SUPPLEMENTARY MATERIAL for "Generalized Low-rank plus Sparse Tensor Estimation by Fast Riemannian Optimization"

8 Application: Poisson Tensor Robust PCA

In this section, we consider the Poisson tensor RPCA model. Suppose we observe $\boldsymbol{\mathcal{Y}} \in \mathbb{N}^{d_1 \times \cdots \times d_m}$ that satisfies

$$\forall \omega \in [d_1] \times \cdots \times [d_m],\ [\boldsymbol{\mathcal{Y}}]_\omega \sim \text{Poisson}(I \exp([\boldsymbol{\mathcal{T}}^*]_\omega + [\boldsymbol{\mathcal{S}}^*]_\omega)) \text{ independently},$$

where $(\boldsymbol{\mathcal{T}}^*, \boldsymbol{\mathcal{S}}^*) \in (\mathbb{U}_{\mathbf{r},\mu_1}, \mathbb{S}_\alpha)$ are the low rank part and sparse part respectively and $I > 0$ is the intensity parameter that is revealed as in [14]. We choose the loss function to be the negative log-likelihood with scaling

$$\mathfrak{L}(\boldsymbol{\mathcal{T}} + \boldsymbol{\mathcal{S}}) = \frac{1}{I} \sum_\omega \left(-[\boldsymbol{\mathcal{Y}}]_\omega [\boldsymbol{\mathcal{T}} + \boldsymbol{\mathcal{S}}]_\omega + I \exp([\boldsymbol{\mathcal{T}} + \boldsymbol{\mathcal{S}}]_\omega)\right).$$

This is an entry-wise loss, and simple calculation shows Assumptions 2 and 3 are satisfied with $\mathbb{B}_2^* = \mathbb{B}_\infty^* = \{\boldsymbol{\mathcal{T}} + \boldsymbol{\mathcal{S}} : \|\boldsymbol{\mathcal{T}} + \boldsymbol{\mathcal{S}}\|_{\ell_\infty} \le \zeta, \boldsymbol{\mathcal{T}} \in \mathbb{M}_{\mathbf{r}}, \boldsymbol{\mathcal{S}} \in \mathbb{S}_{\gamma\alpha}\}$ with $b_{l,\zeta} = e^{-\zeta}, b_{u,\zeta} = e^{\zeta}$. Since the parameter will become trivial in an unbounded set, we impose the following assumption which implies $\|\boldsymbol{\mathcal{T}}^*\|_{\ell_\infty} \le \frac{\zeta}{2}$ and thus $\|\boldsymbol{\mathcal{T}}^* + \boldsymbol{\mathcal{S}}^*\|_{\ell_\infty} \le \zeta$.

Assumption 6. *There exists a small $\zeta > 0$ such that $\|\boldsymbol{\mathcal{S}}^*\|_\infty \le \frac{\zeta}{2}$, $\boldsymbol{\mathcal{T}}^*$ satisfies Assumption 1 with its largest singular value $\bar{\lambda} \le c_m(\kappa_0\mu_1)^{-m}\sqrt{\frac{d^*}{r^*}}\zeta$ where $d^* = d_1 \cdots d_m$ and $r^* = r_1 \cdots r_m$.*

Similar with the binary case, we also need to show $\|\widehat{\boldsymbol{\mathcal{T}}}_l\|_{\ell_\infty}, \|\widehat{\boldsymbol{\mathcal{S}}}_l\|_{\ell_\infty} = O(\zeta)$. These are guaranteed by choosing $\mathsf{k_{pr}} = C_1\zeta$ for some $C_1 > 1$ depending only on $\kappa_0\mu_1, m$ and from Lemma 5.6, when $\kappa_0\mu_1, m = O(1)$ and $\|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}} \le \underline{\lambda}/8$, we have $\|\widehat{\boldsymbol{\mathcal{T}}}_{l+1}\|_{\ell_\infty} = O(\zeta)$. We summarize the result in the following Theorem.

Theorem 8.1. *Let $\gamma > 1, \mathsf{k_{pr}} := C_1\zeta$ be the parameters used in Algorithm 2 for a constant $C_1 > 1$ depending only on $\kappa_0\mu_1$ and m via Lemma 5.6. Suppose Assumptions 1 and 6 hold. Assume $|\Omega^*| \asymp \alpha d^*$, $e^{2\zeta'} \le 0.4(\sqrt{\delta})^{-1}$ for some $\delta \in (0, 1]$ and $\zeta' = (2C_1 + 1)\zeta$, and*

(a) Initialization: $\|\widehat{\boldsymbol{\mathcal{T}}}_0 - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}} \le c_{1,m}\underline{\lambda} \cdot \min\left\{\delta^2\bar{r}^{-1/2}, (\kappa_0^{2m}\bar{r}^{1/2})^{-1}\right\}$, $\|\widehat{\boldsymbol{\mathcal{T}}}_0\|_{\ell_\infty} \le c_{2,m}\zeta$ and $\widehat{\boldsymbol{\mathcal{T}}}_0$ is $(2\mu_1\kappa_0)^2$-incoherent

(b) Signal-to-noise ratio:

$$\underline{\lambda} \cdot \min\left\{\delta^2 \bar{r}^{-1/2}, (\kappa_0^{2m} \bar{r}^{1/2})^{-1}\right\} \geq C_{2,m}\Big(\gamma|\Omega^*| \frac{1+e^{\zeta'}}{e^{-\zeta'}} \cdot e^{\zeta} + \sqrt{(r^* + \bar{d}\bar{r})e^{\zeta}/I}\Big), \text{ and } I \geq Ce^{\zeta}\log(d^*)$$

(c) Sparsity condition: $\alpha \leq c_{3,m} e^{-8\zeta'} (\kappa_0^{4m} \mu_1^{4m} \bar{r}^m)^{-1}$ *and* $\gamma \geq 1 + (4m)^{-1} \cdot e^{8\zeta'}$

where $c_{1,m}, c_{2,m}, c_{3,m}, C_{2,m} > 0$ *are some constants depending on* m *only. If the stepsize* $\beta \in [0.005e^{-3\zeta'}, 0.36e^{-3\zeta'}]$*, after* $l_{\max}$ *iterations, with probability at least* $1 - \frac{2}{d^*}$*,*

$$\|\widehat{\boldsymbol{\mathcal{T}}}_{l_{\max}} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2 \leq (1-\delta^2)^{l_{\max}} \cdot \|\widehat{\boldsymbol{\mathcal{T}}}_0 - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2 + C_{1,\delta} \frac{r^* + \bar{d}\bar{r}}{I/e^{\zeta}} + C_3 e^{2\zeta} \cdot \gamma|\Omega^*|$$

$$\|\widehat{\boldsymbol{\mathcal{S}}}_{l_{\max}} - \boldsymbol{\mathcal{S}}^*\|_{\mathrm{F}}^2 \leq e^{4\zeta'}\big(C_{4,m}\alpha \bar{r}^m (\mu_1\kappa_0)^{4m} + C_{5,m}(\gamma-1)^{-1}\big)\|\widehat{\boldsymbol{\mathcal{T}}}_{l_{\max}} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2 + C_{6,m}e^{2\zeta+2\zeta'} \cdot \gamma|\Omega^*|$$

where $C_3 > 0$ *depends only on* δ, ζ, m *, and* $C_{4,m}, C_{5,m}, C_{6,m} > 0$ *are constants depending only on* m*. Moreover, if* $l_{\max}$ *is chosen large enough such that the second term on RHS of (5.9) dominates and assume* $\kappa_0^{4m}\mu_1^{4m}\bar{r}^m(\bar{r}\bar{d} + r^*) \lesssim_m O(\underline{d}^{m-1})$*, we get with probability at least* $1 - \frac{2}{d^*}$ *that*

$$\begin{aligned}
\|\widehat{\boldsymbol{\mathcal{T}}}_{l_{\max}} - \boldsymbol{\mathcal{T}}^*\|_{\ell_\infty} \leq & C_6\kappa_0^{2m}\mu_1^{2m}(\bar{r}^m/\underline{d}^{m-1})^{1/2}\Big(\frac{r^* + \bar{d}\bar{r}}{I} + \gamma|\Omega^*|\Big)^{1/2} \\
\|\widehat{\boldsymbol{\mathcal{S}}}_{l_{\max}} - \boldsymbol{\mathcal{S}}^*\|_{\ell_\infty} \leq & C_7\kappa_0^{2m}\mu_1^{2m}\bar{r}^{m/2}/\underline{d}^{(m-1)/2} \cdot \Big(\sqrt{(r^* + \bar{d}\bar{r})/I} + |\Omega^*|^{1/2}\Big) + C_8
\end{aligned}$$

where $C_6, C_7, C_8 > 0$ *depend only on* γ, δ, ζ, m*.*

From Theorem 8.1, after a properly chosen $l_{\max}$ iterations, we will obtain $\|\widehat{\boldsymbol{\mathcal{T}}}_{l_{\max}} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2 = O(\frac{r^*+\bar{d}\bar{r}}{I/e^{\zeta}} + e^{2\zeta} \cdot \gamma|\Omega^*|)$. As a special case when $|\Omega^*| = 0$, our result matches the previous result in Poisson tensor PCA in [14] that is rate optimal under the same requirements on the intensity parameter I. When there are outliers, the error for the estimation of $\boldsymbol{\mathcal{T}}^*$ is further influenced by the outliers.

Initialization. We shall adopt the initialization proposed in [14] with slight modification. The theoretical guarantee is summarized in the following lemma.

Lemma 8.2. *Suppose that Assumptions 1 and 6 hold. There exist absolute constants* $c, C > 0$ *such that if* $I \geq C\max\{\bar{d}, \underline{\lambda}^{-2}\sum_{i=1}^m (d_i r_i + d_i^- r_i)\bar{r}\}$*, and the sparsity of* $\boldsymbol{\mathcal{S}}^*$ *satisfies* $|\Omega^*| \leq c\zeta^{-2}\underline{\lambda}^2\bar{r}^{-1}$*, then the output of Algorithm 4 satisfies the initialization requirement in Theorem 8.1 with probability at least* $1 - 1/d^*$*.*

Algorithm 4 Initialization for Poisson RPCA

Set $\widetilde{\boldsymbol{\mathcal{T}}} = \log(\frac{\boldsymbol{\mathcal{Y}}+1/2}{I})$.

Let $\widetilde{\boldsymbol{\mathcal{T}}}_0 = \mathscr{H}^{\mathrm{HO}}_{\mathbf{r}}(\widetilde{\boldsymbol{\mathcal{T}}})$.

Return $\widehat{\boldsymbol{\mathcal{T}}}_0 = \mathsf{Trim}_{\eta,\mathbf{r}}(\widetilde{\boldsymbol{\mathcal{T}}}_0)$ with $\eta = 16\mu_1\|\widetilde{\boldsymbol{\mathcal{T}}}_0\|_{\mathrm{F}}/(7\sqrt{d^*})$.

9 Higher Order Orthogonal Iteration Algorithm

The HOOI algorithm is summarized as follows which is applied for the initialization in Section 5.1 and 5.2.

Algorithm 5 HOOI

Input: $\boldsymbol{\mathcal{Y}} \in \mathbb{R}^{d_1\times\cdots\times d_m}, \mathbf{r} = (r_1, \cdots, r_m)$, maximum number of iteration: $t_{\max}$.

Let $t = 0$, initiate $\widehat{\mathbf{U}}_i^0 = \mathrm{SVD}_{r_i}(\mathcal{M}_i(\boldsymbol{\mathcal{Y}}))$, $i \in [m]$.

for $t = 1, \ldots, t_{\max}$ **do**

  for $i = 1, \ldots, m$ **do**

    $\widehat{\mathbf{U}}_i^t = \mathrm{SVD}_{r_i}\big(\mathcal{M}_i(\boldsymbol{\mathcal{Y}})(\widehat{\mathbf{U}}_m^{t-1} \otimes \cdots \otimes \widehat{\mathbf{U}}_{i+1}^{t-1} \otimes \widehat{\mathbf{U}}_{i-1}^{t} \otimes \cdots \otimes \widehat{\mathbf{U}}_1^{t-1})\big)$

  end for

end for

Output: $\widehat{\mathbf{U}}_i = \widehat{\mathbf{U}}_i^{t_{max}}$, $\widehat{\boldsymbol{\mathcal{T}}} = \boldsymbol{\mathcal{Y}} \times_{i=1}^m \widehat{\mathbf{U}}_i\widehat{\mathbf{U}}_i^T$.

10 When Sparse Component is Absent

In this section, we consider the special case when the sparse component is absent, i.e., $\boldsymbol{\mathcal{S}}^* = \mathbf{0}$. For the exact low-rank tensor model, we observe that many conditions in Section 4 can be relaxed. A major difference is that the spikiness condition is generally not required for exact low-rank model. Consequently, the trimming step in Algorithm 2 is unnecessary. Therefore, it suffices to simply apply the Riemannian gradient descent algorithm to solve for the underlying low-rank tensor $\boldsymbol{\mathcal{T}}^*$. For ease of exposition, the procedure is summarized in Algorithm 6 (largely the same as Algorithm 2).

Algorithm 6 runs fast and guarantees favourable convergence performances under weaker conditions than Theorem 4.1. Indeed, since there is no sparse component, only Assumption 2 is

Algorithm 6 Riemannian Gradient Descent for Exact Low-rank Estimate

Initialization: $\widehat{\boldsymbol{\mathcal{T}}}_0 \in \mathbb{M}_{\mathbf{r}}$ and stepsize $\beta > 0$

for $l = 0, 1, \cdots, l_{\max}$ **do**

$\boldsymbol{\mathcal{G}}_l = \nabla \mathfrak{L}(\widehat{\boldsymbol{\mathcal{T}}}_l)$

$\boldsymbol{\mathcal{W}}_l = \widehat{\boldsymbol{\mathcal{T}}}_l - \beta \mathcal{P}_{\mathbb{T}_l} \boldsymbol{\mathcal{G}}_l$

$\widehat{\boldsymbol{\mathcal{T}}}_{l+1} = \mathscr{H}^{\mathrm{HO}}_{\mathbf{r}}(\boldsymbol{\mathcal{W}}_l)$

end for

Output: $\widehat{\boldsymbol{\mathcal{T}}}_{l_{\max}}$

required to guarantee the convergence of Algorithm 6. Similarly as Section 4, the error of final estimate produced by Algorithm 6 is characterized by the gradient at $\boldsymbol{\mathcal{T}}^*$. With a slightly abuse of notation, denote $\mathsf{Err}_{2\mathbf{r}} = \sup_{\boldsymbol{\mathcal{X}} \in \mathbb{M}_{2\mathbf{r}}, \|\boldsymbol{\mathcal{X}}\|_{\mathrm{F}} \le 1} \langle \nabla \mathfrak{L}(\boldsymbol{\mathcal{T}}^*), \boldsymbol{\mathcal{X}} \rangle$.

Theorem 10.1. *Suppose Assumption 2 holds with* $\boldsymbol{\mathcal{S}}^* = \mathbf{0}$ *and* $\mathbb{B}_2^* = \{\boldsymbol{\mathcal{T}} : \|\boldsymbol{\mathcal{T}} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}} \le c_{0,m}\underline{\lambda}, \boldsymbol{\mathcal{T}} \in \mathbb{M}_{\mathbf{r}}\}$ *for a small constant* $c_{0,m} > 0$ *depending on* m *only, also suppose* $1.5 b_l b_u^{-2} \le 1$ *and* $0.75 b_l b_u^{-1} \ge \delta^{1/2}$ *for some* $\delta \in (0, 1]$ *and the stepsize* $\beta \in [0.4 b_l b_u^{-2}, 1.5 b_l b_u^{-2}]$ *in Algorithm 6. Assume*

(a) Initialization: $\|\widehat{\boldsymbol{\mathcal{T}}}_0 - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}} \le \underline{\lambda} \cdot c_{1,m} \delta \bar{r}^{-1/2}$

(b) Signal-to-noise ratio: $\mathsf{Err}_{2\mathbf{r}}/\underline{\lambda} \le c_{2,m} \delta^2 \bar{r}^{-1/2}$

where $c_{1,m}, c_{2,m} > 0$ *are small constants depending only on* m*. Then for all* $l = 1, \cdots, l_{\max}$*,*

$$\|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2 \le (1 - \delta^2)^l \|\widehat{\boldsymbol{\mathcal{T}}}_0 - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2 + C_\delta \mathsf{Err}_{2\mathbf{r}}^2$$

where $C_\delta > 0$ *is a constant depending only on* δ*. Then after at most* $l_{\max} \asymp \log\left(\underline{\lambda}/\mathsf{Err}_{2\mathbf{r}}\right)$ *iterations (also depends on* $b_l, b_u, m, \bar{r}$ *and* β*), we get*

$$\|\widehat{\boldsymbol{\mathcal{T}}}_{l_{\max}} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}} \le C \cdot \mathsf{Err}_{2\mathbf{r}},$$

where the constant $C > 0$ *depends on only* $b_l, b_u, m, \bar{r}$ *and* β*.*

Note that Theorem 10.1 holds without spikiness condition in contrast with Theorem 4.1. It makes sense for the model has no missing values or sparse corruptions. The assumptions on loss function are also weaker (e.g., no need to be an entry-wise loss or entry-wisely smooth) than those in Theorem 4.1. As a result, Theorem 10.1 is also applicable to the low-rank tensor regression

model among others. See [14, 8, 42] and references therein. The initialization and signal-to-noise conditions are similar to those in Theorem 4.1, e.g., by setting $|\Omega^*| = \alpha = 0$ there. In addition, the error of final estimate depends only on $\mathsf{Err}_{2\mathbf{r}}$. Interestingly, the contraction rate does not depend on the condition number κ_0.

Comparison with existing literature In [14], the authors proposed a general framework for exact low-rank tensor estimation based on regularized jointly gradient descent on the core tensor and associated low-rank factors. Their method is fast and achieves statistical optimality in various models. In contrast, our algorithm is based on Riemannian gradient descent, requires no regularization and also runs fast. An iterative tensor projection algorithm was studied in [44]. But their method only applies to tensor regression. Other notable works focusing only on tensor regression include [47, 50, 15, 33, 22, 26]. A general projected gradient descent algorithm was proposed in [8] for generalized low-rank tensor estimation. For Tucker low-rank tensors, their algorithm is similar to our Algorithm 6 except that they use vanilla gradient $\boldsymbol{\mathcal{G}}_l$ while we use the Riemannian gradient $\mathcal{P}_{\mathbb{T}_l}\boldsymbol{\mathcal{G}}_l$. As explained in Section 3, using the vanilla gradient can cause heavy computation burdens in the subsequent steps. Riemannian gradient descent algorithm for tensor completion was initially proposed by [21]. They focused only on tensor completion model and did not investigate its theoretical guarantees. Recently in [5], the Riemannian gradient descent algorithm is applied for noiseless tensor regression and its convergence analysis is proved.

11 More Numerical Simulations

In Section 11.1, we apply the proposed BIC-type criterion for SG-RPCA and binary tensor learning and demonstrate its effectiveness on synthetic data. Through Section 11.2-11.5, we treat $\mathbf{r}$ and α as given and test the performance of our estimator with respect to different choices of γ. Other algorithmic parameters like μ_1 and k_{pr} are decided as explained in Section 6.

11.1 Performance of BIC-type Criterion

We test the performance of BIC-type criterion (3.3) for SG-RPCA and binary tensor learning. As explained in Section 6, γ is set to 1 and $\mu_1 = 2^m + \log(\bar{d})$. More exactly, the BIC-type criterion

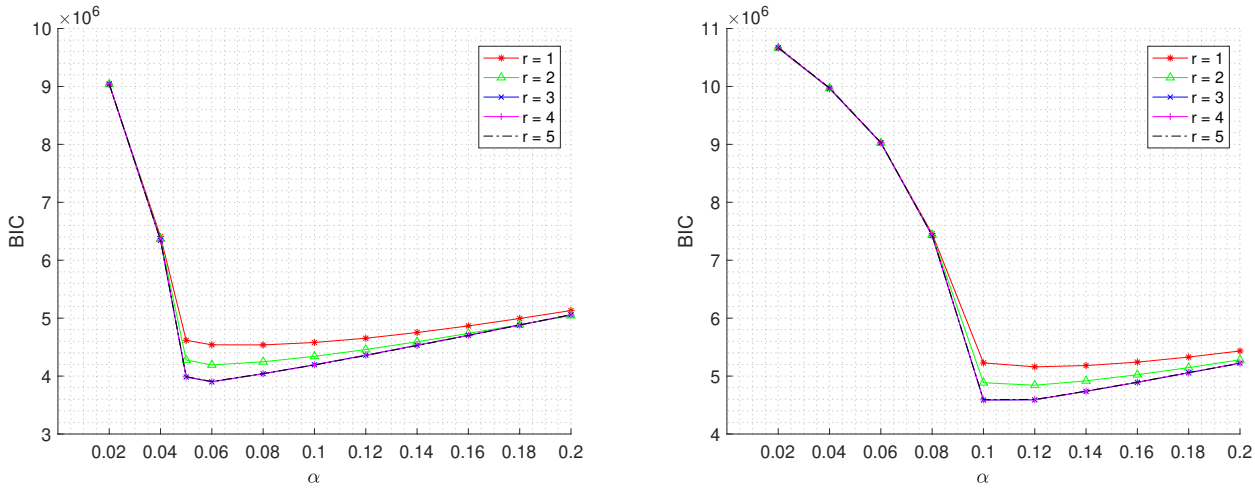


Figure 5: BIC values for SG-RPCA; the true rank $\mathbf{r} = (3,3,3)^\top$. Left: true $\alpha = 0.05$; Right: true $\alpha = 0.1$.

for SG-RPCA (assuming Gaussian noise with equal but unknown variances) is

$$\mathrm{BIC}(\mathbf{r},\alpha) := \big(\|\widehat{\boldsymbol{\mathcal{S}}}_{\mathbf{r},\alpha}\|_{\ell_0} + \sum_{i=1}^m r_i d_i\big)\cdot\log(d^*) + d^*\log(\|\boldsymbol{\mathcal{A}} - \widehat{\boldsymbol{\mathcal{T}}}_{\mathbf{r},\alpha} - \widehat{\boldsymbol{\mathcal{S}}}_{\mathbf{r},\alpha}\|_{\mathrm{F}}^2).$$

The true tensor $\boldsymbol{\mathcal{T}}^* \in \mathbb{R}^{d\times d\times d}$ with $d = 100$ and $\mathbf{r} = (3,3,3)^\top$. We test two true sparsity levels $\alpha \in \{0.05, 0.1\}$. The true tensor $\boldsymbol{\mathcal{T}}^*$ satisfies $\|\boldsymbol{\mathcal{T}}^*\|_{\ell_\infty} = 0.1$, and $\boldsymbol{\mathcal{S}}^*$ is generated as above satisfying $\|\boldsymbol{\mathcal{S}}^*\|_{\ell_\infty} = 4$, and all entries of $\boldsymbol{\mathcal{Z}}^*$ satisfy i.i.d. $\mathrm{N}(0,\sigma_z^2)$ with $\sigma_z = 0.01$. For each $\alpha \in \{0.05, 0.1\}$, we test, in our algorithm, $\mathbf{r} \in \{(1,1,1),(2,2,2),(3,3,3),(4,4,4),(5,5,5)\}$ and $\alpha \in (0.02, 0.2)$. The results are displayed in Figure 5. The BIC-values are sensitive to both $\mathbf{r}$ and α. We note that the BIC-values for $\mathbf{r} > 3$ are strictly larger than that of $\mathbf{r} = (3,3,3)$, but the difference is too small to be spotted in the figures.

For robust binary tensor learning, we also set $\mathbf{r} = (3,3,3)^\top$ and $\boldsymbol{\mathcal{T}}^* \in \mathbb{R}^{d\times d\times d}$ with $d = 100$, and $\boldsymbol{\mathcal{S}}^*$ is generated as above. The true $\alpha \in \{0.005, 0.01\}$. We fix $p(x) = (1+e^{-10x})^{-1}$. The BIC criterion for the binary case is:

$$\mathrm{BIC}(\mathbf{r},\alpha) := \big(\|\widehat{\boldsymbol{\mathcal{S}}}\|_{\ell_0} + \sum_{i=1}^m r_i d_i\big)\cdot\log(d^*) - 2\sum_{\omega}\big([\boldsymbol{\mathcal{A}}]_\omega \log p([\widehat{\boldsymbol{\mathcal{T}}}+\widehat{\boldsymbol{\mathcal{S}}}]_\omega) + \big(1-[\boldsymbol{\mathcal{A}}]_\omega\big)\log\big(1-p([\widehat{\boldsymbol{\mathcal{T}}}+\widehat{\boldsymbol{\mathcal{S}}}]_\omega)\big)\big).$$

For each true $\alpha \in \{0.005, 0.01\}$, we test BIC for $\mathbf{r} \in \{(1,1,1),(2,2,2),(3,3,3),(4,4,4),(5,5,5)\}$ and α varying from 0.001 to 0.015. The results are displayed in Figure 6 showing that BIC is more sensitive to $\mathbf{r}$ and less sensitive to α for a small range. After the true $\mathbf{r}$ is identified, the BIC criterion works reasonably well for selecting α.

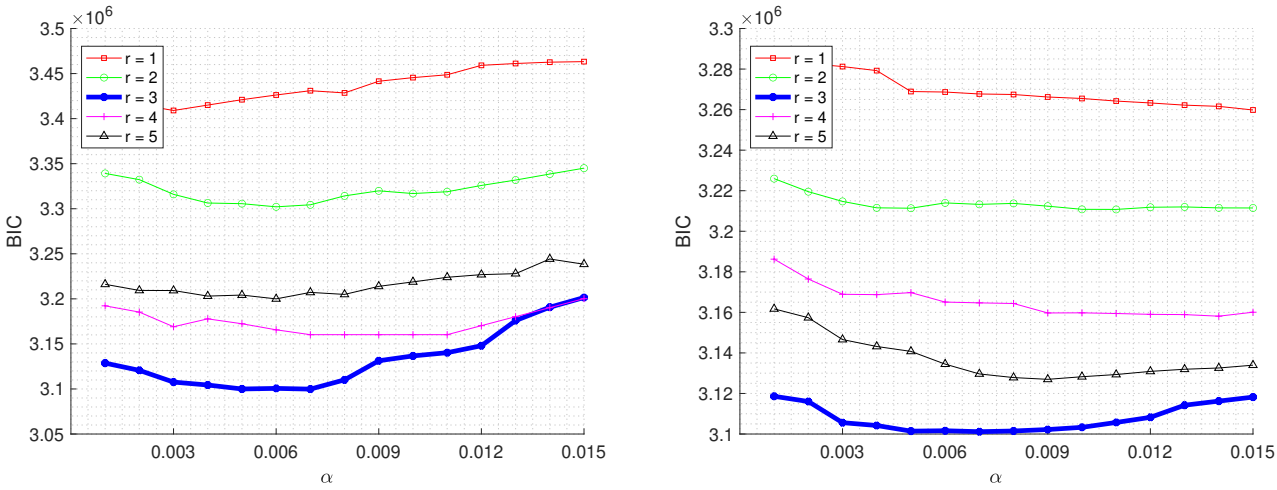


Figure 6: BIC values for binary tensor learning; the true rank $\mathbf{r} = (3, 3, 3)^\top$. Left: true $\alpha = 0.005$; Right: true $\alpha = 0.01$. The BIC curve for $\mathbf{r} = (3, 3, 3)^\top$ is highlighted.

11.2 Tensor Sub-Gaussian Robust PCA

The low-rank tensor $\boldsymbol{\mathcal{T}}^* \in \mathbb{R}^{d\times d\times d}$ with $d = 100$ and Tucker ranks $\mathbf{r} = (2, 2, 2)^\top$ is generated from the HOSVD of a trimmed standard normal tensor. It satisfies the spikiness condition, with high probability, and has singular values $\bar{\lambda} \approx 3$ and $\underline{\lambda} \approx 1$. Given a sparsity level $\alpha \in (0, 1)$, the entries of sparse tensor $\boldsymbol{\mathcal{S}}^*$ are i.i.d. sampled from $\mathrm{Be}(\alpha) \times \mathrm{N}(0, 1)$, which ensures $\boldsymbol{\mathcal{S}}^* \in \mathbb{S}_{O(\alpha)}$ with high probability. This ensures that the non-zero entries of $\boldsymbol{\mathcal{S}}^*$ have typically much larger magnitudes than the entries of $\boldsymbol{\mathcal{T}}^*$. The noise tensor $\boldsymbol{\mathcal{Z}}$ has i.i.d. entries sampled from $\mathrm{N}(0, \sigma_z^2)$. The default choice of γ is 2, $\mathsf{k_{pr}} = \infty$ and μ_1 is set as previously. The convergence performances of $\log(\|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}/\|\boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}})$ by Algorithm 2 are examined and presented in the left panels of Figure 7.

The top-left plot in Figure 7 displays the effects of α on the convergence of Algorithm 2. It shows that the convergence speed of Algorithm 2 is insensitive to α, while the error of final estimates $\widehat{\boldsymbol{\mathcal{T}}}_{l_{\max}}$ is related to α. This is consistent with the claims of Theorem 5.1. In the middle-left plot of Figure 7, we observe that, for a fixed sparsity level α, the error of final estimates grows as the tuning parameter γ becomes larger. The bottom-left plot of Figure 7 shows the convergence of Algorithm 2 for different noise levels. All these plots confirm the fast convergence of our Riemannian gradient descent algorithm. In particular, there are stages during which the log relative error decreases linearly w.r.t. the number of iterations, as proved in Theorem 4.1.

The statistical stability of the final estimates by Algorithm 2 is demonstrated in the right panels

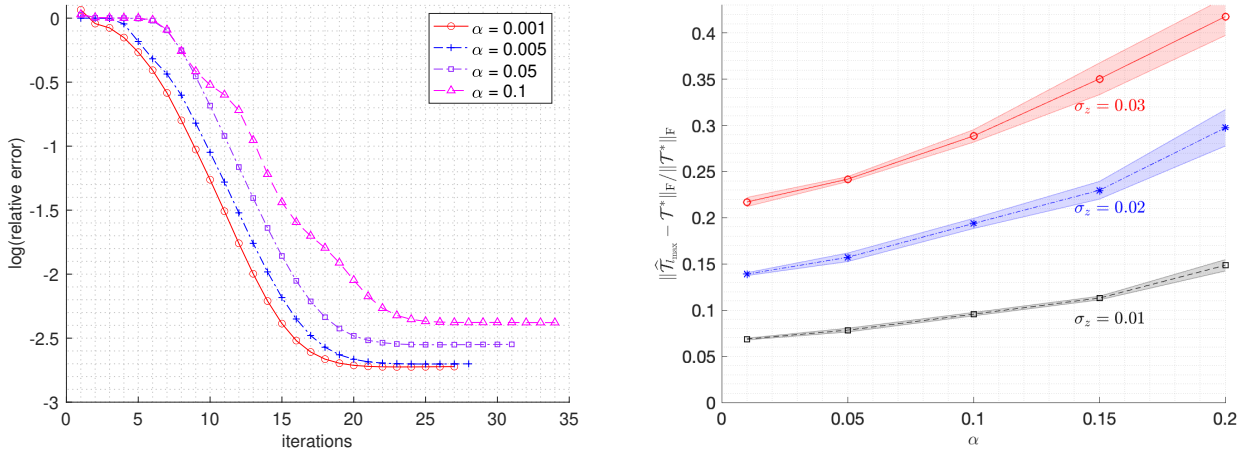
of Figure 7. Each curve represents the average relative error of $\widehat{\boldsymbol{\mathcal{T}}}_{l_{\max}}$ based on 10 simulations, and the error bar shows the confidence region by one empirical standard deviation. Based on these plots, we observe that the standard deviations of $\|\widehat{\boldsymbol{\mathcal{T}}}_{l_{\max}} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}$ grow as the noise level σ_z, the sparsity level α or the tuning parameter γ increases.

11.3 Tensor PCA with Heavy-tailed Noise

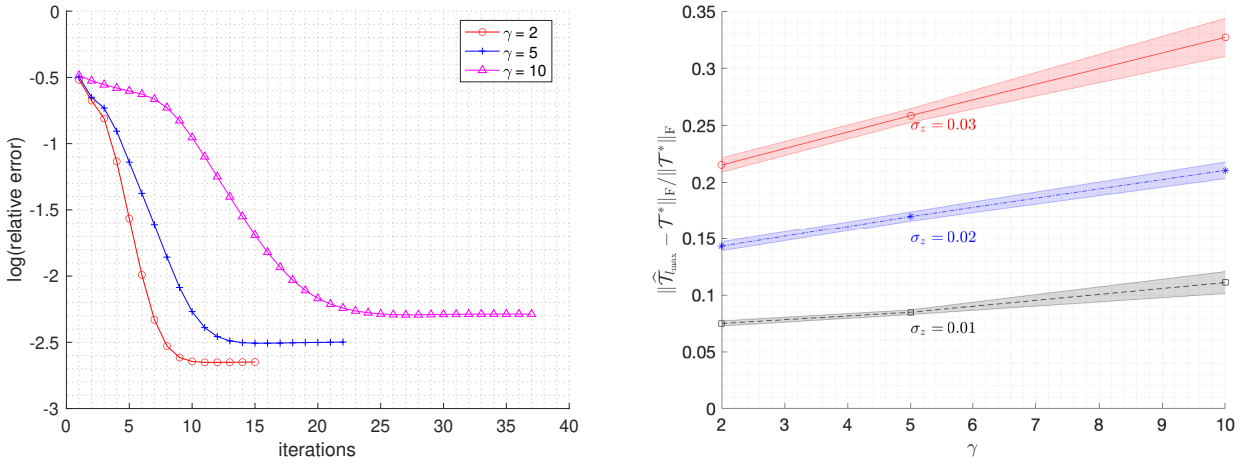
The low-rank tensor $\boldsymbol{\mathcal{T}}^* \in \mathbb{R}^{d\times d\times d}$ with $d = 100$ and Tucker ranks $\mathbf{r} = (2, 2, 2)^\top$ is generated from the HOSVD of a trimmed standard normal tensor, as in Section 11.2. Given a parameter θ, we generate the noisy tensor whose entries are i.i.d. and satisfy the Student-t distribution with degree of freedom θ. But notice here we also apply a global scaling to better control the noise standard deviation. We denote the noisy tensor after scaling by $\boldsymbol{\mathcal{Z}}$. This generated tensor $\boldsymbol{\mathcal{Z}}$ satisfies Assumption 4 with the same parameter θ. Once the parameter θ and global scaling are given, we are able to calculate the variance σ_z^2. The convergence performances of $\log(\|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}/\|\boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}})$ by Algorithm 2 are examined and presented in the upper panels of Figure 8.

In this experiment, we set $\gamma = 2, \mathsf{k_{pr}} = \infty$ and μ_1 as previously. The top-left plot in Figure 8 displays the effects of α on the convergence of Algorithm 2. The case $\alpha = 0$ reduces to the normal Riemannian gradient descent, which cannot output a satisfiable result due to the heavy-tailed noise, even if a warm initialization is provided. This shows the importance of gradient pruning in Algorithm 2. When $\alpha > 0$, the convergence speed of the algorithm is insensitive to α, but the final estimates $\widehat{\boldsymbol{\mathcal{T}}}_{l_{\max}}$ is related to α. In the top-right plot of Figure 8, we observe the error becomes larger as θ decreases (or equivalently, as σ_z^2 increases). All these results match the claim of Theorem 5.4 and confirm the fast convergence of Riemannian gradient descent. And there are indeed stages where the log relative error decreases linearly w.r.t. the number of iterations.

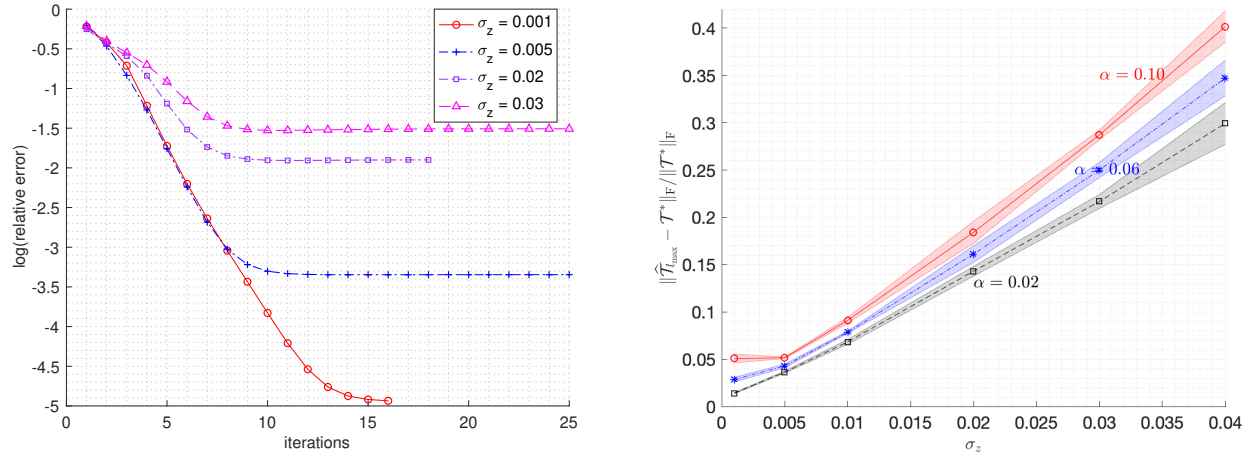
The statistical stability of the final estimates by Algorithm 2 applied to tensor PCA with heavy-tailed noise is demonstrated in the bottom panel of Figure 8. Each curve represents the average relative error of $\widehat{\boldsymbol{\mathcal{T}}}_{l_{\max}}$ based on 5 simulations, and the error bar shows the confidence region by one empirical standard deviation. Based on these plots, we observe that for each fixed θ (or σ_z^2, equivalently), we need to choose α carefully to achieve the best performance. This is reasonable since in the heavy-tail noise setting, we do not know the sparsity of outliers. Also, the figure shows that Algorithm 2 is stable for different α and θ.



(a) Change of sparsity α. Left: $\sigma_z = 0.01, \gamma = 2$; Right: $\gamma = 2$

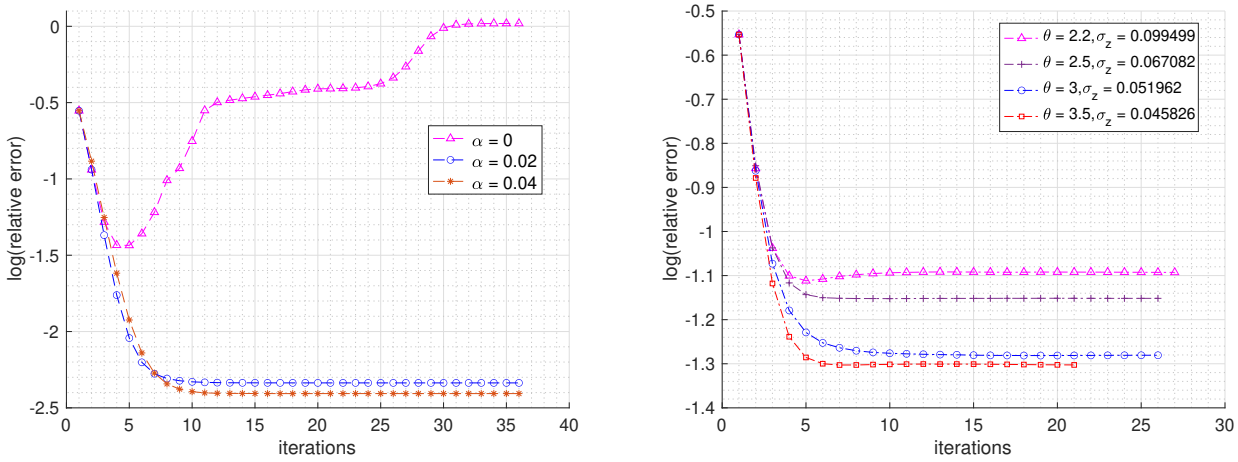


(b) Change of γ. Left: $\alpha = 0.02, \sigma_z = 0.01$; Right: $\alpha = 0.02$

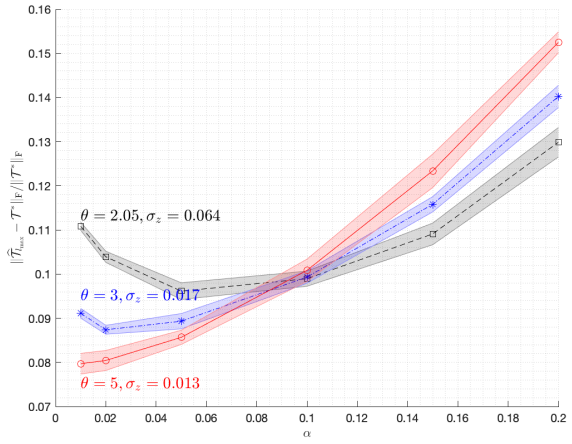


(c) Change of noise size σ_z. Left: $\alpha = 0.02, \gamma = 2$; Right: $\gamma = 2$

Figure 7: Performances of Algorithm 2 for SG-RPCA. The low-rank $\boldsymbol{\mathcal{T}}^*$ has size $d \times d \times d$ with $d = 100$ and has Tucker ranks $\mathbf{r} = (2, 2, 2)^\top$. The relative error on left panels is defined by $\|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}/\|\boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}$. The error bars on the right panels are based on 1 standard deviation from 10 replications. Here the default γ is 2.



(a) Left: Change of α, $\theta = 2.2(\sigma_z = 0.332)$; Right: Change of θ, $\alpha = 0.01$



(b) Change of θ

Figure 8: Performances of Algorithm 2 for tensor PCA with heavy-tailed noise. The low-rank $\boldsymbol{\mathcal{T}}^*$ has size $d \times d \times d$ with $d = 100$ and has Tucker ranks $\mathbf{r} = (2, 2, 2)^\top$. The relative error on upper panels is defined by $\|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}/\|\boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}$. The error bars on the lower panels are based on 1 standard deviation from 5 replications. Here the default choice of γ is 2.

11.4 Binary Tensor Learning

In the binary tensor setting, we generate the low-rank tensor $\boldsymbol{\mathcal{T}}^* \in \mathbb{R}^{d\times d\times d}$ with $d = 100$ and Tucker ranks $\mathbf{r} = (2,2,2)^\top$ from the HOSVD of a trimmed standard normal tensor. But here we did a scaling to $\boldsymbol{\mathcal{T}}^*$ so that the singular value $\bar{\lambda} \approx 300$ and $\underline{\lambda} \approx 100$. Given a sparsity level $\alpha \in (0,1)$, the entries of sparse tensor $\boldsymbol{\mathcal{S}}^*$ are i.i.d. sampled from $\mathrm{Be}(\alpha) \times \mathrm{N}(0,1)$, which ensures $\boldsymbol{\mathcal{S}}^* \in \mathbb{S}_{O(\alpha)}$ with high probability. We generate the tensor $\boldsymbol{\mathcal{T}}^*$ and $\boldsymbol{\mathcal{S}}^*$ in this way in order to meet the requirements of Assumption 5. In the following experiments, we are considering the logistic link function with the scaling parameter σ, i.e., $p(x) = (1+e^{-x/\sigma})^{-1}$. The default choice of γ is 1.1, $\mathsf{k_{pr}} = 1$ and μ_1 is set as previously. The convergence performances of $\log(\|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}/\|\boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}})$ by Algorithm 2 are examined and presented in the top two panels of Figure 9.

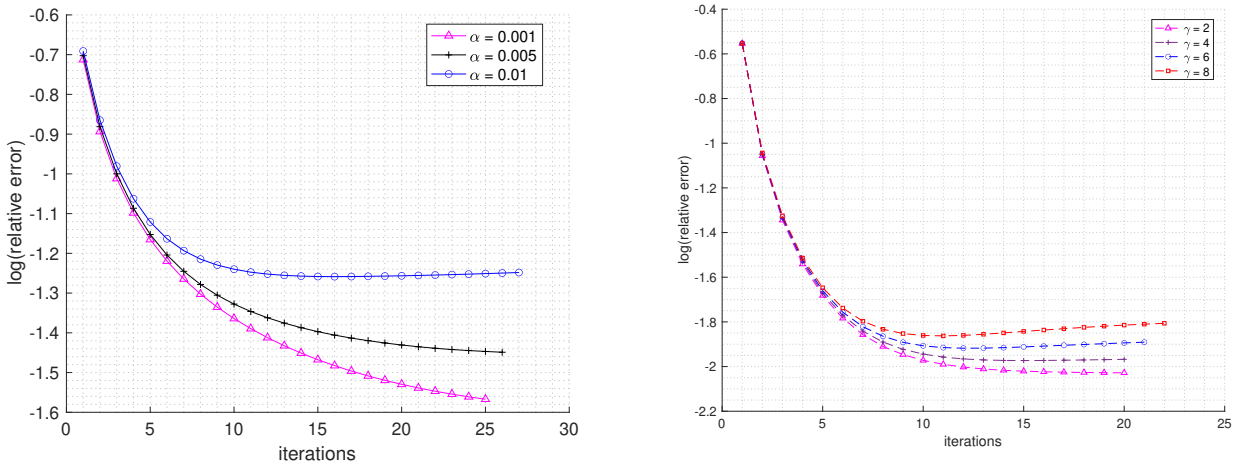
The top-left plot in Figure 9 shows the effect of α on the convergence of Algorithm 2. From the figure, it is clear that the error of final estimates $\widehat{\boldsymbol{\mathcal{T}}}_{l_{\max}}$ is related to α. This again verifies the results in Theorem 5.7. In the top-right plot in Figure 9, we can see the error of the final estimates increases as the parameter γ becomes larger. All these experiments show that Riemannian gradient descent converges fast and there are stages when the log relative error decreases linearly w.r.t. the number of iterations.

The statistical stability of the final estimates by Algorithm 2 is demonstrated in the bottom panel of Figure 9. Each curve represents the average relative error of $\widehat{\boldsymbol{\mathcal{T}}}_{l_{\max}}$ based on 5 simulations, and the error bar shows the confidence region by one empirical standard deviation. From these plots, we observe that the standard deviations of $\|\widehat{\boldsymbol{\mathcal{T}}}_{l_{\max}} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}$ grow as the noise level, the sparsity level α or the tuning parameter γ increases.

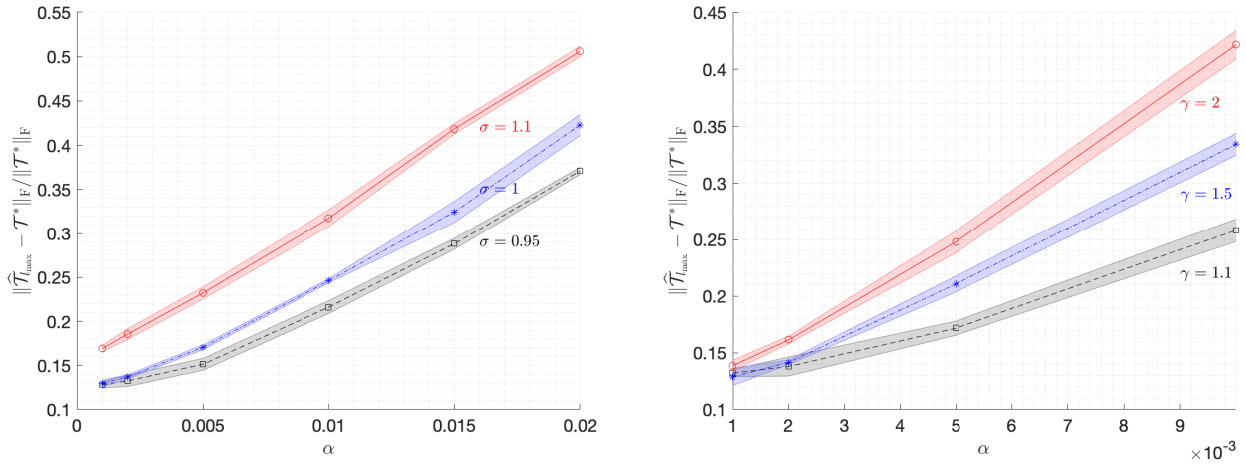
11.5 Tensor Poisson Robust PCA

In the Poisson tensor RPCA case, we generate $\boldsymbol{\mathcal{T}}^* \in \mathbb{R}^{d\times d\times d}$ with $d = 100$ and Tucker rank $\mathbf{r} = (2,2,2)^\top$ such that $\|\boldsymbol{\mathcal{T}}^*\|_{\ell_\infty} = 0.5$. Meanwhile, the sparse outliers $\boldsymbol{\mathcal{S}}^*$ is generated such that all its entries are i.i.d. sampled from $\mathrm{Be}(\alpha) \times \mathrm{N}(0,1)$ and scaled such that $\|\boldsymbol{\mathcal{S}}^*\|_{\ell_\infty} = 0.5$. Throughout the experiments, both ζ and $\mathsf{k_{pr}}$ is set to 0.5, and the default choice of γ is 1.1.

In the first experiment, we fix the intensity $I = 10$ and change the sparsity level. The convergence performances of $\log(\|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}/\|\boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}})$ by Algorithm 2 is displayed in the left panel of



(a) Left: Change of sparsity α, $\sigma = 1, \gamma = 1.1$; Right: Change of γ, $\alpha = 0.001, \sigma = 1$



(b) Left: Change of σ, $\gamma = 1.1$; Right: Change of γ, $\alpha = 0.001$

Figure 9: Performances of Algorithm 2 for binary tensor learning. The low-rank $\boldsymbol{\mathcal{T}}^*$ has size $d \times d \times d$ with $d = 100$ and has Tucker ranks $\mathbf{r} = (2, 2, 2)^\top$. The relative error on left panels is defined by $\|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}} / \|\boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}$. The error bars on the bottom panels are based on 1 standard deviation from 5 replications. The default choice of γ is 1.1.

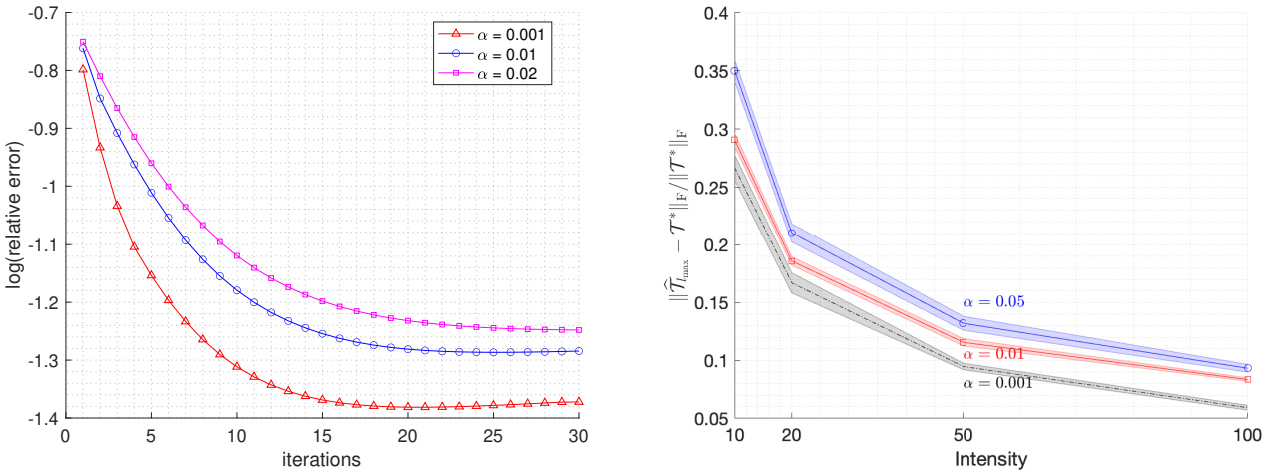


Figure 10: Performance of Algorithm 2 for tensor Poisson RPCA. The Tucker rank of $\boldsymbol{\mathcal{T}}^*$ is $\mathbf{r} = (2, 2, 2)^\top$. Left: Convergence behaviors with different α and I is fixed with $I = 10$; Right: Error bar with each setting repeated 5 i.i.d. times. Here $\gamma = 1.1$ and $\mathsf{k}_{\mathsf{pr}} = 0.5$.

Figure 10. In the second experiment, for different values of α and I, we conduct 5 i.i.d. instances and plot the error bar. The results are displayed in the right panel of Figure 10.

12 Additional result on International Trade Flow Data

We now compare [13]'s method (convex relaxation) and [24]'s method (tubal-tRPCA) with our method in terms of prediction error on the international trade flow dataset. As in Section 7, we analyze the tensor $\log(1 + \boldsymbol{\mathcal{A}})$. We split $\log(1 + \boldsymbol{\mathcal{A}})$ into two parts, namely $\log(1 + \boldsymbol{\mathcal{A}}) =: \boldsymbol{\mathcal{A}}_{\text{train}} + \boldsymbol{\mathcal{A}}_{\text{test}}$, where $\boldsymbol{\mathcal{A}}_{\text{test}}$ is generated by randomly taking 10% of the non-zero entries of $\log(1 + \boldsymbol{\mathcal{A}})$. We then apply tubal-tRPCA with the default parameter the authors provide [3] and convex relaxation with carefully tuned parameters [4]. We use the proposed BIC-type criterion to select the rank and sparsity. As the left panel of Figure 11 suggests, we choose $\mathbf{r} = (3, 3, 3)^\top$, and the right panel of Figure 11 shows the BIC is less sensitive to α for a small range. Therefore we set the rank as $\mathbf{r} = (3, 3, 3)^\top$ and try $\alpha = 0.01, 0.02, 0.03$. The error is measured in terms of the test error $\|[\widehat{\boldsymbol{\mathcal{T}}} + \widehat{\boldsymbol{\mathcal{S}}}]_{\Omega_{\text{test}}} - \boldsymbol{\mathcal{A}}_{\text{test}}\|_{\text{F}}$ and the results are presented in Table 2.

When $\alpha = 0$, all methods perform poorly because the existence of outliers distort the low-rank

[3] Their codes are available at https://github.com/canyilu/tensor-completion-under-linear-transform.

[4] The codes in [13] is not publicly released so we have to tune the parameters by ourselves.

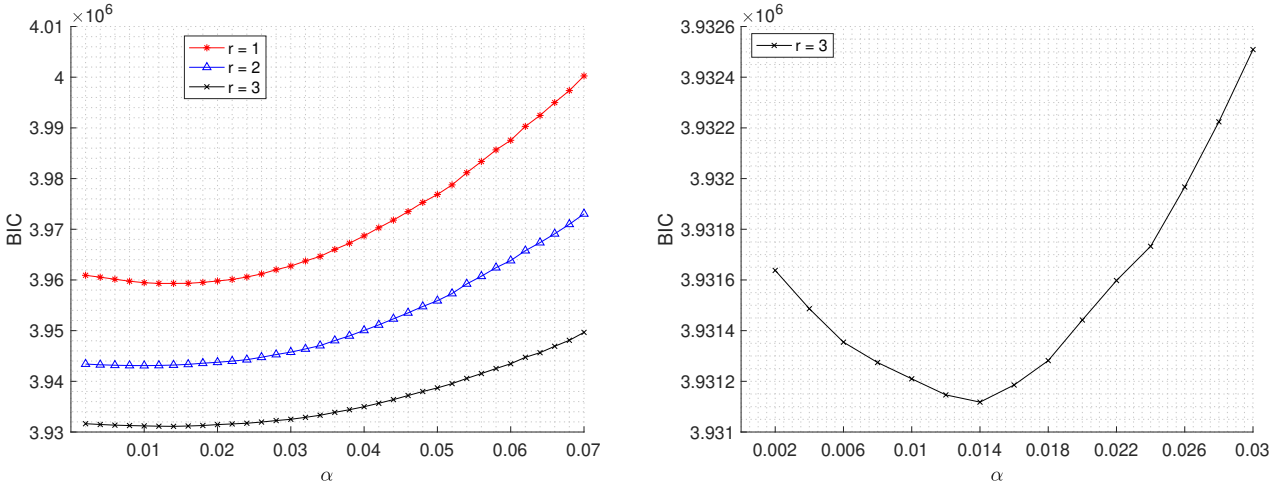


Figure 11: BIC values on the International Trade Flow Data. Left: BIC values for different rank and sparsity; Right: Zoom in on the case $r = 3$.

Method	Convex [13]	tubal-tRPCA [24]	Our method ($\alpha = 0$)	Our method ($\alpha = 0.01$)	Our method ($\alpha = 0.02$)	Our method ($\alpha = 0.03$)
Pred. Error	1892.3	1894.2	1891.2	**693.5**	800.5	980.1

Table 2: Comparison of our method with convex relaxation [13] and tubal-tRPCA [24] in terms of prediction error on the international trade flow data. Our BIC criterion suggests any α between 0.002 and 0.03. We note that our method with $\alpha = 0.003$ yields a prediction error 566.0.

estimate making it ineffective in prediction. Meanwhile, if α is too large, say 0.1, the sparse component might incorrectly absorb useful information from the low-rank component which, as a result, sabotages its prediction accuracy. Fortunately, our method with the BIC suggested α indeed significantly outperforms other methods.

13 Real Data: Statisticians Hypergraph Co-authorship Network

This dataset [16] contains the co-authorship relations of 3607 statisticians based on 3248 papers published in four prestigious statistics journals during 2003-2012. The co-authorship network thus has 3607 nodes and two nodes are connected by an edge if they collaborated on at least one paper.

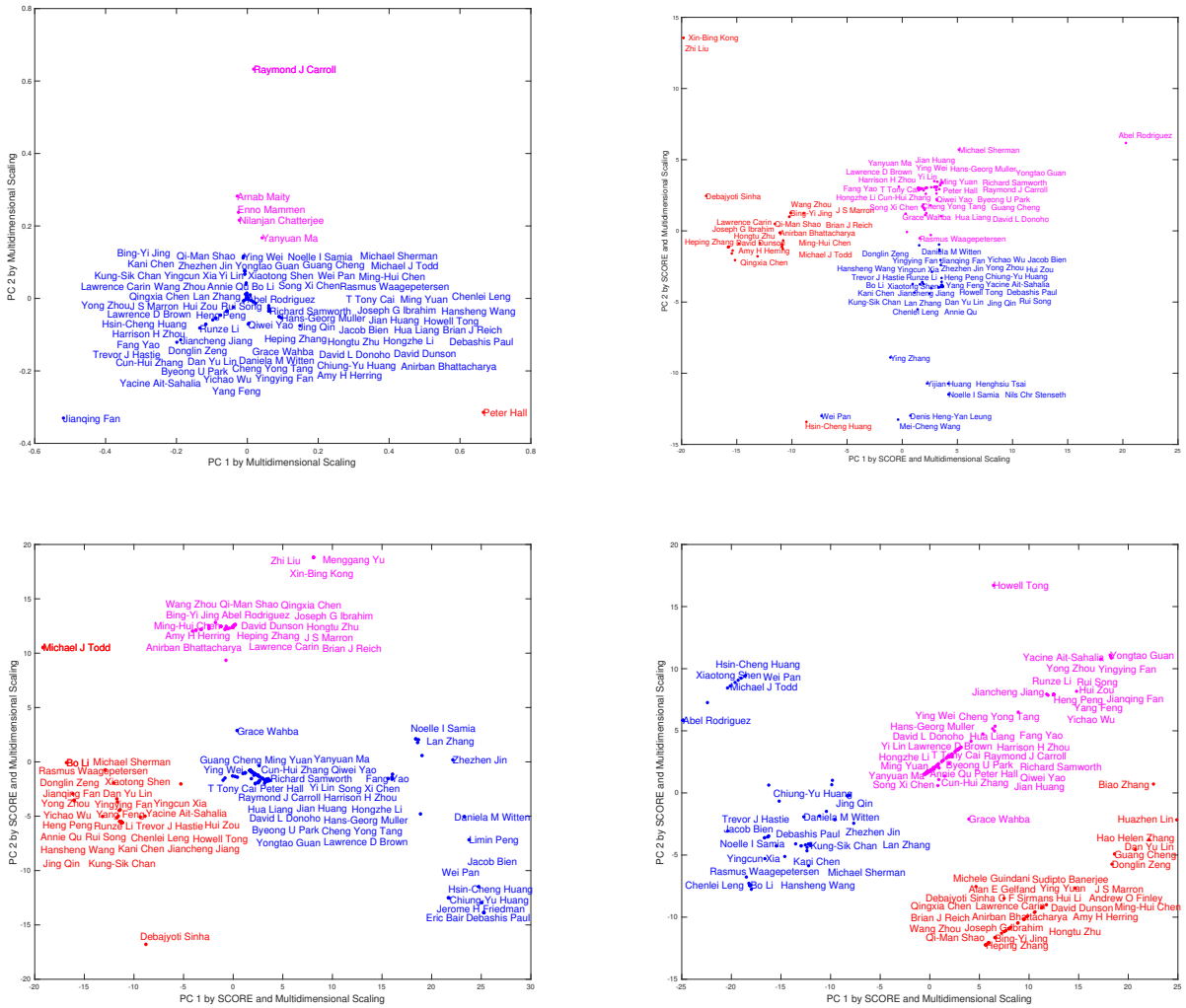
A giant connected component of this network consisting of 236 nodes is seen to be the "High-Dimensional Data Analysis" community. They also carried out community detection analysis to discover substructures in this giant component. See more details in [16].

We analyze the substructures of the giant component by treating it as a hypergraph co-authorship network. These 236 statisticians co-authored 542 papers[5], among which 356 papers have two co-authors, 162 papers have three co-authors and 24 papers have four co-authors. A 3-uniform hypergraph co-authorship network is constructed by, for $i \neq j \neq k$, adding the hyperedge (i, j, k) if the authors i, j, k co-authored at least one paper, and adding the hyperedges (i, i, j) and (i, j, j) if the authors i, j co-authored at least one paper. The hyperedges are *undirected* resulting into a symmetric adjacency tensor $\boldsymbol{\mathcal{A}}$. We adopt the framework from Section 5.1 to learn the latent low-rank tensor $\widehat{\boldsymbol{\mathcal{T}}}$ in $\boldsymbol{\mathcal{A}}$, which is used to detect communities in the giant component. We emphasize that our primary goal is to present the new findings by taking into consideration of higher-order interactions among co-authors and applying novel robust tensor methods. It is not our intention to label an author with a certain community.

The Tucker ranks are set as $(4, 4, 4)$ and sparsity ratio α is varied at $\{0, 10^{-4}, 5 \times 10^{-4}\}$. The number of communities is set at $K = 3$ and the algorithm is initialized by the HOSVD of $\boldsymbol{\mathcal{A}}$. To uncover community structures, we apply spectral clustering to the singular vectors of $\widehat{\boldsymbol{\mathcal{T}}}$. The node degrees are severely heterogeneous with Peter Hall, Jianqing Fan and Raymond Carroll being the top-3 statisticians in terms of # of co-authors. The naive spectral clustering often performs poorly in the existence of heterogeneity, skewing to the high-degree nodes. Indeed, the top-left plot in Figure 12 shows that the naive spectral clustering identifies these three statisticians as the corners in a triangle, and puts Peter Hall in a single community. To mitigate the influence of node heterogeneity, we apply SCORE [17] for community detection, which uses the leading singular vector of $\widehat{\boldsymbol{\mathcal{T}}}$ as normalization.

The community structures found by SCORE are displayed in Figure 12. The top-right plot shows the three clusters identified by SCORE when the sparsity ratio is zero. The three communities are: 1). "North Carolina" group including researchers from Duke University, University of North Carolina and North Carolina State University, together with their close collaborators

[5]There are 328 single-authored papers. They provide no information to co-authorship relations, and are left out in our analysis.



(a) Top-left: $\alpha = 10^{-4}$ and naive spectral clustering; top-right: $\alpha = 0$ and SCORE; bottom-left: $\alpha = 10^{-4}$ and SCORE; bottom-right: $\alpha = 5 \times 10^{-4}$ and SCORE.

Figure 12: Sub-structures detected in the "High-Dimensional Data Analysis" community based on the hypergraph co-authorship network. The Tucker ranks are set as $(4, 4, 4)$ with varied sparsity ratio at $\{0, 10^{-4}, 5 \times 10^{-4}\}$ and the algorithm is initialized by the HOSVD of adjacency tensor $\boldsymbol{\mathcal{A}}$.

such as Debajyoti Sinha, Qi-Man Shao, Bing-Yi Jing, Michael J Todd and etc.; 2). "Carroll-Hall" group including researchers in non-parametric and semi-parametric statistics, functional estimation and high-dimensional statistics, together with collaborators; 3). "Fan and Others" group[6] including *primarily* the researchers collaborating closely with Jianqing Fan or his co-authors, and other researchers who do not *obviously* belong to the first two groups. We note that the fields of researchers in "Fan and Others" group are quite diverse, some of which overlap with those in "Carroll-Hall" group and "North Carolina" group. However, unlike the results in [16], the top-right plot in Figure 12 does not cluster the "Fan and Others" group into either the "North Carolina" group or "Carroll-Hall" group.

We then set the sparsity ratio of $\widehat{\boldsymbol{\mathcal{S}}}$ by $\alpha = 10^{-4}$. The communities identified by SCORE based on the singular vectors of $\widehat{\boldsymbol{\mathcal{T}}}$ are illustrated in the bottom-left plot of Figure 12. Compared with the top-right plot ($\alpha = 0$), the three communities displayed in the bottom-left plot largely remain the same. But the group memberships of some authors do change. Notably, Debajyoti Sinha and Michael J Todd move from the "North Carolina" group to "Fan and Others" group; Abel Rodriguez moves from the "Carroll-Hall" group to "North-Carolina" group; several authors (e.g. Daniela M Witten, Jacob Bien, Pan Wei, Chiung-Yu Huang, Debashis Paul, Zhezhen Jin, Lan Zhang and etc.) move from the "Fan and Others" group to "Carroll-Hall" group; Hsin-Cheng Huang moves from the "North Carolina" group to "Carroll-Hall" group; Rasmus Waggepetersen moves from the "Carroll-Hall" group to "Fan and Others" group. These changes of memberships suggest that these authors may not have strong ties to the "North Carolina", "Carroll-Hall" group or be the co-authors of Jianqing Fan. It may be more reasonable that these authors constitute a separate group.

This indeed happens when the sparsity ratio α increases to a certain level. The bottom-right plot of Figure 12 shows the clustering result of SCORE when $\alpha = 5 \times 10^{-4}$. Compared with the top-right ($\alpha = 0$) and bottom-left ($\alpha = 10^{-4}$) plots, the community structure has a significant change. Indeed, the "Fan and Others" group now splits into a "Fan" group including Jianqing Fan and his co-authors, and an "Others" group including the researchers who do not have obvious ties with "Fan" group. Moreover, the "Fan" group merges into the "Carroll-Hall" group, which

[6]We name it the "Fan and Others" group simply because many researchers in this group are the co-authors of Jianqing Fan. It is not our intention to rank/label the authors.

coincides with the clustering result of SCORE when applied onto the graph co-authorship network (Fig. 6 in [16]). Consequently, we name the three communities in the top-right plot by the "North Carolina", "Carroll-Fan-Hall" and "Others" group. Interestingly, many of the authors in the "Others" group are those whose memberships change when the sparsity ratio α increases from 0 to 10^{-4}. See the top-right and bottom-left plots of Figure 12. In addition, we observe that, as α increases from 10^{-4} to 5×10^{-4}, Donglin Zeng and Dan Yu Lin in the "Fan and Others" group moves to "North Carolina" group. This might be more reasonable since they both work at the University of North Carolina.

14 Proofs of theorems

14.1 Proof of Theorem 4.1

We prove the theorem by induction on $\|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}$ and $\|\widehat{\boldsymbol{\mathcal{S}}}_l - \boldsymbol{\mathcal{S}}^*\|_{\mathrm{F}}$ alternatively. From the initialization condition we have $\|\widehat{\boldsymbol{\mathcal{T}}}_0 - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}} \le c_{1,m} \min\{\frac{\delta^2}{\sqrt{\bar{r}}}, (\kappa_0^{2m}\sqrt{\bar{r}})^{-1}\} \cdot \underline{\lambda}$ and $\widehat{\boldsymbol{\mathcal{T}}}_0 \in \mathbb{B}^*_\infty$ is $(2\mu_1\kappa_0)^2$-incoherent.

Step 1: Bounding $\|\widehat{\boldsymbol{\mathcal{S}}}_l - \boldsymbol{\mathcal{S}}^*\|_{\mathrm{F}}$ for all $l \ge 0$. Suppose we have $\widehat{\boldsymbol{\mathcal{T}}}_l \in \mathbb{B}^*_\infty$ is $(2\mu_1\kappa_0)^2$-incoherent and $\|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}} \le c_{1,m} \min\{\frac{\delta^2}{\sqrt{\bar{r}}}, (\kappa_0^{2m}\sqrt{\bar{r}})^{-1}\} \cdot \underline{\lambda}$.

Now we estimate $\|\widehat{\boldsymbol{\mathcal{S}}}_l - \boldsymbol{\mathcal{S}}^*\|_{\mathrm{F}}$. Denote $\Omega_l = \mathrm{supp}(\widehat{\boldsymbol{\mathcal{S}}}_l)$ and $\Omega^* = \mathrm{supp}(\boldsymbol{\mathcal{S}}^*)$. For $\forall \omega \in \Omega_l$, from the construction of $\widehat{\boldsymbol{\mathcal{S}}}_l$ in Algorithm 1, we have by the definition of Err_∞,

$$|[\nabla \mathfrak{L}(\widehat{\boldsymbol{\mathcal{T}}}_l + \widehat{\boldsymbol{\mathcal{S}}}_l)]_\omega| \le \min_{\|\boldsymbol{\mathcal{X}}\|_{\ell_\infty} \le \mathsf{k_{pr}}} \|\nabla \mathfrak{L}(\boldsymbol{\mathcal{X}})\|_{\ell_\infty} \le \mathsf{Err}_\infty \tag{14.1}$$

From Assumption 3, we get

$$|[\nabla \mathfrak{L}(\widehat{\boldsymbol{\mathcal{T}}}_l + \widehat{\boldsymbol{\mathcal{S}}}_l)]_\omega - [\nabla \mathfrak{L}(\widehat{\boldsymbol{\mathcal{T}}}_l + \boldsymbol{\mathcal{S}}^*)]_\omega| \ge b_l |[\widehat{\boldsymbol{\mathcal{S}}}_l - \boldsymbol{\mathcal{S}}^*]_\omega|. \tag{14.2}$$

Note that to use (14.2), we shall verify the neighborhood condition. From the upper bound of $\|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}$ we have $\|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}} \le \underline{\lambda}/8$, and $\widehat{\boldsymbol{\mathcal{T}}}_l$ is $(2\mu_1\kappa_0)^2$-incoherent. Therefore, from Lemma 15.7, we have:

$$|[\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*]_\omega|^2 \le C_{1,m} \bar{r}^m \underline{d}^{-(m-1)} (\mu_1\kappa_0)^{4m} \|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2.$$

So we have

$$|[\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*]_\omega| \le C_{1,m}\sqrt{\frac{\bar{r}^m}{\underline{d}^{m-1}}}(\mu_1\kappa_0)^{2m}\|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}} \le C_{1,m}\mu_1^{2m}\sqrt{\frac{\bar{r}^{m-1}}{\underline{d}^{m-1}}}\underline{\lambda},$$

where the last inequality is from the upper bound of $\|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}$. As a result, we have

$$\big|[\widehat{\boldsymbol{\mathcal{T}}}_l + \widehat{\boldsymbol{\mathcal{S}}}_l - \boldsymbol{\mathcal{T}}^* - \boldsymbol{\mathcal{S}}^*]_\omega\big| \le \big|[\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*]_\omega\big| + |[\widehat{\boldsymbol{\mathcal{S}}}_l]_\omega| + |[\boldsymbol{\mathcal{S}}^*]_\omega| \le C_{1,m}\mu_1^{2m}\sqrt{\frac{\bar{r}^{m-1}}{\underline{d}^{m-1}}}\underline{\lambda} + \mathsf{k_{pr}} + \|\boldsymbol{\mathcal{S}}^*\|_{\ell_\infty}.$$

Thus, both $\widehat{\boldsymbol{\mathcal{T}}}_l + \widehat{\boldsymbol{\mathcal{S}}}_l$ and $\widehat{\boldsymbol{\mathcal{T}}}_l + \boldsymbol{\mathcal{S}}^*$ belong to the ball $\mathbb{B}^*_\infty$ and thus (14.2) holds.

As a result of (14.1) and (14.2), we get for any $\omega \in \Omega_l$

$$b_l\big|[\widehat{\boldsymbol{\mathcal{S}}}_l - \boldsymbol{\mathcal{S}}^*]_\omega\big| \le \big|[\nabla\mathfrak{L}(\widehat{\boldsymbol{\mathcal{T}}}_l + \boldsymbol{\mathcal{S}}^*)]_\omega\big| + \mathsf{Err}_\infty.$$

Therefore,

$$\begin{aligned}
\|\mathcal{P}_{\Omega_l}(\widehat{\boldsymbol{\mathcal{S}}}_l - \boldsymbol{\mathcal{S}}^*)\|_{\mathrm{F}}^2 &\le \frac{2}{b_l^2}\|\mathcal{P}_{\Omega_l}(\nabla\mathfrak{L}(\widehat{\boldsymbol{\mathcal{T}}}_l + \boldsymbol{\mathcal{S}}^*))\|_{\mathrm{F}}^2 + \frac{2|\Omega_l|}{b_l^2}\mathsf{Err}_\infty^2 \\
&= \frac{2}{b_l^2}\|\mathcal{P}_{\Omega_l}(\nabla\mathfrak{L}(\widehat{\boldsymbol{\mathcal{T}}}_l + \boldsymbol{\mathcal{S}}^*)) - \mathcal{P}_{\Omega_l}(\nabla\mathfrak{L}(\boldsymbol{\mathcal{T}}^* + \boldsymbol{\mathcal{S}}^*)) + \mathcal{P}_{\Omega_l}(\nabla\mathfrak{L}(\boldsymbol{\mathcal{T}}^* + \boldsymbol{\mathcal{S}}^*))\|_{\mathrm{F}}^2 + \frac{2|\Omega_l|}{b_l^2}\mathsf{Err}_\infty^2 \\
&\le \frac{4}{b_l^2}\|\mathcal{P}_{\Omega_l}(\nabla\mathfrak{L}(\widehat{\boldsymbol{\mathcal{T}}}_l + \boldsymbol{\mathcal{S}}^*)) - \mathcal{P}_{\Omega_l}(\nabla\mathfrak{L}(\boldsymbol{\mathcal{T}}^* + \boldsymbol{\mathcal{S}}^*))\|_{\mathrm{F}}^2 + \frac{4}{b_l^2}\|\mathcal{P}_{\Omega_l}(\nabla\mathfrak{L}(\boldsymbol{\mathcal{T}}^* + \boldsymbol{\mathcal{S}}^*))\|_{\mathrm{F}}^2 + \frac{2|\Omega_l|}{b_l^2}\mathsf{Err}_\infty^2 \\
&\le \frac{4b_u^2}{b_l^2}\|\mathcal{P}_{\Omega_l}(\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*)\|_{\mathrm{F}}^2 + \frac{6|\Omega_l|}{b_l^2}\mathsf{Err}_\infty^2, \qquad (14.3)
\end{aligned}$$

where the last inequality is due to $\|\mathcal{P}_{\Omega_l}(\nabla\mathfrak{L}(\boldsymbol{\mathcal{T}}^* + \boldsymbol{\mathcal{S}}^*))\|_{\mathrm{F}}^2 \le |\Omega_l|\mathsf{Err}_\infty^2$ and Assumption 3 since $\widehat{\boldsymbol{\mathcal{T}}}_l + \boldsymbol{\mathcal{S}}^* \in \mathbb{B}^*_\infty$.

From (14.3), Lemma 15.8, we have

$$\|\mathcal{P}_{\Omega_l}(\widehat{\boldsymbol{\mathcal{S}}}_l - \boldsymbol{\mathcal{S}}^*)\|_{\mathrm{F}}^2 \le \frac{C_{2,m}b_u^2}{b_l^2}(\mu_1\kappa_0)^{4m}\bar{r}^m\alpha\|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2 + \frac{6|\Omega_l|}{b_l^2}\mathsf{Err}_\infty^2 \qquad (14.4)$$

here $C_{2,m} > 0$ is an absolute constant depending only on m.

For $\forall\omega = (\omega_1, \ldots, \omega_m) \in \Omega^*\backslash\Omega_l$, we have $|[\widehat{\boldsymbol{\mathcal{S}}}_l - \boldsymbol{\mathcal{S}}^*]_\omega| = |[\boldsymbol{\mathcal{S}}^*]_\omega|$. Since the loss function is entry-wise by Assumption 3, we have $[\nabla\mathfrak{L}(\widehat{\boldsymbol{\mathcal{T}}}_l)]_\omega = [\nabla\mathfrak{L}(\widehat{\boldsymbol{\mathcal{T}}}_l + \widehat{\boldsymbol{\mathcal{S}}}_l)]_\omega$. Clearly, $\widehat{\boldsymbol{\mathcal{T}}}_l$ and $\widehat{\boldsymbol{\mathcal{T}}}_l + \boldsymbol{\mathcal{S}}^*$ both belong to $\mathbb{B}^*_\infty$, by Assumption 3 we get

$$|[\nabla\mathfrak{L}(\widehat{\boldsymbol{\mathcal{T}}}_l)]_\omega - [\nabla\mathfrak{L}(\widehat{\boldsymbol{\mathcal{T}}}_l + \boldsymbol{\mathcal{S}}^*)]_\omega| \ge b_l|[\boldsymbol{\mathcal{S}}^*]_\omega|.$$

Now we bound $|[\widehat{\boldsymbol{\mathcal{S}}}_l - \boldsymbol{\mathcal{S}}^*]_\omega|$ as follows. For any $\omega \in \Omega^* \backslash \Omega_l$,

$$\begin{aligned}
|[\widehat{\boldsymbol{\mathcal{S}}}_l - \boldsymbol{\mathcal{S}}^*]_\omega| = |[\boldsymbol{\mathcal{S}}^*]_\omega| &\leq \frac{1}{b_l}|[\nabla \mathfrak{L}(\widehat{\boldsymbol{\mathcal{T}}}_l)]_\omega - [\nabla \mathfrak{L}(\widehat{\boldsymbol{\mathcal{T}}}_l + \boldsymbol{\mathcal{S}}^*)]_\omega| \\
&\leq \frac{1}{b_l}\Big(|[\nabla \mathfrak{L}(\widehat{\boldsymbol{\mathcal{T}}}_l)]_\omega| + |[\nabla \mathfrak{L}(\widehat{\boldsymbol{\mathcal{T}}}_l + \boldsymbol{\mathcal{S}}^*)]_\omega|\Big) \\
&\leq \frac{1}{b_l}\Big(|[\nabla \mathfrak{L}(\widehat{\boldsymbol{\mathcal{T}}}_l)]_\omega| + |[\nabla \mathfrak{L}(\widehat{\boldsymbol{\mathcal{T}}}_l + \boldsymbol{\mathcal{S}}^*) - \nabla \mathfrak{L}(\boldsymbol{\mathcal{T}}^* + \boldsymbol{\mathcal{S}}^*)]_\omega| + |[\nabla \mathfrak{L}(\boldsymbol{\mathcal{T}}^* + \boldsymbol{\mathcal{S}}^*)]_\omega|\Big) \\
&\leq \frac{1}{b_l}|[\nabla \mathfrak{L}(\widehat{\boldsymbol{\mathcal{T}}}_l)]_\omega| + \frac{b_u}{b_l}|[\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*]_\omega| + \frac{1}{b_l}\mathsf{Err}_\infty,
\end{aligned}$$

where the last inequality is again due to Assumption 3 since $\widehat{\boldsymbol{\mathcal{T}}}_l + \boldsymbol{\mathcal{S}}^* \in \mathbb{B}^*_\infty$. Therefore we have

$$\|\mathcal{P}_{\Omega^* \backslash \Omega_l}(\widehat{\boldsymbol{\mathcal{S}}}_l - \boldsymbol{\mathcal{S}}^*)\|_{\mathrm{F}}^2 \leq \frac{2}{b_l^2}\|\mathcal{P}_{\Omega^* \backslash \Omega_l}(\nabla \mathfrak{L}(\widehat{\boldsymbol{\mathcal{T}}}_l))\|_{\mathrm{F}}^2 + \frac{4b_u^2}{b_l^2}\|\mathcal{P}_{\Omega^* \backslash \Omega_l}(\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*)\|_{\mathrm{F}}^2 + \frac{4}{b_l^2}|\Omega^* \backslash \Omega_l|\mathsf{Err}_\infty^2 \quad (14.5)$$

Since $\omega \in \Omega^* \backslash \Omega_l$, we have

$$|[\nabla \mathfrak{L}(\widehat{\boldsymbol{\mathcal{T}}}_l)]_\omega| \leq \max_{i=1}^m |\mathbf{e}_{\omega_i}^\top \mathcal{M}_i(\nabla \mathfrak{L}(\widehat{\boldsymbol{\mathcal{T}}}_l))|^{(\gamma \alpha d_i^-)} \quad (14.6)$$

Now since we have $\boldsymbol{\mathcal{S}}^* \in \mathbb{S}_\alpha$, we have

$$\begin{aligned}
|[\nabla \mathfrak{L}(\widehat{\boldsymbol{\mathcal{T}}}_l)]_\omega| &\leq \max_{i=1}^m |\mathbf{e}_{\omega_i}^T \mathcal{M}_i(\nabla \mathfrak{L}(\widehat{\boldsymbol{\mathcal{T}}}_l + \boldsymbol{\mathcal{S}}^*))|^{\left((\gamma-1)\alpha d_i^-\right)} \\
&\leq \max_{i=1}^m \left|\mathbf{e}_{\omega_i}^\top \Big(\mathcal{M}_i(\nabla \mathfrak{L}(\widehat{\boldsymbol{\mathcal{T}}}_l + \boldsymbol{\mathcal{S}}^*)) - \mathcal{M}_i(\nabla \mathfrak{L}(\boldsymbol{\mathcal{T}}^* + \boldsymbol{\mathcal{S}}^*))\Big)\right|^{\left((\gamma-1)\alpha d_i^-\right)} + \mathsf{Err}_\infty \quad (14.7)
\end{aligned}$$

Using AM-GM inequality, we have:

$$\begin{aligned}
|[\nabla \mathfrak{L}(\widehat{\boldsymbol{\mathcal{T}}}_l)]_\omega|^2 &\leq 2\max_{i=1}^m \frac{\left\|\mathbf{e}_{\omega_i}^\top \Big(\mathcal{M}_i(\nabla \mathfrak{L}(\widehat{\boldsymbol{\mathcal{T}}}_l + \boldsymbol{\mathcal{S}}^*)) - \mathcal{M}_i(\nabla \mathfrak{L}(\boldsymbol{\mathcal{T}}^* + \boldsymbol{\mathcal{S}}^*))\Big)\right\|_{\mathrm{F}}^2}{(\gamma-1)\alpha d_i^-} + 2\mathsf{Err}_\infty^2 \\
&\leq 2\sum_{i=1}^m \frac{\left\|\mathbf{e}_{\omega_i}^\top \Big(\mathcal{M}_i(\nabla \mathfrak{L}(\widehat{\boldsymbol{\mathcal{T}}}_l + \boldsymbol{\mathcal{S}}^*)) - \mathcal{M}_i(\nabla \mathfrak{L}(\boldsymbol{\mathcal{T}}^* + \boldsymbol{\mathcal{S}}^*))\Big)\right\|_{\mathrm{F}}^2}{(\gamma-1)\alpha d_i^-} + 2\mathsf{Err}_\infty^2 \quad (14.8)
\end{aligned}$$

Now for all fixed $i \in [m]$, for all $\omega_i \in [d_i]$, ω_i appears at most αd_i^- times since $\Omega^* \backslash \Omega_l$ is an α-fraction set. This observation together with (14.8) lead to the following:

$$\begin{aligned}
\|\mathcal{P}_{\Omega^* \backslash \Omega_l}(\nabla \mathfrak{L}(\widehat{\boldsymbol{\mathcal{T}}}_l))\|_{\mathrm{F}}^2 &\leq 2\sum_{i=1}^m \frac{\|\nabla \mathfrak{L}(\widehat{\boldsymbol{\mathcal{T}}}_l + \boldsymbol{\mathcal{S}}^*) - \nabla \mathfrak{L}(\boldsymbol{\mathcal{T}}^* + \boldsymbol{\mathcal{S}}^*)\|_{\mathrm{F}}^2}{\gamma - 1} + 2|\Omega^* \backslash \Omega_l|\mathsf{Err}_\infty^2 \\
&\leq \frac{2mb_u^2}{\gamma - 1}\|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2 + 2|\Omega^* \backslash \Omega_l|\mathsf{Err}_\infty^2. \quad (14.9)
\end{aligned}$$

Therefore together with (14.5) and (14.9) and Lemma 15.8, we have

$$\|\mathcal{P}_{\Omega^*\setminus\Omega_l}(\widehat{\boldsymbol{\mathcal{S}}}_l-\boldsymbol{\mathcal{S}}^*)\|_{\mathrm{F}}^2\leq\left(\frac{4mb_u^2}{b_l^2}\frac{1}{\gamma-1}+C_{4,m}\frac{b_u^2}{b_l^2}(\mu_1\kappa_0)^{4m}\bar{r}^m\alpha\right)\|\widehat{\boldsymbol{\mathcal{T}}}_l-\boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2+\frac{16}{b_l^2}|\Omega^*\setminus\Omega_l|\mathsf{Err}_\infty^2 \tag{14.10}$$

where $C_{4,m}>0$ are constants depending only on m. Now we combine (14.4) and (14.10) and we get

$$\|\widehat{\boldsymbol{\mathcal{S}}}_l-\boldsymbol{\mathcal{S}}^*\|_{\mathrm{F}}^2\leq\left(\frac{4mb_u^2}{b_l^2}\frac{1}{\gamma-1}+C_{5,m}(\mu_1\kappa_0)^{4m}\bar{r}^m\frac{b_u^2}{b_l^2}\alpha\right)\|\widehat{\boldsymbol{\mathcal{T}}}_l-\boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2+\frac{C_1}{b_l^2}|\Omega^*\cup\Omega_l|\mathsf{Err}_\infty^2 \tag{14.11}$$

where $C_{5,m}>0$ depending only on m and $C_1>0$ an absolute constant.

Now if we choose $\alpha\leq(C_{5,m}\kappa_0^{4m}\mu_0^{4m}\bar{r}^m\frac{b_u^4}{b_l^4})^{-1}$ and $\gamma-1\geq 4m\frac{b_u^4}{b_l^4}$ for some sufficient large constants $C_{5,m}>0$ depending only on m, then we have

$$\|\widehat{\boldsymbol{\mathcal{S}}}_l-\boldsymbol{\mathcal{S}}^*\|_{\mathrm{F}}^2\leq\frac{b_l^2}{b_u^2}\|\widehat{\boldsymbol{\mathcal{T}}}_l-\boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2+\frac{C_1}{b_l^2}|\Omega^*\cup\Omega_l|\mathsf{Err}_\infty^2 \tag{14.12}$$

and

$$\|\widehat{\boldsymbol{\mathcal{S}}}_l-\boldsymbol{\mathcal{S}}^*\|_{\mathrm{F}}\leq\frac{b_l}{b_u}\|\widehat{\boldsymbol{\mathcal{T}}}_l-\boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}+\frac{C_1}{b_l}\sqrt{|\Omega^*\cup\Omega_l|}\mathsf{Err}_\infty \tag{14.13}$$

In addition, from the upper bound of $\|\boldsymbol{\mathcal{T}}_l-\boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}$, (14.13) implies that $\|\widehat{\boldsymbol{\mathcal{S}}}_l-\boldsymbol{\mathcal{S}}^*\|_{\mathrm{F}}\leq c_0\underline{\lambda}$ for a small $c_0>0$. This fact is helpful later since it implies that $\widehat{\boldsymbol{\mathcal{T}}}_l+\widehat{\boldsymbol{\mathcal{S}}}_l$ belongs to the ball $\mathbb{B}_2^*$ and thus activates the conditions in Assumption 2.

Step 2: bounding $\|\widehat{\boldsymbol{\mathcal{T}}}_l-\boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2$ for all $l\geq 1$. From previous step, we have verified

$$\|\widehat{\boldsymbol{\mathcal{S}}}_{l-1}-\boldsymbol{\mathcal{S}}^*\|_{\mathrm{F}}\leq\frac{b_l}{b_u}\|\widehat{\boldsymbol{\mathcal{T}}}_l-\boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}+\frac{C_1}{b_l}\sqrt{|\Omega^*\cup\Omega_l|}\mathsf{Err}_\infty\leq c_0\underline{\lambda}. \tag{14.14}$$

And from the Algorithm 2, $\widehat{\boldsymbol{\mathcal{T}}}_l=\mathsf{Trim}_{\zeta_l,\mathbf{r}}(\boldsymbol{\mathcal{W}}_{l-1})$. Now from Lemma 15.6, we get,

$$\begin{aligned}
\|\widehat{\boldsymbol{\mathcal{T}}}_l-\boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2&=\|\mathsf{Trim}_{\zeta_l,\mathbf{r}}(\boldsymbol{\mathcal{W}}_{l-1})-\boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2\\
&\leq\|\boldsymbol{\mathcal{W}}_{l-1}-\boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2+C_m\frac{\sqrt{\bar{r}}}{\underline{\lambda}}\|\boldsymbol{\mathcal{W}}_{l-1}-\boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^3\\
&\leq(1+\frac{\delta}{4})\|\boldsymbol{\mathcal{W}}_{l-1}-\boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2\\
&\leq(1-\delta^2)\|\widehat{\boldsymbol{\mathcal{T}}}_{l-1}-\boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2+6\delta^{-1}\mathsf{Err}_{\mathbf{2r}}+C_1\left(1+b_u+b_u^2\right)b_l^{-2}\left(|\Omega^*|+\gamma\alpha d^*\right)\mathsf{Err}_\infty^2
\end{aligned} \tag{14.15}$$

Notice to use Lemma 15.6, we need to verify $\|\boldsymbol{\mathcal{W}}_{l-1} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}} \leq \underline{\lambda}/8$, which we will check momentarily. Also, from (14.15) and the signal-to-noise ration condition, we get

$$\|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}} \leq c_1 \min\{\delta^2 \bar{r}^{-1/2}, \kappa_0^{-2m} \bar{r}^{-1/2}\} \cdot \underline{\lambda}.$$

On the other hand, from lemma 15.6, we have $\widehat{\boldsymbol{\mathcal{T}}}_l$ is $(2\mu_1\kappa_0)^2$-incoherent. Further, from Lemma 15.7 and the definition of k_∞ we have $\widehat{\boldsymbol{\mathcal{T}}}_l \in \mathbb{B}_\infty^*$. This finishes the induction for the error $\|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}$. Now the only thing we need to check is the upper bound for $\|\boldsymbol{\mathcal{W}}_{l-1} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}$.

Step 2.1: bounding $\|\boldsymbol{\mathcal{W}}_{l-1} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}$. From the Algorithm 2, we have for arbitrary $1 \geq \delta > 0$,

$$\begin{aligned}\|\boldsymbol{\mathcal{W}}_{l-1} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2 &= \|\widehat{\boldsymbol{\mathcal{T}}}_{l-1} - \boldsymbol{\mathcal{T}}^* - \beta \mathcal{P}_{\mathbb{T}_{l-1}}(\boldsymbol{\mathcal{G}}_{l-1} - \boldsymbol{\mathcal{G}}^*) - \beta \mathcal{P}_{\mathbb{T}_{l-1}} \boldsymbol{\mathcal{G}}^*\|_{\mathrm{F}}^2 \\ &\leq (1 + \frac{\delta}{2}) \|\widehat{\boldsymbol{\mathcal{T}}}_{l-1} - \boldsymbol{\mathcal{T}}^* - \beta \mathcal{P}_{\mathbb{T}_{l-1}}(\boldsymbol{\mathcal{G}}_{l-1} - \boldsymbol{\mathcal{G}}^*)\|_{\mathrm{F}}^2 + (1 + \frac{2}{\delta}) \beta^2 \|\mathcal{P}_{\mathbb{T}_{l-1}}(\boldsymbol{\mathcal{G}}^*)\|_{\mathrm{F}}^2 \quad (14.16)\end{aligned}$$

Now we consider the bound for $\|\widehat{\boldsymbol{\mathcal{T}}}_{l-1} - \boldsymbol{\mathcal{T}}^* - \beta \mathcal{P}_{\mathbb{T}_{l-1}}(\boldsymbol{\mathcal{G}}_{l-1} - \boldsymbol{\mathcal{G}}^*)\|_{\mathrm{F}}^2$,

$$\begin{aligned}\|\widehat{\boldsymbol{\mathcal{T}}}_{l-1} - \boldsymbol{\mathcal{T}}^* - \beta \mathcal{P}_{\mathbb{T}_{l-1}}(\boldsymbol{\mathcal{G}}_{l-1} - \boldsymbol{\mathcal{G}}^*)\|_{\mathrm{F}}^2 &= \|\widehat{\boldsymbol{\mathcal{T}}}_{l-1} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2 - 2\beta \langle \widehat{\boldsymbol{\mathcal{T}}}_{l-1} - \boldsymbol{\mathcal{T}}^*, \mathcal{P}_{\mathbb{T}_{l-1}}(\boldsymbol{\mathcal{G}}_{l-1} - \boldsymbol{\mathcal{G}}^*) \rangle \\ &\quad + \beta^2 \|\mathcal{P}_{\mathbb{T}_{l-1}}(\boldsymbol{\mathcal{G}}_{l-1} - \boldsymbol{\mathcal{G}}^*)\|_{\mathrm{F}}^2 \quad (14.17)\end{aligned}$$

The upper bound of $\|\widehat{\boldsymbol{\mathcal{S}}}_{l-1} - \boldsymbol{\mathcal{S}}^*\|_{\mathrm{F}}$ ensures that $\widehat{\boldsymbol{\mathcal{T}}}_{l-1} + \widehat{\boldsymbol{\mathcal{S}}}_{l-1} \in \mathbb{B}_2^*$. Using the smoothness condition in Assumption 2, we get

$$\beta^2 \|\mathcal{P}_{\mathbb{T}_{l-1}}(\boldsymbol{\mathcal{G}}_{l-1} - \boldsymbol{\mathcal{G}}^*)\|_{\mathrm{F}}^2 \leq \beta^2 b_u^2 \|\widehat{\boldsymbol{\mathcal{T}}}_{l-1} + \widehat{\boldsymbol{\mathcal{S}}}_{l-1} - \boldsymbol{\mathcal{T}}^* - \boldsymbol{\mathcal{S}}^*\|_{\mathrm{F}}^2 \quad (14.18)$$

Now we consider the bound for $|\langle \widehat{\boldsymbol{\mathcal{T}}}_{l-1} - \boldsymbol{\mathcal{T}}^*, \mathcal{P}_{\mathbb{T}_{l-1}}(\boldsymbol{\mathcal{G}}_{l-1} - \boldsymbol{\mathcal{G}}^*) \rangle|$. First we have:

$$\langle \widehat{\boldsymbol{\mathcal{T}}}_{l-1} - \boldsymbol{\mathcal{T}}^*, \mathcal{P}_{\mathbb{T}_{l-1}}(\boldsymbol{\mathcal{G}}_{l-1} - \boldsymbol{\mathcal{G}}^*) \rangle = \langle \widehat{\boldsymbol{\mathcal{T}}}_{l-1} - \boldsymbol{\mathcal{T}}^*, \boldsymbol{\mathcal{G}}_{l-1} - \boldsymbol{\mathcal{G}}^* \rangle - \langle \widehat{\boldsymbol{\mathcal{T}}}_{l-1} - \boldsymbol{\mathcal{T}}^*, \mathcal{P}_{\mathbb{T}_{l-1}}^{\perp}(\boldsymbol{\mathcal{G}}_{l-1} - \boldsymbol{\mathcal{G}}^*) \rangle.$$

The estimation of $\langle \widehat{\boldsymbol{\mathcal{T}}}_{l-1} - \boldsymbol{\mathcal{T}}^*, \boldsymbol{\mathcal{G}}_{l-1} - \boldsymbol{\mathcal{G}}^* \rangle$ is as follows:

$$\begin{aligned}\langle \widehat{\boldsymbol{\mathcal{T}}}_{l-1} - \boldsymbol{\mathcal{T}}^*, \boldsymbol{\mathcal{G}}_{l-1} - \boldsymbol{\mathcal{G}}^* \rangle &= \langle \widehat{\boldsymbol{\mathcal{T}}}_{l-1} - \boldsymbol{\mathcal{T}}^* + \widehat{\boldsymbol{\mathcal{S}}}_{l-1} - \boldsymbol{\mathcal{S}}^*, \boldsymbol{\mathcal{G}}_{l-1} - \boldsymbol{\mathcal{G}}^* \rangle - \langle \widehat{\boldsymbol{\mathcal{S}}}_{l-1} - \boldsymbol{\mathcal{S}}^*, \boldsymbol{\mathcal{G}}_{l-1} - \boldsymbol{\mathcal{G}}^* \rangle \\ &\geq b_l \|\widehat{\boldsymbol{\mathcal{T}}}_{l-1} - \boldsymbol{\mathcal{T}}^* + \widehat{\boldsymbol{\mathcal{S}}}_{l-1} - \boldsymbol{\mathcal{S}}^*\|_{\mathrm{F}}^2 - \langle \widehat{\boldsymbol{\mathcal{S}}}_{l-1} - \boldsymbol{\mathcal{S}}^*, \boldsymbol{\mathcal{G}}_{l-1} - \boldsymbol{\mathcal{G}}^* \rangle, \quad (14.19)\end{aligned}$$

where the last inequality follows from Assumption 2. And the estimation of $\langle \widehat{\boldsymbol{\mathcal{T}}}_{l-1} - \boldsymbol{\mathcal{T}}^*, \mathcal{P}_{\mathbb{T}_{l-1}}^{\perp}(\boldsymbol{\mathcal{G}}_{l-1} - \boldsymbol{\mathcal{G}}^*) \rangle$ is as follows:

$$\begin{aligned}|\langle \widehat{\boldsymbol{\mathcal{T}}}_{l-1} - \boldsymbol{\mathcal{T}}^*, \mathcal{P}_{\mathbb{T}_{l-1}}^{\perp}(\boldsymbol{\mathcal{G}}_{l-1} - \boldsymbol{\mathcal{G}}^*) \rangle| &\leq \|\mathcal{P}_{\mathbb{T}_{l-1}}^{\perp}(\widehat{\boldsymbol{\mathcal{T}}}_{l-1} - \boldsymbol{\mathcal{T}}^*)\|_{\mathrm{F}} \|\boldsymbol{\mathcal{G}}_{l-1} - \boldsymbol{\mathcal{G}}^*\|_{\mathrm{F}} \\ &\leq \frac{C_{1,m} b_u}{\underline{\lambda}} \|\widehat{\boldsymbol{\mathcal{T}}}_{l-1} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2 \|\widehat{\boldsymbol{\mathcal{T}}}_{l-1} - \boldsymbol{\mathcal{T}}^* + \widehat{\boldsymbol{\mathcal{S}}}_{l-1} - \boldsymbol{\mathcal{S}}^*\|_{\mathrm{F}} \quad (14.20)\end{aligned}$$

where the last inequality follows from Lemma 15.1. Together with (14.19) and (14.20), we get,

$$\langle \widehat{\boldsymbol{\mathcal{T}}}_{l-1} - \boldsymbol{\mathcal{T}}^*, \mathcal{P}_{\mathbb{T}_{l-1}}(\boldsymbol{\mathcal{G}}_{l-1} - \boldsymbol{\mathcal{G}}^*)\rangle \geq b_l \|\widehat{\boldsymbol{\mathcal{T}}}_{l-1} - \boldsymbol{\mathcal{T}}^* + \widehat{\boldsymbol{\mathcal{S}}}_{l-1} - \boldsymbol{\mathcal{S}}^*\|_{\mathrm{F}}^2 - \langle \widehat{\boldsymbol{\mathcal{S}}}_{l-1} - \boldsymbol{\mathcal{S}}^*, \boldsymbol{\mathcal{G}}_{l-1} - \boldsymbol{\mathcal{G}}^*\rangle - \frac{C_{1,m} b_u}{\underline{\lambda}} \|\widehat{\boldsymbol{\mathcal{T}}}_{l-1} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2 \|\widehat{\boldsymbol{\mathcal{T}}}_{l-1} - \boldsymbol{\mathcal{T}}^* + \widehat{\boldsymbol{\mathcal{S}}}_{l-1} - \boldsymbol{\mathcal{S}}^*\|_{\mathrm{F}} \quad (14.21)$$

Together with (14.18) and (14.21), we get

$$\begin{aligned}\|\widehat{\boldsymbol{\mathcal{T}}}_{l-1} - \boldsymbol{\mathcal{T}}^* - \beta \mathcal{P}_{\mathbb{T}_{l-1}}(\boldsymbol{\mathcal{G}}_{l-1} - \boldsymbol{\mathcal{G}}^*)\|_{\mathrm{F}}^2 \leq & \left(1 + 2\beta b_u \frac{C_{1,m}}{\underline{\lambda}} \|\widehat{\boldsymbol{\mathcal{T}}}_{l-1} - \boldsymbol{\mathcal{T}}^* + \widehat{\boldsymbol{\mathcal{S}}}_{l-1} - \boldsymbol{\mathcal{S}}^*\|_{\mathrm{F}}\right) \|\widehat{\boldsymbol{\mathcal{T}}}_{l-1} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2 \\ & + (\beta^2 b_u^2 - 2\beta b_l) \|\widehat{\boldsymbol{\mathcal{T}}}_{l-1} - \boldsymbol{\mathcal{T}}^* + \widehat{\boldsymbol{\mathcal{S}}}_{l-1} - \boldsymbol{\mathcal{S}}^*\|_{\mathrm{F}}^2 \\ & + 2\beta |\langle \widehat{\boldsymbol{\mathcal{S}}}_{l-1} - \boldsymbol{\mathcal{S}}^*, \boldsymbol{\mathcal{G}}_{l-1} - \boldsymbol{\mathcal{G}}^*\rangle| \end{aligned} \quad (14.22)$$

In order to bound (14.22), we derive separately the bound for each terms.

Bounding $\|\widehat{\boldsymbol{\mathcal{T}}}_{l-1} - \boldsymbol{\mathcal{T}}^* + \widehat{\boldsymbol{\mathcal{S}}}_{l-1} - \boldsymbol{\mathcal{S}}^*\|_{\mathrm{F}}^2$. From the bound for $\|\widehat{\boldsymbol{\mathcal{S}}}_{l-1} - \boldsymbol{\mathcal{S}}^*\|_{\mathrm{F}}$ in (14.14), we get,

$$\begin{aligned}\|\widehat{\boldsymbol{\mathcal{T}}}_{l-1} - \boldsymbol{\mathcal{T}}^* + \widehat{\boldsymbol{\mathcal{S}}}_{l-1} - \boldsymbol{\mathcal{S}}^*\|_{\mathrm{F}}^2 &\leq 2\|\widehat{\boldsymbol{\mathcal{T}}}_{l-1} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2 + 2\|\widehat{\boldsymbol{\mathcal{S}}}_{l-1} - \boldsymbol{\mathcal{S}}^*\|_{\mathrm{F}}^2 \\ &\leq 4\|\widehat{\boldsymbol{\mathcal{T}}}_{l-1} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2 + \frac{C_1}{b_l^2} |\Omega^* \cup \Omega_{l-1}| \mathsf{Err}_\infty^2 \end{aligned} \quad (14.23)$$

Thus,

$$\|\widehat{\boldsymbol{\mathcal{T}}}_{l-1} - \boldsymbol{\mathcal{T}}^* + \widehat{\boldsymbol{\mathcal{S}}}_{l-1} - \boldsymbol{\mathcal{S}}^*\|_{\mathrm{F}} \leq 2\|\widehat{\boldsymbol{\mathcal{T}}}_{l-1} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}} + \frac{C_1}{b_l} \sqrt{|\Omega^* \cup \Omega_{l-1}|} \mathsf{Err}_\infty \quad (14.24)$$

Bounding $|\langle \boldsymbol{\mathcal{G}}_{l-1} - \boldsymbol{\mathcal{G}}^*, \widehat{\boldsymbol{\mathcal{S}}}_{l-1} - \boldsymbol{\mathcal{S}}^*\rangle|$. We first bound $\|\boldsymbol{\mathcal{G}}_{l-1} - \boldsymbol{\mathcal{G}}^*\|_{\mathrm{F}}$ by (14.24):

$$\begin{aligned}\|\boldsymbol{\mathcal{G}}_{l-1} - \boldsymbol{\mathcal{G}}^*\|_{\mathrm{F}} &\leq b_u \|\widehat{\boldsymbol{\mathcal{T}}}_{l-1} - \boldsymbol{\mathcal{T}}^* + \widehat{\boldsymbol{\mathcal{S}}}_{l-1} - \boldsymbol{\mathcal{S}}^*\|_{\mathrm{F}} \\ &\leq 2b_u \|\widehat{\boldsymbol{\mathcal{T}}}_{l-1} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}} + \frac{C_1 b_u}{b_l} \sqrt{|\Omega^* \cup \Omega_{l-1}|} \mathsf{Err}_\infty \end{aligned} \quad (14.25)$$

Now we estimate $|\langle \boldsymbol{\mathcal{G}}_{l-1} - \boldsymbol{\mathcal{G}}^*, \widehat{\boldsymbol{\mathcal{S}}}_{l-1} - \boldsymbol{\mathcal{S}}^*\rangle|$ from (14.14) and (14.25) as follows,

$$\begin{aligned}&|\langle \boldsymbol{\mathcal{G}}_{l-1} - \boldsymbol{\mathcal{G}}^*, \widehat{\boldsymbol{\mathcal{S}}}_{l-1} - \boldsymbol{\mathcal{S}}^*\rangle| \leq \|\boldsymbol{\mathcal{G}}_{l-1} - \boldsymbol{\mathcal{G}}^*\|_{\mathrm{F}} \|\widehat{\boldsymbol{\mathcal{S}}}_{l-1} - \boldsymbol{\mathcal{S}}^*\|_{\mathrm{F}} \\ &\leq (0.02 b_l + 0.01 \beta b_u^2) \|\widehat{\boldsymbol{\mathcal{T}}}_{l-1} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2 + \frac{1}{\beta} \frac{C_1}{b_l^2} |\Omega^* \cup \Omega_{l-1}| \mathsf{Err}_\infty^2 + \frac{C_1 b_u}{b_l^2} |\Omega^* \cup \Omega_{l-1}| \mathsf{Err}_\infty^2 \end{aligned} \quad (14.26)$$

Bounding $|\langle \widehat{\boldsymbol{\mathcal{T}}}_{l-1} - \boldsymbol{\mathcal{T}}^*, \widehat{\boldsymbol{\mathcal{S}}}_{l-1} - \boldsymbol{\mathcal{S}}^* \rangle|$. From (14.14), we have

$$\begin{aligned}
|\langle \widehat{\boldsymbol{\mathcal{T}}}_{l-1} - \boldsymbol{\mathcal{T}}^*, \widehat{\boldsymbol{\mathcal{S}}}_{l-1} - \boldsymbol{\mathcal{S}}^* \rangle| &\leq \|\widehat{\boldsymbol{\mathcal{T}}}_{l-1} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}} \|\widehat{\boldsymbol{\mathcal{S}}}_{l-1} - \boldsymbol{\mathcal{S}}^*\|_{\mathrm{F}} \\
&\leq (0.01 \frac{b_l}{b_u} \|\widehat{\boldsymbol{\mathcal{T}}}_{l-1} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}} + \frac{C_1}{b_l} \sqrt{|\Omega^* \cup \Omega_{l-1}|} \mathsf{Err}_\infty) \|\widehat{\boldsymbol{\mathcal{T}}}_{l-1} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}} \\
&\leq 0.02 \|\widehat{\boldsymbol{\mathcal{T}}}_{l-1} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2 + \frac{C_1}{b_l^2} |\Omega^* \cup \Omega_{l-1}| \mathsf{Err}_\infty^2 \qquad (14.27)
\end{aligned}$$

Now we go back to (14.22) and from (14.23) - (14.27), we get:

$$\begin{aligned}
&\|\widehat{\boldsymbol{\mathcal{T}}}_{l-1} - \boldsymbol{\mathcal{T}}^* - \beta \mathcal{P}_{\mathbb{T}_{l-1}}(\boldsymbol{\mathcal{G}}_{l-1} - \boldsymbol{\mathcal{G}}^*)\|_{\mathrm{F}}^2 \\
&\leq \left(1 - 1.84\beta b_l + 5\beta^2 b_u^2\right) \|\widehat{\boldsymbol{\mathcal{T}}}_{l-1} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2 + C_1(1 + b_u + b_u^2) b_l^{-2} |\Omega^* \cup \Omega_{l-1}| \mathsf{Err}_\infty^2 \qquad (14.28)
\end{aligned}$$

where the condition $\underline{\lambda} \geq C_{1,m} \frac{b_u}{b_l} \|\widehat{\boldsymbol{\mathcal{T}}}_{l-1} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}$ is used in the last step.

By combining (14.17) and (14.28), we get

$$\begin{aligned}
\|\boldsymbol{\mathcal{W}}_{l-1} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2 &= \|\widehat{\boldsymbol{\mathcal{T}}}_{l-1} - \boldsymbol{\mathcal{T}}^* - \beta \mathcal{P}_{\mathbb{T}_{l-1}} \boldsymbol{\mathcal{G}}_{l-1}\|_{\mathrm{F}}^2 \\
&\leq (1 + \frac{\delta}{2}) \|\widehat{\boldsymbol{\mathcal{T}}}_{l-1} - \boldsymbol{\mathcal{T}}^* - \beta \mathcal{P}_{\mathbb{T}_{l-1}}(\boldsymbol{\mathcal{G}}_{l-1} - \boldsymbol{\mathcal{G}}^*)\|_{\mathrm{F}}^2 + (1 + \frac{2}{\delta}) \beta^2 \|\mathcal{P}_{\mathbb{T}_{l-1}}(\boldsymbol{\mathcal{G}}^*)\|_{\mathrm{F}}^2 \\
&\leq (1 + \frac{\delta}{2}) \left(1 - 1.84\beta b_l + 5\beta^2 b_u^2\right) \|\widehat{\boldsymbol{\mathcal{T}}}_{l-1} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2 + (1 + \frac{2}{\delta}) \beta^2 \mathsf{Err}_{\mathbf{2r}}^2 \\
&\quad + C_1 \left(1 + \beta b_u + \beta^2 b_u^2\right) b_u^{-2} |\Omega^* \cup \Omega_{l-1}| \mathsf{Err}_\infty^2 \\
&\leq (1 + \frac{\delta}{2}) \left(1 - 1.84\beta b_l + 5\beta^2 b_u^2\right) \|\widehat{\boldsymbol{\mathcal{T}}}_{l-1} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2 + (1 + \frac{2}{\delta}) \beta^2 \mathsf{Err}_{\mathbf{2r}}^2 \\
&\quad + C_1 \left(1 + \beta b_u + \beta^2 b_u^2\right) \frac{1}{b_u^2} \left(|\Omega^*| + \gamma \alpha d^*\right) \mathsf{Err}_\infty^2 \qquad (14.29)
\end{aligned}$$

where in the second inequality we used

$$\|\mathcal{P}_{\mathbb{T}_{l-1}}(\boldsymbol{\mathcal{G}}^*)\|_{\mathrm{F}} = \sup_{\|\boldsymbol{\mathcal{Y}}\|_{\mathrm{F}}=1} \langle \mathcal{P}_{\mathbb{T}_{l-1}}(\boldsymbol{\mathcal{G}}^*), \boldsymbol{\mathcal{Y}} \rangle = \sup_{\|\boldsymbol{\mathcal{Y}}\|_{\mathrm{F}}=1} \langle \boldsymbol{\mathcal{G}}^*, \mathcal{P}_{\mathbb{T}_{l-1}}(\boldsymbol{\mathcal{Y}}) \rangle \leq \mathsf{Err}_{\mathbf{2r}} \qquad (14.30)$$

since $\mathcal{P}_{\mathbb{T}_{l-1}}(\boldsymbol{\mathcal{Y}}) \in \mathbb{M}_{\mathbf{2r}}$ and in the last inequality we use $|\Omega^* \cup \Omega_{l-1}| \leq |\Omega^*| + |\Omega_{l-1}| \leq |\Omega^*| + \gamma\alpha d^*$.

Now we choose proper $\beta \in [0.005 b_l/(b_u^2), 0.36 b_l/(b_u^2)]$ so $1 - 1.84\beta b_l + 5\beta^2 b_u^2 \leq 1 - \delta$, and we get

$$\|\boldsymbol{\mathcal{W}}_{l-1} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}} \leq (1 - \delta)(1 + \delta/2) \|\widehat{\boldsymbol{\mathcal{T}}}_{l-1} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}} + 3\delta^{-1} \mathsf{Err}_{\mathbf{2r}} + C_1(b_u + 1) b_l^{-1} \sqrt{|\Omega^*| + \alpha\gamma d^*} \mathsf{Err}_\infty \qquad (14.31)$$

where we use the fact that $\beta \leq 1$. From the signal-to-noise ratio condition, we have $3\delta^{-1}\mathsf{Err}_{\mathbf{2r}} + C_1(b_u + 1) b_l^{-1} \sqrt{|\Omega^*| + \alpha\gamma d^*} \mathsf{Err}_\infty \leq \frac{\delta}{4} \frac{\underline{\lambda}}{C_m \sqrt{r}}$. This implies that $\|\boldsymbol{\mathcal{W}}_{l-1} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}} \leq \underline{\lambda}/8$ holds.

14.2 Proof of Theorem 4.3

Let $\widehat{\Omega}$ and Ω^* denote the support of $\widehat{\boldsymbol{\mathcal{S}}}_{l_{\max}}$ and $\boldsymbol{\mathcal{S}}^*$, respectively. By the proof of Theorem 4.1, we have

$$\big|[\widehat{\boldsymbol{\mathcal{S}}}_{l_{\max}} - \boldsymbol{\mathcal{S}}^*]_\omega\big| \le \begin{cases} \frac{b_u}{b_l}\big|[\widehat{\boldsymbol{\mathcal{T}}}_{l_{\max}} - \boldsymbol{\mathcal{T}}^*]_\omega\big| + \frac{2\mathsf{Err}_\infty}{b_l} & , \text{ if } \omega \in \widehat{\Omega} \\ \frac{2b_u}{b_l}\|\widehat{\boldsymbol{\mathcal{T}}}_{l_{\max}} - \boldsymbol{\mathcal{T}}^*\|_{\ell_\infty} + \frac{2\mathsf{Err}_\infty}{b_l} & , \text{ if } \omega \in \Omega^* \setminus \widehat{\Omega} \end{cases}$$

Therefore, we conclude that

$$\|\widehat{\boldsymbol{\mathcal{S}}}_{l_{\max}} - \boldsymbol{\mathcal{S}}^*\|_{\ell_\infty} \le \frac{2b_u}{b_l}\big\|\widehat{\boldsymbol{\mathcal{T}}}_{l_{\max}} - \boldsymbol{\mathcal{T}}^*\big\|_{\ell_\infty} + \frac{2\mathsf{Err}_\infty}{b_l}. \tag{14.32}$$

Now, we can apply Lemma 15.7 and we obtain

$$\|\widehat{\boldsymbol{\mathcal{T}}}_{l_{\max}} - \boldsymbol{\mathcal{T}}^*\|_{\ell_\infty} \le C_{1,m}\bar{r}^{m/2}\underline{d}^{-(m-1)/2}\mu_1^{2m}\kappa_0^{2m}\|\widehat{\boldsymbol{\mathcal{T}}}_{l_{\max}} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}} \tag{14.33}$$

Now, by putting together (14.32), (14.33) and (4.8), we get

$$\|\widehat{\boldsymbol{\mathcal{S}}}_{l_{\max}} - \boldsymbol{\mathcal{S}}^*\|_{\ell_\infty} \le C_{2,m}\kappa_0^{2m}\mu_1^{2m}\Big(\frac{\bar{r}^m}{\underline{d}^{m-1}}\Big)^{1/2} \cdot \big(\mathsf{Err}_{2\mathbf{r}} + (|\Omega^*| + \gamma\alpha d^*)^{1/2}\mathsf{Err}_\infty\big) + \frac{2\mathsf{Err}_\infty}{b_l},$$

where $C_{1,m}$ and $C_{2,m}$ are constants depending only on m. Now since we assume $b_l, b_u = O(1)$, we finish the proof of Theorem 4.3.

14.3 Proof of Theorem 5.1

We first estimate the probability of the following two events.

$$\mathsf{Err}_{2\mathbf{r}} \le C_{0,m}\sigma_z \cdot (\bar{d}\bar{r} + r^*)^{1/2} \tag{14.34}$$

$$\mathsf{Err}_\infty \le C'_{0,m}\sigma_z \log^{1/2}\bar{d} \tag{14.35}$$

for some constants $C_{0,m}, C'_{0,m} > 0$ depending only on m. Notice here the first event (14.34) holds with probability at least $1 - \exp(-c_m\bar{r}\bar{d})$ by Lemma 15.3. And for the second event (14.35), we have from the definition,

$$\mathsf{Err}_\infty = \max\Big\{\|\nabla\mathfrak{L}(\boldsymbol{\mathcal{T}}^* + \boldsymbol{\mathcal{S}}^*)\|_{\ell_\infty}, \min_{\|\boldsymbol{\mathcal{X}}\|_{\ell_\infty}\le\infty}\|\nabla\mathfrak{L}(\boldsymbol{\mathcal{X}})\|_{\ell_\infty}\Big\} = \|\boldsymbol{\mathcal{Z}}\|_{\ell_\infty} \tag{14.36}$$

So we have (14.35) holds with probability at least $1 - 0.5\bar{d}^{-2}$ from Lemma 15.4. Taking union bounds and we get both (14.35) and (14.34) hold with probability at least $1 - \bar{d}^{-2}$. And finally applying Theorem 4.1 and Theorem 4.3 gives the desired result.

14.4 Proof of Lemma 5.2

Denote the event $\mathcal{E}_1 = \{\|\boldsymbol{\mathcal{Z}}\|_{\ell_\infty} \le 2\sqrt{m}\sigma_z\sqrt{\log(\bar{d})}\}$, then from Lemma 15.4, we have $\mathcal{E}_1$ holds with probability at least $1 - 2(d^*)^{-1}$. Now we set $\tau_l = 2\sqrt{m}\sigma_z\sqrt{\log(\bar{d})} + (d^*)^{-1/2}\mu_1\|\boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}$, then under $\mathcal{E}_1$, we have $\|\boldsymbol{\mathcal{T}}^* + \boldsymbol{\mathcal{Z}}\|_{\ell_\infty} \le \tau_l$. From the definition of τ_0, we have $|\tau_0| \le |\boldsymbol{\mathcal{T}}^* + \boldsymbol{\mathcal{Z}}|^{(\lfloor pd^* - |\Omega^*| \rfloor)} \le \tau_l$. Denote $\Omega_1 = \{\omega : |[\boldsymbol{\mathcal{A}}]_\omega| \le \tau_0\}$ From the definition of $\boldsymbol{\mathcal{A}}_0$, we have

$$\begin{aligned}
\|\boldsymbol{\mathcal{A}}_0\|_{\mathrm{F}}^2 &= \sum_{\omega\in\Omega_1} [\boldsymbol{\mathcal{T}}^* + \boldsymbol{\mathcal{S}}^* + \boldsymbol{\mathcal{Z}}]_\omega^2 \\
&\ge \sum_{\omega\in\Omega_1} [\boldsymbol{\mathcal{T}}^* + \boldsymbol{\mathcal{Z}}]_\omega^2 + 2\sum_{\omega\in\Omega_1\cap\Omega^*} [\boldsymbol{\mathcal{S}}^*]_\omega [\boldsymbol{\mathcal{T}}^* + \boldsymbol{\mathcal{Z}}]_\omega \\
&\ge \sum_{\omega\in\Omega_1} [\boldsymbol{\mathcal{T}}^* + \boldsymbol{\mathcal{Z}}]_\omega^2 - 4|\Omega^*|\tau_l^2 \\
&= \|\boldsymbol{\mathcal{T}}^* + \boldsymbol{\mathcal{Z}}\|_{\mathrm{F}}^2 - \sum_{\omega\in\Omega_1^c} [\boldsymbol{\mathcal{T}}^* + \boldsymbol{\mathcal{Z}}]_\omega^2 - 4|\Omega^*|\tau_l^2 \\
&\ge \|\boldsymbol{\mathcal{T}}^* + \boldsymbol{\mathcal{Z}}\|_{\mathrm{F}}^2 - (pd^* + 4|\Omega^*|)\tau_l^2, \qquad (14.37)
\end{aligned}$$

where the penultimate inequality holds since for all ω, $|[\boldsymbol{\mathcal{T}}^* + \boldsymbol{\mathcal{Z}}]_\omega| \le \tau_l$ and for all $\omega \in \Omega_1$, we have $|[\boldsymbol{\mathcal{S}}^*]_\omega| \le |[\boldsymbol{\mathcal{T}}^* + \boldsymbol{\mathcal{Z}}]_\omega| + \tau_0 \le 2\tau_l$. Now we estimate the lower bound for $\|\boldsymbol{\mathcal{T}}^* + \boldsymbol{\mathcal{Z}}\|_{\mathrm{F}}^2$. Since $\boldsymbol{\mathcal{Z}}$ has i.i.d. subgaussian entries, we have $\|\boldsymbol{\mathcal{Z}}\|_{\mathrm{F}}^2 \ge \frac{1}{2}d^*\sigma_z^2$ with probability at least $1 - 2\exp(-cd^*)$ for some absolute constant $c > 0$, and $2\langle\boldsymbol{\mathcal{T}}^*, \boldsymbol{\mathcal{Z}}\rangle \le \frac{1}{2}\|\boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2 + 2\sigma_z^2\log(\bar{d})$ with probability at least $1 - 2(d^*)^{-1}$. Put these altogether, we see

$$\|\boldsymbol{\mathcal{T}}^* + \boldsymbol{\mathcal{Z}}\|_{\mathrm{F}}^2 = \|\boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2 + \|\boldsymbol{\mathcal{Z}}\|_{\mathrm{F}}^2 + 2\langle\boldsymbol{\mathcal{T}}^*, \boldsymbol{\mathcal{Z}}\rangle \ge \frac{1}{2}\|\boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2 + \frac{1}{4}\sigma_z^2 d^*. \qquad (14.38)$$

Combine (14.37) and (14.38), we have

$$\|\boldsymbol{\mathcal{A}}_0\|_{\mathrm{F}}^2 \ge \frac{1}{2}\|\boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2 + \frac{1}{4}\sigma_z^2 d^* - (pd^* + 4|\Omega^*|)\tau_l^2. \qquad (14.39)$$

Therefore with the choice $\tau = 10\sqrt{m}\sqrt{\log(\bar{d})}\mu_1\frac{\|\boldsymbol{\mathcal{A}}_0\|_{\mathrm{F}}}{\sqrt{d^*}}$, we see that $\tau \ge \tau_l$ and $\tau_u := 10\sqrt{m}\sqrt{\log(\bar{d})}\mu_1\tau_l \ge \tau$. With such a choice of τ, since for $\omega \in (\Omega^*)^c$, we have $|[\boldsymbol{\mathcal{T}}^*]_\omega + [\boldsymbol{\mathcal{Z}}]_\omega| \le \tau_l \le \tau$, so we obtain

$$\begin{aligned}
\widetilde{\boldsymbol{\mathcal{A}}} &= \mathcal{P}_{(\Omega^*)^c}(\boldsymbol{\mathcal{A}}) + \mathcal{P}_{\Omega^*}(\widetilde{\boldsymbol{\mathcal{A}}}) = \mathcal{P}_{(\Omega^*)^c}(\boldsymbol{\mathcal{T}}^* + \boldsymbol{\mathcal{Z}}) + \mathcal{P}_{\Omega^*}(\mathsf{Trunc}_\tau(\boldsymbol{\mathcal{A}})) \\
&= \boldsymbol{\mathcal{T}}^* + \boldsymbol{\mathcal{Z}} + \mathcal{P}_{\Omega^*}(\mathsf{Trunc}_\tau(\boldsymbol{\mathcal{A}}) - \boldsymbol{\mathcal{T}}^* - \boldsymbol{\mathcal{Z}}) \\
&=: \boldsymbol{\mathcal{T}}^* + \boldsymbol{\mathcal{Z}} + \mathbb{E},
\end{aligned}$$

where $\boldsymbol{\mathcal{E}} = \mathcal{P}_{\Omega^*}(\mathsf{Trunc}_\tau(\boldsymbol{\mathcal{A}}) - \boldsymbol{\mathcal{T}}^* - \boldsymbol{\mathcal{Z}})$ and the first equality holds since for $\omega \in (\Omega^*)^c$, $|[\boldsymbol{\mathcal{A}}]_\omega| \leq |[\boldsymbol{\mathcal{T}}^*]|_\omega + |[\boldsymbol{\mathcal{Z}}_\omega]| \leq \tau_u$. Under event $\mathcal{E}_1$, we have $\|\boldsymbol{\mathcal{E}}\|_{\mathrm{F}} \leq 2|\Omega^*|^{1/2}\tau_u$.

Now we use bold-face capital letters as shorthand notation for the unfolding of corresponding calligraphic-font bold-face letters, for example, $\mathbf{T}_i^* = \mathcal{M}_i(\boldsymbol{\mathcal{T}}^*), i \in [m]$. We denote $\boldsymbol{\mathcal{X}} = \boldsymbol{\mathcal{T}}^* + \boldsymbol{\mathcal{E}}$. We also denote $\mathbf{U}_i^*$ be the top r_i left singular vectors of $\mathbf{T}_i^*$, $\mathbf{V}_i$ be the top r_i left singular vectors of $\mathbf{X}_i$ and $\widehat{\mathbf{U}}_i^0$ be the top r_i left singular vectors of $\widetilde{\mathbf{A}}_i$.

From Wedin's sinΘ theorem, we have from condition (a),

$$d_c(\mathbf{U}_i^*, \mathbf{V}_i) \leq \frac{C|\Omega^*|^{1/2}\tau_u}{\underline{\lambda}}, \tag{14.40}$$

where $d_c(\mathbf{U}, \mathbf{V}) = \min_{\mathbf{R}\in\mathbb{O}_r} \|\mathbf{U}\mathbf{R} - \mathbf{V}\|$. Meanwhile, from $\|\mathbf{X}_i - \mathbf{T}_i^*\|_{\mathrm{F}} = \|\boldsymbol{\mathcal{E}}\|_{\mathrm{F}} \leq |\Omega^*|^{1/2}\tau_u$, we also have $\sigma_{r_i}(\mathbf{X}_i) \geq \frac{3\underline{\lambda}}{4}$, $\sigma_{r_i+1}(\mathbf{X}_i) \leq \frac{\underline{\lambda}}{4}$ and $\|\mathbf{X}_i\| \leq \frac{5\bar{\lambda}}{4}$.

Since subtracting a multiple of identity matrix does not change the top eigenvectors, in order to bound the distance $d_c(\mathbf{V}_i, \widehat{\mathbf{U}}_i^0)$, we consider $\|\widetilde{\mathbf{A}}_i\widetilde{\mathbf{A}}_i^T - \mathbf{X}_i\mathbf{X}_i^T - \sigma_v^2 d_i^- \mathbf{I}_{d_i}\|$, where σ_v^2 is the variance of the entry of $\boldsymbol{\mathcal{Z}}$ and $d_i^- = d^*/d_i$. In fact, we have

$$\widetilde{\mathbf{A}}_i\widetilde{\mathbf{A}}_i^T - \mathbf{X}_i\mathbf{X}_i^T - \sigma_v^2 d_i^- \mathbf{I}_{d_i} = \mathbf{X}_i\mathbf{Z}_i^T + \mathbf{Z}_i\mathbf{X}_i^T + \mathbf{Z}_i\mathbf{Z}_i^T - \sigma_v^2 d_i^- \mathbf{I}_{d_i}.$$

Now we first consider the operator norm of $\mathbf{X}_i\mathbf{Z}_i^T$ under the event $\mathcal{E}_1$. From Talagrand's concentration inequality, we have

$$\mathbb{P}\left(\Big|\|\mathbf{X}_i\mathbf{Z}_i^T\| - \mathbb{E}\|\mathbf{X}_i\mathbf{Z}_i^T\|\Big| \leq C_m\sqrt{\log(\bar{d})}\sigma_z\|\mathbf{X}_i\| \cdot t \,\Big|\, \mathcal{E}_1\right) \geq 1 - 2\exp(-ct^2).$$

Since $\mathbb{P}(\mathcal{E}_1) \geq 1/2$ and from [34, Theorem 1.1], we have $\mathbb{E}[\|\mathbf{X}_i\mathbf{Z}_i^T\||\mathcal{E}_1] \leq 2\mathbb{E}\|\mathbf{X}_i\mathbf{Z}_i^T\| \leq C\sqrt{d_i}\sigma_z\|\mathbf{X}_i\|$. Therefore setting $t = \sqrt{\log(\bar{d})}$ and the event

$$\mathcal{E}_2^i = \{\|\mathbf{X}_i\mathbf{Z}_i^T\| \leq C_m\sqrt{d_i}\|\mathbf{X}_i\|\sigma_z\}, \quad \mathcal{E}_2 = \cap_{i=1}^m \mathcal{E}_2^i,$$

we know that $\mathbb{P}(\mathcal{E}_2|\mathcal{E}_1) \geq 1 - 2m\bar{d}^{-1}$ and thus $\mathbb{P}(\mathcal{E}_2) \geq (1 - 2m\bar{d}^{-1})(1 - 2(d^*)^{-1})$.

Now we turn to bounding $\|\mathbf{Z}_i\mathbf{Z}_i^T - d_i^-\sigma_z^2\mathbf{I}_{d_i}\|$. From [35, Theorem 4.6.1], we have with probability exceeding $1 - 2\exp(-d_i)$,

$$\|\mathbf{Z}_i\mathbf{Z}_i^T - d_i^-\sigma_z^2\mathbf{I}_{d_i}\| \leq C(d^*)^{1/2}\sigma_z^2.$$

Denote the event $\mathcal{E}_3^i = \{\|\mathbf{Z}_i\mathbf{Z}_i^T - d_i^-\sigma_v^2\mathbf{I}_{d_i}\| \leq C(d^*)^{1/2}\sigma_z^2\}$ and $\mathcal{E}_3 = \cap_{i=1}^m \mathcal{E}_3^i$ and we have $\mathbb{P}(\mathcal{E}_3) \geq 1 - 2\sum_{i=1}^m \exp(-d_i)$. Therefore under the event $\mathcal{E}_2, \mathcal{E}_3$, and from condition (b), we have

$$d_c(\mathbf{V}_i, \widehat{\mathbf{U}}_i^0) \leq \frac{C_m\sqrt{\bar{d}}\sigma_z\bar{\lambda} + C(d^*)^{1/2}\sigma_z^2}{\underline{\lambda}^2}.$$

Together with (14.40), we have

$$d_c(\mathbf{U}_i^*, \widehat{\mathbf{U}}_i^0) \leq \frac{C_m\sqrt{\bar{d}}\sigma_z\bar{\lambda} + C(d^*)^{1/2}\sigma_z^2}{\underline{\lambda}^2} + \frac{C|\Omega^*|^{1/2}\tau_u}{\underline{\lambda}}. \tag{14.41}$$

Denote the event

$$\mathcal{E}_4 = \left\{ \max_{i=1}^m \max_{\|\mathbf{V}_j\|\leq 1, j\neq i} \|\mathbf{Z}_i(\mathbf{V}_{i+1}\otimes\cdots\otimes\mathbf{V}_m\otimes\mathbf{V}_1\otimes\cdots\otimes\mathbf{V}_{i-1})\| \leq C_m(\sqrt{\bar{d}\bar{r}} + \bar{r}^{\frac{m-1}{2}})\sigma_z \right\}.$$

And from [46, Lemma 5], we have $\mathbb{P}(\mathcal{E}_4) \geq 1 - Cm\exp(-c\bar{d})$. For the following we denote

$$\begin{aligned}
\mathbf{X}_1^t &= \mathbf{T}_1^*(\widehat{\mathbf{U}}_2^t \otimes\cdots\otimes \widehat{\mathbf{U}}_m^t) = \mathbf{T}_1^*(\mathcal{P}_{\mathbf{U}_2^*}\widehat{\mathbf{U}}_2^t \otimes\cdots\otimes \mathcal{P}_{\mathbf{U}_m^*}\widehat{\mathbf{U}}_m^t)\\
\mathbf{Z}_1^t &= \mathbf{Z}_1(\widehat{\mathbf{U}}_2^t \otimes\cdots\otimes \widehat{\mathbf{U}}_m^t)\\
\widetilde{\mathbf{A}}_1^t &= \widetilde{\mathbf{A}}_1(\widehat{\mathbf{U}}_2^t \otimes\cdots\otimes \widehat{\mathbf{U}}_m^t),
\end{aligned}$$

where $\mathcal{P}_{\mathbf{U}} = \mathbf{U}\mathbf{U}^T$. We shall denote $L_t = \max_{i=1}^m d_c(\widehat{\mathbf{U}}_i^t, \mathbf{U}_i^*)$. For the base case, from (14.41) and condition (b), we see $L_0 \leq \frac{1}{2}$. Now suppose we have $L_t \leq \frac{1}{2}$.

From the process of HOOI, we have $\widehat{\mathbf{U}}_1^{t+1} = \mathrm{SVD}_{r_1}(\widetilde{\mathbf{A}}_1(\widehat{\mathbf{U}}_2^t \otimes\cdots\otimes \widehat{\mathbf{U}}_m^t))$. And thus we obtain

$$\begin{aligned}
\sigma_{r_1}(\mathbf{X}_1^t) &\geq \sigma_{r_1}(\mathbf{U}_2^* \otimes\cdots\otimes \mathbf{U}_m^*)\cdot\prod_{i=2}^m \sigma_{\min}(\mathbf{U}_i^{*T}\widehat{\mathbf{U}}_i^t)\\
&\geq \sigma_{r_1}(\mathbf{U}_2^* \otimes\cdots\otimes \mathbf{U}_m^*)(1-L_t^2)^{(m-1)/2}\\
&\geq c_m(1-L_t)^2\underline{\lambda},
\end{aligned} \tag{14.42}$$

for some small constant $c_m > 0$ depending only on m, and the last inequality holds since $1-L_t^2 \geq \frac{3}{4}$. We bound $\|\mathbf{Z}_1^t\|$ under the event $\mathcal{E}_4$.

$$\begin{aligned}
\|\mathbf{Z}_1^t\| &= \|\mathbf{Z}_1(\widehat{\mathbf{U}}_2^t \otimes\cdots\otimes \widehat{\mathbf{U}}_m^t)\|\\
&= \|\mathbf{Z}_1\big((\mathcal{P}_{\mathbf{U}_2^*} + \mathcal{P}_{\mathbf{U}_2^*}^\perp)\otimes\cdots\otimes(\mathcal{P}_{\mathbf{U}_m^*} + \mathcal{P}_{\mathbf{U}_m^*}^\perp)\big)(\widehat{\mathbf{U}}_2^t \otimes\cdots\otimes \widehat{\mathbf{U}}_m^t)\|\\
&\leq C_m[(\bar{d})^{1/2} + \bar{r}^{(m-1)/2}]\sigma_z + C_m[(\bar{d}\bar{r})^{1/2} + \bar{r}^{(m-1)/2}]\sigma_z L_t,
\end{aligned} \tag{14.43}$$

where the last inequality holds since $\mathcal{E}_4$ holds and $\|\widehat{\mathbf{U}}_i^{tT}\mathbf{U}_{i\perp}^*\| \leq L_t$. Now since $\widehat{\mathbf{U}}_1^{t+1}$ is the top r_1 left singular vectors of $\widetilde{\mathbf{A}}_1^t$ and $\mathbf{U}_1$ is the top r_1 left singular vectors of $\mathbf{X}_1^t$, from Wedin's $\sin\Theta$ Theorem, we have

$$\begin{aligned}
d_c(\widehat{\mathbf{U}}_1^{t+1}, \mathbf{U}_1) &\leq \frac{C\|\widetilde{\mathbf{A}}_1^t - \mathbf{X}_1^t\|}{\underline{\lambda}} \leq \frac{C(\|\mathbf{E}_1\|_{\mathrm{F}} + \|\mathbf{Z}_1^t\|)}{\underline{\lambda}}\\
&\overset{(14.43)}{\leq} \frac{C|\Omega^*|^{1/2}\tau_u + C_m[(\bar{d})^{1/2} + \bar{r}^{(m-1)/2}]\sigma_z + C_m[(\bar{d}\bar{r})^{1/2} + \bar{r}^{(m-1)/2}]\sigma_z L_t}{\underline{\lambda}}.
\end{aligned}$$

The derivation for $d_c(\widehat{\mathbf{U}}_i^{t+1}, \mathbf{U}_i)$ when $i \geq 2$ is similar to this case and hence

$$L_{t+1} \leq \frac{C|\Omega^*|^{1/2}\tau_u + C_m[(\bar{d})^{1/2} + \bar{r}^{(m-1)/2}]\sigma_z}{\underline{\lambda}} + \frac{C_m[(\bar{d}\bar{r})^{1/2} + \bar{r}^{(m-1)/2}]\sigma_z}{\underline{\lambda}} L_t.$$

From condition (b), we have $C_m[(\bar{d}\bar{r})^{1/2} + \bar{r}^{(m-1)/2}]\sigma_z/\underline{\lambda} \leq 1/2$, so the above inequality implies

$$L_{t_{\max}} \leq (\frac{1}{2})^{t_{\max}} \cdot L_0 + \frac{C|\Omega^*|^{1/2}\tau_u + C_m[(\bar{d})^{1/2} + \bar{r}^{(m-1)/2}]\sigma_z}{\underline{\lambda}}.$$

If we choose $t_{\max} \geq (C_m \log(\bar{d}\kappa_0) \vee 1)$, then

$$L_{t_{\max}} \leq \frac{C|\Omega^*|^{1/2}\tau_u}{\underline{\lambda}} + \frac{C_m[(\bar{d})^{1/2} + \bar{r}^{(m-1)/2}]\sigma_z}{\underline{\lambda}}. \tag{14.44}$$

Set the event $\mathcal{E}_5 = \{\|\boldsymbol{\mathcal{Z}} \times_{i=1}^m \mathcal{P}_{\widehat{\mathbf{U}}_i}\|_{\mathrm{F}} \leq C(r^* + \sum_{i=1}^m d_i r_i)\sigma_z^2\}$. Then from [46, Lemma 5], $\mathbb{P}(\mathcal{E}_5) \geq 1 - \exp(-C\bar{d}\bar{r})$. And we also consider $\|\boldsymbol{\mathcal{T}}^* \times_i \widehat{\mathbf{U}}_{i\perp}^T\|_{\mathrm{F}}$, we consider $i = 1$ for simplicity.

$$\begin{aligned}
\|\boldsymbol{\mathcal{T}}^* \times_1 \widehat{\mathbf{U}}_{1\perp}^T\|_{\mathrm{F}} = \|\widehat{\mathbf{U}}_{1\perp}^T \mathbf{T}_1^*\|_{\mathrm{F}} &\leq \|\mathcal{P}_{\widehat{\mathbf{U}}_{1\perp}} \mathbf{T}_1^*(\widehat{\mathbf{U}}_2^{t_{\max}-1} \otimes \cdots \otimes \widehat{\mathbf{U}}_m^{t_{\max}-1})\|_{\mathrm{F}} \cdot \prod_{i=2}^m \sigma_{\min}^{-1}(\mathbf{U}_i^{*T}\widehat{\mathbf{U}}_i^{t_{\max}-1}) \\
&\leq C_m(\|\boldsymbol{\mathcal{E}}\|_{\mathrm{F}} + \sqrt{r_1}\|\mathbf{Z}_1^{t_{\max}-1}\|) \\
&\leq C_m|\Omega^*|^{1/2}\tau_u + C_m(\sqrt{d_1 r_1} + \sqrt{r^*})\sigma_z,
\end{aligned} \tag{14.45}$$

where the second inequality holds from [46, Lemma 6] and the last inequality holds from (14.43).

Now we are in the right position to bound $\|\widehat{\boldsymbol{\mathcal{T}}} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}$ under $\mathcal{E}_5$.

$$\begin{aligned}
\|\widehat{\boldsymbol{\mathcal{T}}} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}} &= \|\widetilde{\boldsymbol{\mathcal{A}}} \times_{i=1}^m \mathcal{P}_{\widehat{\mathbf{U}}_i} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}} \\
&\leq \|(\widetilde{\boldsymbol{\mathcal{A}}} - \boldsymbol{\mathcal{T}}^*) \times_{i=1}^m \mathcal{P}_{\widehat{\mathbf{U}}_i}\|_{\mathrm{F}} + \|\boldsymbol{\mathcal{T}}^* - \boldsymbol{\mathcal{T}}^* \times_{i=1}^m \mathcal{P}_{\widehat{\mathbf{U}}_i}\|_{\mathrm{F}} \\
&\leq \|\boldsymbol{\mathcal{E}}\|_{\mathrm{F}} + \|\boldsymbol{\mathcal{Z}} \times_{i=1}^m \mathcal{P}_{\widehat{\mathbf{U}}_i}\|_{\mathrm{F}} + \sum_{i=1}^m \|\boldsymbol{\mathcal{T}}^* \times_i \widehat{\mathbf{U}}_{i\perp}^T\|_{\mathrm{F}} \\
&\leq C_m|\Omega^*|^{1/2}\tau_u + C_m(\sqrt{r^*} + \sqrt{\bar{d}\bar{r}})\sigma_z,
\end{aligned} \tag{14.46}$$

where the last inequality follows from (14.45). Finally applying Lemma 15.6 and we get $\widehat{\boldsymbol{\mathcal{T}}}_0$ is $(2\mu_1\kappa_0)^2$-incoherent and $\|\widehat{\boldsymbol{\mathcal{T}}}_0 - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}} \leq 2\|\widehat{\boldsymbol{\mathcal{T}}} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}$. Therefore from condition $(a), (b)$ in Lemma 5.2, the initialization condition (a) in Theorem 5.1 holds.

14.5 Proof of Lemma 5.3

For each $j \in [m]$ and $i \in [d_j]$, we have

$$\|\mathbf{e}_i^\top \mathcal{M}_j(\boldsymbol{\mathcal{S}}_\alpha)\|_{\ell_0} = \sum\nolimits_{\omega:\omega_j=i} \mathbb{1}\big(|[\boldsymbol{\mathcal{Z}}]_\omega| > \alpha\sigma_z\big) = \sum\nolimits_{\omega:\omega_j=i} [\boldsymbol{\mathcal{Y}}]_\omega$$

where $\boldsymbol{\mathcal{Y}} \in \{0,1\}^{d_1 \times \cdots \times d_m}$ having $i.i.d.$ Bernoulli entries and $q := \mathbb{P}([\boldsymbol{\mathcal{Y}}]_\omega = 1) = \mathbb{P}(|[\boldsymbol{\mathcal{Z}}]_\omega| > \alpha\sigma_z) \leq \alpha^{-\theta}$.

Denote $X_{ij} = \sum_{\omega:\omega_j=i} [\boldsymbol{\mathcal{Y}}]_\omega$. By Chernoff bound, if $d_j^- q \geq 3\log(m\bar{d}^3)$, we get

$$\mathbb{P}\Big(X_{ij} - d_j^- q \geq d_j^- q\Big) \leq \exp\big\{-d_j^{-1} q/3\big\} \leq (m\bar{d}^3)^{-1}$$

implying that

$$\mathbb{P}\Big(\bigcap\nolimits_{i,j} \big\{X_{ij} \leq 2d_j^- q\big\}\Big) \geq 1 - m\bar{d}(m\bar{d}^3)^{-1} = 1 - \bar{d}^{-2}. \tag{14.47}$$

On the other hand, if $d_j^- q \leq 3\log(m\bar{d}^3)$, by Chernoff bound, we get

$$\mathbb{P}\Big(X_{ij} \geq 10\log(m\bar{d}^3)\Big) \leq (m\bar{d}^3)^{-1}$$

implying that

$$\mathbb{P}\Big(\bigcap\nolimits_{i,j} \big\{X_{ij} \leq 10\log(m\bar{d}^3)\big\}\Big) \geq 1 - m\bar{d}(m\bar{d}^3)^{-1} = 1 - \bar{d}^{-2}. \tag{14.48}$$

Putting (14.47) and (14.48), since $q \leq \alpha^{-\theta}$, we get

$$\mathbb{P}\Big(\bigcap\nolimits_{i,j} \Big\{X_{ij} \leq \max\big\{10\log(m\bar{d}^3), 2d_j^- \alpha^{-\theta}\big\}\Big\}\Big) \geq 1 - \bar{d}^{-2},$$

which completes the proof.

14.6 Proof of Theorem 5.4

Conditioned on $\mathfrak{E}_1$ defined in Lemma 5.3, Theorem 5.4 is a special case of Theorem 5.1. Indeed, in Theorem 5.1, we replace σ_z with $\alpha\sigma_z$, and $|\Omega^*|\log\bar{d}$ with $\alpha' d^* \asymp \bar{d}\log(m\bar{d})$, then we get Theorem 5.4.

14.7 Proof of Lemma 5.5

From the choice of α in Theorem 5.4, we see that the sparsity of $\boldsymbol{\mathcal{S}}_\alpha$ is bounded by $\alpha' \asymp \frac{\bar{d}}{d^*}\log(m\bar{d}^3)$. Therefore the condition (a) in Lemma 5.2 is satisfied. Now applying Lemma 5.2 and we get the desired result.

14.8 Proof of Lemma 5.6

From Lemma 15.6, we have $\mathsf{Trim}_{\eta,\mathbf{r}}(\boldsymbol{\mathcal{W}})$ is $2\mu_1\kappa_0$-incoherent. Now for all $j \in [m]$,

$$\|\mathcal{M}_j(\mathscr{H}^{\mathrm{HO}}_{\mathbf{r}}(\widetilde{\boldsymbol{\mathcal{W}}}))\| \leq \|\mathcal{M}_j(\widetilde{\boldsymbol{\mathcal{W}}})\| \leq \|\mathcal{M}_j(\boldsymbol{\mathcal{T}}^*)\| + \|\boldsymbol{\mathcal{W}} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}} \leq \frac{9}{8}\bar{\lambda}.$$

So we conclude

$$\|\mathsf{Trim}_{\eta,\mathbf{r}}(\boldsymbol{\mathcal{W}})\|_{\ell_\infty} \leq \frac{9}{8}\bar{\lambda}\prod_{i=1}^{m}(2\mu_1\kappa_0)\sqrt{\frac{r_j}{d_j}} \leq (9\zeta/16)\cdot(\mu_1\kappa_0)^m.$$

where the last inequality follows from the upper bound for $\bar{\lambda}$. This finishes the proof of the lemma.

14.9 Proof of Theorem 5.7

From the choice of ζ' and Lemma 5.6, we know Assumption 2 and 3 hold with parameters $b_{l,\zeta'}$ and $b_{u,\zeta'}$ with respect to the set $\mathbb{B}_2^* = \mathbb{B}_\infty^* = \{\boldsymbol{\mathcal{T}} + \boldsymbol{\mathcal{S}} : \|\boldsymbol{\mathcal{T}} + \boldsymbol{\mathcal{S}}\|_{\ell_\infty} \leq \zeta', \boldsymbol{\mathcal{T}} \in \mathbb{M}_{\mathbf{r}}, \boldsymbol{\mathcal{S}} \in \mathbb{S}_{\gamma\alpha}\}$. Now the proof follows the proof of Theorem 4.1 with slight modification. Since we can now guarantee in each iteration $\widehat{\boldsymbol{\mathcal{T}}}_l + \widehat{\boldsymbol{\mathcal{S}}}_l \in \mathbb{B}_2^* = \mathbb{B}_\infty^*$ from Lemma 5.6 and the choice of $\mathsf{k_{pr}}$, we can use Assumption 3 instead of Assumption 2 when estimating the low rank part. So we only need to estimate Err_∞ and $\mathsf{Err}_{2\mathbf{r}}$. From (5.8), we have $\mathsf{Err}_\infty \leq L_\zeta$. Now we estimate $\mathsf{Err}_{2\mathbf{r}}$. In fact, from the definition of $\mathsf{Err}_{2\mathbf{r}}$, we have

$$\mathsf{Err}_{2\mathbf{r}} = \sup_{\boldsymbol{\mathcal{M}} \in \mathbb{M}_{2\mathbf{r}}, \|\boldsymbol{\mathcal{M}}\|_{\mathrm{F}} \leq 1} \langle \nabla \mathfrak{L}(\boldsymbol{\mathcal{T}}^* + \boldsymbol{\mathcal{S}}^*), \boldsymbol{\mathcal{M}} \rangle.$$

Since for all $\omega \in [d_1] \times \ldots \times [d_m]$, we have $[\nabla \mathfrak{L}(\boldsymbol{\mathcal{T}}^* + \boldsymbol{\mathcal{S}}^*)]_\omega$ is bounded random variable with the upper bound given by L_ζ. So apply Lemma 15.3, we have $\mathsf{Err}_{2\mathbf{r}} \leq C L_\zeta \cdot (\bar{d}\bar{r} + r^*)^{1/2}$ with with probability at least $1 - \bar{d}^{-2}$. Now we plug in the bounds for Err_∞ and $\mathsf{Err}_{2\mathbf{r}}$ to Theorem 4.1 and we get the first part of the theorem. For the ℓ_∞ bound, we apply Theorem 4.3 and Lemma 15.7. And we finish the proof of the theorem.

14.10 Proof of Lemma 5.8

We first introduce some notations. Let $m_0 = \lfloor \frac{m}{2} \rfloor$, and denote $\mathbf{T}^* = (\boldsymbol{\mathcal{T}}^*)^{\langle m_0 \rangle}$, $\mathbf{S}^* = (\boldsymbol{\mathcal{S}}^*)^{\langle m_0 \rangle}$ and $\mathbf{A} = \boldsymbol{\mathcal{A}}^{\langle m_0 \rangle}$, then $\mathbf{T}^*, \mathbf{S}^*, \mathbf{A}$ are matrices of size $d_1 \ldots d_{m_0} \times d_{m_0+1} \ldots d_m =: d_1^* \times d_2^*$. Since $\boldsymbol{\mathcal{T}}^*$ admits the decomposition $\boldsymbol{\mathcal{T}}^* = \boldsymbol{\mathcal{C}}^* \cdot [\![\mathbf{U}_1^*, \cdots, \mathbf{U}_m^*]\!]$, we have $\boldsymbol{\mathcal{T}}^* = (\mathbf{U}_{m_0} \otimes \cdots \otimes \mathbf{U}_1)\boldsymbol{\mathcal{C}}^{\langle m_0 \rangle}(\mathbf{U}_m \otimes \cdots \otimes \mathbf{U}_{m_0+1})^T$ and hence the rank of $\mathbf{T}^*$ is $r = \min\{r_1 \cdots r_{m_0}, r_{m_0+1} \cdots r_m\}$. We denote $\mathbf{M} = \mathbf{T}^* + \mathbf{S}^*$.

Algorithm 7 Initialization for binary tensor

Let $\mathbf{A} = \boldsymbol{\mathcal{A}}^{\langle m_0 \rangle} := \mathsf{reshape}(\boldsymbol{\mathcal{A}}, [d_1 \dots d_{m_0}, d_{m_0+1} \dots d_m])$ with $m_0 = \lfloor \frac{m}{2} \rfloor$ and let $\widehat{\mathbf{M}}$ be the minimizer to (14.49).

$\widehat{\boldsymbol{\mathcal{T}}} = \mathsf{reshape}(\widehat{\mathbf{M}}, [d_1, \dots, d_m])$.

$\widehat{\boldsymbol{\mathcal{T}}}_0 = \mathsf{Trim}_{\eta,\mathbf{r}}(\widehat{\boldsymbol{\mathcal{T}}})$ with $\eta = 16\mu_1 \|\widehat{\boldsymbol{\mathcal{T}}}\|_{\mathrm{F}}/(7\sqrt{d^*})$.

Output: $\widehat{\boldsymbol{\mathcal{T}}}_0$.

Under Assumption 5, we have $\|\mathbf{T}^*\|_{\ell_\infty}, \|\mathbf{S}^*\|_{\ell_\infty} \leq \frac{\zeta}{2}$ and thus $\|\mathbf{M}\|_{\ell_\infty} \leq \zeta$. Now we bound the nuclear norm of $\mathbf{M}$. Using triangle inequality and we have

$$\begin{aligned}
\|\mathbf{M}\|_* &\leq \|\mathbf{T}^*\|_* + \|\mathbf{S}^*\|_* \\
&\leq \frac{\zeta}{2}(rd^*)^{1/2} + \frac{\zeta}{2}|\Omega^*|^{1/2} \min(d_1^*, d_2^*)^{1/2} \\
&= \big(\frac{\zeta}{2} + \frac{\zeta}{2} \cdot \frac{\min(d_1^*, d_2^*)^{1/2}}{(rd^*)^{1/2}} |\Omega^*|^{1/2}\big)(rd^*)^{1/2} \\
&\leq \zeta(rd^*)^{1/2},
\end{aligned}$$

where the last inequality holds since condition (a) holds. Now with a little bit abuse of notation, we consider the following convex program,

$$\min \mathfrak{L}(\mathbf{X}) = -\langle \mathbf{A}, \log(p(X)) \rangle - \langle 1 - \mathbf{A}, \log(1 - p(\mathbf{X})) \rangle, \text{ s.t. } \|\mathbf{X}\|_* \leq \zeta\sqrt{d^* r} \text{ and } \|\mathbf{X}\|_{\ell_\infty} \leq \zeta, \tag{14.49}$$

where the notation $1 - \mathbf{A}$ is the entrywise subtraction, and $p(\mathbf{X})$ is applying p entrywisely to $\mathbf{X}$. Denote $\widehat{\mathbf{M}}$ be the minimizer to (14.49) and apply the Theorem 1 in [10] with the sample size d^* and we get with probability at least $1 - \frac{C}{d_1^* + d_2^*}$,

$$\|\widehat{\mathbf{M}} - \mathbf{M}\|_{\mathrm{F}}^2 \leq C_\zeta [r(d_1^* + d_2^*)d^*]^{1/2}$$

with $C_\zeta = C \cdot \zeta L_\zeta \beta_\zeta$ and $\beta_\zeta = \sup_{|x| \leq \zeta} \frac{p(x)(1-p(x))}{(p'(x))^2}$.

Now we reshape $\widehat{\mathbf{M}}$ back to a tensor, and denote $\widehat{\boldsymbol{\mathcal{T}}} = \mathsf{reshape}(\widehat{\mathbf{M}}, [d_1, \dots, d_m])$. Since $\mathsf{reshape}$ keeps the Frobenius norm unchanged, we have

$$\|\widehat{\boldsymbol{\mathcal{T}}} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}} = \|\widehat{\mathbf{T}} - \mathbf{T}^*\|_{\mathrm{F}} \leq \|\widehat{\mathbf{M}} - \mathbf{M}\|_{\mathrm{F}} + \|\mathbf{S}^*\|_{\mathrm{F}} \leq C_\zeta^{1/2} [r(d_1^* + d_2^*)d^*]^{1/4} + |\Omega^*|^{1/2} \frac{\zeta}{2}.$$

Finally we output $\boldsymbol{\mathcal{T}}_0 = \mathsf{Trim}_{\eta,\mathbf{r}}(\widehat{\boldsymbol{\mathcal{T}}})$ with $\eta = 16\mu_1\|\widehat{\boldsymbol{\mathcal{T}}}\|_{\mathrm{F}}/(7\sqrt{d^*})$, and from Lemma 5.6 and Lemma 15.6, since condition (b) and (c) hold, we get

$$(1)\ \mu(\widehat{\boldsymbol{\mathcal{T}}}_0) \le 2\kappa_0\mu_1;\ (2)\ \|\widehat{\boldsymbol{\mathcal{T}}}_0 - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}} \le 2\|\widehat{\boldsymbol{\mathcal{T}}} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}};\ (3)\ \|\widehat{\boldsymbol{\mathcal{T}}}\|_{\ell_\infty} \le C_m(\mu_1\kappa_0)^m\frac{\sqrt{r^*}}{\sqrt{d^*}}\bar{\lambda}.$$

And together with the upper bound for $\underline{\lambda}$ in Assumption 5, the initialization condition in Theorem 5.7 is satisfied.

14.11 Proof of Theorem 8.1

The proof of this theorem is similar to that of Theorem 5.7. From the choice of ζ' and Lemma 5.6, we know Assumption 2 and 3 hold with parameters $b_{l,\zeta'} = e^{-\zeta'}$ and $b_{u,\zeta'} = e^{\zeta'}$ with respect to the set $\mathbb{B}_2^* = \mathbb{B}_\infty^* = \{\boldsymbol{\mathcal{T}} + \boldsymbol{\mathcal{S}} : \|\boldsymbol{\mathcal{T}} + \boldsymbol{\mathcal{S}}\|_{\ell_\infty} \le \zeta', \boldsymbol{\mathcal{T}} \in \mathbb{M}_{\mathbf{r}}, \boldsymbol{\mathcal{S}} \in \mathbb{S}_{\gamma\alpha}\}$. Now the proof follows the proof of Theorem 4.1 with slight modification. Since we can now guarantee in each iteration $\widehat{\boldsymbol{\mathcal{T}}}_l + \widehat{\boldsymbol{\mathcal{S}}}_l \in \mathbb{B}_2^* = \mathbb{B}_\infty^*$ from Lemma 5.6 and the choice of $\mathsf{k_{pr}}$, we can use Assumption 3 instead of Assumption 2 when estimating the low rank part. So we only need to estimate Err_∞ and $\mathsf{Err}_{\mathbf{2r}}$. From (5.8), we have $\mathsf{Err}_\infty \le \|\nabla\mathfrak{L}(\boldsymbol{\mathcal{T}}^* + \boldsymbol{\mathcal{S}}^*)\|_{\ell_\infty}$. Simple calculation shows

$$\nabla\mathfrak{L}(\boldsymbol{\mathcal{T}}^* + \boldsymbol{\mathcal{S}}^*) = -\frac{1}{I}\boldsymbol{\mathcal{Y}} + \exp(\boldsymbol{\mathcal{T}}^* + \boldsymbol{\mathcal{S}}^*),$$

and notice using a union bound and Poisson's tail bound, when $I \ge Ce^{\zeta}\log(d^*)$, we have with probability exceeding $1 - \frac{1}{d^*}$, $\|\boldsymbol{\mathcal{Y}}\|_{\ell_\infty} \le 10Ie^{\zeta}$. Therefore we have $\mathsf{Err}_\infty \le 11e^{\zeta}$.

The estimation for $\mathsf{Err}_{\mathbf{2r}}$ is given in Theorem 4.3 [14], which states

$$\mathsf{Err}_{\mathbf{2r}} \le C\sqrt{\frac{r^* + m\bar{d}\bar{r}}{I/e^{\zeta}}}$$

with probability exceeding $1 - \frac{1}{d^*}$.

Now we plug in the bounds for Err_∞ and $\mathsf{Err}_{\mathbf{2r}}$ to Theorem 4.1 and we get the first part of the theorem. For the ℓ_∞ bound, we apply Theorem 4.3 and Lemma 15.7. And we finish the proof of the theorem.

14.12 Proof of Lemma 8.2

With slight modification of the proof of Theorem 4.3 in [14], we have

$$\|\widetilde{\boldsymbol{\mathcal{T}}}_0 - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}} \le C\sqrt{\frac{e^{\zeta}}{I}}(\sum_{i=1}^{m}\sqrt{d_i r_i} + \sqrt{d_i^- r_i}) + \|\boldsymbol{\mathcal{S}}^*\|_{\mathrm{F}}$$

under the condition $I \geq Ce^{\zeta}\bar{d}$ with probability exceeding $1 - 1/d^*$. Therefore since we assume $I \geq C_1 \sum_{i=1}^{m}(d_i r_i + d_i^- r_i)\bar{r}\underline{\lambda}^{-2}$ and $|\Omega^*| \leq C\zeta^{-2}\underline{\lambda}^2\bar{r}^{-1}$, we have $\|\widetilde{\boldsymbol{\mathcal{T}}}_0 - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}} \leq c_{1,m}\underline{\lambda} \cdot \min\left\{\delta^2\bar{r}^{-1/2}, (\kappa_0^{2m}\bar{r}^{1/2})^{-1}\right\} \leq \underline{\lambda}/8$. Now we apply Lemma 5.6 and Lemma 15.6 we see

$$(1)\ \mu(\widehat{\boldsymbol{\mathcal{T}}}_0) \leq 2\kappa_0\mu_1;\ (2)\ \|\widehat{\boldsymbol{\mathcal{T}}}_0 - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}} \leq 2\|\widehat{\boldsymbol{\mathcal{T}}} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}};\ (3)\ \|\widehat{\boldsymbol{\mathcal{T}}}\|_{\ell_\infty} \leq C_m(\mu_1\kappa_0)^m \frac{\sqrt{r^*}}{\sqrt{d^*}}\bar{\lambda}.$$

From Assumption 6, we see the initialization requirements in 8.1 is satisfied.

14.13 Proof of Theorem 10.1

We use induction to prove this theorem.

Step 0: Base case. From the initialization, we have $\|\widehat{\boldsymbol{\mathcal{T}}}_0 - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}} \leq c_{1,m}\delta\bar{r}^{-1/2} \cdot \underline{\lambda}$.

Step 1: Estimating $\|\widehat{\boldsymbol{\mathcal{T}}}_{l+1} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}$. We prove this case assuming

$$\|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}} \leq c_{1,m}\delta\bar{r}^{-1/2} \cdot \underline{\lambda}. \tag{14.50}$$

We point out that this also implies $\|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}} \leq c_{1,m}b_l b_u^{-1}\bar{r}^{-1/2} \cdot \underline{\lambda}$ since $\delta \lesssim b_l^2 b_u^{-2}$. In order to use Lemma 15.2, we need to derive an upper bound for $\|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^* - \beta\mathcal{P}_{\mathbb{T}_l}\boldsymbol{\mathcal{G}}_l\|_{\mathrm{F}}$.

Step 1.1: Estimating $\|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^* - \beta\mathcal{P}_{\mathbb{T}_l}\boldsymbol{\mathcal{G}}_l\|_{\mathrm{F}}$. For arbitrary $1 \geq \delta > 0$, we have,

$$\|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^* - \beta\mathcal{P}_{\mathbb{T}_l}\boldsymbol{\mathcal{G}}_l\|_{\mathrm{F}}^2 \leq (1+\delta/2)\|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^* - \beta\mathcal{P}_{\mathbb{T}_l}(\boldsymbol{\mathcal{G}}_l - \boldsymbol{\mathcal{G}}^*)\|_{\mathrm{F}}^2 + (1+2/\delta)\beta^2\|\mathcal{P}_{\mathbb{T}_l}\boldsymbol{\mathcal{G}}^*\|_{\mathrm{F}}^2 \tag{14.51}$$

Now we consider the bound for $\|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^* - \beta\mathcal{P}_{\mathbb{T}_l}(\boldsymbol{\mathcal{G}}_l - \boldsymbol{\mathcal{G}}^*)\|_{\mathrm{F}}^2$.

$$\begin{aligned}\|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^* - \beta\mathcal{P}_{\mathbb{T}_l}(\boldsymbol{\mathcal{G}}_l - \boldsymbol{\mathcal{G}}^*)\|_{\mathrm{F}}^2 &= \|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2 - 2\beta\langle\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*, \mathcal{P}_{\mathbb{T}_l}(\boldsymbol{\mathcal{G}}_l - \boldsymbol{\mathcal{G}}^*)\rangle + \beta^2\|\mathcal{P}_{\mathbb{T}_l}(\boldsymbol{\mathcal{G}}_l - \boldsymbol{\mathcal{G}}^*)\|_{\mathrm{F}}^2 \\ &\leq (1+\beta^2 b_u^2)\|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2 - 2\beta\langle\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*, \mathcal{P}_{\mathbb{T}_l}(\boldsymbol{\mathcal{G}}_l - \boldsymbol{\mathcal{G}}^*)\rangle \end{aligned} \tag{14.52}$$

where the last inequality holds from the Assumption 2 since $\widehat{\boldsymbol{\mathcal{T}}}_l \in \mathbb{B}_2^*$ from (14.50). Also,

$$\begin{aligned}\langle\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*, \mathcal{P}_{\mathbb{T}_l}(\boldsymbol{\mathcal{G}}_l - \boldsymbol{\mathcal{G}}^*)\rangle &= \langle\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*, \boldsymbol{\mathcal{G}}_l - \boldsymbol{\mathcal{G}}^*\rangle - \langle\mathcal{P}_{\mathbb{T}_l}^{\perp}(\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*), \boldsymbol{\mathcal{G}}_l - \boldsymbol{\mathcal{G}}^*\rangle \\ &\geq b_l\|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2 - \frac{C_{1,m}b_u}{\underline{\lambda}}\|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^3 \end{aligned} \tag{14.53}$$

where the last inequality is from Assumption 2, Lemma 15.1 and Cauchy-Schwartz inequality and $C_{1,m} = 2^m - 1$. Together with (14.52) and (14.53), and since we have $\|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}} \leq \frac{0.1b_l}{2b_u C_{1,m}} \cdot \underline{\lambda}$, we get,

$$\begin{aligned}\|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^* - \beta\mathcal{P}_{\mathbb{T}_l}(\boldsymbol{\mathcal{G}}_l - \boldsymbol{\mathcal{G}}^*)\|_{\mathrm{F}}^2 &\leq (1 - 2\beta b_l + \beta^2 b_u^2)\|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2 + \frac{2\beta C_{1,m} b_u}{\underline{\lambda}}\|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^3 \\ &\leq (1 - 1.9\beta b_l + \beta^2 b_u^2)\|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2. \end{aligned} \tag{14.54}$$

Since we have $0.75 b_l b_u^{-1} \geq \delta^{1/2}$, if we choose $\beta \in [0.4 b_l b_u^{-2}, 1.5 b_l b_u^{-2}]$, we have $1 - 1.9\beta b_l + \beta^2 b_u^2 \leq 1 - \delta$.

So from (14.51) and (14.54), we get

$$\|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^* - \beta\mathcal{P}_{\mathbb{T}_l}\boldsymbol{\mathcal{G}}_l\|_{\mathrm{F}}^2 \leq (1 + \frac{\delta}{2})(1 - \delta)\|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2 + (1 + \frac{2}{\delta})\mathsf{Err}_{2\mathbf{r}}^2 \tag{14.55}$$

where in the inequality we use the definition of $\mathsf{Err}_{2\mathbf{r}}$ and that $\beta \leq 1$. Now from the upper bound for $\|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}$ and the signal-to-noise ratio, we verified that $\|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^* - \beta\mathcal{P}_{\mathbb{T}_l}\boldsymbol{\mathcal{G}}_l\|_{\mathrm{F}} \leq \underline{\lambda}/8$ and thus $\sigma_{\max}(\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^* - \beta\mathcal{P}_{\mathbb{T}_l}\boldsymbol{\mathcal{G}}_l) \leq \underline{\lambda}/8$.

Step 1.2: Estimating $\|\widehat{\boldsymbol{\mathcal{T}}}_{l+1} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}$. Now that we verified the condition of Lemma 15.2, from the Algorithm 6, we have,

$$\|\widehat{\boldsymbol{\mathcal{T}}}_{l+1} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2 \leq \|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^* - \beta\mathcal{P}_{\mathbb{T}_l}\boldsymbol{\mathcal{G}}_l\|_{\mathrm{F}}^2 + C_m\frac{\sqrt{\bar{r}}}{\underline{\lambda}}\|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^* - \beta\mathcal{P}_{\mathbb{T}_l}\boldsymbol{\mathcal{G}}_l\|_{\mathrm{F}}^3 \tag{14.56}$$

where $C_m > 0$ is the constant depending only on m as in Lemma 15.2. From (14.55) and the assumption that $\|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}} \lesssim_m \frac{\delta}{\sqrt{\bar{r}}} \cdot \underline{\lambda}$ and $\mathsf{Err}_{2\mathbf{r}} \lesssim_m \frac{\delta^2}{\sqrt{\bar{r}}} \cdot \underline{\lambda}$, we get

$$C_m\frac{\sqrt{\bar{r}}}{\underline{\lambda}}\|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^* - \beta\mathcal{P}_{\mathbb{T}_l}\boldsymbol{\mathcal{G}}_l\|_{\mathrm{F}} \leq \frac{\delta}{4} \tag{14.57}$$

From (14.56), (14.55) and (14.57), we get

$$\|\widehat{\boldsymbol{\mathcal{T}}}_{l+1} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2 \leq (1 + \frac{\delta}{4})\|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^* - \beta\mathcal{P}_{\mathbb{T}_l}\boldsymbol{\mathcal{G}}_l\|_{\mathrm{F}}^2 \leq (1 - \delta^2)\|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2 + \frac{4}{\delta}\mathsf{Err}_{2\mathbf{r}}^2 \tag{14.58}$$

Together with the assumption $\|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}} \lesssim_m \frac{\delta}{\sqrt{\bar{r}}} \cdot \underline{\lambda}$ and $\mathsf{Err}_{2\mathbf{r}} \lesssim_m \frac{\delta^2}{\sqrt{\bar{r}}} \cdot \underline{\lambda}$, we get

$$\|\widehat{\boldsymbol{\mathcal{T}}}_{l+1} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}} \leq c_{1,m}\frac{\delta}{\sqrt{\bar{r}}} \cdot \underline{\lambda}, \tag{14.59}$$

which completes the induction and completes the proof.

15 Technical Lemmas

Lemma 15.1. *Suppose $\mathbb{T}_l$ is the tangent space at the point $\widehat{\boldsymbol{\mathcal{T}}}_l$, then we have*

$$\|\mathcal{P}_{\mathbb{T}_l}^{\perp}\boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}} \leq \frac{2^m-1}{\underline{\lambda}}\|\boldsymbol{\mathcal{T}}^* - \widehat{\boldsymbol{\mathcal{T}}}_l\|_{\mathrm{F}}^2.$$

Proof. See ([5], Lemma 5.2). □

Lemma 15.2. *Let $\boldsymbol{\mathcal{T}}^* = \boldsymbol{\mathcal{S}}^* \cdot (\mathbf{V}_1^*, \ldots, \mathbf{V}_m^*)$ be the tensor with Tucker rank $\mathbf{r} = (r_1, \ldots, r_m)$. Let $\boldsymbol{\mathcal{D}} \in \mathbb{R}^{d_1 \times \ldots \times d_m}$ be a perturbation tensor such that $\underline{\lambda} \geq 8\sigma_{\max}(\boldsymbol{\mathcal{D}})$, where $\sigma_{\max}(\boldsymbol{\mathcal{D}}) = \max_{i=1}^m \|\mathcal{M}_i(\boldsymbol{\mathcal{D}})\|$. Then we have*

$$\|\mathscr{H}_{\mathbf{r}}^{\mathrm{HO}}(\boldsymbol{\mathcal{T}}^* + \boldsymbol{\mathcal{D}}) - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}} \leq \|\boldsymbol{\mathcal{D}}\|_{\mathrm{F}} + C_m\frac{\sqrt{\bar{r}}\|\boldsymbol{\mathcal{D}}\|_{\mathrm{F}}^2}{\underline{\lambda}}$$

where $C_m > 0$ is an absolute constant depending only on m.

Proof. Without loss of generality, we only prove the Lemma in the case $m = 3$. First notice that

$$\mathscr{H}_{\mathbf{r}}^{\mathrm{HO}}(\boldsymbol{\mathcal{T}}^* + \boldsymbol{\mathcal{D}}) = (\boldsymbol{\mathcal{T}}^* + \boldsymbol{\mathcal{D}}) \cdot [\![\mathcal{P}_{\mathbf{U}_1}, \mathcal{P}_{\mathbf{U}_2}, \mathcal{P}_{\mathbf{U}_3}]\!],$$

where $\mathbf{U}_i$ are leading r_i left singular vectors of $\mathcal{M}_i(\boldsymbol{\mathcal{T}}^* + \boldsymbol{\mathcal{D}})$ and $\mathcal{P}_{\mathbf{U}_i} = \mathbf{U}_i\mathbf{U}_i^\top$.

First from ([39], Theorem 1), we have for all $i \in [m]$

$$\mathcal{P}_{\mathbf{U}_i} - \mathcal{P}_{\mathbf{V}_i^*} = \mathcal{S}_{i,1} + \sum_{j\geq 2}\mathcal{S}_{i,j},$$

where $\mathcal{S}_{i,j} = \mathcal{S}_{\mathcal{M}_i(\boldsymbol{\mathcal{T}}^*),j}(\mathcal{M}_i(\boldsymbol{\mathcal{D}}))$ and specially $\mathcal{S}_{i,1} = (\mathcal{M}_i(\boldsymbol{\mathcal{T}}^*)^\top)^\dagger(\mathcal{M}_i(\boldsymbol{\mathcal{D}}))^\top\mathcal{P}_{\mathbf{V}_i^*}^{\perp} + \mathcal{P}_{\mathbf{V}_i^*}^{\perp}\mathcal{M}_i(\boldsymbol{\mathcal{D}})(\mathcal{M}_i(\boldsymbol{\mathcal{T}}^*))^\dagger$. The explicit form of $S_{i,j}$ can be found in [39, Theorem 1]. Here, we denote $\mathbf{A}^\dagger$ the pseudo-inverse of $\mathbf{A}$, i.e., $\mathbf{A}^\dagger = \mathbf{R}\boldsymbol{\Sigma}^{-1}\mathbf{L}^\top$ if $\mathbf{A}$ has a thin-SVD as $\mathbf{A} = \mathbf{L}\boldsymbol{\Sigma}\mathbf{R}^\top$. With a little abuse of notations, we write $(\mathbf{A}^\dagger)^k = \mathbf{R}\boldsymbol{\Sigma}^{-k}\mathbf{L}^\top$ for any positive integer $k \geq 1$.

For the sake of brevity, we denote $\mathbf{S}_i = \sum_{j\geq 1}\mathcal{S}_{i,j}$. By the definition of $\mathcal{S}_{i,j}$, we have the bound $\|\mathcal{S}_{i,j}\| \leq \left(\frac{4\sigma_{\max}(\boldsymbol{\mathcal{D}})}{\underline{\lambda}}\right)^j$. We get the upper bound for $\|\mathbf{S}_i\|$ as follows,

$$\|\mathbf{S}_i\| = \|\sum_{j\geq 1}\mathcal{S}_{i,j}\| \leq \frac{4\sigma_{\max}(\boldsymbol{\mathcal{D}})}{\underline{\lambda} - 4\sigma_{\max}(\boldsymbol{\mathcal{D}})} \leq \frac{8\sigma_{\max}(\boldsymbol{\mathcal{D}})}{\underline{\lambda}} \tag{15.1}$$

So we have,

$$\begin{aligned}\boldsymbol{\mathcal{T}}^*\cdot[\![\mathcal{P}_{\mathbf{U}_1},\mathcal{P}_{\mathbf{U}_2},\mathcal{P}_{\mathbf{U}_3}]\!] &= \boldsymbol{\mathcal{T}}^*\cdot[\![\mathcal{P}_{\mathbf{V}_1^*}+\mathbf{S}_1,\mathcal{P}_{\mathbf{V}_2^*}+\mathbf{S}_2,\mathcal{P}_{\mathbf{V}_3^*}+\mathbf{S}_3]\!]\\ =&\boldsymbol{\mathcal{T}}^*\cdot[\![\mathcal{P}_{\mathbf{V}_1^*},\mathcal{P}_{\mathbf{V}_2^*},\mathcal{P}_{\mathbf{V}_3^*}]\!] \qquad (15.2)\\ &+\boldsymbol{\mathcal{T}}^*\cdot[\![\mathbf{S}_1,\mathcal{P}_{\mathbf{V}_2^*},\mathcal{P}_{\mathbf{V}_3^*}]\!]+\boldsymbol{\mathcal{T}}^*\cdot[\![\mathcal{P}_{\mathbf{V}_1^*},\mathbf{S}_2,\mathcal{P}_{\mathbf{V}_3^*}]\!]+\boldsymbol{\mathcal{T}}^*\cdot[\![\mathcal{P}_{\mathbf{V}_1^*},\mathcal{P}_{\mathbf{V}_2^*},\mathbf{S}_3]\!]\\ &+\boldsymbol{\mathcal{T}}^*\cdot[\![\mathbf{S}_1,\mathbf{S}_2,\mathcal{P}_{\mathbf{V}_3^*}]\!]+\boldsymbol{\mathcal{T}}^*\cdot[\![\mathcal{P}_{\mathbf{V}_1^*},\mathbf{S}_2,\mathbf{S}_3]\!]+\boldsymbol{\mathcal{T}}^*\cdot[\![\mathbf{S}_1,\mathcal{P}_{\mathbf{V}_2^*},\mathbf{S}_3]\!]\\ &+\boldsymbol{\mathcal{T}}^*\cdot[\![\mathbf{S}_1,\mathbf{S}_2,\mathbf{S}_3]\!] \qquad (15.3)\end{aligned}$$

We now bound each of $\|\boldsymbol{\mathcal{T}}^*\cdot[\![\mathbf{S}_1,\mathbf{S}_2,\mathcal{P}_{\mathbf{V}_3^*}]\!]\|_{\mathrm{F}}$, $\|\boldsymbol{\mathcal{T}}^*\cdot[\![\mathcal{P}_{\mathbf{V}_1^*},\mathbf{S}_2,\mathbf{S}_3]\!]\|_{\mathrm{F}}$ and $\|\boldsymbol{\mathcal{T}}^*\cdot[\![\mathbf{S}_1,\mathcal{P}_{\mathbf{V}_2^*},\mathbf{S}_3]\!]\|_{\mathrm{F}}$. Without loss of generality, we only prove the bound of the first term.

$$\mathcal{M}_1\big(\boldsymbol{\mathcal{T}}^*\cdot[\![\mathbf{S}_1,\mathbf{S}_2,\mathcal{P}_{\mathbf{V}_3^*}]\!]\big)=\mathbf{S}_1\mathcal{M}_1(\boldsymbol{\mathcal{T}}^*)\left(\mathcal{P}_{\mathbf{V}_3}^*\otimes\mathbf{S}_2\right)^\top \qquad (15.4)$$

Write

$$\begin{aligned}\mathbf{S}_1\mathcal{M}_1(\boldsymbol{\mathcal{T}}^*)&=\left(\mathcal{S}_{1,1}+\sum_{j\geq 2}\mathcal{S}_{1,j}\right)\mathcal{M}_1(\boldsymbol{\mathcal{T}}^*)\\ &=\mathcal{P}_{\mathbf{V}_1^*}^{\perp}\mathcal{M}_1(\boldsymbol{\mathcal{D}})\left(\mathcal{M}_1(\boldsymbol{\mathcal{T}}^*)\right)^\dagger\mathcal{M}_1(\boldsymbol{\mathcal{T}}^*)+\sum_{j\geq 2}\mathcal{S}_{1,j}\mathcal{M}_1(\boldsymbol{\mathcal{T}}^*)\\ &=\mathcal{M}_1\big(\boldsymbol{\mathcal{D}}\cdot[\![\mathcal{P}_{\mathbf{V}_1^*}^{\perp},\mathcal{P}_{\mathbf{V}_2^*},\mathcal{P}_{\mathbf{V}_3^*}]\!]\big)+\sum_{j\geq 2}\mathcal{S}_{1,j}\mathcal{M}_1(\boldsymbol{\mathcal{T}}^*) \qquad (15.5)\end{aligned}$$

where we used the fact $\mathcal{P}_{\mathbf{V}_1^*}^{\perp}\mathcal{M}_1(\boldsymbol{\mathcal{T}}^*)=\mathbf{0}$.

Thus we obtain an upper bound for $\|\mathbf{S}_1\mathcal{M}_1(\boldsymbol{\mathcal{T}}^*)\|$ as follows

$$\|\mathbf{S}_1\mathcal{M}_1(\boldsymbol{\mathcal{T}}^*)\|\leq\sigma_{\max}(\boldsymbol{\mathcal{D}})+\underline{\lambda}\sum_{j\geq 2}\left(\frac{4\sigma_{\max}(\boldsymbol{\mathcal{D}})}{\underline{\lambda}}\right)^j\leq 4\sigma_{\max}(\boldsymbol{\mathcal{D}}), \qquad (15.6)$$

where the first inequality is due to the explicit form of $\mathcal{S}_{1,j}$. See [39, Theorem 1].

So from (15.4) and (15.6), we get

$$\|\boldsymbol{\mathcal{T}}^*\cdot[\![\mathbf{S}_1,\mathbf{S}_2,\mathcal{P}_{\mathbf{V}_3^*}]\!]\|_{\mathrm{F}}\leq\|\mathbf{S}_1\mathcal{M}_1(\boldsymbol{\mathcal{T}}^*)\|_{\mathrm{F}}\cdot\|\mathcal{P}_{\mathbf{V}_3}^*\otimes\mathbf{S}_2\|\leq C_1\sqrt{\bar{r}}\frac{\sigma_{\max}(\boldsymbol{\mathcal{D}})^2}{\underline{\lambda}} \qquad (15.7)$$

where $C_1>0$ is an absolute constant.

Now we consider the linear terms $\boldsymbol{\mathcal{T}}^*\cdot[\![\mathbf{S}_1,\mathcal{P}_{\mathbf{V}_2^*},\mathcal{P}_{\mathbf{V}_3^*}]\!]$, $\boldsymbol{\mathcal{T}}^*\cdot[\![\mathcal{P}_{\mathbf{V}_1^*},\mathbf{S}_2,\mathcal{P}_{\mathbf{V}_3^*}]\!]$ and $\boldsymbol{\mathcal{T}}^*\cdot[\![\mathcal{P}_{\mathbf{V}_1^*},\mathcal{P}_{\mathbf{V}_2^*},\mathbf{S}_3]\!]$. Clearly, we have

$$\begin{aligned}\mathcal{M}_1\left(\boldsymbol{\mathcal{T}}^*\cdot[\![\mathbf{S}_1,\mathcal{P}_{\mathbf{V}_2^*},\mathcal{P}_{\mathbf{V}_3^*}]\!]\right)&=\mathbf{S}_1\mathcal{M}_1(\boldsymbol{\mathcal{T}}^*)\\ \mathcal{M}_2\left(\boldsymbol{\mathcal{T}}^*\cdot[\![\mathcal{P}_{\mathbf{V}_1^*},\mathbf{S}_2,\mathcal{P}_{\mathbf{V}_3^*}]\!]\right)&=\mathbf{S}_2\mathcal{M}_2(\boldsymbol{\mathcal{T}}^*)\\ \mathcal{M}_3\left(\boldsymbol{\mathcal{T}}^*\cdot[\![\mathcal{P}_{\mathbf{V}_1^*},\mathcal{P}_{\mathbf{V}_2^*},\mathbf{S}_3]\!]\right)&=\mathbf{S}_3\mathcal{M}_3(\boldsymbol{\mathcal{T}}^*), \qquad (15.8)\end{aligned}$$

whose explicit representations are already studied in eq. (15.5). As a result, we can write

$$\begin{aligned}
&\boldsymbol{\mathcal{T}}^* \cdot [\![\mathbf{S}_1, \mathcal{P}_{\mathbf{V}_2^*}, \mathcal{P}_{\mathbf{V}_3^*}]\!] + \boldsymbol{\mathcal{T}}^* \cdot [\![\mathcal{P}_{\mathbf{V}_1^*}, \mathbf{S}_2, \mathcal{P}_{\mathbf{V}_3^*}]\!] + \boldsymbol{\mathcal{T}}^* \cdot [\![\mathcal{P}_{\mathbf{V}_1^*}, \mathcal{P}_{\mathbf{V}_2^*}, \mathbf{S}_3]\!] \\
&\quad = \boldsymbol{\mathcal{D}} \cdot [\![\mathcal{P}^{\perp}_{\mathbf{V}_1^*}, \mathcal{P}_{\mathbf{V}_2^*}, \mathcal{P}_{\mathbf{V}_3^*}]\!] + \boldsymbol{\mathcal{D}} \cdot [\![\mathcal{P}_{\mathbf{V}_1^*}, \mathcal{P}^{\perp}_{\mathbf{V}_2^*}, \mathcal{P}_{\mathbf{V}_3^*}]\!] + \boldsymbol{\mathcal{D}} \cdot [\![\mathcal{P}_{\mathbf{V}_1^*}, \mathcal{P}_{\mathbf{V}_2^*}, \mathcal{P}^{\perp}_{\mathbf{V}_3^*}]\!] \\
&\qquad + \sum_{j \geq 2} \Big(\mathcal{M}_1(\boldsymbol{\mathcal{T}}^*) \cdot [\![\mathcal{S}_{1,j}, \mathcal{P}_{\mathbf{V}_2^*}, \mathcal{P}_{\mathbf{V}_3^*}]\!] + \mathcal{M}_2(\boldsymbol{\mathcal{T}}^*) \cdot [\![\mathcal{P}_{\mathbf{V}_1^*}, \mathcal{S}_{2,j}, \mathcal{P}_{\mathbf{V}_3^*}]\!] + \mathcal{M}_3(\boldsymbol{\mathcal{T}}^*) \cdot [\![\mathcal{P}_{\mathbf{V}_1^*}, \mathcal{P}_{\mathbf{V}_2^*}, \mathcal{S}_{3,j}]\!] \Big).
\end{aligned} \tag{15.9}$$

Now we bound $\boldsymbol{\mathcal{D}} \cdot [\![\mathcal{P}_{\mathbf{U}_1}, \mathcal{P}_{\mathbf{U}_2}, \mathcal{P}_{\mathbf{U}_3}]\!]$ as follows

$$\begin{aligned}
\boldsymbol{\mathcal{D}} \cdot [\![\mathcal{P}_{\mathbf{U}_1}, \mathcal{P}_{\mathbf{U}_2}, \mathcal{P}_{\mathbf{U}_3}]\!] &= \boldsymbol{\mathcal{D}} \cdot [\![\mathcal{P}_{\mathbf{V}_1^*} + \mathbf{S}_1, \mathcal{P}_{\mathbf{V}_2^*} + \mathbf{S}_2, \mathcal{P}_{\mathbf{V}_3^*} + \mathbf{S}_3]\!] \\
&= \boldsymbol{\mathcal{D}} \cdot [\![\mathcal{P}_{\mathbf{V}_1^*}, \mathcal{P}_{\mathbf{V}_2^*}, \mathcal{P}_{\mathbf{V}_3^*}]\!] \\
&\quad + \boldsymbol{\mathcal{D}} \cdot [\![\mathbf{S}_1, \mathcal{P}_{\mathbf{V}_2^*}, \mathcal{P}_{\mathbf{V}_3^*}]\!] + \boldsymbol{\mathcal{D}} \cdot [\![\mathcal{P}_{\mathbf{V}_1^*}, \mathbf{S}_2, \mathcal{P}_{\mathbf{V}_3^*}]\!] + \boldsymbol{\mathcal{D}} \cdot [\![\mathcal{P}_{\mathbf{V}_1^*}, \mathcal{P}_{\mathbf{V}_2^*}, \mathbf{S}_3]\!] \\
&\quad + \boldsymbol{\mathcal{D}} \cdot [\![\mathbf{S}_1, \mathbf{S}_2, \mathcal{P}_{\mathbf{V}_3^*}]\!] + \boldsymbol{\mathcal{D}} \cdot [\![\mathcal{P}_{\mathbf{V}_1^*}, \mathbf{S}_2, \mathbf{S}_3]\!] + \boldsymbol{\mathcal{D}} \cdot [\![\mathbf{S}_1, \mathcal{P}_{\mathbf{V}_2^*}, \mathbf{S}_3]\!] \\
&\quad + \boldsymbol{\mathcal{D}} \cdot [\![\mathbf{S}_1, \mathbf{S}_2, \mathbf{S}_3]\!]
\end{aligned} \tag{15.10}$$

Similarly as proving the bound (15.7), we can show

$$\max\big\{ \|\boldsymbol{\mathcal{D}} \cdot [\![\mathbf{S}_1, \mathcal{P}_{\mathbf{V}_2^*}, \mathcal{P}_{\mathbf{V}_3^*}]\!]\|_{\mathrm{F}}, \|\boldsymbol{\mathcal{D}} \cdot [\![\mathbf{S}_1, \mathbf{S}_2, \mathcal{P}_{\mathbf{V}_3^*}]\!]\|_{\mathrm{F}}, \|\boldsymbol{\mathcal{D}} \cdot [\![\mathbf{S}_1, \mathbf{S}_2, \mathbf{S}_3]\!]\|_{\mathrm{F}} \big\} \leq C_1 \sqrt{\bar{r}} \frac{\sigma_{\max}(\boldsymbol{\mathcal{D}})^2}{\underline{\lambda}} \tag{15.11}$$

where $C_1 > 0$ is an absolute constant.

Finally, by (15.5), (15.7), (15.9) and (15.11), we have

$$\begin{aligned}
&\big\| (\boldsymbol{\mathcal{T}}^* + \boldsymbol{\mathcal{D}}) \cdot [\![\mathcal{P}_{\mathbf{U}_1}, \mathcal{P}_{\mathbf{U}_2}, \mathcal{P}_{\mathbf{U}_3}]\!] - \boldsymbol{\mathcal{T}}^* \big\|_{\mathrm{F}} \\
&\quad \leq \big\| \boldsymbol{\mathcal{D}} \cdot [\![\mathcal{P}^{\perp}_{\mathbf{V}_1^*}, \mathcal{P}_{\mathbf{V}_2^*}, \mathcal{P}_{\mathbf{V}_3^*}]\!] + \boldsymbol{\mathcal{D}} \cdot [\![\mathcal{P}_{\mathbf{V}_1^*}, \mathcal{P}^{\perp}_{\mathbf{V}_2^*}, \mathcal{P}_{\mathbf{V}_3^*}]\!] + \boldsymbol{\mathcal{D}} \cdot [\![\mathcal{P}_{\mathbf{V}_1^*}, \mathcal{P}_{\mathbf{V}_2^*}, \mathcal{P}^{\perp}_{\mathbf{V}_3^*}]\!] + \boldsymbol{\mathcal{D}} \cdot [\![\mathcal{P}_{\mathbf{V}_1^*}, \mathcal{P}_{\mathbf{V}_2^*}, \mathcal{P}_{\mathbf{V}_3^*}]\!] \big\|_{\mathrm{F}} \\
&\qquad + C_1 \frac{\sqrt{\bar{r}} \sigma_{\max}(\boldsymbol{\mathcal{D}})^2}{\underline{\lambda}} \\
&\quad \leq \|\boldsymbol{\mathcal{D}}\|_{\mathrm{F}} + C_2 \frac{\sqrt{\bar{r}} \sigma_{\max}(\boldsymbol{\mathcal{D}})^2}{\underline{\lambda}}
\end{aligned} \tag{15.12}$$

where $C_1, C_2 > 0$ are absolute constants ($C_{2,m} = 16m + 2^{m+1}$ in the case of general m). This finishes the proof of the lemma. □

Lemma 15.3. *Assume all the entries of $\boldsymbol{\mathcal{Z}} \in \mathbb{R}^{d_1 \times \ldots \times d_m}$ are independent mean-zero random variables with bounded Orlicz-ψ_2 norm:*

$$\|[\boldsymbol{\mathcal{Z}}]_\omega\|_{\psi_2} = \sup_{q \geq 1} (\mathbb{E}|[\boldsymbol{\mathcal{Z}}]_\omega|^q)^{1/q} / q^{1/2} \leq \sigma_z$$

Then there exists some constants $C_m, c_m > 0$ depending only on m such that

$$\sup_{\boldsymbol{\mathcal{M}} \in \mathbb{M}_{2\mathbf{r}}, \|\boldsymbol{\mathcal{M}}\|_F \leq 1} \langle \boldsymbol{\mathcal{Z}}, \boldsymbol{\mathcal{M}} \rangle \leq C_m \sigma_z \left(r^* + \sum_{i=1}^{m} d_i r_i \right)^{1/2}$$

with probability at least $1 - \exp(-c_m \sum_{i=1}^{m} d_i r_i)$, where $r^* = r_1 \dots r_m$.

Proof. See the proof of ([14], Lemma D.5). □

Lemma 15.4 (Maximum of sub-Gaussian)**.** *Let $Z_1, \dots, Z_N$ be N random variables such that $\mathbb{E} \exp\{tZ_i\} \leq \exp\{t^2 \sigma_z^2 / 2\}$ for all $i \in [N]$. Then*

$$\mathbb{P}(\max_{1 \leq i \leq N} |Z_i| > t) \leq 2N \exp(-\frac{t^2}{2\sigma_z^2}).$$

Proof. The claim follows from the following two facts:

$$\mathbb{P}(\max_{1 \leq i \leq N} Z_i > t) \leq \mathbb{P}(\cup_{1 \leq i \leq N} \{Z_i > t\}) \leq N \mathbb{P}(Z_i > t) \leq N \exp(-\frac{t^2}{2\sigma_z^2}),$$

and

$$\max_{1 \leq i \leq N} |Z_i| = \max_{1 \leq i \leq 2N} Z_i$$

with $Z_{N+i} = -Z_i$ for $i \in [N]$. □

Lemma 15.5 (Spikiness implies incoherence)**.** *Let $\boldsymbol{\mathcal{T}}^* \in \mathbb{M}_{\mathbf{r}}$ satisfies Assumption 1 with parameter μ_1. Then we have:*

$$\mu(\boldsymbol{\mathcal{T}}^*) \leq \mu_1 \kappa_0.$$

where $\mu(\boldsymbol{\mathcal{T}}^*)$ is the incoherence parameter of $\boldsymbol{\mathcal{T}}^*$ and κ_0 is the condition number of $\boldsymbol{\mathcal{T}}^*$.

Proof. Denote $\boldsymbol{\mathcal{T}}^* = \boldsymbol{\mathcal{C}}^* \cdot [\![\mathbf{U}_1, \dots, \mathbf{U}_m]\!]$. Now we check the incoherence condition of $\boldsymbol{\mathcal{T}}^*$. For all $i \in [d_j]$ and $j \in [m]$,

$$\|\mathbf{e}_i^\top \mathcal{M}_j(\boldsymbol{\mathcal{T}}^*)\|_{\ell_2} = \|\mathbf{e}_i \mathbf{U}_j \mathcal{M}_j(\boldsymbol{\mathcal{C}}^*)\|_{\ell_2} \geq \|\mathbf{e}_i^\top \mathbf{U}_j\|_{\ell_2} \cdot \underline{\lambda} \geq \|\mathbf{e}_i^\top \mathbf{U}_j\|_{\ell_2} \frac{\|\boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}}{\sqrt{r_j} \kappa_0}.$$

On the other hand, we have

$$\|\mathbf{e}_i^\top \mathcal{M}_j(\boldsymbol{\mathcal{T}}^*)\|_{\ell_2} \leq \sqrt{d_j^-} \|\boldsymbol{\mathcal{T}}^*\|_{\ell_\infty} \leq \mu_1 \|\boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}} \frac{1}{\sqrt{d_j}},$$

where the last inequality is due to the spikiness condition $\boldsymbol{\mathcal{T}}^*$ satisfies. Together with these two inequalities, we have

$$\|\mathbf{e}_i^\top \mathbf{U}_j\|_{\ell_2} \leq \sqrt{\frac{r_j}{d_j}} \mu_1 \kappa_0.$$

And this finishes the proof of the lemma. □

Lemma 15.6. *Let $\boldsymbol{\mathcal{T}}^* \in \mathbb{M}_{\mathbf{r}}$ satisfies Assumption 1 with parameter μ_1. Suppose that $\boldsymbol{\mathcal{W}}$ satisfies $\|\boldsymbol{\mathcal{W}} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}} \leq \frac{\underline{\lambda}}{8}$, then we have $\mathsf{Trim}_{\zeta,\mathbf{r}}(\boldsymbol{\mathcal{W}})$ is $(2\mu_1\kappa_0)^2$-incoherent if we choose $\zeta = \frac{16}{7}\mu_1\frac{\|\boldsymbol{\mathcal{W}}\|_{\mathrm{F}}}{\sqrt{d^*}}$. Also, it satisfies*

$$\|\mathsf{Trim}_{\zeta,\mathbf{r}}(\boldsymbol{\mathcal{W}}) - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}} \leq \|\boldsymbol{\mathcal{W}} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}} + \frac{C_m\sqrt{\bar{r}}\|\boldsymbol{\mathcal{W}} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2}{\underline{\lambda}},$$

where $C_m > 0$ depends only on m.

Proof. Notice $\mathsf{Trim}_{\zeta,\mathbf{r}}(\boldsymbol{\mathcal{W}}) = \mathscr{H}^{\mathrm{HO}}_{\mathbf{r}}(\widetilde{\boldsymbol{\mathcal{W}}})$, where $\widetilde{\boldsymbol{\mathcal{W}}}$ is the entrywise truncation of $\boldsymbol{\mathcal{W}}$ with the thresholding $\zeta/2$. To check the incoherence of $\mathscr{H}^{\mathrm{HO}}_{\mathbf{r}}(\widetilde{\boldsymbol{\mathcal{W}}})$, denote $\widetilde{\mathbf{U}}_j$ the top-r_j left singular vectors of $\mathcal{M}_j(\widetilde{\boldsymbol{\mathcal{W}}})$, and $\widetilde{\boldsymbol{\Lambda}}_j$ the $r_j \times r_j$ diagonal matrix containing the top-r_j singular values of $\mathcal{M}_j(\widetilde{\boldsymbol{\mathcal{W}}})$. Then, there exist a $\widetilde{\mathbf{V}}_j \in \mathbb{R}^{d_j^- \times r_j}$ satisfying $\widetilde{\mathbf{V}}_j^\top\widetilde{\mathbf{V}}_j = \mathbf{I}_{r_j}$ such that

$$\widetilde{\mathbf{U}}_j\widetilde{\boldsymbol{\Lambda}}_j = \mathcal{M}_j(\widetilde{\boldsymbol{\mathcal{W}}})\widetilde{\mathbf{V}}_j.$$

Now we can also bound the ℓ_∞-norm of $\boldsymbol{\mathcal{T}}^*$:

$$\|\boldsymbol{\mathcal{T}}^*\|_{\ell_\infty} \leq \mu_1\frac{\|\boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}}{\sqrt{d^*}} \leq \mu_1\frac{\|\boldsymbol{\mathcal{W}}\|_{\mathrm{F}} + \|\boldsymbol{\mathcal{T}}^* - \boldsymbol{\mathcal{W}}\|_{\mathrm{F}}}{\sqrt{d^*}} \leq \mu_1\frac{\|\boldsymbol{\mathcal{W}}\|_{\mathrm{F}} + \|\boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}/8}{\sqrt{d^*}}.$$

This together with the definition of ζ, we have:

$$\mu_1\frac{\|\boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}}{\sqrt{d^*}} \leq 8/7 \cdot \mu_1\frac{\|\boldsymbol{\mathcal{W}}\|_{\mathrm{F}}}{\sqrt{d^*}} = \zeta/2.$$

And thus $\|\boldsymbol{\mathcal{T}}^*\|_{\ell_\infty} \leq \zeta/2$. Then for all $i \in [d_j]$,

$$\|\mathbf{e}_i^\top\widetilde{\mathbf{U}}_j\|_{\ell_2} = \|\mathbf{e}_i^\top\mathcal{M}_j(\widetilde{\boldsymbol{\mathcal{W}}})\widetilde{\mathbf{V}}_j\widetilde{\boldsymbol{\Lambda}}_j^{-1}\|_{\ell_2} \leq \frac{\|\mathbf{e}_i^\top\mathcal{M}_j(\widetilde{\boldsymbol{\mathcal{W}}})\|_{\ell_2}}{\lambda_{r_j}(\widetilde{\boldsymbol{\Lambda}}_j)} \leq \frac{\zeta/2 \cdot (d_j^-)^{1/2}}{7/8 \cdot \lambda_{r_j}(\mathcal{M}_j(\boldsymbol{\mathcal{T}}^*))}.$$

where the last inequality is due to $\|\widetilde{\boldsymbol{\mathcal{W}}} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}} \leq \|\boldsymbol{\mathcal{W}} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}} \leq \underline{\lambda}/8$ since $\|\boldsymbol{\mathcal{T}}^*\|_{\ell_\infty} \leq \zeta/2$ and $\|\widetilde{\boldsymbol{\mathcal{W}}}\|_{\ell_\infty} \leq \zeta/2$. Meanwhile,

$$\|\boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}} \leq \sqrt{r_j}\kappa_0\lambda_{r_j}(\mathcal{M}_j(\boldsymbol{\mathcal{T}}^*)).$$

There for the $\zeta = \frac{16}{7}\mu_1\frac{\|\boldsymbol{\mathcal{W}}\|_{\mathrm{F}}}{\sqrt{d^*}}$, we have for all $j \in [m]$

$$\max_{i\in[d_j]}\|\mathbf{e}_i\widetilde{\mathbf{U}}_j\|_{\ell_2} \leq \frac{64}{49}\mu_1\kappa_0\frac{\|\boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}} + \underline{\lambda}/8}{\|\boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}}\sqrt{\frac{r_j}{d_j}} \leq 2\mu_1\kappa_0\sqrt{\frac{r_j}{d_j}}.$$

where the second last inequality is from $\|\boldsymbol{\mathcal{W}}\|_{\mathrm{F}} \leq \|\boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}} + \|\boldsymbol{\mathcal{W}} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}$ and the last inequality is from $\|\boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}} \geq \underline{\lambda}$.

The second claim follows from the fact that $\|\widetilde{\boldsymbol{\mathcal{W}}} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}} \leq \|\boldsymbol{\mathcal{W}} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}} \leq \underline{\lambda}/8$, and from Lemma 15.2,

$$\begin{aligned}\|\mathsf{Trim}_{\zeta,\mathbf{r}}(\boldsymbol{\mathcal{W}}) - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}} = \|\widetilde{\boldsymbol{\mathcal{W}}} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}} &\leq \|\widetilde{\boldsymbol{\mathcal{W}}} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}} + C_m \frac{\sqrt{\bar{r}}\|\widetilde{\boldsymbol{\mathcal{W}}} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2}{\underline{\lambda}} \\ &\leq \|\boldsymbol{\mathcal{W}} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}} + C_m \frac{\sqrt{\bar{r}}\|\boldsymbol{\mathcal{W}} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2}{\underline{\lambda}}\end{aligned}$$

This finishes the proof of the lemma. □

We introduce some notations for the following lemmas. Denote by $\widehat{\boldsymbol{\mathcal{T}}}_l = \boldsymbol{\mathcal{C}}_l \cdot (\mathbf{U}_1, \ldots, \mathbf{U}_m)$, $\boldsymbol{\mathcal{T}}^* = \boldsymbol{\mathcal{C}}^* \cdot (\mathbf{U}_1^*, \ldots, \mathbf{U}_m^*)$.

$$\mathbf{R}_i = \arg\min_{\mathbf{R} \in \mathbb{O}_{r_i}} \|\mathbf{U}_i - \mathbf{U}_i^* \mathbf{R}\|_{\mathrm{F}}, i \in [m] \tag{15.13}$$

If we let $\mathbf{U}_i^{*T}\mathbf{U}_i = \mathbf{L}_i \mathbf{S}_i \mathbf{W}_i^\top$ be the SVD of $\mathbf{U}_i^{*T}\mathbf{U}_i$, then the closed form of $\mathbf{R}_i$ is given by $\mathbf{R}_i = \mathbf{L}_i \mathbf{W}_i^\top$. And we rewrite

$$\boldsymbol{\mathcal{T}}^* = \boldsymbol{\mathcal{S}}^* \cdot (\mathbf{V}_1^*, \cdots, \mathbf{V}_m^*)$$

where $\boldsymbol{\mathcal{S}}^* = \boldsymbol{\mathcal{C}}^* \cdot (\mathbf{R}_1^\top, \cdots, \mathbf{R}_m^\top)$ and $\mathbf{V}_i^* = \mathbf{U}_i^* \mathbf{R}_i, i \in [m]$. So $\mathbf{V}_i^*$ is also μ_0-incoherent.

Lemma 15.7 (Entry-wise estimation of $|[\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*]_\omega|$)**.** *Suppose $\boldsymbol{\mathcal{T}}^*$ satisfies Assumption 1. Under the assumptions that $\widehat{\boldsymbol{\mathcal{T}}}_l$ is $(2\mu_1\kappa_0)^2$-incoherent and $\|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}} \leq \frac{\underline{\lambda}}{16m\bar{r}^{1/2}\kappa_0}$, then we have*

$$|[\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*]_\omega|^2 \leq C_m \bar{r}^m \underline{d}^{-(m-1)} (\mu_1\kappa_0)^{4m} \|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2,$$

where $C_m = 2^{4m+1}(m+1)$.

Proof. First we have

$$\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^* = (\boldsymbol{\mathcal{C}}_l - \boldsymbol{\mathcal{S}}^*) \cdot (\mathbf{U}_1, \cdots, \mathbf{U}_m) + \sum_{i=1}^m \boldsymbol{\mathcal{S}}^* \cdot (\mathbf{V}_1^*, \ldots, \mathbf{V}_{i-1}^*, \mathbf{U}_i - \mathbf{V}_i^*, \mathbf{U}_{i+1}, \ldots, \mathbf{U}_m) \tag{15.14}$$

From Lemma 15.5, we get $\boldsymbol{\mathcal{T}}^*$ is $\mu_1^2\kappa_0^2$-incoherent. So we have for all $\omega = (\omega_1, \ldots, \omega_m) \in [d_1] \times \ldots \times [d_m]$

$$\begin{aligned}|[\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*]_\omega| &\leq \|\boldsymbol{\mathcal{C}}_l - \boldsymbol{\mathcal{S}}^*\|_{\mathrm{F}} \prod_{i=1}^m \|(\mathbf{U}_i)_{\omega_i:}\| + \sum_{i=1}^m \|\boldsymbol{\mathcal{S}}^*\|_{\mathrm{F}} \|(\mathbf{U}_i - \mathbf{V}_i^*)_{\omega_i:}\| \prod_{k=1}^{i-1} \|(\mathbf{V}_k^*)_{\omega_k:}\| \prod_{k=i+1}^m \|(\mathbf{U}_k)_{\omega_k:}\| \\ &\leq \sqrt{\frac{r^*}{d^*}} (2\mu_1\kappa_0)^{2m} \|\boldsymbol{\mathcal{C}}_l - \boldsymbol{\mathcal{S}}^*\|_{\mathrm{F}} + (2\mu_1\kappa_0)^{2m-2} \sqrt{\frac{\bar{r}^{m-1}}{\underline{d}^{m-1}}} \|\boldsymbol{\mathcal{S}}^*\|_{\mathrm{F}} \sum_{i=1}^m \|(\mathbf{U}_i - \mathbf{V}_i^*)_{\omega_i:}\|\end{aligned}$$

where $r^* = \prod_{i=1}^m r_i, d^* = \prod_{i=1}^m d_i$ and $\bar{r} = \max_{i=1}^m r_i, \underline{d} = \min_{i=1}^m d_i$. From AG–GM inequality, we have

$$|[\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*]_\omega|^2 \leq (m+1)(2\mu_1\kappa_0)^{4m}\frac{r^*}{d^*}\|\boldsymbol{\mathcal{C}}_l - \boldsymbol{\mathcal{S}}^*\|_{\mathrm{F}}^2 + (m+1)(2\mu_1\kappa_0)^{4m-4}\frac{\bar{r}^{m-1}}{\underline{d}^{m-1}}\|\boldsymbol{\mathcal{S}}^*\|_{\mathrm{F}}^2\sum_{i=1}^m\|(\mathbf{U}_i - \mathbf{V}_i^*)_{\omega_i:}\|^2 \tag{15.15}$$

$$\leq (m+1)\bar{r}^m\underline{d}^{-(m-1)}(2\mu_1\kappa_0)^{4m}\left(\|\boldsymbol{\mathcal{C}}_l - \boldsymbol{\mathcal{S}}^*\|_{\mathrm{F}}^2 + \underline{\lambda}^2\sum_{i=1}^m\|\mathbf{U}_i - \mathbf{V}_i^*\|_{\mathrm{F}}^2\right)$$

$$\leq 2(m+1)\bar{r}^m\underline{d}^{-(m-1)}(2\mu_1\kappa_0)^{4m}\|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2$$

where the last inequality is from Lemma 15.9, and this finishes the proof of the lemma. □

Lemma 15.8 (Estimation of $\|\mathcal{P}_\Omega(\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*)\|_{\mathrm{F}}^2$)**.** *Let Ω be the α-fraction set. Suppose $\boldsymbol{\mathcal{T}}^*$ satisfies Assumption 1. Under the assumptions that $\widehat{\boldsymbol{\mathcal{T}}}_l$ is $(2\mu_1\kappa_0)^2$-incoherent and $\|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}} \leq \frac{\underline{\lambda}}{16m\bar{r}^{1/2}\kappa_0}$, we have*

$$\|\mathcal{P}_\Omega(\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*)\|_{\mathrm{F}}^2 \leq C_m(\mu_1\kappa_0)^{4m}\bar{r}^m\alpha\|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2,$$

where $C_m = 2^{4m+1}(m+1)$.

Proof. First from (15.15) in Lemma 15.7, we have

$$|[\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*]_\omega|^2 \leq (m+1)(2\mu_1\kappa_0)^{4m}\frac{r^*}{d^*}\|\boldsymbol{\mathcal{C}}_l - \boldsymbol{\mathcal{S}}^*\|_{\mathrm{F}}^2 + (m+1)(2\mu_1\kappa_0)^{4m-4}\frac{\bar{r}^{m-1}}{\underline{d}^{m-1}}\|\boldsymbol{\mathcal{S}}^*\|_{\mathrm{F}}^2\sum_{i=1}^m\|(\mathbf{U}_i - \mathbf{V}_i^*)_{\omega_i:}\|^2.$$

Since Ω is an α-fraction set, we have

$$\|\mathcal{P}_\Omega(\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*)\|_{\mathrm{F}}^2 = \sum_{\omega\in\Omega}[\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*]_\omega^2$$

$$\leq (m+1)(2\mu_1\kappa_0)^{4m}\alpha r^*\|\boldsymbol{\mathcal{C}}_l - \boldsymbol{\mathcal{S}}^*\|_{\mathrm{F}}^2 + (m+1)(2\mu_1\kappa_0)^{4m-4}\alpha\bar{r}^{m-1}\|\boldsymbol{\mathcal{S}}^*\|_{\mathrm{F}}^2\sum_{i=1}^m\|\mathbf{U}_i - \mathbf{V}_i^*\|_{\mathrm{F}}^2$$

$$\leq (m+1)(2\mu_1\kappa_0)^{4m}\alpha r^*\|\boldsymbol{\mathcal{C}}_l - \boldsymbol{\mathcal{S}}^*\|_{\mathrm{F}}^2 + (m+1)(2\mu_1\kappa_0)^{4m-4}\alpha\bar{r}^m\overline{\lambda}^2\sum_{i=1}^m\|\mathbf{U}_i - \mathbf{V}_i^*\|_{\mathrm{F}}^2$$

$$\leq (m+1)(2\mu_1\kappa_0)^{4m}\bar{r}^m\alpha\left(\|\boldsymbol{\mathcal{C}}_l - \boldsymbol{\mathcal{S}}^*\|_{\mathrm{F}}^2 + \underline{\lambda}^2\sum_{i=1}^m\|\mathbf{U}_i - \mathbf{V}_i^*\|_{\mathrm{F}}^2\right) \tag{15.16}$$

Now we invoke Lemma 15.9, and we get

$$\|\mathcal{P}_\Omega(\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*)\|_{\mathrm{F}}^2 \leq 2(m+1)(2\mu_1\kappa_0)^{4m}\bar{r}^m\alpha\|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2,$$

which finishes the proof of the lemma. □

Lemma 15.9 (Estimation of $\|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2$)**.** *Let $\widehat{\boldsymbol{\mathcal{T}}}_l = \boldsymbol{\mathcal{C}}_l \cdot (\mathbf{U}_1, \cdots, \mathbf{U}_m)$ be the l-th step value in Algorithm 2 and let $\boldsymbol{\mathcal{T}}^* = \boldsymbol{\mathcal{S}}^* \cdot (\mathbf{V}_1^*, \cdots, \mathbf{V}_m^*)$. Suppose $\widehat{\boldsymbol{\mathcal{T}}}_l$ satisfies $\|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}} \leq \frac{\underline{\lambda}}{16m\bar{r}^{1/2}\kappa_0}$. Then we have the following estimation for $\|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2$:*

$$\|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2 \geq 0.5\|\boldsymbol{\mathcal{C}}_l - \boldsymbol{\mathcal{S}}^*\|_{\mathrm{F}}^2 + 0.5\underline{\lambda}^2 \sum_{i=1}^m \|\mathbf{U}_i - \mathbf{V}_i^*\|_{\mathrm{F}}^2.$$

Proof. First we have

$$\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^* = (\boldsymbol{\mathcal{C}}_l - \boldsymbol{\mathcal{S}}^*) \cdot (\mathbf{U}_1, \cdots, \mathbf{U}_m) + \sum_{i=1}^m \boldsymbol{\mathcal{S}}^* \cdot (\mathbf{V}_1^*, \ldots, \mathbf{V}_{i-1}^*, \mathbf{U}_i - \mathbf{V}_i^*, \mathbf{U}_{i+1}, \ldots, \mathbf{U}_m) \quad (15.17)$$

Notice that we have

$$\|\boldsymbol{\mathcal{S}}^* \cdot (\mathbf{V}_1^*, \ldots, \mathbf{V}_{i-1}^*, \mathbf{U}_i - \mathbf{V}_i^*, \mathbf{U}_{i+1}, \ldots, \mathbf{U}_m)\|_{\mathrm{F}}^2 = \|(\mathbf{U}_i - \mathbf{V}_i^*)\mathcal{M}_i(\boldsymbol{\mathcal{S}}^*)\|_{\mathrm{F}}^2 \quad (15.18)$$

Denote $\boldsymbol{\mathcal{X}}_i = \boldsymbol{\mathcal{S}}^* \cdot (\mathbf{V}_1^*, \ldots, \mathbf{V}_{i-1}^*, \mathbf{U}_i - \mathbf{V}_i^*, \mathbf{U}_{i+1}, \ldots, \mathbf{U}_m)$, then we have

$$\begin{aligned}
\|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2 &= \|\boldsymbol{\mathcal{C}}_l - \boldsymbol{\mathcal{S}}^*\|_{\mathrm{F}}^2 + \sum_{i=1}^m \|(\mathbf{U}_i - \mathbf{V}_i^*)\mathcal{M}_i(\boldsymbol{\mathcal{S}}^*)\|_{\mathrm{F}}^2 + 2\sum_{i<j}\langle \boldsymbol{\mathcal{X}}_i, \boldsymbol{\mathcal{X}}_j\rangle + 2\sum_{i=1}^m \langle (\boldsymbol{\mathcal{C}}_l - \boldsymbol{\mathcal{S}}^*) \cdot (\mathbf{U}_1, \cdots, \mathbf{U}_m) \\
&\geq \|\boldsymbol{\mathcal{C}}_l - \boldsymbol{\mathcal{S}}^*\|_{\mathrm{F}}^2 + \sum_{i=1}^m \underline{\lambda}^2 \|\mathbf{U}_i - \mathbf{V}_i^*\|_{\mathrm{F}}^2 + 2\sum_{i<j}\langle \boldsymbol{\mathcal{X}}_i, \boldsymbol{\mathcal{X}}_j\rangle + 2\sum_{i=1}^m \langle (\boldsymbol{\mathcal{C}}_l - \boldsymbol{\mathcal{S}}^*) \cdot (\mathbf{U}_1, \cdots, \mathbf{U}_m), \boldsymbol{\mathcal{X}}_i\rangle
\end{aligned} \quad (15.19)$$

Notice that $\mathcal{M}_i(\boldsymbol{\mathcal{X}}_i) = (\mathbf{U}_i - \mathbf{V}_i^*)\mathcal{M}_i(\boldsymbol{\mathcal{S}}^*)\big(\mathbf{U}_m \otimes \mathbf{U}_{i+1} \otimes \mathbf{V}_{i-1} \otimes \mathbf{V}_1\big)^\top$. So we have the estimation of $|\langle (\boldsymbol{\mathcal{C}}_l - \boldsymbol{\mathcal{S}}^*) \cdot (\mathbf{U}_1, \cdots, \mathbf{U}_m), \boldsymbol{\mathcal{X}}_i\rangle|$ is as follows:

$$\begin{aligned}
|\langle (\boldsymbol{\mathcal{C}}_l - \boldsymbol{\mathcal{S}}^*) \cdot (\mathbf{U}_1, \cdots, \mathbf{U}_m), \boldsymbol{\mathcal{X}}_i\rangle| &= |\langle \mathcal{M}_i\big((\boldsymbol{\mathcal{C}}_l - \boldsymbol{\mathcal{S}}^*) \cdot (\mathbf{U}_1, \cdots, \mathbf{U}_m)\big), \mathcal{M}_i(\boldsymbol{\mathcal{X}}_i)\rangle| \\
&\leq \|(\mathbf{U}_i - \mathbf{V}_i^*)^\top \mathbf{U}_i\| \|\boldsymbol{\mathcal{C}}_l - \boldsymbol{\mathcal{S}}^*\|_{\mathrm{F}} \|\boldsymbol{\mathcal{S}}^*\|_{\mathrm{F}} \\
&\leq \sqrt{\bar{r}}\bar{\lambda} \|\mathbf{U}_i^\top (\mathbf{U}_i - \mathbf{V}_i^*)\|_{\mathrm{F}} \|\boldsymbol{\mathcal{C}}_l - \boldsymbol{\mathcal{S}}^*\|_{\mathrm{F}}
\end{aligned} \quad (15.20)$$

Now we estimate $\|\mathbf{U}_i^\top (\mathbf{U}_i - \mathbf{V}_i^*)\|_{\mathrm{F}}$ by plugging in the closed form of $\mathbf{V}_i^*$ as in (15.13)

$$\|\mathbf{U}_i^\top (\mathbf{U}_i - \mathbf{V}_i^*)\|_{\mathrm{F}} = \|\mathbf{I} - \mathbf{S}_i\|_{\mathrm{F}} \leq \|\mathbf{I} - \mathbf{S}_i^2\|_{\mathrm{F}} = \|\mathbf{U}_{i\perp}^{*T}\mathbf{U}_i\|_{\mathrm{F}}^2 \leq \|\mathbf{U}_i - \mathbf{U}_i^*\mathbf{R}_i\|_{\mathrm{F}}^2 \quad (15.21)$$

From Wedin' sinΘ Theorem, we have for $i \in [m]$

$$\|\mathbf{U}_i - \mathbf{V}_i^*\|_{\mathrm{F}} \leq \|\mathbf{U}_i - \mathbf{U}_i^*\|_{\mathrm{F}} \leq \frac{\sqrt{2}\|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}}{\underline{\lambda} - \|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*\|} \leq \frac{2\sqrt{2}\|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}}{\underline{\lambda}} \leq \frac{1}{4m\bar{r}^{1/2}\kappa_0} \quad (15.22)$$

where the second last inequality is from $\|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*\| \leq \underline{\lambda}/2$ and the last inequality is from $\|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}} \leq \frac{\underline{\lambda}}{16m\bar{r}^{1/2}\kappa_0}$. Then from (15.20) and (15.22), we have

$$|\langle(\boldsymbol{\mathcal{C}}_l - \boldsymbol{\mathcal{S}}^*) \cdot (\mathbf{U}_1, \cdots, \mathbf{U}_m), \boldsymbol{\mathcal{X}}_i\rangle| \leq \frac{1}{8m^2}\|\boldsymbol{\mathcal{C}}_l - \boldsymbol{\mathcal{S}}^*\|_{\mathrm{F}}^2 + \frac{1}{8}\underline{\lambda}^2\|\mathbf{U}_i - \mathbf{V}_i^*\|_{\mathrm{F}}^2 \tag{15.23}$$

The estimation of $|\langle\boldsymbol{\mathcal{X}}_i, \boldsymbol{\mathcal{X}}_j\rangle|(i < j)$ is as follows. From (15.22), we have

$$\begin{aligned}
|\langle\boldsymbol{\mathcal{X}}_i, \boldsymbol{\mathcal{X}}_j\rangle| &= |\langle\mathcal{M}_i(\boldsymbol{\mathcal{S}}^*)\mathbf{M}_{i,j}, (\mathbf{U}_i - \mathbf{V}_i^*)^\top\mathbf{V}_i^*\mathcal{M}_i(\boldsymbol{\mathcal{S}}^*)\rangle| \\
&\leq \bar{\lambda}\|\boldsymbol{\mathcal{S}}^*\|_{\mathrm{F}}\|\mathbf{M}_{i,j}\|\|(\mathbf{U}_i - \mathbf{V}_i^*)^\top\mathbf{V}_i^*\|_{\mathrm{F}} \\
&\leq \bar{\lambda}\|\boldsymbol{\mathcal{S}}^*\|_{\mathrm{F}}\|(\mathbf{U}_i - \mathbf{V}_i^*)^\top\mathbf{V}_i^*\|_{\mathrm{F}}\|(\mathbf{U}_j - \mathbf{V}_j^*)^\top\mathbf{V}_j^*\|_{\mathrm{F}} \\
&\overset{(a)}{\leq} \sqrt{\bar{r}}\bar{\lambda}^2\|\mathbf{U}_i - \mathbf{V}_i^*\|_{\mathrm{F}}^2\|\mathbf{U}_j - \mathbf{V}_j^*\|_{\mathrm{F}}^2 \\
&\overset{(b)}{\leq} \frac{1}{16m^2}\underline{\lambda}^2\|\mathbf{U}_i - \mathbf{V}_i^*\|_{\mathrm{F}}\|\mathbf{U}_j - \mathbf{V}_j^*\|_{\mathrm{F}} \\
&\leq \frac{1}{32m^2}\underline{\lambda}^2\|\mathbf{U}_i - \mathbf{V}_i^*\|_{\mathrm{F}}^2 + \frac{1}{32m^2}\underline{\lambda}^2\|\mathbf{U}_j - \mathbf{V}_j^*\|_{\mathrm{F}}^2
\end{aligned} \tag{15.24}$$

where $\mathbf{M}_{i,j} = \mathbf{I}\otimes\ldots\otimes\mathbf{I}\otimes\mathbf{U}_j^\top(\mathbf{U}_j - \mathbf{V}_j^*)\otimes\mathbf{U}_{j-1}^\top\mathbf{V}_{j-1}^*\otimes\ldots\mathbf{U}_{i+1}^\top\mathbf{V}_{i+1}^*\otimes\mathbf{I}\otimes\ldots\otimes\mathbf{I}$, (a) holds because of (15.21), (b) holds because of (15.22).

As a result of (15.19), (15.23) and (15.24), we have

$$\|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2 \geq 0.5\|\boldsymbol{\mathcal{C}}_l - \boldsymbol{\mathcal{S}}^*\|_{\mathrm{F}}^2 + 0.5\underline{\lambda}^2\sum_{i=1}^m\|\mathbf{U}_i - \mathbf{V}_i^*\|_{\mathrm{F}}^2$$

which finishes the proof of the lemma. □