i,i}^{(3)}$ equal to -1 with the remaining one equal to 1.

Proof. See Appendix E.5. ∎

In a nutshell, Lemma E.2 confirms local strong convexity and smoothness of the noiseless regularized loss $g_{\mathsf{clean}}(\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W})$, provided that (282) holds and that the matrices $\boldsymbol{U}, \boldsymbol{V}$ and $\boldsymbol{W}$ are sufficiently close to the ground truth in every single row. This is similar to the property of the symmetric counterpart $f_{\mathsf{clean}}(\boldsymbol{U})$ (cf. Lemma 5.1 in Appendix 5.1), except that we need to deal with asymmetric tensor factors as well as additional regularization terms. In particular, when $d_1 \asymp d_2 \asymp d_3 \asymp d$ and $\mu, r \asymp 1$, Condition (282) reduces to $p \gg d^{-3/2} \log^3 d$ and $r \ll \sqrt{d}$, which resembles (42) in Lemma 5.1 derived for the symmetric case.

Having established the preceding local geometric properties of g_{clean}, one can then argue similarly as in Lemmas 5.3 and 5.6 in Appendix 5.1 to prove that gradient descent converges linearly, as long as it is provided with an initial estimate satisfying the condition (285). Here, we emphasize that the regularization term (277), which essentially balances the sizes of the three tensor factors, is crucial for the local strong convexity and smoothness of g_{clean} to hold.

2. Guaranteeing a reasonably good initialization. Another crucial ingredient lies in guaranteeing an initial estimate with sufficiently good accuracy. Recall that our initialization scheme consists of two stages: (1) subspace estimation, and (2) retrieval of individual tensor factors.

- The subspace estimation part remains largely unchanged: we shall unfold the observed tensor along the 1-st mode to estimate the subspace spanned by tensor factors $\{\boldsymbol{u}_i^\star\}_{i=1}^r$, and the $\ell_{2,\infty}$ subspace estimation accuracy can be established by invoking the main theorems of our companion paper [CLC+20].

- Regarding the retrieval of individual tensor factors, the key observation is that: the random vector $\boldsymbol{g}^\tau$ we generate satisfies

$$\boldsymbol{T}^\star \times_1 \mathcal{P}_{\boldsymbol{U}^\star}(\boldsymbol{g}^\tau) = \sum_{i=1}^r \lambda_i^\star \langle \overline{\boldsymbol{u}}_i^\star, \mathcal{P}_{\boldsymbol{U}^\star}(\boldsymbol{g}^\tau) \rangle \overline{\boldsymbol{v}}_i^\star \overline{\boldsymbol{w}}_i^{\star\top} = \sum_{i=1}^r \lambda_i^\star \langle \overline{\boldsymbol{u}}_i^\star, \boldsymbol{g}^\tau \rangle \overline{\boldsymbol{v}}_i^\star \overline{\boldsymbol{w}}_i^{\star\top},$$

where $\mathcal{P}_{\boldsymbol{U}^\star}$ is the projection onto the subspace spanned by $\{\boldsymbol{u}_i^\star\}_{i=1}^r$, and

$$\overline{\boldsymbol{u}}_i^\star = \frac{1}{\|\boldsymbol{u}_i^\star\|_2} \boldsymbol{u}_i^\star, \qquad \overline{\boldsymbol{v}}_i^\star = \frac{1}{\|\boldsymbol{v}_i^\star\|_2} \boldsymbol{v}_i^\star, \qquad and \qquad \overline{\boldsymbol{w}}_i^\star = \frac{1}{\|\boldsymbol{w}_i^\star\|_2} \boldsymbol{w}_i^\star. \tag{286}$$

Given a sufficiently accurate subspace estimate $\boldsymbol{U}$ for $\boldsymbol{U}^\star$ obtained in the subspace estimation stage, for each $1 \leq i \leq r$, there exists at least a point $1 \leq \tau \leq L$ such that the spectral gap of the population version of $\boldsymbol{T} \times_1 \boldsymbol{\theta}^\tau = \boldsymbol{T} \times_1 \mathcal{P}_{\boldsymbol{U}}(\boldsymbol{g}^\tau)$ (with respect to the sampling and noise) is large enough and that the perturbation is sufficiently small. As a result, the top left (resp. right) singular vector $\boldsymbol{\nu}^{(2)}$ (resp. $\boldsymbol{\nu}^{(3)}$) of $\boldsymbol{T} \times_1 \boldsymbol{\theta}^\tau$ is close to the (normalized) tensor factor $\overline{\boldsymbol{v}}_i^\star$ (resp. $\overline{\boldsymbol{w}}_i^\star$) both in the ℓ_2 and ℓ_∞ norm (similar to Lemmas 5.13-5.14 in Section 5.3 for the symmetric case). In turn, this further allows us to reliably estimate $\overline{\boldsymbol{u}}_i^\star$ and the magnitude $\lambda_i^\star$, in a way similar to what we have done in Lemma 5.16 in Section 5.3. By repeating the procedures with random restarts and invoking a pruning procedure similar to the symmetric case, we can hope to recover all tensor factors with high probability.

Following the above strategies, one could adapt the proofs of Theorems 5.9-5.10 in Section 5.3 to show that: our initial estimates $\{\boldsymbol{u}_i, \boldsymbol{v}_i, \boldsymbol{w}_i\}_{i=1}^r$ are all exceedingly close to the ground truth in the entrywise sense (up to global permutation and global signs). This in turn confirms that the algorithm will enter a locally strongly convex and smooth region as characterized in Lemma E.2, thus leading to our performance guarantees for the entire algorithm. Once again, while the analysis ideas for the asymmetric case bear much resemblance to the symmetric counterpart, a complete proof has to be fairly long due to more clumsy notation compared to the symmetric case; for the sake of brevity, we do not provide the full proof here.

E.5   Proof of Lemma E.2

Fix arbitrary matrices $\boldsymbol{X} = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_r] \in \mathbb{R}^{d_1 \times r}$, $\boldsymbol{Y} = [\boldsymbol{y}_1, \cdots, \boldsymbol{y}_r] \in \mathbb{R}^{d_2 \times r}$ and $\boldsymbol{Z} = [\boldsymbol{z}_1, \cdots, \boldsymbol{z}_r] \in \mathbb{R}^{d_3 \times r}$. Direct computation reveals that

$$\mathsf{vec}\left(\begin{bmatrix} \boldsymbol{X} \\ \boldsymbol{Y} \\ \boldsymbol{Z} \end{bmatrix}\right)^\top \nabla^2 g_{\mathsf{clean}}(\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}) \, \mathsf{vec}\left(\begin{bmatrix} \boldsymbol{X} \\ \boldsymbol{Y} \\ \boldsymbol{Z} \end{bmatrix}\right)$$

$$= \frac{1}{3p} \left\| \mathcal{P}_\Omega \Big(\sum_{i=1}^r \boldsymbol{x}_i \otimes \boldsymbol{v}_i \otimes \boldsymbol{w}_i + \boldsymbol{u}_i \otimes \boldsymbol{y}_i \otimes \boldsymbol{w}_i + \boldsymbol{u}_i \otimes \boldsymbol{v}_i \otimes \boldsymbol{z}_i \Big) \right\|_{\mathrm{F}}^2$$

$$+ \frac{2}{3p} \Big\langle \sum_{i=1}^r \boldsymbol{x}_i \otimes \boldsymbol{y}_i \otimes \boldsymbol{w}_i + \boldsymbol{x}_i \otimes \boldsymbol{v}_i \otimes \boldsymbol{z}_i + \boldsymbol{u}_i \otimes \boldsymbol{y}_i \otimes \boldsymbol{z}_i, \mathcal{P}_\Omega \Big(\sum_{i=1}^r \boldsymbol{u}_i \otimes \boldsymbol{v}_i \otimes \boldsymbol{w}_i - \boldsymbol{T}^\star \Big) \Big\rangle$$

$$+ \frac{1}{3} \sum_{i=1}^r \alpha_i \big(\langle \boldsymbol{x}_i, \boldsymbol{u}_i \rangle - \langle \boldsymbol{y}_i, \boldsymbol{v}_i \rangle \big)^2 + \frac{1}{6} \sum_{i=1}^r \alpha_i \big(\|\boldsymbol{u}_i\|_2^2 - \|\boldsymbol{v}_i\|_2^2 \big) \big(\|\boldsymbol{x}_i\|_2^2 - \|\boldsymbol{y}_i\|_2^2 \big)$$

$$+ \frac{1}{3} \sum_{i=1}^r \alpha_i \big(\langle \boldsymbol{x}_i, \boldsymbol{u}_i \rangle - \langle \boldsymbol{z}_i, \boldsymbol{w}_i \rangle \big)^2 + \frac{1}{6} \sum_{i=1}^r \alpha_i \big(\|\boldsymbol{u}_i\|_2^2 - \|\boldsymbol{w}_i\|_2^2 \big) \big(\|\boldsymbol{x}_i\|_2^2 - \|\boldsymbol{z}_i\|_2^2 \big)$$

$$+ \frac{1}{3} \sum_{i=1}^r \alpha_i \big(\langle \boldsymbol{y}_i, \boldsymbol{v}_i \rangle - \langle \boldsymbol{z}_i, \boldsymbol{w}_i \rangle \big)^2 + \frac{1}{6} \sum_{i=1}^r \alpha_i \big(\|\boldsymbol{v}_i\|_2^2 - \|\boldsymbol{w}_i\|_2^2 \big) \big(\|\boldsymbol{y}_i\|_2^2 - \|\boldsymbol{z}_i\|_2^2 \big).$$

For notational convenience, we shall define

$$\boldsymbol{\Delta}^U = \big[\boldsymbol{\Delta}_1^U, \cdots, \boldsymbol{\Delta}_r^U \big] := \boldsymbol{U} \boldsymbol{S}^{(1)} - \boldsymbol{U}^\star \in \mathbb{R}^{d_1 \times r},$$

85

and define $\boldsymbol{\Delta}^{\boldsymbol{V}} \in \mathbb{R}^{d_2 \times r}$ and $\boldsymbol{\Delta}^{\boldsymbol{W}} \in \mathbb{R}^{d_3 \times r}$ in an analogous manner. Without loss of generality, it is assumed that $\boldsymbol{S}^{(1)} = \boldsymbol{I}_r$ and $\boldsymbol{S}^{(2)} = \boldsymbol{S}^{(3)} = -\boldsymbol{I}_r$.

With the above notation in place, one can decompose

$$
\mathsf{vec}\left(\begin{bmatrix} \boldsymbol{X} \\ \boldsymbol{Y} \\ \boldsymbol{Z} \end{bmatrix}\right)^{\top} \nabla^2 g_{\mathsf{clean}}(\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}) \, \mathsf{vec}\left(\begin{bmatrix} \boldsymbol{X} \\ \boldsymbol{Y} \\ \boldsymbol{Z} \end{bmatrix}\right) = \sum_{i=1}^{4} \beta_i
$$

where $\beta_i, 1 \le i \le 4$ are given respectively by

$$
\beta_1 := \frac{1}{3} \Big\| \sum_{i=1}^{r} \boldsymbol{x}_i \otimes (-\boldsymbol{v}_i^\star) \otimes (-\boldsymbol{w}_i^\star) + \boldsymbol{u}_i^\star \otimes \boldsymbol{y}_i \otimes (-\boldsymbol{w}_i^\star) + \boldsymbol{u}_i^\star \otimes (-\boldsymbol{v}_i^\star) \otimes \boldsymbol{z}_i \Big\|_{\mathrm{F}}^2
$$
$$
+ \frac{1}{3} \sum_{i=1}^{r} \lambda_i^{\star 2/3} \Big[\big(\langle \boldsymbol{x}_i, \boldsymbol{u}_i^\star \rangle - \langle \boldsymbol{y}_i, -\boldsymbol{v}_i^\star \rangle \big)^2 + \big(\langle \boldsymbol{x}_i, \boldsymbol{u}_i^\star \rangle - \langle \boldsymbol{z}_i, -\boldsymbol{w}_i^\star \rangle \big)^2 + \big(\langle \boldsymbol{y}_i, -\boldsymbol{v}_i^\star \rangle - \langle \boldsymbol{z}_i, -\boldsymbol{w}_i^\star \rangle \big)^2 \Big], \quad (287\text{a})
$$

$$
\beta_2 := \frac{1}{3p} \Big\| \mathcal{P}_\Omega \Big(\sum_{i=1}^{r} \boldsymbol{x}_i \otimes \boldsymbol{v}_i \otimes \boldsymbol{w}_i + \boldsymbol{u}_i \otimes \boldsymbol{y}_i \otimes \boldsymbol{w}_i + \boldsymbol{u}_i \otimes \boldsymbol{v}_i \otimes \boldsymbol{z}_i \Big) \Big\|_{\mathrm{F}}^2
$$
$$
- \frac{1}{3} \Big\| \sum_{i=1}^{r} \boldsymbol{x}_i \otimes (-\boldsymbol{v}_i^\star) \otimes (-\boldsymbol{w}_i^\star) + \boldsymbol{u}_i^\star \otimes \boldsymbol{y}_i \otimes (-\boldsymbol{w}_i^\star) + \boldsymbol{u}_i^\star \otimes (-\boldsymbol{v}_i^\star) \otimes \boldsymbol{z}_i \Big\|_{\mathrm{F}}^2, \quad (287\text{b})
$$

$$
\beta_3 := \frac{2}{3p} \Big\langle \sum_{i=1}^{r} \boldsymbol{x}_i \otimes \boldsymbol{y}_i \otimes \boldsymbol{w}_i + \boldsymbol{x}_i \otimes \boldsymbol{v}_i \otimes \boldsymbol{z}_i + \boldsymbol{u}_i \otimes \boldsymbol{y}_i \otimes \boldsymbol{z}_i, \mathcal{P}_\Omega \Big(\sum_{i=1}^{r} \boldsymbol{u}_i \otimes \boldsymbol{v}_i \otimes \boldsymbol{w}_i - \boldsymbol{T}^\star \Big) \Big\rangle, \quad (287\text{c})
$$

$$
\beta_4 := \frac{1}{3} \sum_{i=1}^{r} \big(\alpha_i - \lambda_i^{\star 2/3} \big) \Big[\big(\langle \boldsymbol{x}_i, \boldsymbol{u}_i^\star \rangle - \langle \boldsymbol{y}_i, -\boldsymbol{v}_i^\star \rangle \big)^2 + \big(\langle \boldsymbol{x}_i, \boldsymbol{u}_i^\star \rangle - \langle \boldsymbol{z}_i, -\boldsymbol{w}_i^\star \rangle \big)^2 + \big(\langle \boldsymbol{y}_i, -\boldsymbol{v}_i^\star \rangle - \langle \boldsymbol{z}_i, -\boldsymbol{w}_i^\star \rangle \big)^2 \Big]
$$
$$
+ \frac{1}{3} \sum_{i=1}^{r} \alpha_i \Big[\big(\langle \boldsymbol{x}_i, \boldsymbol{\Delta}_i^{\boldsymbol{U}} \rangle - \langle \boldsymbol{y}_i, \boldsymbol{\Delta}_i^{\boldsymbol{V}} \rangle \big)^2 + \big(\langle \boldsymbol{x}_i, \boldsymbol{\Delta}_i^{\boldsymbol{U}} \rangle - \langle \boldsymbol{z}_i, \boldsymbol{\Delta}_i^{\boldsymbol{W}} \rangle \big)^2 + \big(\langle \boldsymbol{y}_i, \boldsymbol{\Delta}_i^{\boldsymbol{V}} \rangle - \langle \boldsymbol{z}_i, \boldsymbol{\Delta}_i^{\boldsymbol{W}} \rangle \big)^2 \Big]
$$
$$
+ \frac{2}{3} \sum_{i=1}^{r} \alpha_i \Big[\big(\langle \boldsymbol{x}_i, \boldsymbol{u}_i^\star \rangle - \langle \boldsymbol{y}_i, -\boldsymbol{v}_i^\star \rangle \big) \big(\langle \boldsymbol{x}_i, \boldsymbol{\Delta}_i^{\boldsymbol{U}} \rangle - \langle \boldsymbol{y}_i, \boldsymbol{\Delta}_i^{\boldsymbol{V}} \rangle \big) + \big(\langle \boldsymbol{x}_i, \boldsymbol{u}_i^\star \rangle - \langle \boldsymbol{z}_i, -\boldsymbol{w}_i^\star \rangle \big) \big(\langle \boldsymbol{x}_i, \boldsymbol{\Delta}_i^{\boldsymbol{U}} \rangle - \langle \boldsymbol{z}_i, \boldsymbol{\Delta}_i^{\boldsymbol{W}} \rangle \big)
$$
$$
+ \big(\langle \boldsymbol{y}_i, -\boldsymbol{v}_i^\star \rangle - \langle \boldsymbol{z}_i, -\boldsymbol{w}_i^\star \rangle \big) \big(\langle \boldsymbol{y}_i, \boldsymbol{\Delta}_i^{\boldsymbol{V}} \rangle - \langle \boldsymbol{z}_i, \boldsymbol{\Delta}_i^{\boldsymbol{W}} \rangle \big) \Big]
$$
$$
+ \frac{1}{6} \sum_{i=1}^{r} \alpha_i \Big[\big(\|\boldsymbol{u}_i\|_2^2 - \|\boldsymbol{v}_i\|_2^2 \big) \big(\|\boldsymbol{x}_i\|_2^2 - \|\boldsymbol{y}_i\|_2^2 \big) + \big(\|\boldsymbol{u}_i\|_2^2 - \|\boldsymbol{w}_i\|_2^2 \big) \big(\|\boldsymbol{x}_i\|_2^2 - \|\boldsymbol{z}_i\|_2^2 \big)
$$
$$
+ \big(\|\boldsymbol{v}_i\|_2^2 - \|\boldsymbol{w}_i\|_2^2 \big) \big(\|\boldsymbol{y}_i\|_2^2 - \|\boldsymbol{z}_i\|_2^2 \big) \Big]. \quad (287\text{d})
$$

In what follows, we shall demonstrate that β_1 is the dominant term, with the remaining terms being negligible compared to β_1. Here, we note that the proof idea is almost identical to that of the symmetric case (cf. Lemma 5.1 in Appendix 5.1). For the sake of conciseness, we will focus only on the part where the symmetric and the asymmetric cases differ, and omit the proof details when their analyses are similar.

Bounding β_1 Let us first expand

$$
\beta_1 = \frac{1}{3} \Big\| \sum_{i=1}^{r} \boldsymbol{x}_i \otimes \boldsymbol{v}_i^\star \otimes \boldsymbol{w}_i^\star - \boldsymbol{u}_i^\star \otimes \boldsymbol{y}_i \otimes \boldsymbol{w}_i^\star - \boldsymbol{u}_i^\star \otimes \boldsymbol{v}_i^\star \otimes \boldsymbol{z}_i \Big\|_{\mathrm{F}}^2
$$
$$
+ \frac{1}{3} \sum_{i=1}^{r} \lambda_i^{\star 2/3} \Big[\big(\langle \boldsymbol{x}_i, \boldsymbol{u}_i^\star \rangle + \langle \boldsymbol{y}_i, \boldsymbol{v}_i^\star \rangle \big)^2 + \big(\langle \boldsymbol{x}_i, \boldsymbol{u}_i^\star \rangle + \langle \boldsymbol{z}_i, \boldsymbol{w}_i^\star \rangle \big)^2 + \big(\langle \boldsymbol{y}_i, \boldsymbol{v}_i^\star \rangle - \langle \boldsymbol{z}_i, \boldsymbol{w}_i^\star \rangle \big)^2 \Big]
$$
$$
= \underbrace{\frac{1}{3} \Big\| \sum_{i=1}^{r} \boldsymbol{x}_i \otimes \boldsymbol{v}_i^\star \otimes \boldsymbol{w}_i^\star \Big\|_{\mathrm{F}}^2 + \frac{1}{3} \Big\| \sum_{i=1}^{r} \boldsymbol{u}_i^\star \otimes \boldsymbol{y}_i \otimes \boldsymbol{w}_i^\star \Big\|_{\mathrm{F}}^2 + \frac{1}{3} \Big\| \sum_{i=1}^{r} \boldsymbol{u}_i^\star \otimes \boldsymbol{v}_i^\star \otimes \boldsymbol{z}_i \Big\|_{\mathrm{F}}^2}_{=: \gamma_1}
$$

$$+ \frac{2}{3} \sum_{i=1}^{r} \lambda_i^{\star 2/3} \langle \boldsymbol{x}_i, \boldsymbol{u}_i^\star \rangle^2 + \frac{2}{3} \sum_{i=1}^{r} \lambda_i^{\star 2/3} \langle \boldsymbol{y}_i, \boldsymbol{v}_i^\star \rangle^2 + \frac{2}{3} \sum_{i=1}^{r} \lambda_i^{\star 2/3} \langle \boldsymbol{z}_i, \boldsymbol{w}_i^\star \rangle^2$$

$$\underbrace{\phantom{+ \frac{2}{3} \sum_{i=1}^{r} \lambda_i^{\star 2/3} \langle \boldsymbol{x}_i, \boldsymbol{u}_i^\star \rangle^2 + \frac{2}{3} \sum_{i=1}^{r} \lambda_i^{\star 2/3} \langle \boldsymbol{y}_i, \boldsymbol{v}_i^\star \rangle^2 + \frac{2}{3}}}_{=: \gamma_2}$$

$$+ -\frac{2}{3} \Big\langle \sum_{i=1}^{r} \boldsymbol{x}_i \otimes \boldsymbol{v}_i^\star \otimes \boldsymbol{w}_i^\star, \sum_{i=1}^{r} \boldsymbol{u}_i^\star \otimes \boldsymbol{y}_i \otimes \boldsymbol{w}_i^\star \Big\rangle + \frac{2}{3} \sum_{i=1}^{r} \lambda_i^{\star 2/3} \langle \boldsymbol{x}_i, \boldsymbol{u}_i^\star \rangle \langle \boldsymbol{y}_i, \boldsymbol{v}_i^\star \rangle$$

$$\underbrace{\phantom{+ -\frac{2}{3} \Big\langle \sum_{i=1}^{r} \boldsymbol{x}_i \otimes \boldsymbol{v}_i^\star \otimes \boldsymbol{w}_i^\star, \sum_{i=1}^{r} \boldsymbol{u}_i^\star \otimes \boldsymbol{y}_i \otimes \boldsymbol{w}_i^\star \Big\rangle}}_{=: \gamma_3}$$

$$+ -\frac{2}{3} \Big\langle \sum_{i=1}^{r} \boldsymbol{x}_i \otimes \boldsymbol{v}_i^\star \otimes \boldsymbol{w}_i^\star, \sum_{i=1}^{r} \boldsymbol{u}_i^\star \otimes \boldsymbol{v}_i^\star \otimes \boldsymbol{z}_i \Big\rangle + \frac{2}{3} \sum_{i=1}^{r} \lambda_i^{\star 2/3} \langle \boldsymbol{x}_i, \boldsymbol{u}_i^\star \rangle \langle \boldsymbol{z}_i, \boldsymbol{w}_i^\star \rangle$$

$$\underbrace{\phantom{+ -\frac{2}{3} \Big\langle \sum_{i=1}^{r} \boldsymbol{x}_i \otimes \boldsymbol{v}_i^\star \otimes \boldsymbol{w}_i^\star, \sum_{i=1}^{r} \boldsymbol{u}_i^\star \otimes \boldsymbol{v}_i^\star \otimes \boldsymbol{z}_i \Big\rangle}}_{=: \gamma_4}$$

$$+ \frac{2}{3} \Big\langle \sum_{i=1}^{r} \boldsymbol{u}_i^\star \otimes \boldsymbol{y}_i \otimes \boldsymbol{w}_i^\star, \sum_{i=1}^{r} \boldsymbol{u}_i^\star \otimes \boldsymbol{v}_i^\star \otimes \boldsymbol{z}_i \Big\rangle - \frac{2}{3} \sum_{i=1}^{r} \lambda_i^{\star 2/3} \langle \boldsymbol{y}_i, \boldsymbol{v}_i^\star \rangle \langle \boldsymbol{z}_i, \boldsymbol{w}_i^\star \rangle,$$

$$\underbrace{\phantom{+ \frac{2}{3} \Big\langle \sum_{i=1}^{r} \boldsymbol{u}_i^\star \otimes \boldsymbol{y}_i \otimes \boldsymbol{w}_i^\star, \sum_{i=1}^{r} \boldsymbol{u}_i^\star \otimes \boldsymbol{v}_i^\star \otimes \boldsymbol{z}_i \Big\rangle}}_{=: \gamma_5}$$

leaving us with five terms to control.

1. Let us begin with γ_1. Observe that we can express

$$\Big\| \sum_{i=1}^{r} \boldsymbol{x}_i \otimes \boldsymbol{v}_i^\star \otimes \boldsymbol{w}_i^\star \Big\|_{\mathrm{F}} = \big\| \boldsymbol{X} \boldsymbol{H}^\top \big\|_{\mathrm{F}}, \qquad \text{with} \quad \boldsymbol{H} = [\boldsymbol{v}_1^\star \otimes \boldsymbol{w}_1^\star, \cdots, \boldsymbol{v}_r^\star \otimes \boldsymbol{w}_r^\star] \in \mathbb{R}^{d^2 \times r}$$

Arguing similarly as in the proof of Lemma D.1, one can derive

$$\big| \sigma_{\min}(\boldsymbol{H}) - \lambda_{\min}^{\star 2/3} \big| \ll \lambda_{\min}^{\star 2/3} \qquad \text{and} \qquad \big| \sigma_{\max}(\boldsymbol{H}) - \lambda_{\max}^{\star 2/3} \big| \ll \lambda_{\min}^{\star 2/3},$$

provided that $r \ll d_{\min}/\mu$. This leads to the following inequalities

$$\frac{19}{20} \lambda_{\min}^{\star 2/3} \|\boldsymbol{X}\|_{\mathrm{F}} \le \sigma_{\min}(\boldsymbol{H}) \|\boldsymbol{X}\|_{\mathrm{F}} \le \Big\| \sum_{i=1}^{r} \boldsymbol{x}_i \otimes \boldsymbol{v}_i^\star \otimes \boldsymbol{w}_i^\star \Big\|_{\mathrm{F}} \le \sigma_{\max}(\boldsymbol{H}) \|\boldsymbol{X}\|_{\mathrm{F}} \le \frac{11}{10} \lambda_{\max}^{\star 2/3} \|\boldsymbol{X}\|_{\mathrm{F}}.$$

Clearly, the same argument reveals that

$$\frac{19}{20} \lambda_{\min}^{\star 2/3} \|\boldsymbol{Y}\|_{\mathrm{F}} \le \Big\| \sum_{i=1}^{r} \boldsymbol{u}_i^\star \otimes \boldsymbol{y}_i \otimes \boldsymbol{w}_i^\star \Big\|_{\mathrm{F}} \le \frac{11}{10} \lambda_{\max}^{\star 2/3} \|\boldsymbol{Y}\|_{\mathrm{F}},$$

$$\frac{19}{20} \lambda_{\min}^{\star 2/3} \|\boldsymbol{Z}\|_{\mathrm{F}} \le \Big\| \sum_{i=1}^{r} \boldsymbol{u}_i^\star \otimes \boldsymbol{v}_i^\star \otimes \boldsymbol{z}_i \Big\|_{\mathrm{F}} \le \frac{11}{10} \lambda_{\max}^{\star 2/3} \|\boldsymbol{Z}\|_{\mathrm{F}}.$$

Combining these bounds, we reach

$$\gamma_1 \le \frac{3}{2} \lambda_{\max}^{\star 4/3} \left(\|\boldsymbol{X}\|_{\mathrm{F}}^2 + \|\boldsymbol{Y}\|_{\mathrm{F}}^2 + \|\boldsymbol{Z}\|_{\mathrm{F}}^2 \right).$$

2. We now move on to γ_2. Recall our assumption that $\|\boldsymbol{u}_i^\star\|_2 = \|\boldsymbol{v}_i^\star\|_2 = \|\boldsymbol{w}_i^\star\|_2 = \lambda_i^{\star 1/3}$ for all $1 \le i \le r$. It follows from the Cauchy-Schwartz inequality that

$$0 \le \gamma_2 \le \frac{2}{3} \sum_{i=1}^{r} \lambda_i^{\star 2/3} \|\boldsymbol{x}_i\|^2 \|\boldsymbol{u}_i^\star\|_2^2 + \frac{2}{3} \sum_{i=1}^{r} \lambda_i^{\star 2/3} \|\boldsymbol{y}_i\|^2 \|\boldsymbol{v}_i^\star\|_2^2 + \frac{2}{3} \sum_{i=1}^{r} \lambda_i^{\star 2/3} \|\boldsymbol{z}_i\|_2^2 \|\boldsymbol{w}_i^\star\|_2^2$$

$$\le \frac{2}{3} \lambda_{\max}^{\star 4/3} \left(\|\boldsymbol{X}\|_{\mathrm{F}}^2 + \|\boldsymbol{Y}\|_{\mathrm{F}}^2 + \|\boldsymbol{Z}\|_{\mathrm{F}}^2 \right).$$

3. Turning to γ_3, one can straightforwardly bound

$$|\gamma_3| \overset{\text{(i)}}{=} \frac{2}{3} \Big| \sum_{i_1 \ne i_2} \langle \boldsymbol{x}_{i_1}, \boldsymbol{u}_{i_2}^\star \rangle \langle \boldsymbol{v}_{i_1}^\star, \boldsymbol{y}_{i_2} \rangle \langle \boldsymbol{w}_{i_1}^\star, \boldsymbol{w}_{i_2}^\star \rangle \Big|$$

$$\overset{\text{(ii)}}{\leq} \max_{i_1 \neq i_2} \left| \langle \boldsymbol{w}_{i_1}^\star, \boldsymbol{w}_{i_2}^\star \rangle \right| \Big(\sum_{i=1}^r \|\boldsymbol{x}_i\|_2 \|\boldsymbol{v}_i^\star\|_2 \Big) \Big(\sum_{i=1}^r \|\boldsymbol{u}_i^\star\|_2 \|\boldsymbol{y}_i\|_2 \Big)$$

$$\overset{\text{(iii)}}{\leq} \max_{i_1 \neq i_2} \left| \langle \boldsymbol{w}_{i_1}^\star, \boldsymbol{w}_{i_2}^\star \rangle \right| \|\boldsymbol{U}^\star\|_{\mathrm{F}} \|\boldsymbol{V}^\star\|_{\mathrm{F}} \|\boldsymbol{X}\|_{\mathrm{F}} \|\boldsymbol{Y}\|_{\mathrm{F}}$$

$$\overset{\text{(iv)}}{\leq} r \sqrt{\frac{\mu}{d}} \lambda_{\max}^{\star 4/3} \|\boldsymbol{X}\|_{\mathrm{F}} \|\boldsymbol{Y}\|_{\mathrm{F}}$$

$$\overset{\text{(v)}}{\ll} \lambda_{\max}^{\star 4/3} \|\boldsymbol{X}\|_{\mathrm{F}} \|\boldsymbol{Y}\|_{\mathrm{F}}.$$

Here, we have used the fact that $\|\boldsymbol{w}_i^\star\|_2 = \lambda_i^{\star 1/3}$ in (i); the inequalities (ii) and (iii) arise from the Cauchy-Schwartz inequality; (iv) follows from (280c); (v) holds as long as $r \ll \sqrt{d_{\min}/\mu}$. In a similar manner, one can easily verify that

$$|\gamma_4| \ll \lambda_{\max}^{\star 4/3} \|\boldsymbol{X}\|_{\mathrm{F}} \|\boldsymbol{Z}\|_{\mathrm{F}} \qquad \text{and} \qquad |\gamma_5| \ll \lambda_{\max}^{\star 4/3} \|\boldsymbol{Y}\|_{\mathrm{F}} \|\boldsymbol{Z}\|_{\mathrm{F}}.$$

It then follows from the AM-GM inequality that

$$|\gamma_3| + |\gamma_4| + |\gamma_5| \ll \lambda_{\max}^{\star 4/3} \left(\|\boldsymbol{X}\|_{\mathrm{F}}^2 + \|\boldsymbol{Y}\|_{\mathrm{F}}^2 + \|\boldsymbol{Z}\|_{\mathrm{F}}^2 \right).$$

4. Putting the above bounds together allows us to bound β_1 as follows

$$\beta_1 \leq \gamma_1 + \gamma_2 + |\gamma_3| + |\gamma_4| + |\gamma_5| \leq \frac{7}{2} \lambda_{\max}^{\star 4/3} \left(\|\boldsymbol{X}\|_{\mathrm{F}}^2 + \|\boldsymbol{Y}\|_{\mathrm{F}}^2 + \|\boldsymbol{Z}\|_{\mathrm{F}}^2 \right);$$

$$\beta_1 \geq \gamma_1 - |\gamma_3| - |\gamma_4| - |\gamma_5| \geq \frac{9}{10} \lambda_{\min}^{\star 4/3} \left(\|\boldsymbol{X}\|_{\mathrm{F}}^2 + \|\boldsymbol{Y}\|_{\mathrm{F}}^2 + \|\boldsymbol{Z}\|_{\mathrm{F}}^2 \right).$$

Bounding β_2 To control β_2, we note that β_2 involves two quantities: (1) the deviation of $\frac{1}{p} \left\| \mathcal{P}_\Omega \Big(\sum_{i=1}^r \boldsymbol{x}_i \otimes \right.$ $\boldsymbol{v}_i \otimes \boldsymbol{w}_i + \boldsymbol{u}_i \otimes \boldsymbol{y}_i \otimes \boldsymbol{w}_i + \boldsymbol{u}_i \otimes \boldsymbol{v}_i \otimes \boldsymbol{z}_i \Big) \left. \right\|_{\mathrm{F}}^2$ from its expectation with respect to Ω; (2) the distance between $\sum_{i=1}^r \boldsymbol{x}_i \otimes \boldsymbol{v}_i \otimes \boldsymbol{w}_i + \boldsymbol{u}_i \otimes \boldsymbol{y}_i \otimes \boldsymbol{w}_i + \boldsymbol{u}_i \otimes \boldsymbol{v}_i \otimes \boldsymbol{z}_i$ and $\sum_{i=1}^r \boldsymbol{x}_i \otimes \boldsymbol{v}_i^\star \otimes \boldsymbol{w}_i - \boldsymbol{u}_i^\star \otimes \boldsymbol{y}_i \otimes \boldsymbol{w}_i^\star - \boldsymbol{u}_i^\star \otimes \boldsymbol{v}_i^\star \otimes \boldsymbol{z}_i$. The first term can be shown to be exceedingly small under our sample size condition with the help of [YZ16, Lemma 5], while the second term is also guaranteed to be sufficiently small by the assumptions of the error of $\boldsymbol{U}, \boldsymbol{V}$ and $\boldsymbol{W}$ in (285). Therefore, one can apply an analogous argument for α_1 and α_2 in the proof of Lemma 5.1 in Appendix A to derive: with probability at least $1 - O(d_{\min}^{-10})$ one has

$$\left| \beta_2 \right| \ll \lambda_{\min}^{\star 4/3} \left(\|\boldsymbol{X}\|_{\mathrm{F}}^2 + \|\boldsymbol{Y}\|_{\mathrm{F}}^2 + \|\boldsymbol{Z}\|_{\mathrm{F}}^2 \right),$$

provided the sampling rate exceeds $p \gg \frac{\mu^2 r^2 d_{\max} \log^2 d_{\max}}{d_1 d_2 d_3}$.

Bounding β_3 By the definition of the operator norm, one can bound

$$|\beta_3| \lesssim \left\| p^{-1} \mathcal{P}_\Omega \Big(\sum_{i=1}^r \boldsymbol{u}_i \otimes \boldsymbol{v}_i \otimes \boldsymbol{w}_i - \boldsymbol{T}^\star \Big) \right\| \cdot \sum_{i=1}^r \Big(\|\boldsymbol{w}_i\|_2 \|\boldsymbol{x}_i\|_2 \|\boldsymbol{y}_i\|_2 + \|\boldsymbol{v}_i\|_2 \|\boldsymbol{x}_i\|_2 \|\boldsymbol{z}_i\|_2 + \|\boldsymbol{u}_i\|_2 \|\boldsymbol{y}_i\|_2 \|\boldsymbol{z}_i\|_2 \Big)$$

$$\lesssim \lambda_{\max}^{2/3} \left\| \sum_{i=1}^r \boldsymbol{u}_i \otimes \boldsymbol{v}_i \otimes \boldsymbol{w}_i - \boldsymbol{T}^\star \right\| \left\| p^{-1} \mathcal{P}_\Omega \big(\mathbf{1}_{d_1} \otimes \mathbf{1}_{d_2} \otimes \mathbf{1}_{d_3} \big) \right\| \left(\|\boldsymbol{X}\|_{\mathrm{F}}^2 + \|\boldsymbol{Y}\|_{\mathrm{F}}^2 + \|\boldsymbol{Z}\|_{\mathrm{F}}^2 \right).$$

One can easily adapt the proof of Lemma D.2 to show that, with probability at with probability at least $1 - O(d_{\min}^{-10})$,

$$\left\| \big(p^{-1} \mathcal{P}_\Omega - \mathcal{I} \big) \big(\mathbf{1}_{d_1} \otimes \mathbf{1}_{d_2} \otimes \mathbf{1}_{d_3} \big) \right\| \lesssim \frac{\log^3 d_{\max}}{p} + \sqrt{\frac{d_{\max} \log^5 d_{\max}}{p}} \lesssim \sqrt{d_1 d_2 d_3} = \left\| \mathbf{1}_{d_1} \otimes \mathbf{1}_{d_2} \otimes \mathbf{1}_{d_3} \right\|$$

holds as long as $p \gg \max \left\{ \frac{\log^3 d_{\max}}{\sqrt{d_1 d_2 d_3}}, \frac{d_{\max} \log^5 d_{\max}}{d_1 d_2 d_3} \right\}$. We can then adapt the proof for bounding α_3 in the proof of Lemma 5.1 in Appendix A to derive that with probability at least $1 - O(d_{\min}^{-10})$,

$$\left| \beta_3 \right| \ll \lambda_{\min}^{\star 4/3} \left(\|\boldsymbol{X}\|_{\mathrm{F}}^2 + \|\boldsymbol{Y}\|_{\mathrm{F}}^2 + \|\boldsymbol{Z}\|_{\mathrm{F}}^2 \right).$$

Bounding β_4 It remains to control β_4. By symmetry, it suffices to consider the following terms:

$$\gamma_1 := \sum_{i=1}^{r} \big(\alpha_i - \lambda_i^{\star 2/3}\big)\big(\langle \boldsymbol{x}_i, \boldsymbol{u}_i^\star \rangle + \langle \boldsymbol{y}_i, \boldsymbol{v}_i^\star \rangle\big)^2,$$

$$\gamma_2 := \sum_{i=1}^{r} \alpha_i\big(\langle \boldsymbol{x}_i, \boldsymbol{\Delta}_i^{\boldsymbol{U}} \rangle - \langle \boldsymbol{y}_i, \boldsymbol{\Delta}_i^{\boldsymbol{V}} \rangle\big)^2,$$

$$\gamma_3 := \sum_{i=1}^{r} \alpha_i\big(\langle \boldsymbol{x}_i, \boldsymbol{u}_i^\star \rangle + \langle \boldsymbol{y}_i, \boldsymbol{v}_i^\star \rangle\big)\big(\langle \boldsymbol{x}_i, \boldsymbol{\Delta}_i^{\boldsymbol{U}} \rangle - \langle \boldsymbol{y}_i, \boldsymbol{\Delta}_i^{\boldsymbol{V}} \rangle\big),$$

$$\gamma_4 := \sum_{i=1}^{r} \alpha_i\big(\|\boldsymbol{u}_i\|_2^2 - \|\boldsymbol{v}_i\|_2^2\big)\big(\|\boldsymbol{x}_i\|_2^2 - \|\boldsymbol{y}_i\|_2^2\big).$$

In the following, we shall bound these four terms separately. By the assumptions that $\boldsymbol{S}^{(1)} = \boldsymbol{I}_r$, $\boldsymbol{S}^{(2)} = \boldsymbol{S}^{(3)} = -\boldsymbol{I}_r$, $\delta \ll 1$ and $\kappa \asymp 1$, one has

$$\|\boldsymbol{\Delta}_i^{\boldsymbol{U}}\|_2 \leq \|\boldsymbol{U} - \boldsymbol{U}^\star\|_{\mathrm{F}} \leq \delta\,\|\boldsymbol{U}^\star\|_{\mathrm{F}} \leq \delta\sqrt{r}\lambda_{\max}^{\star 1/3} \ll \lambda_{\min}^{\star 1/3} \leq \|\boldsymbol{u}_i^\star\|_2; \tag{288a}$$

$$\|\boldsymbol{\Delta}_i^{\boldsymbol{V}}\|_2 \leq \|-\boldsymbol{V} - \boldsymbol{V}^\star\|_{\mathrm{F}} \leq \delta\,\|\boldsymbol{V}^\star\|_{\mathrm{F}} \leq \delta\sqrt{r}\lambda_{\max}^{\star 1/3} \ll \lambda_{\min}^{\star 1/3} \leq \|\boldsymbol{v}_i^\star\|_2; \tag{288b}$$

$$\|\boldsymbol{\Delta}_i^{\boldsymbol{W}}\|_2 \leq \|-\boldsymbol{W} - \boldsymbol{W}^\star\|_{\mathrm{F}} \leq \delta\,\|\boldsymbol{W}^\star\|_{\mathrm{F}} \leq \delta\sqrt{r}\lambda_{\max}^{\star 1/3} \ll \lambda_{\min}^{\star 1/3} \leq \|\boldsymbol{w}_i^\star\|_2. \tag{288c}$$

In addition, for each $1 \leq i \leq r$, one has

$$\frac{9}{10}\lambda_{\min}^{\star 2/3} \leq \lambda_i - \big|\alpha_i - \lambda_i^{\star 2/3}\big| \leq \alpha_i \leq \lambda_i + \big|\alpha_i - \lambda_i^{\star 2/3}\big| \leq \frac{11}{10}\lambda_{\max}^{\star 2/3} \tag{289}$$

by virtue of the condition that $\big|\alpha_i - \lambda_i^{\star 2/3}\big| \ll \lambda_{\min}^{\star 2/3}$.

1. Let us start with γ_1. By the condition that $\max_i \big|\alpha_i - \lambda_i^{\star 2/3}\big| \ll \lambda_{\min}^{\star 2/3}$, it is easily seen that

$$|\gamma_1| \lesssim \sum_{i=1}^{r} \big|\alpha_i - \lambda_i^{\star 2/3}\big|\big(\langle \boldsymbol{x}_i, \boldsymbol{u}_i^\star \rangle^2 + \langle \boldsymbol{y}_i, \boldsymbol{v}_i^\star \rangle^2\big) \leq \max_i \big|\alpha_i - \lambda_i^{\star 2/3}\big| \sum_{i=1}^{r} \big(\|\boldsymbol{x}_i\|_2^2\|\boldsymbol{u}_i^\star\|_2^2 + \|\boldsymbol{y}_i\|_2^2\|\boldsymbol{v}_i^\star\|_2^2\big)$$

$$\leq \max_i \big|\alpha_i - \lambda_i^{\star 2/3}\big|\big(\|\boldsymbol{u}_i^\star\|_2^2 + \|\boldsymbol{v}_i^\star\|_2^2\big)\big(\|\boldsymbol{X}\|_{\mathrm{F}}^2 + \|\boldsymbol{Y}\|_{\mathrm{F}}^2\big) \ll \lambda_{\min}^{\star 4/3}\big(\|\boldsymbol{X}\|_{\mathrm{F}}^2 + \|\boldsymbol{Y}\|_{\mathrm{F}}^2\big),$$

   with the proviso that $\kappa \asymp 1$.

2. As for γ_2, one can bound

$$|\gamma_2| \lesssim \sum_{i=1}^{r} \alpha_i\big(\langle \boldsymbol{x}_i, \boldsymbol{\Delta}_i^{\boldsymbol{U}} \rangle^2 - \langle \boldsymbol{y}_i, \boldsymbol{\Delta}_i^{\boldsymbol{V}} \rangle^2\big) \leq \max_i \alpha_i \sum_{i=1}^{r} \big(\|\boldsymbol{x}_i\|_2^2\|\boldsymbol{\Delta}_i^{\boldsymbol{U}}\|_2^2 + \|\boldsymbol{y}_i\|_2^2\|\boldsymbol{\Delta}_i^{\boldsymbol{V}}\|_2^2\big)$$

$$\leq \max_i \alpha_i\big(\|\boldsymbol{\Delta}_i^{\boldsymbol{U}}\|_2^2 + \|\boldsymbol{\Delta}_i^{\boldsymbol{V}}\|_2^2\big)\big(\|\boldsymbol{X}\|_{\mathrm{F}}^2 + \|\boldsymbol{Y}\|_{\mathrm{F}}^2\big) \ll \lambda_{\min}^{\star 4/3}\big(\|\boldsymbol{X}\|_{\mathrm{F}}^2 + \|\boldsymbol{Y}\|_{\mathrm{F}}^2\big).$$

   Here, we use (288a), (288b), (289) and $\kappa \asymp 1$ in the last step.

3. Turning to γ_3, one can develop a similar bound as follows

$$|\gamma_3| \leq \sqrt{\sum_{i=1}^{r} \alpha_i\big(\langle \boldsymbol{x}_i, \boldsymbol{u}_i^\star \rangle + \langle \boldsymbol{y}_i, \boldsymbol{v}_i^\star \rangle\big)^2} \cdot \sqrt{\sum_{i=1}^{r} \alpha_i\big(\langle \boldsymbol{x}_i, \boldsymbol{\Delta}_i^{\boldsymbol{U}} \rangle - \langle \boldsymbol{y}_i, \boldsymbol{\Delta}_i^{\boldsymbol{V}} \rangle\big)^2}$$

$$\lesssim \max_i \alpha_i \sqrt{\big(\|\boldsymbol{u}_i^\star\|_2^2 + \|\boldsymbol{v}_i^\star\|_2^2\big)\big(\|\boldsymbol{X}\|_{\mathrm{F}}^2 + \|\boldsymbol{Y}\|_{\mathrm{F}}^2\big)} \cdot \sqrt{\big(\|\boldsymbol{\Delta}_i^{\boldsymbol{U}}\|_2^2 + \|\boldsymbol{\Delta}_i^{\boldsymbol{V}}\|_2^2\big)\big(\|\boldsymbol{X}\|_{\mathrm{F}}^2 + \|\boldsymbol{Y}\|_{\mathrm{F}}^2\big)}$$

$$\ll \lambda_{\min}^{\star 4/3}\big(\|\boldsymbol{X}\|_{\mathrm{F}}^2 + \|\boldsymbol{Y}\|_{\mathrm{F}}^2\big).$$

4. With regards to γ_4, we can expand

$$|\gamma_4| = \sum_{i=1}^{r} \alpha_i \left(\|\boldsymbol{u}_i^\star + \boldsymbol{\Delta}_i^{\boldsymbol{U}}\|_2^2 - \|-\boldsymbol{v}_i^\star + \boldsymbol{\Delta}_i^{\boldsymbol{V}}\|_2^2 \right) \left(\|\boldsymbol{x}_i\|_2^2 - \|\boldsymbol{y}_i\|_2^2 \right)$$

$$= \sum_{i=1}^{r} \alpha_i \left(\|\boldsymbol{u}_i^\star\|_2^2 + 2\langle \boldsymbol{u}_i^\star, \boldsymbol{\Delta}_i^{\boldsymbol{U}} \rangle + \|\boldsymbol{\Delta}_i^{\boldsymbol{U}}\|_2^2 - \|\boldsymbol{v}_i^\star\|_2^2 + 2\langle \boldsymbol{v}_i^\star, \boldsymbol{\Delta}_i^{\boldsymbol{V}} \rangle - \|\boldsymbol{\Delta}_i^{\boldsymbol{V}}\|_2^2 \right) \left(\|\boldsymbol{x}_i\|_2^2 - \|\boldsymbol{y}_i\|_2^2 \right)$$

$$= \sum_{i=1}^{r} \alpha_i \left(2\langle \boldsymbol{u}_i^\star, \boldsymbol{\Delta}_i^{\boldsymbol{U}} \rangle + \|\boldsymbol{\Delta}_i^{\boldsymbol{U}}\|_2^2 + 2\langle \boldsymbol{v}_i^\star, \boldsymbol{\Delta}_i^{\boldsymbol{V}} \rangle - \|\boldsymbol{\Delta}_i^{\boldsymbol{V}}\|_2^2 \right) \left(\|\boldsymbol{x}_i\|_2^2 - \|\boldsymbol{y}_i\|_2^2 \right),$$

where the last step follows from the assumption $\|\boldsymbol{u}_i^\star\|_2 = \|\boldsymbol{v}_i^\star\|_2 = \|\boldsymbol{w}_i^\star\|_2$. It follows that

$$\left| \sum_{i=1}^{r} \alpha_i \left(\|\boldsymbol{u}_i\|_2^2 - \|\boldsymbol{v}_i\|_2^2 \right) \left(\|\boldsymbol{x}_i\|_2^2 - \|\boldsymbol{y}_i\|_2^2 \right) \right| \lesssim \sum_{i=1}^{r} \alpha_i \left| \langle \boldsymbol{u}_i^\star, \boldsymbol{\Delta}_i^{\boldsymbol{U}} \rangle + \|\boldsymbol{\Delta}_i^{\boldsymbol{U}}\|_2^2 + \langle \boldsymbol{v}_i^\star, \boldsymbol{\Delta}_i^{\boldsymbol{V}} \rangle + \|\boldsymbol{\Delta}_i^{\boldsymbol{V}}\|_2^2 \right| \left(\|\boldsymbol{x}_i\|_2^2 + \|\boldsymbol{y}_i\|_2^2 \right)$$

$$\leq \max_i \alpha_i \left(\|\boldsymbol{u}_i^\star\|_2 \|\boldsymbol{\Delta}_i^{\boldsymbol{U}}\|_2 + \|\boldsymbol{v}_i^\star\|_2 \|\boldsymbol{\Delta}_i^{\boldsymbol{V}}\|_2 \right) \sum_{i=1}^{r} \left(\|\boldsymbol{x}_i\|_2^2 + \|\boldsymbol{y}_i\|_2^2 \right)$$

$$\ll \lambda_{\min}^{\star 4/3} \left(\|\boldsymbol{X}\|_{\mathrm{F}}^2 + \|\boldsymbol{Y}\|_{\mathrm{F}}^2 \right),$$

where we have used the conditions that $\max_i \alpha_i \lesssim \lambda_{\max}^{\star 2/3}$ and $\kappa \asymp 1$.

5. It is not hard to check that similar bounds also hold for the remaining terms in β_4. As a result, we arrive at

$$|\beta_4| \ll \lambda_{\min}^{\star 4/3} \left(\|\boldsymbol{X}\|_{\mathrm{F}}^2 + \|\boldsymbol{Y}\|_{\mathrm{F}}^2 + \|\boldsymbol{Z}\|_{\mathrm{F}}^2 \right).$$

Combining the previous bounds on $\beta_1, \beta_2, \beta_3$ and β_4 Putting the above estimates together, we conclude that with probability at least $1 - O(d_{\min}^{-10})$, one has

$$\mathsf{vec}\left(\begin{bmatrix} \boldsymbol{X} \\ \boldsymbol{Y} \\ \boldsymbol{Z} \end{bmatrix} \right)^{\top} \nabla^2 g_{\mathsf{clean}}(\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}) \, \mathsf{vec}\left(\begin{bmatrix} \boldsymbol{X} \\ \boldsymbol{Y} \\ \boldsymbol{Z} \end{bmatrix} \right) \geq \beta_1 - |\beta_2| - |\beta_3| - |\beta_4| \geq \frac{1}{2} \lambda_{\min}^{\star 4/3} \left(\|\boldsymbol{X}\|_{\mathrm{F}}^2 + \|\boldsymbol{Y}\|_{\mathrm{F}}^2 + \|\boldsymbol{Z}\|_{\mathrm{F}}^2 \right)$$

$$\mathsf{vec}\left(\begin{bmatrix} \boldsymbol{X} \\ \boldsymbol{Y} \\ \boldsymbol{Z} \end{bmatrix} \right)^{\top} \nabla^2 g_{\mathsf{clean}}(\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}) \, \mathsf{vec}\left(\begin{bmatrix} \boldsymbol{X} \\ \boldsymbol{Y} \\ \boldsymbol{Z} \end{bmatrix} \right) \leq \beta_1 + |\beta_2| + |\beta_3| + |\beta_4| \leq 4\lambda_{\max}^{\star 4/3} \left(\|\boldsymbol{X}\|_{\mathrm{F}}^2 + \|\boldsymbol{Y}\|_{\mathrm{F}}^2 + \|\boldsymbol{Z}\|_{\mathrm{F}}^2 \right)$$

as claimed.

References

[AFWZ17] E. Abbe, J. Fan, K. Wang, and Y. Zhong. Entrywise eigenvector analysis of random matrices with low expected rank. *arXiv:1709.09565*, 2017.

[AGH+14] A. Anandkumar, R. Ge, D. Hsu, S. M. Kakade, and M. Telgarsky. Tensor decompositions for learning latent variable models. *The Journal of Machine Learning Research*, 15(1):2773–2832, 2014.

[AGJ14] A. Anandkumar, R. Ge, and M. Janzamin. Guaranteed non-orthogonal tensor decomposition via alternating rank-1 updates. *arXiv preprint arXiv:1402.5180*, 2014.

[AGJ15] A. Anandkumar, R. Ge, and M. Janzamin. Learning overcomplete latent variable models through tensor methods. In *Proceedings of the Conference on Learning Theory*, pages 36–112, 2015.

[AGJ17]   A. Anandkumar, R. Ge, and M. Janzamin. Analyzing tensor power method dynamics in over-complete regime. *The Journal of Machine Learning Research*, 18(1):752–791, 2017.

[AW17]   M. Ashraphijuo and X. Wang. Fundamental conditions for low-CP-rank tensor completion. *The Journal of Machine Learning Research*, 18(1):2116–2145, 2017.

[BM16]   B. Barak and A. Moitra. Noisy tensor completion via the sum-of-squares hierarchy. In *Conference on Learning Theory*, pages 417–445, 2016.

[Bub15]   S. Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

[CC14]   Y. Chen and Y. Chi. Robust spectral compressed sensing via structured matrix completion. *IEEE Transactions on Information Theory*, 60(10):6576 – 6601, 2014.

[CC17]   Y. Chen and E. J. Candès. Solving random quadratic systems of equations is nearly as easy as solving linear systems. *Comm. Pure Appl. Math.*, 70(5):822–883, 2017.

[CC18]   Y. Chen and Y. Chi. Harnessing structures in big data via guaranteed low-rank matrix estimation: Recent theory and fast algorithms via convex and nonconvex optimization. *IEEE Signal Processing Magazine*, 35(4):14–31, July 2018.

[CCF$^+$19]   Y. Chen, Y. Chi, J. Fan, C. Ma, and Y. Yan. Noisy matrix completion: Understanding statistical guarantees for convex relaxation via nonconvex optimization. *arXiv preprint arXiv:1902.07698, accepted to SIAM Journal on Optimization*, 2019.

[CCFM19]   Y. Chen, Y. Chi, J. Fan, and C. Ma. Gradient descent with random initialization: Fast global convergence for nonconvex phase retrieval. *Mathematical Programming*, 176(1-2):5–37, July 2019.

[CFMW19]   Y. Chen, J. Fan, C. Ma, and K. Wang. Spectral method and regularized MLE are both optimal for top-K ranking. *Annals of Statistics*, 47(4):2204–2235, August 2019.

[CFMY19]   Y. Chen, J. Fan, C. Ma, and Y. Yan. Inference and uncertainty quantification for noisy matrix completion. *arXiv:1906.04159, accepted to the Proceedings of the National Academy of Sciences (PNAS)*, 2019.

[CKK$^+$97]   C. A. Cocosco, V. Kollokian, R. K.-S. Kwan, G. B. Pike, and A. C. Evans. Brainweb: Online interface to a 3d mri simulated brain database. In *NeuroImage*. Citeseer, 1997.

[CLC19]   Y. Chi, Y. Lu, and Y. Chen. Nonconvex optimization meets low-rank matrix factorization: An overview. *IEEE Transactions on Signal Processing*, 67(20):5239–5269, 2019.

[CLC$^+$20]   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. *The Annals of Statistics,* to appear, 2020.

[CLL19]   J. Chen, D. Liu, and X. Li. Nonconvex rectangular matrix completion via gradient descent without $\ell_{2,\infty}$ regularization. *arXiv:1901.06116v1*, 2019.

[CLPC19]   C. Cai, G. Li, H. V. Poor, and Y. Chen. Nonconvex low-rank tensor completion from noisy data. In *Advances in Neural Information Processing Systems*, pages 1863–1874, 2019.

[CLS15]   E. J. Candes, X. Li, and M. Soltanolkotabi. Phase retrieval via wirtinger flow: Theory and algorithms. *IEEE Transactions on Information Theory*, 61(4):1985–2007, 2015.

[CR09]   E. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772, April 2009.

[CRY19]   H. Chen, G. Raskutti, and M. Yuan. Non-convex projected gradient descent for generalized low-rank tensor regression. *The Journal of Machine Learning Research*, 20(1):172–208, 2019.

[CW15]   Y. Chen and M. J. Wainwright. Fast low-rank estimation by projected gradient descent: General statistical and algorithmic guarantees. *arXiv:1509.03025*, 2015.

[CZA⁺17]   J. Y. Cheng, T. Zhang, M. T. Alley, M. Uecker, M. Lustig, J. M. Pauly, and S. S. Vasanawala. Comprehensive multi-dimensional MRI for the simultaneous assessment of cardiopulmonary anatomy and physiology. *Scientific reports*, 7(1):5330, 2017.

[DBBG19]   D. Driggs, S. Becker, and J. Boyd-Graber. Tensor robust principal component analysis: Better recovery with atomic norm regularization. *arXiv preprint arXiv:1901.10991*, 2019.

[DC18]   L. Ding and Y. Chen. The leave-one-out approach for matrix completion: Primal and dual analysis. *arXiv preprint arXiv:1803.07554*, 2018.

[DR16]   M. A. Davenport and J. Romberg. An overview of low-rank matrix recovery from incomplete observations. *IEEE Journal of Selected Topics in Signal Processing*, 10(4):608–622, 2016.

[EAHK13]   G. Ely, S. Aeron, N. Hao, and M. E. Kilmer. 5D and 4D pre-stack seismic data completion using tensor nuclear norm (TNN). In *SEG Technical Program Expanded Abstracts 2013*, pages 3639–3644. Society of Exploration Geophysicists, 2013.

[EK15]   N. El Karoui. On the impact of predictor geometry on the performance on high-dimensional ridge-regularized generalized robust regression estimators. *Probability Theory and Related Fields*, pages 1–81, 2015.

[FL19]   V. F. Farias and A. A. Li. Learning preferences with side information. *Management Science*, 65(7):3131–3149, 2019.

[GBW18]   D. Gilboa, S. Buchanan, and J. Wright. Efficient dictionary learning with gradient descent. *arXiv preprint arXiv:1809.10313*, 2018.

[GHJY15]   R. Ge, F. Huang, C. Jin, and Y. Yuan. Escaping from saddle points online stochastic gradient for tensor decomposition. In *Conference on Learning Theory*, pages 797–842, 2015.

[GM17]   R. Ge and T. Ma. On the optimization landscape of tensor decompositions. *arXiv preprint arXiv:1706.05598*, 2017.

[GQ14]   D. Goldfarb and Z. Qin. Robust low-rank tensor recovery: Models and algorithms. *SIAM Journal on Matrix Analysis and Applications*, 35(1):225–253, 2014.

[Gro11]   D. Gross. Recovering low-rank matrices from few coefficients in any basis. *IEEE Transactions on Information Theory*, 57(3):1548–1566, March 2011.

[GRY11]   S. Gandy, B. Recht, and I. Yamada. Tensor completion and low-n-rank tensor recovery via convex optimization. *Inverse Problems*, 27(2):025010, 2011.

[HL13]   C. J. Hillar and L.-H. Lim. Most tensor problems are np-hard. *Journal of the ACM (JACM)*, 60(6):45, 2013.

[HMGW15]   B. Huang, C. Mu, D. Goldfarb, and J. Wright. Provable models for robust low-rank tensor completion. *Pacific Journal of Optimization*, 11(2):339–364, 2015.

[HSSS16]   S. B. Hopkins, T. Schramm, J. Shi, and D. Steurer. Fast spectral algorithms from sum-of-squares proofs: tensor decomposition and planted sparse vectors. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 178–191. ACM, 2016.

[HWW⁺19]   B. Hao, B. Wang, P. Wang, J. Zhang, J. Yang, and W. W. Sun. Sparse tensor additive regression. *arXiv preprint arXiv:1904.00479*, 2019.

[HZC20]   B. Hao, A. Zhang, and G. Cheng. Sparse and low-rank tensor estimation via cubic sketchings. *IEEE Transactions on Information Theory*, 2020.

[JHZ+16]   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. *Information Sciences*, 326:243–257, 2016.

[JO14]   P. Jain and S. Oh. Provable tensor factorization with missing data. In *Advances in Neural Information Processing Systems*, pages 1431–1439, 2014.

[KB09]   T. G. Kolda and B. W. Bader. Tensor decompositions and applications. *SIAM review*, 51(3):455–500, 2009.

[KBHH13]   M. E. Kilmer, K. Braman, N. Hao, and R. C. Hoover. Third-order tensors as operators on matrices: A theoretical and computational framework with applications in imaging. *SIAM Journal on Matrix Analysis and Applications*, 34(1):148–172, 2013.

[KM16]   H. Kasai and B. Mishra. Low-rank tensor completion: a riemannian manifold preconditioning approach. In *International Conference on Machine Learning*, pages 1012–1021, 2016.

[KMO10a]   R. H. Keshavan, A. Montanari, and S. Oh. Matrix completion from a few entries. *IEEE transactions on information theory*, 56(6):2980–2998, 2010.

[KMO10b]   R. H. Keshavan, A. Montanari, and S. Oh. Matrix completion from noisy entries. *J. Mach. Learn. Res.*, 11:2057–2078, 2010.

[KOKC13]   H.-J. Kim, E. Ollila, V. Koivunen, and C. Croux. Robust and sparse estimation of tensor decompositions. In *2013 IEEE Global Conference on Signal and Information Processing*, pages 965–968. IEEE, 2013.

[Kol01]   T. G. Kolda. Orthogonal tensor decompositions. *SIAM Journal on Matrix Analysis and Applications*, 23(1):243–255, 2001.

[Kol11]   V. Koltchinskii. *Oracle inequalities in empirical risk minimization and sparse recovery problems*, volume 2033 of *Lecture Notes in Mathematics*. 2011.

[KS13]   A. Krishnamurthy and A. Singh. Low-rank matrix and tensor completion via adaptive sampling. In *Advances in Neural Information Processing Systems*, pages 836–844, 2013.

[KSS13]   N. Kreimer, A. Stanton, and M. D. Sacchi. Tensor completion based on nuclear norm minimization for 5d seismic data reconstruction. *Geophysics*, 78(6):V273–V284, 2013.

[LFC+16]   C. Lu, J. Feng, Y. Chen, W. Liu, Z. Lin, and S. Yan. Tensor robust principal component analysis: Exact recovery of corrupted low-rank tensors via convex optimization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5249–5257, 2016.

[LMWY13]   J. Liu, P. Musialski, P. Wonka, and J. Ye. Tensor completion for estimating missing values in visual data. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):208–220, 2013.

[Lou14]   K. Lounici. High-dimensional covariance matrix estimation with missing observations. *Bernoulli*, 20(3):1029–1058, 2014.

[LSC+14]   Y. Liu, F. Shang, H. Cheng, J. Cheng, and H. Tong. Factor matrix trace norm minimization for low-rank tensor completion. In *Proceedings of the 2014 SIAM International Conference on Data Mining*, pages 866–874. SIAM, 2014.

[LT17]   Q. Li and G. Tang. Convex and nonconvex geometries of symmetric tensor factorization. In *2017 51st Asilomar Conference on Signals, Systems, and Computers*, pages 305–309. IEEE, 2017.

[LYX17]   X. Li, Y. Ye, and X. Xu. Low-rank tensor completion with total variation for visual data inpainting. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[LZT19]   Q. Li, Z. Zhu, and G. Tang. The non-convex geometry of low-rank matrix optimization. *Information and Inference: A Journal of the IMA*, 8(1):51–96, 2019.

[MHWG14] C. Mu, B. Huang, J. Wright, and D. Goldfarb. Square deal: Lower bounds and improved relaxations for tensor recovery. In *International conference on machine learning*, pages 73–81, 2014.

[MP20]   V. V. Mišić and G. Perakis. Data analytics in operations management: A review. *Manufacturing & Service Operations Management*, 22(1):158–169, 2020.

[MS18]   A. Montanari and N. Sun. Spectral algorithms for tensor completion. *Communications on Pure and Applied Mathematics*, 71(11):2381–2425, 2018.

[MWCC17] C. Ma, K. Wang, Y. Chi, and Y. Chen. Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval, matrix completion and blind deconvolution. *arXiv preprint arXiv:1711.10467, accepted to Foundations of Computational Mathematics*, 2017.

[NDT15]   N. H. Nguyen, P. Drineas, and T. D. Tran. Tensor sparsification via a bound on the spectral norm of random tensors. *Information and Inference: A Journal of the IMA*, 4(3):195–229, 2015.

[Paw19]   C. Pawlowski. *Machine learning for problems with missing and uncertain data with applications to personalized medicine*. PhD thesis, Massachusetts Institute of Technology, 2019.

[PS17]   A. Potechin and D. Steurer. Exact tensor completion with sum-of-squares. In *Conference on Learning Theory*, pages 1619–1673, 2017.

[PW19]   A. Pananjady and M. J. Wainwright. Value function estimation in markov reward processes: Instance-dependent ℓ_∞-bounds for policy evaluation. *arXiv preprint arXiv:1909.08749*, 2019.

[RM14]   E. Richard and A. Montanari. A statistical model for tensor PCA. In *Advances in Neural Information Processing Systems*, pages 2897–2905, 2014.

[RPP13]   B. Romera-Paredes and M. Pontil. A new convex relaxation for tensor completion. In *Advances in Neural Information Processing Systems*, pages 2967–2975, 2013.

[RSS17]   H. Rauhut, R. Schneider, and Ž. Stojanac. Low rank tensor recovery via iterative hard thresholding. *Linear Algebra and its Applications*, 523:220–262, 2017.

[SDLF+17] 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 Transactions on Signal Processing*, 65(13):3551–3582, 2017.

[SHKM14]   O. Semerci, N. Hao, M. E. Kilmer, and E. L. Miller. Tensor-based formulation and nuclear norm regularization for multienergy computed tomography. *IEEE Transactions on Image Processing*, 23(4):1678–1693, 2014.

[SL16]   R. Sun and Z.-Q. Luo. Guaranteed matrix completion via non-convex factorization. *IEEE Transactions on Information Theory*, 62(11):6535–6579, 2016.

[SLLC17]   W. W. Sun, J. Lu, H. Liu, and G. Cheng. Provable sparse tensor decomposition. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3):899–916, 2017.

[SN16]   S. R. Soroushmehr and K. Najarian. Transforming big data into computational models for personalized medicine and health care. *Dialogues in clinical neuroscience*, 18(3):339, 2016.

[Ste16]   M. Steinlechner. Riemannian optimization for high-dimensional tensor completion. *SIAM Journal on Scientific Computing*, 38(5):S461–S484, 2016.

[SY19]   D. Shah and C. L. Yu. Iterative collaborative filtering for sparse noisy tensor estimation. *arXiv preprint arXiv:1908.01241*, 2019.

[TBS+16]   S. Tu, R. Boczar, M. Simchowitz, M. Soltanolkotabi, and B. Recht. Low-rank solutions of linear matrix equations via procrustes flow. In *International Conference on Machine Learning*, pages 964–973, 2016.

[THK10]   R. Tomioka, K. Hayashi, and H. Kashima. Estimation of low-rank tensors via convex optimization. *arXiv preprint arXiv:1010.0789*, 2010.

[TS15]   G. Tang and P. Shah. Guaranteed tensor decomposition: A moment approach. In *International Conference on Machine Learning*, pages 1491–1500, 2015.

[TV19]   Y. S. Tan and R. Vershynin. Online stochastic gradient descent with arbitrary initialization solves non-smooth, non-convex phase retrieval. *arXiv preprint arXiv:1910.12837*, 2019.

[Ver10]   R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

[Ver18]   R. Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press, 2018.

[WAA16]   W. Wang, V. Aggarwal, and S. Aeron. Tensor completion by alternating minimization under the tensor train (tt) model. *arXiv preprint arXiv:1609.05587*, 2016.

[XHYS15]   Y. Xu, R. Hao, W. Yin, and Z. Su. Parallel matrix factorization for low-rank tensor completion. *Inverse Problems & Imaging*, 9(2):601–624, 2015.

[XY13]   Y. Xu and W. Yin. A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion. *SIAM Journal on imaging sciences*, 6(3):1758–1789, 2013.

[XY17]   D. Xia and M. Yuan. On polynomial time methods for exact low rank tensor completion. *arXiv preprint arXiv:1702.06980*, 2017.

[XYZ17]   D. Xia, M. Yuan, and C.-H. Zhang. Statistically optimal and computationally efficient low rank tensor completion from noisy entries. *arXiv preprint arXiv:1711.04934*, 2017.

[Yao18]   Q. Yao. Scalable tensor completion with nonconvex regularization. *arXiv preprint arXiv:1807.08725*, 2018.

[YLW+17]   J. Ying, H. Lu, Q. Wei, J.-F. Cai, D. Guo, J. Wu, Z. Chen, and X. Qu. Hankel matrix nuclear norm regularized tensor completion for n-dimensional exponential signals. *IEEE Transactions on Signal Processing*, 65(14):3702–3717, 2017.

[YPCC16]   X. Yi, D. Park, Y. Chen, and C. Caramanis. Fast algorithms for robust PCA via gradient descent. In *NIPS*, pages 4152–4160, 2016.

[YZ16]   M. Yuan and C.-H. Zhang. On tensor completion via nuclear norm minimization. *Foundations of Computational Mathematics*, 16(4):1031–1068, 2016.

[YZ17]   M. Yuan and C.-H. Zhang. Incoherent tensor norms and their applications in higher order tensor completion. *IEEE Transactions on Information Theory*, 63(10):6753–6766, 2017.

[ZA17]   Z. Zhang and S. Aeron. Exact tensor completion using t-svd. *IEEE Trans. Signal Processing*, 65(6):1511–1526, 2017.

[ZB18]   Y. Zhong and N. Boumal. Near-optimal bound for phase synchronization. *SIAM Journal on Optimization*, 2018.

[Zha19]   A. Zhang. Cross: Efficient low-rank tensor completion. *The Annals of Statistics*, 47(2):936–964, 2019.

[ZKOM18] A. M. Zoubir, V. Koivunen, E. Ollila, and M. Muma. *Robust statistics for signal processing.* Cambridge University Press, 2018.

[ZL16] Q. Zheng and J. Lafferty. Convergence analysis for rectangular matrix completion using Burer-Monteiro factorization and gradient descent. *arXiv:1605.07051*, 2016.

[ZX18] A. Zhang and D. Xia. Tensor SVD: Statistical and computational limits. *IEEE Transactions on Information Theory*, 64(11):7311–7338, 2018.