SUPPLEMENTARY MATERIAL FOR A HYBRID FRAMEWORK COMBINING AUTOREGRESSION AND COMMON FACTORS FOR MATRIX TIME SERIES

Zhiyun Fan[†], Xiaoyu Zhang[‡], and Di Wang[†]

† *Shanghai Jiao Tong University and* ‡ *Tongji University*

Supplementary Material

This is the supplementary material for the main paper. We present detailed proofs of the theorems in the main text. Lemmas, their proofs, along with other technical tools, are also provided. Appendix S1 summarizes the entire procedure for the proposed model. Appendix S2 provides the detailed expressions of the gradients used in the proposed algorithm. Appendix S3 establishes the computational convergence of the algorithm via a deterministic analysis. Appendix S4 derives the statistical convergence rates of the final estimators, along with verification of the RSC/RSS condition and bounding of the deviation term and initialization error via a stochastic analysis. Finally, Appendix S5 establishes the selection consistency of the rank selection method.

S1 Modeling Procedure

First, we provide a summary of the entire procedure for constructing the MARCF model on matrix-valued time series data. This is presented in the following Figure S.1 to offer a clear and step-by-step guidance for the modeling process.

Input:

centered and standardized stationary data $\mathcal{D}$; $\bar{r}_1, \bar{r}_2$; step size η; λ_0, λ_1, λ_2, and b

Shared convex first-stage initialization:

1. Estimate the nuclear-norm penalized VAR model with λ_0 to obtain $\widehat{\mathbf{A}}^{\mathrm{VAR}}$.
2. Make NKP approximation to obtain $\widehat{\mathbf{A}}_1^{\mathrm{KP}}$ and $\widehat{\mathbf{A}}_2^{\mathrm{KP}}$.

Rank selection:

1. Compute the truncated SVD on $\widehat{\mathbf{A}}_1^{\mathrm{KP}}$ and $\widehat{\mathbf{A}}_2^{\mathrm{KP}}$

   with $\bar{r}_1, \bar{r}_2$ to obtain $\widehat{\mathbf{A}}_1^{\mathrm{RR}}(\bar{r}_1)$ and $\widehat{\mathbf{A}}_2^{\mathrm{RR}}(\bar{r}_2)$;
2. Compute the ridge-type ratios in Section 3.3 and select $\widehat{r}_1$, and $\widehat{r}_2$;
3. Truncate and obtain $\widehat{\mathbf{A}}_1^{\mathrm{RR}}(\widehat{r}_1)$ and $\widehat{\mathbf{A}}_2^{\mathrm{RR}}(\widehat{r}_2)$.

Common dimension selection:

1. For every possible pair of (d_1, d_2):
   - initialize $\mathbf{\Theta}^{(0)}$ from $\widehat{\mathbf{A}}_1^{\mathrm{RR}}(\widehat{r}_1)$ and $\widehat{\mathbf{A}}_2^{\mathrm{RR}}(\widehat{r}_2)$;
   - run Algorithm 1 with step size η

     and parameters λ_1, λ_2, and b;
2. Compute BIC and select $\widehat{d}_1, \widehat{d}_2$.

Final estimation:

Initialize $\mathbf{\Theta}^{(0)}$ with $\widehat{r}_1, \widehat{r}_2, \widehat{d}_1, \widehat{d}_2$ and run Algorithm 1.

Output:

estimates $\widehat{\mathbf{\Theta}} = (\widehat{\mathbf{C}}_i, \widehat{\mathbf{R}}_i, \widehat{\mathbf{P}}_i, \widehat{\mathbf{D}}_i)_{i=1}^2$

Figure S.1: Modeling procedure of MARCF model.

S2 Detailed Expressions of Gradients in Algorithm 1

In this appendix, we provide detailed expressions of the gradients in Algorithm 1 which are omitted in the main article.

The objective of Algorithm 1 is to minimize $\overline{\mathcal{L}}$ in (3.11) with known r_1, r_2 and d_1, d_2:

$$\widehat{\mathbf{\Theta}} := \arg\min \overline{\mathcal{L}}(\mathbf{\Theta}; \lambda_1, \lambda_2, b) = \arg\min \left\{ \mathcal{L}(\mathbf{\Theta}) + \frac{\lambda_1}{4}\mathcal{R}_1(\mathbf{\Theta}) + \frac{\lambda_2}{4}\mathcal{R}_2(\mathbf{\Theta}; b) \right\}.$$

To derive the expressions of the gradients, we first define some useful notations. For any matrix $\mathbf{M} \in \mathbb{R}^{p_1 \times p_1}$ and $\mathbf{N} \in \mathbb{R}^{p_2 \times p_2}$, since all entries of $\mathbf{M} \otimes \mathbf{N}$ and $\text{vec}(\mathbf{M})\text{vec}(\mathbf{N})^\top$ are the same up to a permutation, we define the entry-wise re-arrangement operator $\mathcal{P}$ such that

$$\mathcal{P}(\mathbf{M} \otimes \mathbf{N}) = \text{vec}(\mathbf{M})\text{vec}(\mathbf{N})^\top,$$

where the dimensions of the matrices are omitted for simplicity. The inverse operator $\mathcal{P}^{-1}$ is defined such that

$$\mathcal{P}^{-1}(\text{vec}(\mathbf{M})\text{vec}(\mathbf{N})^\top) = \mathbf{M} \otimes \mathbf{N},$$

where we use the notation $\text{vec}(\cdot)$ and $\text{mat}(\cdot)$ to denote the vectorization operator and its inverse matricization operator (with dimensions omitted for brevity), respectively.

The proposed MARCF model is given by

$$\mathbf{Y}_t = \mathbf{A}_1 \mathbf{Y}_{t-1} \mathbf{A}_2^\top + \mathbf{E}_t, \quad t = 1, \dots, T,$$

where $\mathbf{A}_i = [\mathbf{C}_i \ \mathbf{R}_i]\mathbf{D}_i[\mathbf{C}_i \ \mathbf{P}_i]^\top$, for $i = 1, 2$. The matrix time series can be vectorized and the model be rewritten as

$$\mathbf{y}_t = (\mathbf{A}_2 \otimes \mathbf{A}_1)\mathbf{y}_{t-1} + \mathbf{e}_t, \quad t = 1, \dots, T,$$

where $\mathbf{y}_t := \text{vec}(\mathbf{Y}_t)$ and $\mathbf{e}_t := \text{vec}(\mathbf{E}_t)$. We denote $\mathbf{A} := \mathbf{A}_2 \otimes \mathbf{A}_1$ to be the uniquely identifiable parameter matrix, and we also let $\mathbf{Y} := [\mathbf{y}_T, \mathbf{y}_{T-1}, ..., \mathbf{y}_1]$ and $\mathbf{X} := [\mathbf{y}_{T-1}, \mathbf{y}_{T-2}, ..., \mathbf{y}_0]$.

Then, the loss function $\mathcal{L}$ in (3.1) can be viewed as a function of $\mathbf{A}$:

$$\widetilde{\mathcal{L}}(\mathbf{A}) := \mathcal{L}(\mathbf{\Theta}) = \frac{1}{2T}\left\|\mathbf{Y}-\mathbf{A}\mathbf{X}\right\|_{\mathrm{F}}^2 = \frac{1}{2T}\left\|\mathbf{Y}-(\mathbf{A}_2\otimes\mathbf{A}_1)\mathbf{X}\right\|_{\mathrm{F}}^2.$$

First, the gradient of $\widetilde{\mathcal{L}}$ with respect to $\mathbf{A}$ is

$$\nabla\widetilde{\mathcal{L}}(\mathbf{A}) := \frac{\partial\widetilde{\mathcal{L}}}{\partial\mathbf{A}} = -\frac{1}{T}\sum_{t=1}^{T}(\mathbf{y}_t-\mathbf{A}\mathbf{y}_{t-1})\mathbf{y}_{t-1}^\top = -\frac{1}{T}(\mathbf{Y}-\mathbf{A}\mathbf{X})\mathbf{X}^\top.$$

Let $\mathbf{a}_i = \mathrm{vec}(\mathbf{A}_i)$ and $\mathbf{B} := \mathrm{vec}(\mathbf{A}_2)\otimes\mathrm{vec}(\mathbf{A}_1)^\top = \mathbf{a}_2\mathbf{a}_1^\top$. Obviously, $\mathbf{A} = \mathcal{P}^{-1}(\mathbf{B})$ and $\mathbf{B} = \mathcal{P}(\mathbf{A})$. Define $\mathcal{L}_2(\mathbf{B}) = \widetilde{\mathcal{L}}(\mathbf{A})$. Based on the entry-wise matrix permutation operators $\mathcal{P}$ and $\mathcal{P}^{-1}$, we have

$$\begin{aligned}
\nabla\mathcal{L}_2(\mathbf{B}) &= \mathcal{P}(\nabla\widetilde{\mathcal{L}}(\mathbf{A})) = \mathcal{P}(-(\mathbf{Y}-\mathbf{A}\mathbf{X})\mathbf{X}^\top/T),\\
\nabla_{\mathbf{a}_1}\mathcal{L}_2 &= \mathcal{P}(-(\mathbf{Y}-\mathbf{A}\mathbf{X})\mathbf{X}^\top/T)^\top\mathbf{a}_2,\\
\nabla_{\mathbf{a}_2}\mathcal{L}_2 &= \mathcal{P}(-(\mathbf{Y}-\mathbf{A}\mathbf{X})\mathbf{X}^\top/T)\mathbf{a}_1,\\
\nabla_{\mathbf{A}_1}\widetilde{\mathcal{L}} &= \mathrm{mat}(\mathcal{P}(-(\mathbf{Y}-\mathbf{A}\mathbf{X})\mathbf{X}^\top/T)^\top\mathrm{vec}(\mathbf{A}_2))\\
&= \mathrm{mat}(\mathcal{P}(\nabla\widetilde{\mathcal{L}}(\mathbf{A}))^\top\mathrm{vec}(\mathbf{A}_2)),\\
\text{and } \nabla_{\mathbf{A}_2}\widetilde{\mathcal{L}} &= \mathrm{mat}(\mathcal{P}(-(\mathbf{Y}-\mathbf{A}\mathbf{X})\mathbf{X}^\top/T)\mathrm{vec}(\mathbf{A}_1))\\
&= \mathrm{mat}(\mathcal{P}(\nabla\widetilde{\mathcal{L}}(\mathbf{A}))\mathrm{vec}(\mathbf{A}_1)).
\end{aligned}$$

The gradients of $\widetilde{\mathcal{L}}$ with respect to $\mathbf{C}_i$, $\mathbf{R}_i$, $\mathbf{P}_i$, and $\mathbf{D}_i$ are:

$$\begin{aligned}
\nabla_{\mathbf{C}_i}\widetilde{\mathcal{L}} &= \nabla_{\mathbf{A}_i}\widetilde{\mathcal{L}}(\mathbf{C}_i\mathbf{D}_{i,11}^\top+\mathbf{P}_i\mathbf{D}_{i,12}^\top) + \nabla_{\mathbf{A}_i}\widetilde{\mathcal{L}}^\top(\mathbf{C}_i\mathbf{D}_{i,11}+\mathbf{R}_i\mathbf{D}_{i,21}),\\
\nabla_{\mathbf{R}_i}\widetilde{\mathcal{L}} &= \nabla_{\mathbf{A}_i}\widetilde{\mathcal{L}}[\mathbf{C}_i\ \mathbf{P}_i][\mathbf{D}_{i,21}\ \mathbf{D}_{i,22}]^\top,\\
\nabla_{\mathbf{P}_i}\widetilde{\mathcal{L}} &= \nabla_{\mathbf{A}_i}\widetilde{\mathcal{L}}^\top[\mathbf{C}_i\ \mathbf{R}_i][\mathbf{D}_{i,12}^\top\ \mathbf{D}_{i,22}^\top]^\top,\\
\text{and } \nabla_{\mathbf{D}_i}\widetilde{\mathcal{L}} &= [\mathbf{C}_i\ \mathbf{R}_i]^\top\nabla_{\mathbf{A}_i}\widetilde{\mathcal{L}}[\mathbf{C}_i\ \mathbf{P}_i].
\end{aligned}$$

Denote the following block structure of $\mathbf{D}_i$ as

$$\mathbf{D}_i = \begin{pmatrix}\mathbf{D}_{i,11} & \mathbf{D}_{i,12}\\ \mathbf{D}_{i,21} & \mathbf{D}_{i,22}\end{pmatrix},$$

where $\mathbf{D}_{i,11} \in \mathbb{R}^{d_i \times d_i}, \mathbf{D}_{i,12} \in \mathbb{R}^{d_i \times (r_i - d_i)}, \mathbf{D}_{i,21} \in \mathbb{R}^{(r_i-d_i) \times d_i}, \mathbf{D}_{i,22} \in \mathbb{R}^{(r_i-d_i)\times(r_i-d_i)}$ are block matrices in $\mathbf{D}_i$, for $i = 1, 2$. Finally, combined with the partial gradients of the regularization terms, the partial gradients of $\overline{\mathcal{L}}$ are given by:

$$
\begin{aligned}
\nabla_{\mathbf{C}_i}\overline{\mathcal{L}} =& \nabla_{\mathbf{A}_i}\widetilde{\mathcal{L}}(\mathbf{C}_i\mathbf{D}_{i,11}^\top + \mathbf{P}_i\mathbf{D}_{i,12}^\top) + \nabla_{\mathbf{A}_i}\widetilde{\mathcal{L}}^\top(\mathbf{C}_i\mathbf{D}_{i,11} + \mathbf{R}_i\mathbf{D}_{i,21}) \\
&+ (-1)^{i+1}\lambda_1\left(\|\mathbf{A}_1\|_\mathrm{F}^2 - \|\mathbf{A}_2\|_\mathrm{F}^2\right)\left(\mathbf{A}_i[\mathbf{C}_i\ \mathbf{P}_i][\mathbf{D}_{i,11}\ \mathbf{D}_{i,12}]^\top + \mathbf{A}_i^\top[\mathbf{C}_i\ \mathbf{R}_i][\mathbf{D}_{i,11}^\top\ \mathbf{D}_{i,21}^\top]^\top\right) \\
&+ \lambda_2\left(2\mathbf{C}_i(\mathbf{C}_i^\top\mathbf{C}_i - b^2\mathbf{I}_{d_i}) + \mathbf{R_i}\mathbf{R}_i^\top\mathbf{C}_i + \mathbf{P_i}\mathbf{P}_i^\top\mathbf{C}_i\right), \\
\nabla_{\mathbf{R}_i}\overline{\mathcal{L}} =& \nabla_{\mathbf{A}_i}\widetilde{\mathcal{L}}[\mathbf{C}_i\ \mathbf{P}_i][\mathbf{D}_{i,21}\ \mathbf{D}_{i,22}]^\top + (-1)^{i+1}\lambda_1\left(\|\mathbf{A}_1\|_\mathrm{F}^2 - \|\mathbf{A}_2\|_\mathrm{F}^2\right)\mathbf{A}_i[\mathbf{C}_i\ \mathbf{P}_i][\mathbf{D}_{i,21}\ \mathbf{D}_{i,22}]^\top \\
&+ \lambda_2\left(\mathbf{R}_i(\mathbf{R}_i^\top\mathbf{R}_i - b^2\mathbf{I}_{r_i-d_i}) + \mathbf{C}_i\mathbf{C}_i^\top\mathbf{R}_i\right), \\
\nabla_{\mathbf{P}_i}\overline{\mathcal{L}} =& \nabla_{\mathbf{A}_i}\widetilde{\mathcal{L}}^\top[\mathbf{C}_i\ \mathbf{R}_i][\mathbf{D}_{i,12}^\top\ \mathbf{D}_{i,22}^\top]^\top + (-1)^{i+1}\lambda_1\left(\|\mathbf{A}_1\|_\mathrm{F}^2 - \|\mathbf{A}_2\|_\mathrm{F}^2\right)\mathbf{A}_i^\top[\mathbf{C}_i\ \mathbf{R}_i][\mathbf{D}_{i,12}^\top\ \mathbf{D}_{i,22}^\top]^\top \\
&+ \lambda_2\left(\mathbf{P}_i(\mathbf{P}_i^\top\mathbf{P}_i - b^2\mathbf{I}_{r_i-d_i}) + \mathbf{C_i}\mathbf{C}_i^\top\mathbf{P}_i\right), \\
\nabla_{\mathbf{D}_i}\overline{\mathcal{L}} =& [\mathbf{C}_i\ \mathbf{R}_i]^\top\nabla_{\mathbf{A}_i}\widetilde{\mathcal{L}}[\mathbf{C}_i\ \mathbf{P}_i] + (-1)^{i+1}\lambda_1\left(\|\mathbf{A}_1\|_\mathrm{F}^2 - \|\mathbf{A}_2\|_\mathrm{F}^2\right)[\mathbf{C}_i\ \mathbf{R}_i]^\top\mathbf{A}_i[\mathbf{C}_i\ \mathbf{P}_i].
\end{aligned}
$$

S3 Computational Convergence Analysis

In this appendix, we present the proof of Theorem 1, i.e., the computational convergence of regularized gradient descent algorithm proposed in the main article. To make it easy to read, it is divided into several steps. Auxiliary lemmas and their proofs are presented in Appendix S3.2.

S3.1 Proof of Theorem 1

Step 1. **(Notations, conditions, and proof outline)**

We begin by introducing some important notations and conditions required for the convergence analysis. Other notations not mentioned in this step are inherited from Appendix S2. We then provide an outline of the proof, highlighting the key intermediate results and main ideas.

To begin, we restate the definitions of the measures quantifying the estimation error and the statistical error. Throughout this article, the true values of all parameters are defined by letters with an asterisk superscript, i.e., $\{\mathbf{C}_i^*, \mathbf{R}_i^*, \mathbf{P}_i^*, \mathbf{D}_i^*, \mathbf{A}_i^*\}_{i=1}^2$. Let $\mathbb{O}^{n_1\times n_2}$ be the set of $n_1 \times n_2$ matrices with orthonormal columns. When $n_1 = n_2 = n$, we use $\mathbb{O}^n$ for short. For the j-th iterate, we quantify the combined estimation errors up to optimal rotations defined in (2.9) as

$$\begin{aligned}\operatorname{dist}(\boldsymbol{\Theta}^{(j)}, \boldsymbol{\Theta}^*)^2 = \min_{\substack{\mathbf{O}_{i,r},\mathbf{O}_{i,p}\in\mathbb{O}^{r_i-d_i}\\ \mathbf{O}_{i,c}\in\mathbb{O}^{d_i}}} \sum_{i=1,2} \Big\{ &\|\mathbf{C}_i^{(j)} - \mathbf{C}_i^*\mathbf{O}_{i,c}\|_{\mathrm{F}}^2 + \|\mathbf{R}_i^{(j)} - \mathbf{R}_i^*\mathbf{O}_{i,r}\|_{\mathrm{F}}^2 \\ &+\|\mathbf{P}_i^{(j)} - \mathbf{P}_i^*\mathbf{O}_{i,p}\|_{\mathrm{F}}^2 + \|\mathbf{D}_i^{(j)} - \operatorname{diag}(\mathbf{O}_{i,c}, \mathbf{O}_{i,r})^\top \mathbf{D}_i^* \operatorname{diag}(\mathbf{O}_{i,c}, \mathbf{O}_{i,p})\|_{\mathrm{F}}^2 \Big\},\end{aligned} \tag{S3.1}$$

and the corresponding optimal rotations as $\mathbf{O}_{i,c}^{(j)}, \mathbf{O}_{i,r}^{(j)}, \mathbf{O}_{i,p}^{(j)}$, for $i = 1, 2$. For simplicity, we let $\mathbf{O}_{i,u} := \operatorname{diag}(\mathbf{O}_{i,c}, \mathbf{O}_{i,r})$, $\mathbf{O}_{i,v} := \operatorname{diag}(\mathbf{O}_{i,c}, \mathbf{O}_{i,p})$, and use $\operatorname{dist}_{(j)}^2$ to represent $\operatorname{dist}(\boldsymbol{\Theta}^{(j)}, \boldsymbol{\Theta}^*)^2$.

In addition, as defined in Definition 2, the statistical error is quantified by

$$\xi(r_1, r_2, d_1, d_2) := \sup_{\substack{[\mathbf{C}_i\ \mathbf{R}_i] \in \mathbb{O}^{p_i \times r_i}, \\ [\mathbf{C}_i\ \mathbf{P}_i] \in \mathbb{O}^{p_i \times r_i}, \\ \mathbf{D}_i \in \mathbb{R}^{r_i \times r_i}, \|\mathbf{D}_i\|_{\mathrm{F}}=1}} \left\langle \nabla \widetilde{\mathcal{L}}(\mathbf{A}^*), [\mathbf{C}_2\ \mathbf{R}_2]\mathbf{D}_2[\mathbf{C}_2\ \mathbf{P}_2]^\top \otimes [\mathbf{C}_1\ \mathbf{R}_1]\mathbf{D}_1[\mathbf{C}_1\ \mathbf{P}_1]^\top \right\rangle,$$

where $\mathbf{A}^* = \mathbf{A}_2^* \otimes \mathbf{A}_1^*$ and $\widetilde{\mathcal{L}}(\mathbf{A})$ is referred to as the least-squares loss function with respect to the identifiable Kronecker product type parameter $\mathbf{A} = \mathbf{A}_2 \otimes \mathbf{A}_1$.

Next, we state three conditions required for convergence analysis. The first condition stipulates that $\widetilde{\mathcal{L}}(\mathbf{A})$ satisfies the RSC/RSS condition defined in Definition 1. Specifically, we assume there exist constants $0 < \alpha \leq \beta$ such that for any pair of identifiable parameter matrices $\mathbf{A} = \mathbf{A}_2 \otimes \mathbf{A}_1$ and $\mathbf{A}' = \mathbf{A}_2' \otimes \mathbf{A}_1'$ admitting the MARCF structural decomposition, where $\mathbf{A}_i = [\mathbf{C}_i\ \mathbf{R}_i]\mathbf{D}_i[\mathbf{C}_i\ \mathbf{P}_i]^\top$ and $\mathbf{A}_i' = [\mathbf{C}_i'\ \mathbf{R}_i']\mathbf{D}_i'[\mathbf{C}_i'\ \mathbf{P}_i']^\top$, the following inequalities hold:

$$\begin{aligned} &(\text{RSC}) \quad \frac{\alpha}{2}\|\mathbf{A} - \mathbf{A}'\|_{\mathrm{F}}^2 \leq \mathcal{L}(\mathbf{A}) - \mathcal{L}(\mathbf{A}') - \langle \nabla \mathcal{L}(\mathbf{A}'), \mathbf{A} - \mathbf{A}' \rangle, \\ &(\text{RSS}) \quad \mathcal{L}(\mathbf{A}) - \mathcal{L}(\mathbf{A}') - \langle \nabla \mathcal{L}(\mathbf{A}'), \mathbf{A} - \mathbf{A}' \rangle \leq \frac{\beta}{2}\|\mathbf{A} - \mathbf{A}'\|_{\mathrm{F}}^2, \end{aligned}$$

The condition is a premise of Theorem 1. As in Nesterov (2004), it implies that

$$\mathcal{L}(\mathbf{A}') - \mathcal{L}(\mathbf{A}) \geq \langle \nabla \mathcal{L}(\mathbf{A}), \mathbf{A}' - \mathbf{A} \rangle + \frac{1}{2\beta}\|\nabla \mathcal{L}(\mathbf{A}') - \nabla \mathcal{L}(\mathbf{A})\|_{\mathrm{F}}^2.$$

Combining this inequality with the RSC condition, we have

$$\langle \nabla \mathcal{L}(\mathbf{A}) - \nabla \mathcal{L}(\mathbf{A}'), \mathbf{A} - \mathbf{A}' \rangle \geq \frac{\alpha}{2}\|\mathbf{A} - \mathbf{A}'\|_{\mathrm{F}}^2 + \frac{1}{2\beta}\|\nabla \mathcal{L}(\mathbf{A}) - \nabla \mathcal{L}(\mathbf{A}')\|_{\mathrm{F}}^2, \tag{S3.2}$$

which is equivalent to the restricted correlated gradient (RCG) condition in Han et al. (2022).

The second and the third conditions ensure that the estimates remain within a small neighborhood of the true values during all iterations. They are established inductively in the proof. Let $\underline{\sigma} := \min\{\sigma_{r_1}(\mathbf{A}_1^*), \sigma_{r_2}(\mathbf{A}_2^*)\}$ be the smallest value among all the non-zero singular values of $\mathbf{A}_1^*$ and $\mathbf{A}_2^*$. Define $\phi := \|\mathbf{A}_1^*\|_{\mathrm{F}} = \|\mathbf{A}_2^*\|_{\mathrm{F}}$ and $\kappa := \phi/\underline{\sigma}$. With the notations, the second condition is specified as

$$\operatorname{dist}_{(j)}^2 \le \frac{C_D \alpha \phi^{2/3}}{\beta \kappa^2}, \tag{S3.3}$$

and then

$$\operatorname{dist}_{(j)}^2 \le \frac{C_D \phi^{2/3}}{\kappa^2} \le C_D \phi^{2/3}, \text{ for all } j = 0, 1, 2, \ldots$$

where C_D is a positive constant small enough to ensure local convergence.

For the decomposition of $\mathbf{A}_i^*$, we require $[\mathbf{C}_i^* \ \mathbf{R}_i^*]^\top [\mathbf{C}_i^* \ \mathbf{R}_i^*] = b^2 \mathbf{I}_{r_i}$, $[\mathbf{C}_i^* \ \mathbf{P}_i^*]^\top [\mathbf{C}_i^* \ \mathbf{P}_i^*] = b^2 \mathbf{I}_{r_i}$, for $i = 1, 2$. For simplicity, we consider $b = \phi^{1/3}$, though our proof can be readily extended to the case where $b \asymp \phi^{1/3}$.

The third condition is:

$$\begin{gathered} \left\| \left[\mathbf{C}_i^{(j)} \ \mathbf{R}_i^{(j)} \right] \right\|_{\mathrm{op}} \le (1 + c_b) b, \quad \left\| \left[\mathbf{C}_i^{(j)} \ \mathbf{P}_i^{(j)} \right] \right\|_{\mathrm{op}} \le (1 + c_b) b, \\ \left\| \mathbf{D}_i^{(j)} \right\|_{\mathrm{F}} \le \frac{(1 + c_b)\phi}{b^2}, \quad \forall i = 1, 2, \text{ and } j = 0, 1, 2, \ldots \end{gathered} \tag{S3.4}$$

By sub-multiplicativity, $\left\| \mathbf{A}_1^{(j)} \right\|_{\mathrm{F}} \le (1 + c_a)\phi$ and $\left\| \mathbf{A}_2^{(j)} \right\|_{\mathrm{F}} \le (1 + c_a)\phi$ with c_b and c_a being two positive constants. We assume that $c_b \le 0.01$ and then $c_a \le 0.04$. In fact, the constant 0.01 reflects the accuracy of the initial estimate, which can be replaced by any small positive numbers.

Finally, we give an outline of the proof of Theorem 1, which proceeds by induction. Assuming that (S3.3) and (S3.4) hold for $\mathbf{\Theta}^{(j)}$, based on RSC and RSS conditions, we derive a recursive relationship between $\operatorname{dist}_{(j+1)}^2$ and $\operatorname{dist}_{(j)}^2$ with an extra statistical error term. Then, we prove that (S3.3) and (S3.4) hold for $\mathbf{\Theta}^{(j+1)}$ and verify the conditions for $\mathbf{\Theta}^{(0)}$, thereby finishing the inductive argument. Finally, built on the recursive relationship, we establish the three upper bounds presented in Theorem 1.

Specifically, in Steps 2-4, we focus on the update at $(j+1)$-th iterate to establish a recursive inequality between $\operatorname{dist}_{(j+1)}^2$ and $\operatorname{dist}_{(j)}^2$, given (S3.2), (S3.3), and (S3.4). In Step

2, for $\mathbf{R}_1$, we use the rule $\mathbf{R}_1^{(j+1)} = \mathbf{R}_1^{(j)} - \eta \nabla_{\mathbf{R}_1} \overline{\mathcal{L}}$ and upper bound the estimation error of $\mathbf{R}^{(j+1)}$ as

$$\min_{\mathbf{O}_{1,r} \in \mathbb{O}^{r_1 - d_1}} \left\| \mathbf{R}_1^{(j+1)} - \mathbf{R}_1^* \mathbf{O}_{1,r} \right\|_{\mathrm{F}}^2 \leq \min_{\mathbf{O}_{1,r} \in \mathbb{O}^{r_1 - d_1}} \left\| \mathbf{R}_1^{(j)} - \mathbf{R}_1^* \mathbf{O}_{1,r} \right\|_{\mathrm{F}}^2$$
$$- 2\eta Q_{\mathbf{R}_1,1} - 2\lambda_1 \eta G_{\mathbf{R}_1} - 2\lambda_2 \eta T_{\mathbf{R}_1} + \eta^2 Q_{\mathbf{R}_1,2}.$$

Similarly, we do the same for $\mathbf{P}_i^{(j)}, \mathbf{C}_i^{(j)}$ and $\mathbf{D}_i^{(j)}$ and sum them up to obtain the following first-stage upper bound, which is an informal version of (S3.11).

$$\mathrm{dist}_{(j+1)}^2 \leq \mathrm{dist}_{(j)}^2 + \eta^2 Q_2 - 2\eta Q_1 - 2\lambda_1 \eta G - 2\lambda_2 \eta T.$$

In Steps 3.1-3.4, we give further the lower bounds for Q_1, G, and T, whose coefficients are -2η, and the upper bound for Q_2, whose coefficient is η^2. They lead to an intermediate upper bound in (S3.17), whose right hand side is negatively correlated to the estimation error of $\mathbf{A}_2^{(j)} \otimes \mathbf{A}_1^{(j)}$. Thus, in Step 3.5, we construct a lower bound with respect to $\mathrm{dist}_{(j)}^2$ and the regularization terms, $\mathcal{R}_1(\mathbf{\Theta}^{(j)})$ and $\mathcal{R}_2(\mathbf{\Theta}^{(j)})$. Plugging it into the intermediate bound, we have the second-stage upper bound of $\mathrm{dist}_{(j+1)}^2$ as follows, which is an informal version of (S3.22)

$$\mathrm{dist}_{(j+1)}^2 \leq (1 - 2\eta\rho_1)\,\mathrm{dist}_{(j)}^2 + \rho_2 \xi^2$$
$$+ \rho_3 \left\| \nabla \widetilde{\mathcal{L}}(\mathbf{A}^{(j)}) - \nabla \widetilde{\mathcal{L}}(\mathbf{A}^*) \right\|_{\mathrm{F}}^2 + \rho_4 \mathcal{R}_2\left(\mathbf{\Theta}^{(j)}\right) + \rho_5 \mathcal{R}_1\left(\mathbf{\Theta}^{(j)}\right).$$

In the above inequality, $\{\rho_i\}_{i=1}^5$ are coefficients related to η, λ_1, λ_2, α, β, ϕ, and κ.

Next, in Step 4, we impose some sufficient conditions on η, λ_1, λ_2, α, and β for algorithm convergence. Then, we obtain the final recursive relationship (S3.23):

$$\mathrm{dist}_{(j+1)}^2 \leq (1 - C_0 \eta_0 \alpha \beta^{-1} \kappa^{-2})\mathrm{dist}_{(j)}^2 + C \eta_0 \kappa^2 \alpha^{-1} \beta^{-1} \phi^{-10/3} \xi^2.$$

To complete the inductive argument, in Step 5, we first verify the conditions (S3.3) and (S3.4) for $\mathbf{\Theta}^{(0)}$. Given that (S3.3) is an assumption of Theorem 1, it suffices to verify (S3.4).

Then, we prove that both conditions hold for $\boldsymbol{\Theta}^{(j+1)}$ and complete the induction. Finally, in Step 6, we establish the three upper bounds presented in Theorem 1 based on the recursive relationship and the approximate equivalence of the three metrics, thereby completing the whole proof.

Step 2. (**Upper bound of** $\mathrm{dist}^2_{(j+1)} - \mathrm{dist}^2_{(j)}$)

By definition,

$$
\begin{aligned}
&\mathrm{dist}^2_{(j+1)} \\
&= \sum_{i=1,2} \Big\{ \|\mathbf{R}_i^{(j+1)} - \mathbf{R}_i^* \mathbf{O}_{i,r}^{(j+1)}\|_{\mathrm{F}}^2 + \|\mathbf{P}_i^{(j+1)} - \mathbf{P}_i^* \mathbf{O}_{i,p}^{(j+1)}\|_{\mathrm{F}}^2 + \|\mathbf{C}_i^{(j+1)} - \mathbf{C}_i^* \mathbf{O}_{i,c}^{(j+1)}\|_{\mathrm{F}}^2 \\
&\quad + \|\mathbf{D}_i^{(j+1)} - \mathbf{O}_{i,u}^{(j+1)\top} \mathbf{D}_i^* \mathbf{O}_{i,v}^{(j+1)}\|_{\mathrm{F}}^2 \Big\} \\
&\le \sum_{i=1,2} \Big\{ \|\mathbf{R}_i^{(j+1)} - \mathbf{R}_i^* \mathbf{O}_{i,r}^{(j)}\|_{\mathrm{F}}^2 + \|\mathbf{P}_i^{(j+1)} - \mathbf{P}_i^* \mathbf{O}_{i,p}^{(j)}\|_{\mathrm{F}}^2 + \|\mathbf{C}_i^{(j+1)} - \mathbf{C}_i^* \mathbf{O}_{i,c}^{(j)}\|_{\mathrm{F}}^2 \\
&\quad + \|\mathbf{D}_i^{(j+1)} - \mathbf{O}_{i,u}^{(j)\top} \mathbf{D}_i^* \mathbf{O}_{i,v}^{(j)}\|_{\mathrm{F}}^2 \Big\}.
\end{aligned}
$$

In the following substeps, we derive upper bounds for the errors of $\mathbf{C}, \mathbf{R}, \mathbf{P}$ and $\mathbf{D}$ separately. Then, we combine them to give a first-stage upper bound (S3.11).

Step 2.1 (Upper bounds for the errors of $\mathbf{R}_i$ and $\mathbf{P}_i$)

By definition, we have

$$
\begin{aligned}
&\|\mathbf{R}_1^{(j+1)} - \mathbf{R}_1^*\mathbf{O}_{1,r}^{(j)}\|_{\mathrm{F}}^2 \\
=&\Big\|\mathbf{R}_1^{(j)} - \mathbf{R}_1^*\mathbf{O}_{1,r}^{(j)} - \eta\Big(\nabla_{\mathbf{A}_1}\widetilde{\mathcal{L}}(\mathbf{A}^{(j)})[\mathbf{C}_1^{(j)}\ \mathbf{P}_1^{(j)}][\mathbf{D}_{1,21}^{(j)}\ \mathbf{D}_{1,22}^{(j)}]^\top \\
&+\lambda_1\left(\left\|\mathbf{A}_1^{(j)}\right\|_{\mathrm{F}}^2 - \left\|\mathbf{A}_2^{(j)}\right\|_{\mathrm{F}}^2\right)\mathbf{A}_1^{(j)}[\mathbf{C}_1^{(j)}\ \mathbf{P}_1^{(j)}][\mathbf{D}_{1,21}^{(j)}\ \mathbf{D}_{1,22}^{(j)}]^\top \\
&+\lambda_2\mathbf{R}_1^{(j)}(\mathbf{R}_1^{(j)\top}\mathbf{R}_1^{(j)} - b^2\mathbf{I}_{r_1-d_1}) + \lambda_2\mathbf{C}_1^{(j)}\mathbf{C}_1^{(j)\top}\mathbf{R}_1^{(j)}\Big)\Big\|_{\mathrm{F}}^2 \\
=&\|\mathbf{R}_1^{(j)} - \mathbf{R}_1^*\mathbf{O}_{1,r}^{(j)}\|_{\mathrm{F}}^2 + \eta^2\Big\|\nabla_{\mathbf{A}_1}\widetilde{\mathcal{L}}(\mathbf{A}^{(j)})[\mathbf{C}_1^{(j)}\ \mathbf{P}_1^{(j)}][\mathbf{D}_{1,21}^{(j)}\ \mathbf{D}_{1,22}^{(j)}]^\top \\
&+\lambda_1\left(\left\|\mathbf{A}_1^{(j)}\right\|_{\mathrm{F}}^2 - \left\|\mathbf{A}_2^{(j)}\right\|_{\mathrm{F}}^2\right)\mathbf{A}_1^{(j)}[\mathbf{C}_1^{(j)}\ \mathbf{P}_1^{(j)}][\mathbf{D}_{1,21}^{(j)}\ \mathbf{D}_{1,22}^{(j)}]^\top \\
&+\lambda_2\mathbf{R}_1^{(j)}(\mathbf{R}_1^{(j)\top}\mathbf{R}_1^{(j)} - b^2\mathbf{I}_{r_1-d_1}) + \lambda_2\mathbf{C}_1^{(j)}\mathbf{C}_1^{(j)\top}\mathbf{R}_1^{(j)}\Big\|_{\mathrm{F}}^2 \\
&-2\eta\Big\langle\mathbf{R}_1^{(j)} - \mathbf{R}_1^*\mathbf{O}_{1,r}^{(j)}, \nabla_{\mathbf{A}_1}\widetilde{\mathcal{L}}(\mathbf{A}^{(j)})[\mathbf{C}_1^{(j)}\ \mathbf{P}_1^{(j)}][\mathbf{D}_{1,21}^{(j)}\ \mathbf{D}_{1,22}^{(j)}]^\top\Big\rangle \\
&-2\lambda_1\eta\left\langle\mathbf{R}_1^{(j)} - \mathbf{R}_1^*\mathbf{O}_{1,r}^{(j)}, \left(\left\|\mathbf{A}_1^{(j)}\right\|_{\mathrm{F}}^2 - \left\|\mathbf{A}_2^{(j)}\right\|_{\mathrm{F}}^2\right)\mathbf{A}_1^{(j)}[\mathbf{C}_1^{(j)}\ \mathbf{P}_1^{(j)}][\mathbf{D}_{1,21}^{(j)}\ \mathbf{D}_{1,22}^{(j)}]^\top\right\rangle \\
&-2\lambda_2\eta\Big\langle\mathbf{R}_1^{(j)} - \mathbf{R}_1^*\mathbf{O}_{1,r}^{(j)}, \mathbf{R}_1^{(j)}(\mathbf{R}_1^{(j)\top}\mathbf{R}_1^{(j)} - b^2\mathbf{I}_{r_1-d_1})\Big\rangle \\
&-2\lambda_2\eta\Big\langle\mathbf{R}_1^{(j)} - \mathbf{R}_1^*\mathbf{O}_{1,r}^{(j)}, \mathbf{C}_1^{(j)}\mathbf{C}_1^{(j)\top}\mathbf{R}_1^{(j)}\Big\rangle \\
:=&\left\|\mathbf{R}_1^{(j)} - \mathbf{R}_1^*\mathbf{O}_{1,r}^{(j)}\right\|_{\mathrm{F}}^2 + \eta^2 I_{\mathbf{R}_1,2} - 2\eta I_{\mathbf{R}_1,1}.
\end{aligned}
\tag{S3.5}
$$

For $I_{\mathbf{R}_1,2}$ (the second term in (S3.5)), by Cauchy's inequality,

$$
\begin{aligned}
I_{\mathbf{R}_1,2} =&\Big\|\nabla_{\mathbf{A}_1}\widetilde{\mathcal{L}}(\mathbf{A}^{(j)})[\mathbf{C}_1^{(j)}\ \mathbf{P}_1^{(j)}][\mathbf{D}_{1,21}^{(j)}\ \mathbf{D}_{1,22}^{(j)}]^\top \\
&+\lambda_1\left(\left\|\mathbf{A}_1^{(j)}\right\|_{\mathrm{F}}^2 - \left\|\mathbf{A}_2^{(j)}\right\|_{\mathrm{F}}^2\right)\mathbf{A}_1^{(j)}[\mathbf{C}_1^{(j)}\ \mathbf{P}_1^{(j)}][\mathbf{D}_{1,21}^{(j)}\ \mathbf{D}_{1,22}^{(j)}]^\top \\
&+\lambda_2\mathbf{R}_1^{(j)}(\mathbf{R}_1^{(j)\top}\mathbf{R}_1^{(j)} - b^2\mathbf{I}_{r_1-d_1}) + \lambda_2\mathbf{C}_1^{(j)}\mathbf{C}_1^{(j)\top}\mathbf{R}_1^{(j)}\Big\|_{\mathrm{F}}^2 \\
\leq&4\|\nabla_{\mathbf{A}_1}\widetilde{\mathcal{L}}(\mathbf{A}^{(j)})[\mathbf{C}_1^{(j)}\ \mathbf{P}_1^{(j)}][\mathbf{D}_{1,21}^{(j)}\ \mathbf{D}_{1,22}^{(j)}]^\top\|_{\mathrm{F}}^2 \\
&+4\lambda_1^2\left(\left\|\mathbf{A}_1^{(j)}\right\|_{\mathrm{F}}^2 - \left\|\mathbf{A}_2^{(j)}\right\|_{\mathrm{F}}^2\right)^2\left\|\mathbf{A}_1^{(j)}[\mathbf{C}_1^{(j)}\ \mathbf{P}_1^{(j)}][\mathbf{D}_{1,21}^{(j)}\ \mathbf{D}_{1,22}^{(j)}]^\top\right\|_{\mathrm{F}}^2 \\
&+4\lambda_2^2\|\mathbf{R}_1^{(j)}(\mathbf{R}_1^{(j)\top}\mathbf{R}_1^{(j)} - b^2\mathbf{I}_{r_1-d_1})\|_{\mathrm{F}}^2 + 4\lambda_2^2\|\mathbf{C}_1^{(j)}\mathbf{C}_1^{(j)\top}\mathbf{R}_1^{(j)}\|_{\mathrm{F}}^2,
\end{aligned}
$$

where the first term in the RHS of $I_{\mathbf{R}_1,2}$ can be bounded by

$$\begin{aligned}
&\|\nabla_{\mathbf{A}_1}\widetilde{\mathcal{L}}(\mathbf{A}^{(j)})[\mathbf{C}_1^{(j)}\ \mathbf{P}_1^{(j)}][\mathbf{D}_{1,21}^{(j)}\ \mathbf{D}_{1,22}^{(j)}]^\top\|_{\mathrm{F}}^2\\
=&\|\mathrm{mat}(\mathcal{P}(\nabla\widetilde{\mathcal{L}}(\mathbf{A}^{(j)}))^\top\mathrm{vec}(\mathbf{A}_2^{(j)}))[\mathbf{C}_1^{(j)}\ \mathbf{P}_1^{(j)}][\mathbf{D}_{1,21}^{(j)}\ \mathbf{D}_{1,22}^{(j)}]^\top\|_{\mathrm{F}}^2\\
\leq&2\|\mathrm{mat}(\mathcal{P}(\nabla\widetilde{\mathcal{L}}(\mathbf{A}^*))^\top\mathrm{vec}(\mathbf{A}_2^{(j)}))[\mathbf{C}_1^{(j)}\ \mathbf{P}_1^{(j)}][\mathbf{D}_{1,21}^{(j)}\ \mathbf{D}_{1,22}^{(j)}]^\top\|_{\mathrm{F}}^2\\
&+2\|\mathrm{mat}(\mathcal{P}(\nabla\widetilde{\mathcal{L}}(\mathbf{A}^{(j)})-\nabla\widetilde{\mathcal{L}}(\mathbf{A}^*))^\top\mathrm{vec}(\mathbf{A}_2^{(j)}))[\mathbf{C}_1^{(j)}\ \mathbf{P}_1^{(j)}][\mathbf{D}_{1,21}^{(j)}\ \mathbf{D}_{1,22}^{(j)}]^\top\|_{\mathrm{F}}^2.
\end{aligned}$$

By duality of the Frobenius norm and definition of ξ, we have

$$\begin{aligned}
&\left\|\mathrm{mat}(\mathcal{P}(\nabla\widetilde{\mathcal{L}}(\mathbf{A}^*))^\top\mathrm{vec}(\mathbf{A}_2^{(j)}))[\mathbf{C}_1^{(j)}\ \mathbf{P}_1^{(j)}][\mathbf{D}_{1,21}^{(j)}\ \mathbf{D}_{1,22}^{(j)}]^\top\right\|_{\mathrm{F}}\\
=&\sup_{\|\mathbf{W}\|_{\mathrm{F}}=1}\left\langle\nabla\widetilde{\mathcal{L}}(\mathbf{A}^*),\mathbf{A}_2^{(j)}\otimes[\mathbf{C}_1^{(j)}\ \mathbf{W}]\begin{bmatrix}\mathbf{0}&\mathbf{0}\\\mathbf{D}_{1,21}^{(j)}&\mathbf{D}_{1,22}^{(j)}\end{bmatrix}[\mathbf{C}_1^{(j)}\ \mathbf{P}_1^{(j)}]^\top\right\rangle\\
\leq&\|\mathbf{A}_2^{(j)}\|_{\mathrm{F}}\cdot\|[\mathbf{D}_{1,21}^{(j)}\ \mathbf{D}_{1,22}^{(j)}]\|_{\mathrm{op}}\cdot\|[\mathbf{C}_1^{(j)}\ \mathbf{P}_1^{(j)}]\|_{\mathrm{op}}\cdot\xi(r_1,r_2,d_1,d_2).
\end{aligned}$$

Thus, the first term can be bounded by

$$\begin{aligned}
&\|\nabla_{\mathbf{A}_1}\widetilde{\mathcal{L}}(\mathbf{A}^{(j)})[\mathbf{C}_1^{(j)}\ \mathbf{P}_1^{(j)}][\mathbf{D}_{1,21}^{(j)}\ \mathbf{D}_{1,22}^{(j)}]^\top\|_{\mathrm{F}}^2\\
\leq&2\|\mathrm{mat}(\mathcal{P}(\nabla\widetilde{\mathcal{L}}(\mathbf{A}^*))^\top\mathrm{vec}(\mathbf{A}_2^{(j)}))[\mathbf{C}_1^{(j)}\ \mathbf{P}_1^{(j)}][\mathbf{D}_{1,21}^{(j)}\ \mathbf{D}_{1,22}^{(j)}]^\top\|_{\mathrm{F}}^2\\
&+2\|\mathrm{mat}(\mathcal{P}(\nabla\widetilde{\mathcal{L}}(\mathbf{A}^{(j)})-\nabla\widetilde{\mathcal{L}}(\mathbf{A}^*))^\top\mathrm{vec}(\mathbf{A}_2^{(j)}))\|_{\mathrm{F}}^2\cdot\|[\mathbf{C}_1^{(j)}\ \mathbf{P}_1^{(j)}]\|_{\mathrm{op}}^2\cdot\|[\mathbf{D}_{1,21}^{(j)}\ \mathbf{D}_{1,22}^{(j)}]\|_{\mathrm{F}}^2\\
\leq&4b^{-2}\phi^4\cdot[\xi^2(r_1,r_2,d_1,d_2)+\|\nabla\widetilde{\mathcal{L}}(\mathbf{A}^{(j)})-\nabla\widetilde{\mathcal{L}}(\mathbf{A}^*)\|_{\mathrm{F}}^2].
\end{aligned}$$

The second term can be bounded by

$$\begin{aligned}
&\lambda_1^2\left(\left\|\mathbf{A}_1^{(j)}\right\|_{\mathrm{F}}^2-\left\|\mathbf{A}_2^{(j)}\right\|_{\mathrm{F}}^2\right)^2\left\|\mathbf{A}_1^{(j)}[\mathbf{C}_1^{(j)}\ \mathbf{P}_1^{(j)}][\mathbf{D}_{1,21}^{(j)}\ \mathbf{D}_{1,22}^{(j)}]^\top\right\|_{\mathrm{F}}^2\\
&\leq 2\lambda_1^2b^{-2}\phi^4\left(\left\|\mathbf{A}_1^{(j)}\right\|_{\mathrm{F}}^2-\left\|\mathbf{A}_2^{(j)}\right\|_{\mathrm{F}}^2\right)^2.
\end{aligned}$$

The third term can be bounded by

$$\begin{aligned}
&\lambda_2^2\|\mathbf{R}_1^{(j)}(\mathbf{R}_1^{(j)\top}\mathbf{R}_1^{(j)}-b^2\mathbf{I}_{r_1-d_1})\|_{\mathrm{F}}^2\leq\lambda_2^2\|\mathbf{R}_1^{(j)}\|_{\mathrm{op}}^2\|\mathbf{R}_1^{(j)\top}\mathbf{R}_1^{(j)}-b^2\mathbf{I}_{r_1-d_1}\|_{\mathrm{F}}^2\\
\leq&2\lambda_2^2b^2\|\mathbf{R}_1^{(j)\top}\mathbf{R}_1^{(j)}-b^2\mathbf{I}_{r_1-d_1}\|_{\mathrm{F}}^2,
\end{aligned}$$

and the fourth term can be bounded by

$$\lambda_2^2\|\mathbf{C}_1^{(j)}\mathbf{C}_1^{(j)\top}\mathbf{R}_1^{(j)}\|_{\mathrm{F}}^2 \leq 2\lambda_2^2 b^2\|\mathbf{C}_1^{(j)\top}\mathbf{R}_1^{(j)}\|_{\mathrm{F}}^2.$$

Combining these four upper bounds, we have

$$\begin{aligned}
I_{\mathbf{R}_1,2} \leq & 16b^{-2}\phi^4\left(\xi^2(r_1,r_2,d_1,d_2)+\|\nabla\widetilde{\mathcal{L}}(\mathbf{A}^{(j)})-\nabla\widetilde{\mathcal{L}}(\mathbf{A}^*)\|_{\mathrm{F}}^2\right)\\
&+8\lambda_2^2 b^2\left(\|\mathbf{R}_1^{(j)\top}\mathbf{R}_1^{(j)}-b^2\mathbf{I}_{r_1-d_1}\|_{\mathrm{F}}^2+\|\mathbf{C}_1^{(j)\top}\mathbf{R}_1^{(j)}\|_{\mathrm{F}}^2\right)\\
&+8\lambda_1^2 b^{-2}\phi^4\left(\left\|\mathbf{A}_1^{(j)}\right\|_{\mathrm{F}}^2-\left\|\mathbf{A}_2^{(j)}\right\|_{\mathrm{F}}^2\right)^2\\
:= & Q_{\mathbf{R}_1,2}.
\end{aligned} \tag{S3.6}$$

For $I_{\mathbf{R}_1,1}$ defined in (S3.5), rewrite its first term:

$$\begin{aligned}
&\left\langle\mathbf{R}_1^{(j)}-\mathbf{R}_1^*\mathbf{O}_{1,r}^{(j)},\nabla_{\mathbf{A}_1}\widetilde{\mathcal{L}}(\mathbf{A}^{(j)})[\mathbf{C}_1^{(j)}\ \mathbf{P}_1^{(j)}][\mathbf{D}_{1,21}^{(j)}\ \mathbf{D}_{1,22}^{(j)}]^\top\right\rangle\\
=&\left\langle\mathbf{R}_1^{(j)}[\mathbf{D}_{1,21}^{(j)}\ \mathbf{D}_{1,22}^{(j)}][\mathbf{C}_1^{(j)}\ \mathbf{P}_1^{(j)}]^\top-\mathbf{R}_1^*\mathbf{O}_{1,r}^{(j)}[\mathbf{D}_{1,21}^{(j)}\ \mathbf{D}_{1,22}^{(j)}][\mathbf{C}_1^{(j)}\ \mathbf{P}_1^{(j)}]^\top,\nabla_{\mathbf{A}_1}\widetilde{\mathcal{L}}(\mathbf{A}^{(j)})\right\rangle\\
:=&\left\langle\mathbf{A}_{1,r}^{(j)},\nabla_{\mathbf{A}_1}\widetilde{\mathcal{L}}(\mathbf{A}^{(j)})\right\rangle\\
=&\left\langle\mathbf{A}_{1,r}^{(j)},\mathrm{mat}(\mathcal{P}(\nabla\widetilde{\mathcal{L}}(\mathbf{A}^{(j)}))^\top\mathrm{vec}(\mathbf{A}_2^{(j)}))\right\rangle\\
=&\left\langle\mathrm{vec}(\mathbf{A}_{1,r}^{(j)}),\mathcal{P}(\nabla\widetilde{\mathcal{L}}(\mathbf{A}^{(j)}))^\top\mathrm{vec}(\mathbf{A}_2^{(j)})\right\rangle\\
=&\left\langle\mathrm{vec}(\mathbf{A}_2^{(j)})\mathrm{vec}(\mathbf{A}_{1,r}^{(j)})^\top,\mathcal{P}(\nabla\widetilde{\mathcal{L}}(\mathbf{A}^{(j)}))\right\rangle\\
=&\left\langle\mathbf{A}_2^{(j)}\otimes\mathbf{A}_{1,r}^{(j)},\nabla\widetilde{\mathcal{L}}(\mathbf{A}^{(j)})\right\rangle\\
:=&Q_{\mathbf{R}_1,1}
\end{aligned}$$

For the second term of $I_{\mathbf{R}_1,1}$, define

$$G_{\mathbf{R}_1}:=\left\langle\mathbf{R}_1^{(j)}-\mathbf{R}_1^*\mathbf{O}_{1,r}^{(j)},\left(\left\|\mathbf{A}_1^{(j)}\right\|_{\mathrm{F}}^2-\left\|\mathbf{A}_2^{(j)}\right\|_{\mathrm{F}}^2\right)\mathbf{A}_1^{(j)}[\mathbf{C}_1^{(j)}\ \mathbf{P}_1^{(j)}][\mathbf{D}_{1,21}^{(j)}\ \mathbf{D}_{1,22}^{(j)}]^\top\right\rangle. \tag{S3.7}$$

For the third and fourth terms of $I_{\mathbf{R}_1,1}$, define

$$T_{\mathbf{R}_1}:=\left\langle\mathbf{R}_1^{(j)}-\mathbf{R}_1^*\mathbf{O}_{1,r}^{(j)},\mathbf{R}_1^{(j)}(\mathbf{R}_1^{(j)\top}\mathbf{R}_1^{(j)}-b^2\mathbf{I}_{r_1-d_1})+\mathbf{C}_1^{(j)}\mathbf{C}_1^{(j)\top}\mathbf{R}_1^{(j)}\right\rangle.$$

Therefore, we can rewrite $I_{\mathbf{R}_1,1}$ as

$$I_{\mathbf{R}_1,1} = Q_{\mathbf{R}_1,1} + \lambda_1 G_{\mathbf{R}_1} + \lambda_2 T_{\mathbf{R}_1}.$$

Combining the bound for the $I_{\mathbf{R}_1,2}$ in (S3.6), we have

$$\|\mathbf{R}_1^{(j+1)} - \mathbf{R}_1^*\mathbf{O}_{1,r}^{(j)}\|_{\mathrm{F}}^2 - \|\mathbf{R}_1^{(j)} - \mathbf{R}_1^*\mathbf{O}_{1,r}^{(j)}\|_{\mathrm{F}}^2 \leq -2\eta Q_{\mathbf{R}_1,1} - 2\lambda_1\eta G_{\mathbf{R}_1} - 2\lambda_2\eta T_{\mathbf{R}_1} + \eta^2 Q_{\mathbf{R}_1,2}.$$

Similarly, for $\mathbf{R}_2^{(j+1)}$, $\mathbf{P}_1^{(j+1)}$, and $\mathbf{P}_2^{(j+1)}$, we can define the similar quantities and show that

$$\begin{aligned}
&\|\mathbf{R}_2^{(j+1)} - \mathbf{R}_2^*\mathbf{O}_{2,r}^{(j)}\|_{\mathrm{F}}^2 - \|\mathbf{R}_2^{(j)} - \mathbf{R}_2^*\mathbf{O}_{2,r}^{(j)}\|_{\mathrm{F}}^2 \leq -2\eta Q_{\mathbf{R}_2,1} - 2\lambda_1\eta G_{\mathbf{R}_2} - 2\lambda_2\eta T_{\mathbf{R}_2} + \eta^2 Q_{\mathbf{R}_2,2},\\
&\|\mathbf{P}_1^{(j+1)} - \mathbf{P}_1^*\mathbf{O}_{1,p}^{(j)}\|_{\mathrm{F}}^2 - \|\mathbf{P}_1^{(j)} - \mathbf{P}_1^*\mathbf{O}_{1,p}^{(j)}\|_{\mathrm{F}}^2 \leq -2\eta Q_{\mathbf{P}_1,1} - 2\lambda_1\eta G_{\mathbf{P}_1} - 2\lambda_2\eta T_{\mathbf{P}_1} + \eta^2 Q_{\mathbf{P}_1,2},\\
\text{and } &\|\mathbf{P}_2^{(j+1)} - \mathbf{P}_2^*\mathbf{O}_{2,p}^{(j)}\|_{\mathrm{F}}^2 - \|\mathbf{P}_2^{(j)} - \mathbf{P}_2^*\mathbf{O}_{2,p}^{(j)}\|_{\mathrm{F}}^2 \leq -2\eta Q_{\mathbf{P}_2,1} - 2\lambda_1\eta G_{\mathbf{P}_2} - 2\lambda_2\eta T_{\mathbf{P}_2} + \eta^2 Q_{\mathbf{P}_2,2}.
\end{aligned} \tag{S3.8}$$

Step 2.2 (Upper bound for the errors of $\mathbf{C}_i$)

For $\mathbf{C}_1$, note that

$$\nabla_{\mathbf{C}_i}\widetilde{\mathcal{L}}^{(j)} = \nabla_{\mathbf{A}_i}\widetilde{\mathcal{L}}^{(j)}(\mathbf{C}_i^{(j)}\mathbf{D}_{i,11}^{(j)\top} + \mathbf{P}_i^{(j)}\mathbf{D}_{i,12}^{(j)\top}) + \nabla_{\mathbf{A}_i}\widetilde{\mathcal{L}}^{(j)\top}(\mathbf{C}_i^{(j)}\mathbf{D}_{i,11}^{(j)} + \mathbf{R}_i^{(j)}\mathbf{D}_{i,21}^{(j)}),$$

and then,

$$
\begin{aligned}
&\left\|\mathbf{C}_1^{(j+1)} - \mathbf{C}_1^{*}\mathbf{O}_{1,c}^{(j)}\right\|_{\mathrm{F}}^2 \\
=&\left\|\mathbf{C}_1^{(j)} - \mathbf{C}_1^{*}\mathbf{O}_{1,c}^{(j)} - \eta\Big\{\nabla_{\mathbf{C}_1}\mathcal{L}^{(j)} + \lambda_1\left(\left\|\mathbf{A}_1^{(j)}\right\|_{\mathrm{F}}^2 - \left\|\mathbf{A}_2^{(j)}\right\|_{\mathrm{F}}^2\right)\mathbf{A}_1^{(j)}[\mathbf{C}_1^{(j)}\ \mathbf{P}_1^{(j)}][\mathbf{D}_{1,11}^{(j)}\ \mathbf{D}_{1,12}^{(j)}]^\top\right. \\
&\quad + \lambda_1\left(\left\|\mathbf{A}_1^{(j)}\right\|_{\mathrm{F}}^2 - \left\|\mathbf{A}_2^{(j)}\right\|_{\mathrm{F}}^2\right)\mathbf{A}_1^{(j)\top}[\mathbf{C}_1^{(j)}\ \mathbf{R}_1^{(j)}][\mathbf{D}_{1,11}^{(j)\top}\ \mathbf{D}_{1,21}^{(j)\top}]^\top \\
&\quad \left.+2\lambda_2\mathbf{C}_1^{(j)}\left(\mathbf{C}_1^{(j)\top}\mathbf{C}_1^{(j)} - b^2\mathbf{I}_{d_1}\right) + \lambda_2\mathbf{R}_1^{(j)}\mathbf{R}_1^{(j)\top}\mathbf{C}_1^{(j)} + \lambda_2\mathbf{P}_1^{(j)}\mathbf{P}_1^{(j)\top}\mathbf{C}_1^{(j)}\Big\}\right\|_{\mathrm{F}}^2 \\
=&\left\|\mathbf{C}_1^{(j)} - \mathbf{C}_1^{*}\mathbf{O}_{1,c}^{(j)}\right\|_{\mathrm{F}}^2 \\
&+\eta^2\left\|\nabla_{\mathbf{C}_1}\mathcal{L}^{(j)} + \lambda_1\left(\left\|\mathbf{A}_1^{(j)}\right\|_{\mathrm{F}}^2 - \left\|\mathbf{A}_2^{(j)}\right\|_{\mathrm{F}}^2\right)\mathbf{A}_1^{(j)}[\mathbf{C}_1^{(j)}\ \mathbf{P}_1^{(j)}][\mathbf{D}_{1,11}^{(j)}\ \mathbf{D}_{1,12}^{(j)}]^\top\right. \\
&\qquad + \lambda_1\left(\left\|\mathbf{A}_1^{(j)}\right\|_{\mathrm{F}}^2 - \left\|\mathbf{A}_2^{(j)}\right\|_{\mathrm{F}}^2\right)\mathbf{A}_1^{(j)\top}[\mathbf{C}_1^{(j)}\ \mathbf{R}_1^{(j)}][\mathbf{D}_{1,11}^{(j)\top}\ \mathbf{D}_{1,21}^{(j)\top}]^\top \\
&\qquad \left.+ 2\lambda_2\mathbf{C}_1^{(j)}\left(\mathbf{C}_1^{(j)\top}\mathbf{C}_1^{(j)} - b^2\mathbf{I}_{d_1}\right) + \lambda_2\mathbf{R}_1^{(j)}\mathbf{R}_1^{(j)\top}\mathbf{C}_1^{(j)} + \lambda_2\mathbf{P}_1^{(j)}\mathbf{P}_1^{(j)\top}\mathbf{C}_1^{(j)}\right\|_{\mathrm{F}}^2 \\
&-2\eta\left\langle \mathbf{C}_1^{(j)} - \mathbf{C}_1^{*}\mathbf{O}_{1,c}^{(j)}, \nabla_{\mathbf{C}_1}\mathcal{L}^{(j)}\right\rangle \\
&-2\lambda_1\eta\left\langle \mathbf{C}_1^{(j)} - \mathbf{C}_1^{*}\mathbf{O}_{1,c}^{(j)}, \left(\left\|\mathbf{A}_1^{(j)}\right\|_{\mathrm{F}}^2 - \left\|\mathbf{A}_2^{(j)}\right\|_{\mathrm{F}}^2\right)\mathbf{A}_1^{(j)}[\mathbf{C}_1^{(j)}\ \mathbf{P}_1^{(j)}][\mathbf{D}_{1,11}^{(j)}\ \mathbf{D}_{1,12}^{(j)}]^\top\right\rangle \\
&-2\lambda_1\eta\left\langle \mathbf{C}_1^{(j)} - \mathbf{C}_1^{*}\mathbf{O}_{1,c}^{(j)}, \left(\left\|\mathbf{A}_1^{(j)}\right\|_{\mathrm{F}}^2 - \left\|\mathbf{A}_2^{(j)}\right\|_{\mathrm{F}}^2\right)\mathbf{A}_1^{(j)\top}[\mathbf{C}_1^{(j)}\ \mathbf{R}_1^{(j)}][\mathbf{D}_{1,11}^{(j)\top}\ \mathbf{D}_{1,21}^{(j)\top}]^\top\right\rangle \\
&-2\lambda_2\eta\left\langle \mathbf{C}_1^{(j)} - \mathbf{C}_1^{*}\mathbf{O}_{1,c}^{(j)}, 2\mathbf{C}_1^{(j)}\left(\mathbf{C}_1^{(j)\top}\mathbf{C}_1^{(j)} - b^2\mathbf{I}_{d_1}\right)\right\rangle \\
&-2\lambda_2\eta\left\langle \mathbf{C}_1^{(j)} - \mathbf{C}_1^{*}\mathbf{O}_{1,c}^{(j)}, \mathbf{R}_1^{(j)}\mathbf{R}_1^{(j)\top}\mathbf{C}_1^{(j)} + \mathbf{P}_1^{(j)}\mathbf{P}_1^{(j)\top}\mathbf{C}_1^{(j)}\right\rangle \\
:=&\left\|\mathbf{C}_1^{(j)} - \mathbf{C}_1^{*}\mathbf{O}_{1,c}^{(j)}\right\|_{\mathrm{F}}^2 + \eta^2 I_{\mathbf{C}_1,2} - 2\eta I_{\mathbf{C}_1,1}.
\end{aligned}
\tag{S3.9}
$$

For $I_{\mathbf{C}_1,2}$ in (S3.9),

$$
\begin{aligned}
I_{\mathbf{C}_1,2} \leq& 5\left\|\nabla_{\mathbf{C}_1}\mathcal{L}^{(j)}\right\|_{\mathrm{F}}^2 + 40\lambda_1^2 b^{-2}\phi^4\left(\left\|\mathbf{A}_1^{(j)}\right\|_{\mathrm{F}}^2 - \left\|\mathbf{A}_2^{(j)}\right\|_{\mathrm{F}}^2\right)^2 \\
&+ 40\lambda_2^2 b^2\left\|\mathbf{C}_1^{(j)\top}\mathbf{C}_1^{(j)} - b^2\mathbf{I}_{d_1}\right\|_{\mathrm{F}}^2 + 10\lambda_2^2 b^2\left\|\mathbf{R}_1^{(j)\top}\mathbf{C}_1^{(j)}\right\|_{\mathrm{F}}^2 + 10\lambda_2^2 b^2\left\|\mathbf{P}_1^{(j)\top}\mathbf{C}_1^{(j)}\right\|_{\mathrm{F}}^2.
\end{aligned}
$$

The first term of the RHS can be bounded as

$$
\begin{aligned}
&\left\|\nabla_{\mathbf{C}_1}\mathcal{L}^{(j)}\right\|_{\mathrm{F}}^2\\
\leq{}& 2\left\|\nabla_{\mathbf{A}_1}\widetilde{\mathcal{L}}(\mathbf{A}^{(j)})\left[\mathbf{C}_1^{(j)}\ \mathbf{P}_1^{(j)}\right]\left[\mathbf{D}_{1,11}^{(j)}\ \mathbf{D}_{1,12}^{(j)}\right]^\top\right\|_{\mathrm{F}}^2\\
&+2\left\|\nabla_{\mathbf{A}_1}\widetilde{\mathcal{L}}(\mathbf{A}^{(j)})^\top\left[\mathbf{C}_1^{(j)}\ \mathbf{R}_1^{(j)}\right]\left[\mathbf{D}_{1,11}^{(j)\top}\ \mathbf{D}_{1,21}^{(j)\top}\right]^\top\right\|_{\mathrm{F}}^2\\
={}& 2\left\|\mathrm{mat}\left(\mathcal{P}\left(\nabla\widetilde{\mathcal{L}}\left(\mathbf{A}^{(j)}\right)\right)^\top\mathrm{vec}(\mathbf{A}_2^{(j)})\right))\left[\mathbf{C}_1^{(j)}\ \mathbf{P}_1^{(j)}\right]\left[\mathbf{D}_{1,11}^{(j)}\ \mathbf{D}_{1,12}^{(j)}\right]^\top\right\|_{\mathrm{F}}^2\\
&+2\left\|\mathrm{mat}\left(\mathcal{P}\left(\nabla\widetilde{\mathcal{L}}\left(\mathbf{A}^{(j)}\right)\right)^\top\mathrm{vec}(\mathbf{A}_2^{(j)})\right)^\top\left[\mathbf{C}_1^{(j)}\ \mathbf{R}_1^{(j)}\right]\left[\mathbf{D}_{1,11}^{(j)\top}\ \mathbf{D}_{1,21}^{(j)\top}\right]^\top\right\|_{\mathrm{F}}^2\\
\leq{}& 4\left\|\mathrm{mat}\left(\mathcal{P}\left(\nabla\widetilde{\mathcal{L}}\left(\mathbf{A}^{*}\right)\right)^\top\mathrm{vec}(\mathbf{A}_2^{(j)})\right)\left[\mathbf{C}_1^{(j)}\ \mathbf{P}_1^{(j)}\right]\left[\mathbf{D}_{1,11}^{(j)}\ \mathbf{D}_{1,12}^{(j)}\right]^\top\right\|_{\mathrm{F}}^2\\
&+4\left\|\mathrm{mat}\left(\mathcal{P}\left(\nabla\widetilde{\mathcal{L}}\left(\mathbf{A}^{*}\right)\right)^\top\mathrm{vec}(\mathbf{A}_2^{(j)})\right))^\top\left[\mathbf{C}_1^{(j)}\ \mathbf{R}_1^{(j)}\right]\left[\mathbf{D}_{1,11}^{(j)\top}\ \mathbf{D}_{1,21}^{(j)\top}\right]^\top\right\|_{\mathrm{F}}^2\\
&+4\left\|\mathrm{mat}\left(\mathcal{P}\left(\nabla\widetilde{\mathcal{L}}(\mathbf{A}^{(j)})-\nabla\widetilde{\mathcal{L}}(\mathbf{A}^{*})\right)^\top\mathrm{vec}(\mathbf{A}_2^{(j)})\right)\left[\mathbf{C}_1^{(j)}\ \mathbf{P}_1^{(j)}\right]\left[\mathbf{D}_{1,11}^{(j)}\ \mathbf{D}_{1,12}^{(j)}\right]^\top\right\|_{\mathrm{F}}^2\\
&+4\left\|\mathrm{mat}\left(\mathcal{P}\left(\nabla\widetilde{\mathcal{L}}(\mathbf{A}^{(j)})-\nabla\widetilde{\mathcal{L}}(\mathbf{A}^{*})\right)^\top\mathrm{vec}(\mathbf{A}_2^{(j)})\right)^\top\left[\mathbf{C}_1^{(j)}\ \mathbf{R}_1^{(j)}\right]\left[\mathbf{D}_{1,11}^{(j)\top}\ \mathbf{D}_{1,21}^{(j)\top}\right]^\top\right\|_{\mathrm{F}}^2\\
\leq{}& 16b^{-2}\phi^4\xi^2(r_1,r_2,d_1,d_2)+16b^{-2}\phi^4\left\|\nabla\widetilde{\mathcal{L}}(\mathbf{A}^{(j)})-\nabla\widetilde{\mathcal{L}}(\mathbf{A}^{*})\right\|_{\mathrm{F}}^2\\
={}& 16b^{-2}\phi^4\left(\xi^2(r_1,r_2,d_1,d_2)+\left\|\nabla\widetilde{\mathcal{L}}(\mathbf{A}^{(j)})-\nabla\widetilde{\mathcal{L}}(\mathbf{A}^{*})\right\|_{\mathrm{F}}^2\right).
\end{aligned}
$$

Thus, $I_{\mathbf{C}_1,2}$ can be upper bounded by

$$
\begin{aligned}
I_{\mathbf{C}_1,2}\leq{}&80b^{-2}\phi^4\left(\xi^2(r_1,r_2,d_1,d_2)+\left\|\nabla\widetilde{\mathcal{L}}(\mathbf{A}^{(j)})-\nabla\widetilde{\mathcal{L}}(\mathbf{A}^{*})\right\|_{\mathrm{F}}^2\right)\\
&+40\lambda_1^2b^{-2}\phi^4\left(\left\|\mathbf{A}_1^{(j)}\right\|_{\mathrm{F}}^2-\left\|\mathbf{A}_2^{(j)}\right\|_{\mathrm{F}}^2\right)^2\\
&+40\lambda_2^2b^2\left\|\mathbf{C}_1^{(j)\top}\mathbf{C}_1^{(j)}-b^2\mathbf{I}_{d_1}\right\|_{\mathrm{F}}^2+10\lambda_2^2b^2\left\|\mathbf{R}_1^{(j)\top}\mathbf{C}_1^{(j)}\right\|_{\mathrm{F}}^2+10\lambda_2^2b^2\left\|\mathbf{P}_1^{(j)\top}\mathbf{C}_1^{(j)}\right\|_{\mathrm{F}}^2\\
:={}&Q_{\mathbf{C}_1,2}.
\end{aligned}
\tag{S3.10}
$$

For $I_{\mathbf{C}_1,1}$ in (S3.9), similarly to $\mathbf{R}_1$ step, we rewrite its first term as the following:

$$\begin{aligned}
&\left\langle \mathbf{C}_1^{(j)} - \mathbf{C}_1^{*}\mathbf{O}_{1,c}^{(j)}, \nabla_{\mathbf{C}_1}\mathcal{L}^{(j)} \right\rangle \\
=&\left\langle \mathbf{R}_1^{(j)}\mathbf{D}_{1,21}^{(j)}\mathbf{C}_1^{(j)\top} - \mathbf{R}_1^{(j)}\mathbf{D}_{1,21}^{(j)}\mathbf{O}_{1,c}^{(j)\top}\mathbf{C}_1^{*\top}, \nabla_{\mathbf{A}_1}\widetilde{\mathcal{L}}\left(\mathbf{A}^{(j)}\right) \right\rangle \\
+&\left\langle \mathbf{C}_1^{(j)}\mathbf{D}_{1,12}^{(j)}\mathbf{P}_1^{(j)\top} - \mathbf{C}_1^{*}\mathbf{O}_{1,c}^{(j)}\mathbf{D}_{1,12}^{(j)}\mathbf{P}_1^{(j)\top}, \nabla_{\mathbf{A}_1}\widetilde{\mathcal{L}}\left(\mathbf{A}^{(j)}\right) \right\rangle \\
+&\left\langle \mathbf{C}_1^{(j)}\mathbf{D}_{1,11}^{(j)}\mathbf{C}_1^{(j)\top} - \mathbf{C}_1^{*}\mathbf{O}_{1,c}^{(j)}\mathbf{D}_{1,11}^{(j)}\mathbf{C}_1^{(j)\top}, \nabla_{\mathbf{A}_1}\widetilde{\mathcal{L}}\left(\mathbf{A}^{(j)}\right) \right\rangle \\
+&\left\langle \mathbf{C}_1^{(j)}\mathbf{D}_{1,11}^{(j)}\mathbf{C}_1^{(j)\top} - \mathbf{C}_1^{(j)}\mathbf{D}_{1,11}^{(j)}\mathbf{O}_{1,c}^{(j)\top}\mathbf{C}_1^{*\top}, \nabla_{\mathbf{A}_1}\widetilde{\mathcal{L}}\left(\mathbf{A}^{(j)}\right) \right\rangle \\
:=&\left\langle \mathbf{A}_{\mathbf{C}_1}^{(j)}, \nabla_{\mathbf{A}_1}\widetilde{\mathcal{L}}\left(\mathbf{A}^{(j)}\right) \right\rangle \\
=&\left\langle \mathbf{A}_2^{(j)} \otimes \mathbf{A}_{\mathbf{C}_1}^{(j)}, \nabla\widetilde{\mathcal{L}}(\mathbf{A}^{(j)}) \right\rangle \\
:=&Q_{\mathbf{C}_1,1}.
\end{aligned}$$

For the last three terms, denote

$$\begin{aligned}
G_{\mathbf{C}_1} :=& \left\langle \mathbf{C}_1^{(j)} - \mathbf{C}_1^{*}\mathbf{O}_{1,c}^{(j)}, \left(\left\|\mathbf{A}_1^{(j)}\right\|_{\mathrm{F}}^2 - \left\|\mathbf{A}_2^{(j)}\right\|_{\mathrm{F}}^2 \right) \mathbf{A}_1^{(j)}[\mathbf{C}_1^{(j)}\ \mathbf{P}_1^{(j)}][\mathbf{D}_{1,11}^{(j)}\ \mathbf{D}_{1,12}^{(j)}]^\top \right\rangle \\
&+ \left\langle \mathbf{C}_1^{(j)} - \mathbf{C}_1^{*}\mathbf{O}_{1,c}^{(j)}, \left(\left\|\mathbf{A}_1^{(j)}\right\|_{\mathrm{F}}^2 - \left\|\mathbf{A}_2^{(j)}\right\|_{\mathrm{F}}^2 \right) \mathbf{A}_1^{(j)\top}[\mathbf{C}_1^{(j)}\ \mathbf{R}_1^{(j)}][\mathbf{D}_{1,11}^{(j)\top}\ \mathbf{D}_{1,21}^{(j)\top}]^\top \right\rangle
\end{aligned}$$

and

$$T_{\mathbf{C}_1} := \left\langle \mathbf{C}_1^{(j)} - \mathbf{C}_1^{*}\mathbf{O}_{1,c}^{(j)}, 2\mathbf{C}_1^{(j)}\left(\mathbf{C}_1^{(j)\top}\mathbf{C}_1^{(j)} - b^2\mathbf{I}_{d_1}\right) + \mathbf{R}_1^{(j)}\mathbf{R}_1^{(j)\top}\mathbf{C}_1^{(j)} + \mathbf{P}_1^{(j)}\mathbf{P}_1^{(j)\top}\mathbf{C}_1^{(j)} \right\rangle.$$

Therefore, we can rewrite $I_{\mathbf{C}_1,1}$ as

$$I_{\mathbf{C}_1,1} = Q_{\mathbf{C}_1,1} + \lambda_1 G_{\mathbf{C}_1} + \lambda_2 T_{\mathbf{C}_1}.$$

Combining these bounds for $I_{\mathbf{C}_1,2}$ in (S3.10), we have

$$\left\|\mathbf{C}_1^{(j+1)} - \mathbf{C}_1^{*}\mathbf{O}_{1,c}^{(j)}\right\|_{\mathrm{F}}^2 - \left\|\mathbf{C}_1^{(j)} - \mathbf{C}_1^{*}\mathbf{O}_{1,c}^{(j)}\right\|_{\mathrm{F}}^2 \le -2\eta Q_{\mathbf{C}_1,1} - 2\lambda_1\eta G_{\mathbf{C}_1} - 2\lambda_2\eta T_{\mathbf{C}_1} + \eta^2 Q_{\mathbf{C}_1,2}.$$

Similarly, we also have

$$\left\|\mathbf{C}_2^{(j+1)} - \mathbf{C}_2^{*}\mathbf{O}_{2,c}^{(j)}\right\|_{\mathrm{F}}^2 - \left\|\mathbf{C}_2^{(j)} - \mathbf{C}_2^{*}\mathbf{O}_{2,c}^{(j)}\right\|_{\mathrm{F}}^2 \le -2\eta Q_{\mathbf{C}_2,1} - 2\lambda_1\eta G_{\mathbf{C}_2} - 2\lambda_2\eta T_{\mathbf{C}_2} + \eta^2 Q_{\mathbf{C}_2,2}.$$

Step 2.3 (Upper bound for the errors of $\mathbf{D}_i$)

$$\begin{aligned}
&\left\|\mathbf{D}_1^{(j+1)} - \mathbf{O}_{1,u}^{(j)\top}\mathbf{D}_1^*\mathbf{O}_{1,v}^{(j)}\right\|_{\mathrm{F}}^2 \\
=&\left\|\mathbf{D}_1^{(j)} - \mathbf{O}_{1,u}^{(j)\top}\mathbf{D}_1^*\mathbf{O}_{1,v}^{(j)} - \eta\left\{\left[\mathbf{C}_1^{(j)}\ \mathbf{R}_1^{(j)}\right]^\top \nabla_{\mathbf{A}_1}\widetilde{\mathcal{L}}(\mathbf{A}^{(j)})\left[\mathbf{C}_1^{(j)}\ \mathbf{P}_1^{(j)}\right]\right.\right. \\
&\left.\left.+\lambda_1\left(\left\|\mathbf{A}_1^{(j)}\right\|_{\mathrm{F}}^2 - \left\|\mathbf{A}_2^{(j)}\right\|_{\mathrm{F}}^2\right)[\mathbf{C}_1^{(j)}\ \mathbf{R}_1^{(j)}]^\top\mathbf{A}_1^{(j)}[\mathbf{C}_1^{(j)}\ \mathbf{P}_1^{(j)}]\right\}\right\|_{\mathrm{F}}^2 \\
=&\left\|\mathbf{D}_1^{(j)} - \mathbf{O}_{1,u}^{(j)\top}\mathbf{D}_1^*\mathbf{O}_{1,v}^{(j)}\right\|_{\mathrm{F}}^2 \\
&+\eta^2\left\|\left[\mathbf{C}_1^{(j)}\ \mathbf{R}_1^{(j)}\right]^\top \nabla_{\mathbf{A}_1}\widetilde{\mathcal{L}}(\mathbf{A}^{(j)})\left[\mathbf{C}_1^{(j)}\ \mathbf{P}_1^{(j)}\right]\right. \\
&\qquad\left.+\ \lambda_1\left(\left\|\mathbf{A}_1^{(j)}\right\|_{\mathrm{F}}^2 - \left\|\mathbf{A}_2^{(j)}\right\|_{\mathrm{F}}^2\right)[\mathbf{C}_1^{(j)}\ \mathbf{R}_1^{(j)}]^\top\mathbf{A}_1^{(j)}[\mathbf{C}_1^{(j)}\ \mathbf{P}_1^{(j)}]\right\|_{\mathrm{F}}^2 \\
&-2\eta\left\langle \mathbf{D}_1^{(j)} - \mathbf{O}_{1,u}^{(j)\top}\mathbf{D}_1^*\mathbf{O}_{1,v}^{(j)}, \left[\mathbf{C}_1^{(j)}\ \mathbf{R}_1^{(j)}\right]^\top \nabla_{\mathbf{A}_1}\widetilde{\mathcal{L}}(\mathbf{A}^{(j)})\left[\mathbf{C}_1^{(j)}\ \mathbf{P}_1^{(j)}\right]\right\rangle \\
&-2\lambda_1\eta\left\langle \mathbf{D}_1^{(j)} - \mathbf{O}_{1,u}^{(j)\top}\mathbf{D}_1^*\mathbf{O}_{1,v}^{(j)}, \left(\left\|\mathbf{A}_1^{(j)}\right\|_{\mathrm{F}}^2 - \left\|\mathbf{A}_2^{(j)}\right\|_{\mathrm{F}}^2\right)\left[\mathbf{C}_1^{(j)}\ \mathbf{R}_1^{(j)}\right]^\top \mathbf{A}_1^{(j)}\left[\mathbf{C}_1^{(j)}\ \mathbf{P}_1^{(j)}\right]\right\rangle \\
:=&\left\|\mathbf{D}_1^{(j)} - \mathbf{O}_{1,u}^{(j)\top}\mathbf{D}_1^*\mathbf{O}_{1,v}^{(j)}\right\|_{\mathrm{F}}^2 + \eta^2 I_{\mathbf{D}_1,2} - 2\eta I_{\mathbf{D}_1,1}.
\end{aligned}$$

For $I_{\mathbf{D}_1,2}$:

$$I_{\mathbf{D}_1,2} \leq 2\left\|\left[\mathbf{C}_1^{(j)}\ \mathbf{R}_1^{(j)}\right]^\top \nabla_{\mathbf{A}_1}\widetilde{\mathcal{L}}(\mathbf{A}^{(j)})\left[\mathbf{C}_1^{(j)}\ \mathbf{P}_1^{(j)}\right]\right\|_{\mathrm{F}}^2 + 4\lambda_1^2 b^4\phi^2\left(\left\|\mathbf{A}_1^{(j)}\right\|_{\mathrm{F}}^2 - \left\|\mathbf{A}_2^{(j)}\right\|_{\mathrm{F}}^2\right)^2.$$

For the first term,

$$\begin{aligned}
&\left\|\left[\mathbf{C}_1^{(j)}\ \mathbf{R}_1^{(j)}\right]^\top \nabla_{\mathbf{A}_1}\widetilde{\mathcal{L}}(\mathbf{A}^{(j)})\left[\mathbf{C}_1^{(j)}\ \mathbf{P}_1^{(j)}\right]\right\|_{\mathrm{F}}^2 \\
=&\left\|\left[\mathbf{C}_1^{(j)}\ \mathbf{R}_1^{(j)}\right]^\top \operatorname{mat}\left(\mathcal{P}\left(\nabla\widetilde{\mathcal{L}}\left(\mathbf{A}^{(j)}\right)\right)^\top \operatorname{vec}(\mathbf{A}_2^{(j)})\right))\left[\mathbf{C}_1^{(j)}\ \mathbf{P}_1^{(j)}\right]\right\|_{\mathrm{F}}^2 \\
\leq&2\left\|\left[\mathbf{C}_1^{(j)}\ \mathbf{R}_1^{(j)}\right]^\top \operatorname{mat}\left(\mathcal{P}\left(\nabla\widetilde{\mathcal{L}}\left(\mathbf{A}^*\right)\right)^\top \operatorname{vec}(\mathbf{A}_2^{(j)})\right))\left[\mathbf{C}_1^{(j)}\ \mathbf{P}_1^{(j)}\right]\right\|_{\mathrm{F}}^2 \\
&+2\left\|\left[\mathbf{C}_1^{(j)}\ \mathbf{R}_1^{(j)}\right]^\top \operatorname{mat}\left(\mathcal{P}\left[\nabla\widetilde{\mathcal{L}}\left(\mathbf{A}^{(j)}\right) - \nabla\widetilde{\mathcal{L}}(\mathbf{A}^*)\right]^\top \operatorname{vec}(\mathbf{A}_2^{(j)})\right)\left[\mathbf{C}_1^{(j)}\ \mathbf{P}_1^{(j)}\right]\right\|_{\mathrm{F}}^2 \\
\leq&4b^4\phi^2\left(\xi^2(r_1,r_2,d_1,d_2) + \left\|\nabla\widetilde{\mathcal{L}}(\mathbf{A}^{(j)}) - \nabla\widetilde{\mathcal{L}}(\mathbf{A}^*)\right\|_{\mathrm{F}}^2\right)
\end{aligned}$$

Then $I_{\mathbf{D}_1,2}$ can be upper bounded as

$$
\begin{aligned}
&I_{\mathbf{D}_1,2}\\
\leq& 8b^4\phi^2\left(\xi^2(r_1,r_2,d_1,d_2)+\left\|\nabla\widetilde{\mathcal{L}}(\mathbf{A}^{(j)})-\nabla\widetilde{\mathcal{L}}(\mathbf{A}^*)\right\|_{\mathrm{F}}^2\right)+4\lambda_1^2b^4\phi^2\left(\left\|\mathbf{A}_1^{(j)}\right\|_{\mathrm{F}}^2-\left\|\mathbf{A}_2^{(j)}\right\|_{\mathrm{F}}^2\right)^2\\
:=& Q_{\mathbf{D}_1,2}
\end{aligned}
$$

For $I_{\mathbf{D}_1,1}$, similarly we define

$$
\begin{aligned}
&\left\langle \mathbf{D}_1^{(j)}-\mathbf{O}_{1,u}^{(j)\top}\mathbf{D}_1^*\mathbf{O}_{1,v}^{(j)},\left[\mathbf{C}_1^{(j)}\ \mathbf{R}_1^{(j)}\right]^\top\nabla_{\mathbf{A}_1}\widetilde{\mathcal{L}}(\mathbf{A}^{(j)})\left[\mathbf{C}_1^{(j)}\ \mathbf{P}_1^{(j)}\right]\right\rangle\\
=&\left\langle \mathbf{A}_1^{(j)}-\left[\mathbf{C}_1^{(j)}\ \mathbf{R}_1^{(j)}\right]\mathbf{O}_{1,u}^{(j)\top}\mathbf{D}_1^*\mathbf{O}_{1,v}^{(j)}\left[\mathbf{C}_1^{(j)}\ \mathbf{P}_1^{(j)}\right]^\top,\nabla_{\mathbf{A}_1}\widetilde{\mathcal{L}}(\mathbf{A}^{(j)})\right\rangle\\
:=&\left\langle \mathbf{A}_{\mathbf{D}_1}^{(j)},\nabla_{\mathbf{A}_1}\widetilde{\mathcal{L}}(\mathbf{A}^{(j)})\right\rangle\\
=&\left\langle \mathbf{A}_2^{(j)}\otimes\mathbf{A}_{\mathbf{D}_1}^{(j)},\nabla\widetilde{\mathcal{L}}(\mathbf{A}^{(j)})\right\rangle\\
:=&Q_{\mathbf{D}_1,1},
\end{aligned}
$$

and

$$
G_{\mathbf{D}_1}:=\left\langle \mathbf{D}_1^{(j)}-\mathbf{O}_{1,u}^{(j)\top}\mathbf{D}_1^*\mathbf{O}_{1,v}^{(j)},\left(\left\|\mathbf{A}_1^{(j)}\right\|_{\mathrm{F}}^2-\left\|\mathbf{A}_2^{(j)}\right\|_{\mathrm{F}}^2\right)\left[\mathbf{C}_1^{(j)}\ \mathbf{R}_1^{(j)}\right]^\top\mathbf{A}_1^{(j)}\left[\mathbf{C}_1^{(j)}\ \mathbf{P}_1^{(j)}\right]\right\rangle
$$

Hence, we have

$$
\left\|\mathbf{D}_1^{(j+1)}-\mathbf{O}_{1,u}^{(j)\top}\mathbf{D}_1^*\mathbf{O}_{1,v}^{(j)}\right\|_{\mathrm{F}}^2-\left\|\mathbf{D}_1^{(j)}-\mathbf{O}_{1,u}^{(j)\top}\mathbf{D}_1^*\mathbf{O}_{1,v}^{(j)}\right\|_{\mathrm{F}}^2\leq-2\eta Q_{\mathbf{D}_1,1}-2\lambda_1\eta G_{\mathbf{D}_1}+\eta^2Q_{\mathbf{D}_1,2},
$$

and

$$
\left\|\mathbf{D}_2^{(j+1)}-\mathbf{O}_{2,u}^{(j)\top}\mathbf{D}_2^*\mathbf{O}_{2,v}^{(j)}\right\|_{\mathrm{F}}^2-\left\|\mathbf{D}_2^{(j)}-\mathbf{O}_{2,u}^{(j)\top}\mathbf{D}_2^*\mathbf{O}_{2,v}^{(j)}\right\|_{\mathrm{F}}^2\leq-2\eta Q_{\mathbf{D}_2,1}-2\lambda_1\eta G_{\mathbf{D}_2}+\eta^2Q_{\mathbf{D}_2,2}.
$$

Combining the pieces above,

$$\begin{aligned}\operatorname{dist}_{(j+1)}^{2} \leq \operatorname{dist}_{(j)}^{2} &+ \eta^{2} \sum_{i=1}^{2}\left(Q_{\mathrm{D}_{i}, 2}+Q_{\mathrm{R}_{i}, 2}+Q_{\mathrm{P}_{i}, 2}+Q_{\mathrm{C}_{i}, 2}\right) \\ &- 2 \eta \sum_{i=1}^{2}\left(Q_{\mathrm{D}_{i}, 1}+Q_{\mathrm{R}_{i}, 1}+Q_{\mathrm{P}_{i}, 1}+Q_{\mathrm{C}_{i}, 1}\right) \\ &- 2 \lambda_{1} \eta \sum_{i=1}^{2}\left(G_{D_{i}}+G_{R_{i}}+G_{P_{i}}+G_{C_{i}}\right) \\ &- 2 \lambda_{2} \eta \sum_{i=1}^{2}\left(T_{R_{i}}+T_{P_{i}}+T_{C_{i}}\right).\end{aligned} \tag{S3.11}$$

Step 3. (**Recursive relationship between** $\operatorname{dist}_{(j+1)}^{2}$ **and** $\operatorname{dist}_{(j)}^{2}$)

In this step, we will derive upper or lower bounds of $Q_{\cdot,1}, Q_{\cdot,2}, G$ and T terms, and finally obtain an upper bound as in (S3.22).

Step 3.1 (Lower bound for $\sum_{i=1}^{2}\left(Q_{\mathbf{D}_{i}, 1}+Q_{\mathbf{R}_{i}, 1}+Q_{\mathbf{P}_{i}, 1}+Q_{\mathbf{C}_{i}, 1}\right)$)

By definition,

$$\begin{aligned}&\sum_{i=1}^{2}\left(Q_{\mathbf{D}_{i}, 1}+Q_{\mathbf{R}_{i}, 1}+Q_{\mathbf{P}_{i}, 1}+Q_{\mathbf{C}_{i}, 1}\right) \\ =&\left\langle\begin{array}{c}\mathbf{A}_{2}^{(j)} \otimes\left(\mathbf{A}_{\mathbf{D}_{1}}^{(j)}+\mathbf{A}_{\mathbf{R}_{1}}^{(j)}+\mathbf{A}_{\mathbf{P}_{1}}^{(j)}+\mathbf{A}_{\mathbf{C}_{1}}^{(j)}\right) \\ +\left(\mathbf{A}_{\mathbf{D}_{2}}^{(j)}+\mathbf{A}_{\mathbf{R}_{2}}^{(j)}+\mathbf{A}_{\mathbf{P}_{2}}^{(j)}+\mathbf{A}_{\mathbf{C}_{2}}^{(j)}\right) \otimes \mathbf{A}_{1}^{(j)}\end{array}, \nabla \widetilde{\mathcal{L}}\left(\mathbf{A}^{(j)}\right)\right\rangle .\end{aligned} \tag{S3.12}$$

Noting that

$$\begin{aligned}&\mathbf{A}_{\mathbf{D}_{1}}^{(j)}+\mathbf{A}_{\mathbf{R}_{1}}^{(j)}+\mathbf{A}_{\mathbf{P}_{1}}^{(j)}+\mathbf{A}_{\mathbf{C}_{1}}^{(j)} \\ =&3 \mathbf{A}_{1}^{(j)}-\left[\mathbf{C}_{1}^{(j)}\ \mathbf{R}_{1}^{(j)}\right] \mathbf{O}_{1, u}^{(j) \top} \mathbf{D}_{1}^{*} \mathbf{O}_{1, v}^{(j)}\left[\mathbf{C}_{1}^{(j)}\ \mathbf{P}_{1}^{(j)}\right]^{\top}-\left[\mathbf{C}_{1}^{(j)}\ \mathbf{R}_{1}^{(j)}\right] \mathbf{D}_{1}^{(j)} \mathbf{O}_{1, v}^{(j) \top}\left[\mathbf{C}_{1}^{*}\ \mathbf{P}_{1}^{*}\right]^{\top} \\ &-\left[\mathbf{C}_{1}^{*}\ \mathbf{R}_{1}^{*}\right] \mathbf{O}_{1, u}^{(j)} \mathbf{D}_{1}^{(j)}\left[\mathbf{C}_{1}^{(j)}\ \mathbf{P}_{1}^{(j)}\right]^{\top},\end{aligned}$$

we define

$$\mathbf{H}_{1}^{(j)}:=\mathbf{A}_{2}^{(j)} \otimes\left(\mathbf{A}_{\mathbf{D}_{1}}^{(j)}+\mathbf{A}_{\mathbf{R}_{1}}^{(j)}+\mathbf{A}_{\mathbf{P}_{1}}^{(j)}+\mathbf{A}_{\mathbf{C}_{1}}^{(j)}\right)+\left(\mathbf{A}_{\mathbf{D}_{2}}^{(j)}+\mathbf{A}_{\mathbf{R}_{2}}^{(j)}+\mathbf{A}_{\mathbf{P}_{2}}^{(j)}+\mathbf{A}_{\mathbf{C}_{2}}^{(j)}\right) \otimes \mathbf{A}_{1}^{(j)},$$

which contains all the first order perturbation terms. By Lemma S.1, we know that

$$\mathbf{H}_1^{(j)} = \mathbf{A}_2^{(j)} \otimes \mathbf{A}_1^{(j)} - \mathbf{A}_2^* \otimes \mathbf{A}_1^* + \mathbf{H}^{(j)}.$$

Hence, (S3.12) can be simplified as $\left\langle \mathbf{A}_2^{(j)} \otimes \mathbf{A}_1^{(j)} - \mathbf{A}_2^* \otimes \mathbf{A}_1^* + \mathbf{H}^{(j)}, \nabla\widetilde{\mathcal{L}}\left(\mathbf{A}^{(j)}\right) \right\rangle$. With conditions (S3.4) (S3.3) and Lemma S.1, we have

$$\left\|\mathbf{H}^{(j)}\right\|_{\mathrm{F}} \leq C_h \phi^{4/3} \mathrm{dist}_{(j)}^2,$$

where C_h is a constant of moderate size.

Then, for (S3.12), with RCG condition (S3.2), we have

$$\begin{aligned}
&\left\langle \begin{array}{l} \mathbf{A}_2^{(j)} \otimes \left(\mathbf{A}_{\mathbf{D}_1}^{(j)} + \mathbf{A}_{\mathbf{R}_1}^{(j)} + \mathbf{A}_{\mathbf{P}_1}^{(j)} + \mathbf{A}_{\mathbf{C}_1}^{(j)}\right) \\ + \left(\mathbf{A}_{\mathbf{D}_2}^{(j)} + \mathbf{A}_{\mathbf{R}_2}^{(j)} + \mathbf{A}_{\mathbf{P}_2}^{(j)} + \mathbf{A}_{\mathbf{C}_2}^{(j)}\right) \otimes \mathbf{A}_1^{(j)} \end{array}, \nabla\widetilde{\mathcal{L}}\left(\mathbf{A}^{(j)}\right) \right\rangle \\
=&\left\langle \mathbf{A}_2^{(j)} \otimes \mathbf{A}_1^{(j)} - \mathbf{A}_2^* \otimes \mathbf{A}_1^* + \mathbf{H}^{(j)}, \nabla\widetilde{\mathcal{L}}\left(\mathbf{A}^{(j)}\right) \right\rangle \\
=&\left\langle \mathbf{A}_2^{(j)} \otimes \mathbf{A}_1^{(j)} - \mathbf{A}_2^* \otimes \mathbf{A}_1^*, \nabla\widetilde{\mathcal{L}}(\mathbf{A}^{(j)}) - \nabla\widetilde{\mathcal{L}}(\mathbf{A}^*) \right\rangle + \left\langle \mathbf{H}^{(j)}, \nabla\widetilde{\mathcal{L}}(\mathbf{A}^{(j)}) - \nabla\widetilde{\mathcal{L}}(\mathbf{A}^*) \right\rangle \\
&+ \left\langle \mathbf{A}_2^{(j)} \otimes \mathbf{A}_1^{(j)} - \mathbf{A}_2^* \otimes \mathbf{A}_1^* + \mathbf{H}^{(j)}, \nabla\widetilde{\mathcal{L}}\left(\mathbf{A}^*\right) \right\rangle \\
\geq&\frac{\alpha}{2}\left\|\mathbf{A}_2^{(j)} \otimes \mathbf{A}_1^{(j)} - \mathbf{A}_2^* \otimes \mathbf{A}_1^*\right\|_F^2 + \frac{1}{2\beta}\left\|\nabla\widetilde{\mathcal{L}}(\mathbf{A}^{(j)}) - \nabla\widetilde{\mathcal{L}}(\mathbf{A}^*)\right\|_F^2 \\
&- \left\|\mathbf{H}^{(j)}\right\|_{\mathrm{F}}\left\|\nabla\widetilde{\mathcal{L}}(\mathbf{A}^{(j)}) - \nabla\widetilde{\mathcal{L}}(\mathbf{A}^*)\right\|_{\mathrm{F}} \\
&- \left|\left\langle \mathbf{A}_2^{(j)} \otimes \left(\mathbf{A}_{\mathbf{D}_1}^{(j)} + \mathbf{A}_{\mathbf{R}_1}^{(j)} + \mathbf{A}_{\mathbf{P}_1}^{(j)} + \mathbf{A}_{\mathbf{C}_1}^{(j)}\right), \nabla\widetilde{\mathcal{L}}\left(\mathbf{A}^*\right) \right\rangle\right| \\
&- \left|\left\langle \left(\mathbf{A}_{\mathbf{D}_2}^{(j)} + \mathbf{A}_{\mathbf{R}_2}^{(j)} + \mathbf{A}_{\mathbf{P}_2}^{(j)} + \mathbf{A}_{\mathbf{C}_2}^{(j)}\right) \otimes \mathbf{A}_1^{(j)}, \nabla\widetilde{\mathcal{L}}\left(\mathbf{A}^*\right) \right\rangle\right|.
\end{aligned} \tag{S3.13}$$

For the fourth term in (S3.13), plugging in $b=\phi^{1/3}$:

$$\begin{aligned}
&\left|\left\langle \mathbf{A}_2^{(j)} \otimes \left(\mathbf{A}_{\mathbf{D}_1}^{(j)}+\mathbf{A}_{\mathbf{R}_1}^{(j)}+\mathbf{A}_{\mathbf{P}_1}^{(j)}+\mathbf{A}_{\mathbf{C}_1}^{(j)}\right), \nabla\widetilde{\mathcal{L}}\left(\mathbf{A}^*\right)\right\rangle\right| \\
\leq&\left|\left\langle \mathbf{A}_2^{(j)} \otimes \left(\mathbf{A}_1^{(j)} - \left[\mathbf{C}_1^* \ \mathbf{R}_1^*\right]\mathbf{O}_{1,u}^{(j)}\mathbf{D}_1^{(j)}\left[\mathbf{C}_1^{(j)} \ \mathbf{P}_1^{(j)}\right]^\top\right), \nabla\widetilde{\mathcal{L}}\left(\mathbf{A}^*\right)\right\rangle\right| \\
&+\left|\left\langle \mathbf{A}_2^{(j)} \otimes \left(\mathbf{A}_1^{(j)} - \left[\mathbf{C}_1^{(j)} \ \mathbf{R}_1^{(j)}\right]\mathbf{D}_1^{(j)}\mathbf{O}_{1,v}^{(j)\top}\left[\mathbf{C}_1^* \ \mathbf{P}_1^*\right]^\top\right), \nabla\widetilde{\mathcal{L}}\left(\mathbf{A}^*\right)\right\rangle\right| \\
&+\left|\left\langle \mathbf{A}_2^{(j)} \otimes \left(\mathbf{A}_1^{(j)} - \left[\mathbf{C}_1^{(j)} \ \mathbf{R}_1^{(j)}\right]\mathbf{O}_{1,u}^{(j)\top}\mathbf{D}_1^*\mathbf{O}_{1,v}^{(j)}\left[\mathbf{C}_1^{(j)} \ \mathbf{P}_1^{(j)}\right]^\top\right), \nabla\widetilde{\mathcal{L}}\left(\mathbf{A}^*\right)\right\rangle\right| \\
\leq&\xi(r_1,r_2,d_1,d_2)\left(\left\|\mathbf{A}_2^{(j)}\right\|_{\mathrm{F}}\left\|\mathbf{D}_1^{(j)}\right\|_{\mathrm{op}}\cdot\left\|\left[\mathbf{C}_1^{(j)} \ \mathbf{P}_1^{(j)}\right]\right\|_{\mathrm{op}}\cdot\left\|\left[\mathbf{C}_1^{(j)} \ \mathbf{R}_1^{(j)}\right]-\left[\mathbf{C}_1^* \ \mathbf{R}_1^*\right]\mathbf{O}_{1,u}^{(j)}\right\|_{\mathrm{F}}\right) \\
&+\xi(r_1,r_2,d_1,d_2)\left(\left\|\mathbf{A}_2^{(j)}\right\|_{\mathrm{F}}\left\|\mathbf{D}_1^{(j)}\right\|_{\mathrm{op}}\cdot\left\|\left[\mathbf{C}_1^{(j)} \ \mathbf{R}_1^{(j)}\right]\right\|_{\mathrm{op}}\cdot\left\|\left[\mathbf{C}_1^{(j)} \ \mathbf{P}_1^{(j)}\right]-\left[\mathbf{C}_1^* \ \mathbf{P}_1^*\right]\mathbf{O}_{1,v}^{(j)}\right\|_{\mathrm{F}}\right) \\
&+\xi(r_1,r_2,d_1,d_2)\left(\left\|\mathbf{A}_2^{(j)}\right\|_{\mathrm{F}}\left\|\left[\mathbf{C}_1^{(j)} \ \mathbf{R}_1^{(j)}\right]\right\|_{\mathrm{op}}\cdot\left\|\left[\mathbf{C}_1^{(j)} \ \mathbf{P}_1^{(j)}\right]\right\|_{\mathrm{op}}\cdot\left\|\mathbf{D}_1^{(j)}-\mathbf{O}_{1,u}^{(j)\top}\mathbf{D}_1^*\mathbf{O}_{1,v}^{(j)}\right\|_{\mathrm{F}}\right) \\
\leq&\left(2b^2+4\phi/b\right)\phi\xi(r_1,r_2,d_1,d_2)\mathrm{dist}_{(j)} \\
=&6\phi^{5/3}\xi(r_1,r_2,d_1,d_2)\mathrm{dist}_{(j)}.
\end{aligned}$$

Applying the same analysis on the last term in (S3.13), the last two terms in (S3.13) can be upper bounded by

$$\begin{aligned}
&\left|\left\langle \mathbf{A}_2^{(j)} \otimes \left(\mathbf{A}_{\mathbf{D}_1}^{(j)}+\mathbf{A}_{\mathbf{R}_1}^{(j)}+\mathbf{A}_{\mathbf{P}_1}^{(j)}+\mathbf{A}_{\mathbf{C}_1}^{(j)}\right), \nabla\widetilde{\mathcal{L}}\left(\mathbf{A}^*\right)\right\rangle\right| \\
&+\left|\left\langle \left(\mathbf{A}_{\mathbf{D}_2}^{(j)}+\mathbf{A}_{\mathbf{R}_2}^{(j)}+\mathbf{A}_{\mathbf{P}_2}^{(j)}+\mathbf{A}_{\mathbf{C}_2}^{(j)}\right)\otimes\mathbf{A}_1^{(j)}, \nabla\widetilde{\mathcal{L}}\left(\mathbf{A}^*\right)\right\rangle\right| \\
\leq&6\phi^{5/3}\xi(r_1,r_2,d_1,d_2)\mathrm{dist}_{(j)}+6\phi^{5/3}\xi(r_1,r_2,d_1,d_2)\mathrm{dist}_{(j)} \\
=&12\phi^{5/3}\xi(r_1,r_2,d_1,d_2)\mathrm{dist}_{(j)} \\
\leq&36c\phi^{10/3}\mathrm{dist}_{(j)}^2+\frac{1}{c}\xi(r_1,r_2,d_1,d_2)^2, \quad \forall c>0.
\end{aligned}$$

The last inequality stems from the fact that $x^2+y^2\geq 2xy$.

For the third term in (S3.13), we know from Lemma S.1 that $\left\|\mathbf{H}^{(j)}\right\|_{\mathrm{F}}\leq C_h\phi^{4/3}\mathrm{dist}_{(j)}^2$, and with condition (S3.3), $\text{dist}_{(j)}^2 \le C_D\phi^{2/3}\alpha\beta^{-1}\kappa^{-2}$, and

$$\begin{aligned}
&\left\|\mathbf{H}^{(j)}\right\|_{\mathrm{F}}\left\|\nabla\widetilde{\mathcal{L}}(\mathbf{A}^{(j)})-\nabla\widetilde{\mathcal{L}}(\mathbf{A}^*)\right\|_{\mathrm{F}}\\
\le&\frac{1}{4\beta}\left\|\nabla\widetilde{\mathcal{L}}(\mathbf{A}^{(j)})-\nabla\widetilde{\mathcal{L}}(\mathbf{A}^*)\right\|_{\mathrm{F}}^2+\beta\left\|\mathbf{H}^{(j)}\right\|_{\mathrm{F}}^2\\
\le&\frac{1}{4\beta}\left\|\nabla\widetilde{\mathcal{L}}(\mathbf{A}^{(j)})-\nabla\widetilde{\mathcal{L}}(\mathbf{A}^*)\right\|_{\mathrm{F}}^2+\beta(C_h^2\phi^{8/3}\text{dist}_{(j)}^2)\text{dist}_{(j)}^2\\
\le&\frac{1}{4\beta}\left\|\nabla\widetilde{\mathcal{L}}(\mathbf{A}^{(j)})-\nabla\widetilde{\mathcal{L}}(\mathbf{A}^*)\right\|_{\mathrm{F}}^2+\frac{C_H\alpha\phi^{10/3}}{\kappa^2}\text{dist}_{(j)}^2,
\end{aligned}$$

where $C_H = C_h^2 C_D$ is a small positive constant. Consequently, putting together the bounds, we have:

$$\begin{aligned}
&\left\langle \begin{matrix}\mathbf{A}_2^{(j)}\otimes\left(\mathbf{A}_{\mathbf{D}_1}^{(j)}+\mathbf{A}_{\mathbf{R}_1}^{(j)}+\mathbf{A}_{\mathbf{P}_1}^{(j)}+\mathbf{A}_{\mathbf{C}_1}^{(j)}\right)\\ +\left(\mathbf{A}_{\mathbf{D}_2}^{(j)}+\mathbf{A}_{\mathbf{R}_2}^{(j)}+\mathbf{A}_{\mathbf{P}_2}^{(j)}+\mathbf{A}_{\mathbf{C}_2}^{(j)}\right)\otimes\mathbf{A}_1^{(j)}\end{matrix}, \nabla\widetilde{\mathcal{L}}\left(\mathbf{A}^{(j)}\right)\right\rangle\\
\ge&\frac{\alpha}{2}\left\|\mathbf{A}_2^{(j)}\otimes\mathbf{A}_1^{(j)}-\mathbf{A}_2^*\otimes\mathbf{A}_1^*\right\|_{\mathrm{F}}^2+\frac{1}{4\beta}\left\|\nabla\widetilde{\mathcal{L}}(\mathbf{A}^{(j)})-\nabla\widetilde{\mathcal{L}}(\mathbf{A}^*)\right\|_{\mathrm{F}}^2\\
&-\frac{C_H\alpha\phi^{10/3}}{\kappa^2}\text{dist}_{(j)}^2-36c\phi^{10/3}\text{dist}_{(j)}^2-\frac{1}{c}\xi(r_1,r_2,d_1,d_2)^2.
\end{aligned}$$

Step 3.2. (Lower bound for $\sum_{i=1}^2\left(G_{\mathbf{D}_i}+G_{\mathbf{R}_i}+G_{\mathbf{P}_i}+G_{\mathbf{C}_i}\right)$)

Recall the definitions in (S3.7) and (S3.8). It can be easily verified that, by adding the terms together, we have

$$\begin{aligned}
&G_{\mathbf{D}_1}+G_{\mathbf{R}_1}+G_{\mathbf{P}_1}+G_{\mathbf{C}_1}\\
=&\left\langle\begin{matrix}\mathbf{A}_1^{(j)}-\left[\mathbf{C}_1^*\ \mathbf{R}_1^*\right]\mathbf{O}_{1,u}\mathbf{D}_1^{(j)}\left[\mathbf{C}_1^{(j)}\ \mathbf{P}_1^{(j)}\right]^\top\\ +\mathbf{A}_1^{(j)}-\left[\mathbf{C}_1^{(j)}\ \mathbf{R}_1^{(j)}\right]\mathbf{O}_{1,u}^\top\mathbf{D}_1^*\mathbf{O}_{1,v}\left[\mathbf{C}_1^{(j)}\ \mathbf{P}_1^{(j)}\right]^\top\\ +\mathbf{A}_1^{(j)}-\left[\mathbf{C}_1^{(j)}\ \mathbf{R}_1^{(j)}\right]\mathbf{D}_1^{(j)}\mathbf{O}_{1,v}^\top\left[\mathbf{C}_1^*\ \mathbf{P}_1^*\right]^\top\end{matrix}, \left(\left\|\mathbf{A}_1^{(j)}\right\|_{\mathrm{F}}^2-\left\|\mathbf{A}_2^{(j)}\right\|_{\mathrm{F}}^2\right)\mathbf{A}_1^{(j)}\right\rangle \qquad \text{(S3.14)}\\
:=&\left\langle\mathbf{H}_{\mathbf{A}_1,1}, \left(\left\|\mathbf{A}_1^{(j)}\right\|_{\mathrm{F}}^2-\left\|\mathbf{A}_2^{(j)}\right\|_{\mathrm{F}}^2\right)\mathbf{A}_1^{(j)}\right\rangle.
\end{aligned}$$

The left side of the inner product, $\mathbf{H}_{\mathbf{A}_1,1}$, contains the first order perturbation terms with respect to $\mathbf{A}_1$. By Lemma A.1 in Wang et al. (2023), it is exactly $\mathbf{A}_1^{(j)}-\mathbf{A}_1^*+\mathbf{H}_{\mathbf{A}_1,2}$, where $\mathbf{H}_{\mathbf{A}_1,2}$ comprises all the second and third-order perturbation terms. Applying it with

$b = \phi^{1/3}$, we have $\|\mathbf{H}_{\mathbf{A}_1,2}\|_{\mathrm{F}} \le 5\phi^{1/3}\mathrm{dist}^2_{(j)}$. Similarly for $\mathbf{A}_2$, we have

$$
\begin{aligned}
&G_{\mathbf{D}_2} + G_{\mathbf{R}_2} + G_{\mathbf{P}_2} + G_{\mathbf{C}_2} \\
&= \left\langle \begin{matrix} \mathbf{A}_2^{(j)} - [\mathbf{C}_2^* \ \mathbf{R}_2^*]\, \mathbf{O}_{2,u} \mathbf{D}_2^{(j)} \left[\mathbf{C}_2^{(j)} \ \mathbf{P}_2^{(j)}\right]^\top \\ + \mathbf{A}_2^{(j)} - \left[\mathbf{C}_2^{(j)} \ \mathbf{R}_2^{(j)}\right] \mathbf{O}_{2,u}^\top \mathbf{D}_2^* \mathbf{O}_{2,v} \left[\mathbf{C}_2^{(j)} \ \mathbf{P}_2^{(j)}\right]^\top \\ + \mathbf{A}_2^{(j)} - \left[\mathbf{C}_2^{(j)} \ \mathbf{R}_2^{(j)}\right] \mathbf{D}_2^{(j)} \mathbf{O}_{2,v}^\top \left[\mathbf{C}_2^* \ \mathbf{P}_2^*\right]^\top \end{matrix}, - \left(\left\|\mathbf{A}_1^{(j)}\right\|_{\mathrm{F}}^2 - \left\|\mathbf{A}_2^{(j)}\right\|_{\mathrm{F}}^2 \right) \mathbf{A}_2^{(j)} \right\rangle \\
&:= \left\langle \mathbf{H}_{\mathbf{A}_2,1}, - \left(\left\|\mathbf{A}_1^{(j)}\right\|_{\mathrm{F}}^2 - \left\|\mathbf{A}_2^{(j)}\right\|_{\mathrm{F}}^2 \right) \mathbf{A}_2^{(j)} \right\rangle \\
&= \left\langle \mathbf{A}_2^{(j)} - \mathbf{A}_2^* + \mathbf{H}_{\mathbf{A}_2,2}, - \left(\left\|\mathbf{A}_1^{(j)}\right\|_{\mathrm{F}}^2 - \left\|\mathbf{A}_2^{(j)}\right\|_{\mathrm{F}}^2 \right) \mathbf{A}_2^{(j)} \right\rangle,
\end{aligned}
\tag{S3.15}
$$

and $\|\mathbf{H}_{\mathbf{A}_2,2}\|_{\mathrm{F}} \le 5\phi^{1/3}\mathrm{dist}^2_{(j)}$.

Now we define

$$
\mathbf{Z} = \begin{pmatrix} \mathrm{vec}(\mathbf{A}_1^{(j)}) \\ \mathrm{vec}(\mathbf{A}_2^{(j)}) \end{pmatrix}, \widetilde{\mathbf{Z}} = \begin{pmatrix} \mathrm{vec}(\mathbf{A}_1^{(j)}) \\ -\mathrm{vec}(\mathbf{A}_2^{(j)}) \end{pmatrix}, \mathbf{Z}^* = \begin{pmatrix} \mathrm{vec}(\mathbf{A}_1^*) \\ \mathrm{vec}(\mathbf{A}_2^*) \end{pmatrix}, \widetilde{\mathbf{Z}}^* = \begin{pmatrix} \mathrm{vec}(\mathbf{A}_1^*) \\ -\mathrm{vec}(\mathbf{A}_2^*) \end{pmatrix}.
$$

Then, one can show that

$$
\left\|\mathbf{A}_1^{(j)}\right\|_{\mathrm{F}}^2 - \left\|\mathbf{A}_2^{(j)}\right\|_{\mathrm{F}}^2 = \mathbf{Z}^\top \widetilde{\mathbf{Z}} = \widetilde{\mathbf{Z}}^\top \mathbf{Z}.
$$

Vectorizing the matrices and putting (S3.14) and (S3.15) together, we have

$$
\begin{aligned}
&\sum_{i=1}^{2} \left(G_{\mathbf{D}_i} + G_{\mathbf{R}_i} + G_{\mathbf{P}_i} + G_{\mathbf{C}_i} \right) \\
&= \left\langle \mathbf{Z} - \mathbf{Z}^*, \widetilde{\mathbf{Z}} \widetilde{\mathbf{Z}}^\top \mathbf{Z} \right\rangle \\
&+ \left\langle \mathbf{H}_{\mathbf{A}_1,2}, \left(\left\|\mathbf{A}_1^{(j)}\right\|_{\mathrm{F}}^2 - \left\|\mathbf{A}_2^{(j)}\right\|_{\mathrm{F}}^2 \right) \mathbf{A}_1^{(j)} \right\rangle + \left\langle \mathbf{H}_{\mathbf{A}_2,2}, - \left(\left\|\mathbf{A}_1^{(j)}\right\|_{\mathrm{F}}^2 - \left\|\mathbf{A}_2^{(j)}\right\|_{\mathrm{F}}^2 \right) \mathbf{A}_2^{(j)} \right\rangle.
\end{aligned}
$$

The second and third terms can be lower bounded by

$$
\begin{aligned}
&\left\langle \mathbf{H}_{\mathbf{A}_1,2}, \left(\left\|\mathbf{A}_1^{(j)}\right\|_{\mathrm{F}}^2 - \left\|\mathbf{A}_2^{(j)}\right\|_{\mathrm{F}}^2 \right) \mathbf{A}_1^{(j)} \right\rangle + \left\langle \mathbf{H}_{\mathbf{A}_2,2}, - \left(\left\|\mathbf{A}_1^{(j)}\right\|_{\mathrm{F}}^2 - \left\|\mathbf{A}_2^{(j)}\right\|_{\mathrm{F}}^2 \right) \mathbf{A}_2^{(j)} \right\rangle \\
&\ge - \left| \left\|\mathbf{A}_1^{(j)}\right\|_{\mathrm{F}}^2 - \left\|\mathbf{A}_2^{(j)}\right\|_{\mathrm{F}}^2 \right| \left(\left\|\mathbf{A}_1^{(j)}\right\|_{\mathrm{F}} \|\mathbf{H}_{\mathbf{A}_1,2}\|_{\mathrm{F}} + \left\|\mathbf{A}_2^{(j)}\right\|_{\mathrm{F}} \|\mathbf{H}_{\mathbf{A}_2,2}\|_{\mathrm{F}} \right) \\
&\ge - 20\phi^{4/3}\mathrm{dist}^2_{(j)} \left| \left\|\mathbf{A}_1^{(j)}\right\|_{\mathrm{F}}^2 - \left\|\mathbf{A}_2^{(j)}\right\|_{\mathrm{F}}^2 \right|.
\end{aligned}
$$

For the first term, noting that $\mathbf{Z}^{*\top}\widetilde{\mathbf{Z}}^{*} = \widetilde{\mathbf{Z}}^{*\top}\mathbf{Z}^{*} = 0$ and $\widetilde{\mathbf{Z}}^{\top}\mathbf{Z}^{*} = \mathbf{Z}^{\top}\widetilde{\mathbf{Z}}^{*}$, we have

$$
\begin{aligned}
&\left\langle \mathbf{Z} - \mathbf{Z}^{*}, \widetilde{\mathbf{Z}}\widetilde{\mathbf{Z}}^{\top}\mathbf{Z} \right\rangle \\
=&\left\langle \widetilde{\mathbf{Z}}^{\top}\mathbf{Z} - \widetilde{\mathbf{Z}}^{\top}\mathbf{Z}^{*}, \widetilde{\mathbf{Z}}^{\top}\mathbf{Z} \right\rangle \\
=&\frac{1}{2}\left\|\widetilde{\mathbf{Z}}^{\top}\mathbf{Z}\right\|_{\mathrm{F}}^{2} + \frac{1}{2}\left\langle \widetilde{\mathbf{Z}}^{\top}\mathbf{Z} - 2\widetilde{\mathbf{Z}}^{\top}\mathbf{Z}^{*}, \widetilde{\mathbf{Z}}^{\top}\mathbf{Z} \right\rangle \\
=&\frac{1}{2}\left\|\widetilde{\mathbf{Z}}^{\top}\mathbf{Z}\right\|_{\mathrm{F}}^{2} + \frac{1}{2}\left\langle \widetilde{\mathbf{Z}}^{\top}(\mathbf{Z} - \mathbf{Z}^{*}), \widetilde{\mathbf{Z}}^{\top}\mathbf{Z} \right\rangle + \frac{1}{2}\left\langle -\widetilde{\mathbf{Z}}^{\top}\mathbf{Z}^{*}, \widetilde{\mathbf{Z}}^{\top}\mathbf{Z} \right\rangle \\
=&\frac{1}{2}\left\|\widetilde{\mathbf{Z}}^{\top}\mathbf{Z}\right\|_{\mathrm{F}}^{2} + \frac{1}{2}\left\langle \widetilde{\mathbf{Z}}^{\top}(\mathbf{Z} - \mathbf{Z}^{*}), \widetilde{\mathbf{Z}}^{\top}\mathbf{Z} \right\rangle + \frac{1}{2}\left\langle \mathbf{Z}^{*\top}\widetilde{\mathbf{Z}}^{*} - \mathbf{Z}^{\top}\widetilde{\mathbf{Z}}^{*}, \widetilde{\mathbf{Z}}^{\top}\mathbf{Z} \right\rangle \\
=&\frac{1}{2}\left\|\widetilde{\mathbf{Z}}^{\top}\mathbf{Z}\right\|_{\mathrm{F}}^{2} + \frac{1}{2}\left\langle (\mathbf{Z} - \mathbf{Z}^{*})^{\top}(\widetilde{\mathbf{Z}} - \widetilde{\mathbf{Z}}^{*}), \widetilde{\mathbf{Z}}^{\top}\mathbf{Z} \right\rangle \\
\geq&\frac{1}{2}\left\|\widetilde{\mathbf{Z}}^{\top}\mathbf{Z}\right\|_{\mathrm{F}}^{2} - \frac{1}{2}\left\|\mathbf{Z} - \mathbf{Z}^{*}\right\|_{\mathrm{F}}\left\|\widetilde{\mathbf{Z}} - \widetilde{\mathbf{Z}}^{*}\right\|_{\mathrm{F}}\left\|\widetilde{\mathbf{Z}}^{\top}\mathbf{Z}\right\|_{\mathrm{F}}
\end{aligned}
$$

Note that $\|\mathbf{Z} - \mathbf{Z}^{*}\|_{\mathrm{F}}^{2} = \left\|\widetilde{\mathbf{Z}} - \widetilde{\mathbf{Z}}^{*}\right\|_{\mathrm{F}}^{2} = \left\|\mathbf{A}_{1}^{(j)} - \mathbf{A}_{1}^{*}\right\|_{\mathrm{F}}^{2} + \left\|\mathbf{A}_{2}^{(j)} - \mathbf{A}_{2}^{*}\right\|_{\mathrm{F}}^{2}$. By Lemma S.3 with $c_b \leq 0.01$, we have $\left\|\mathbf{A}_{1}^{(j)} - \mathbf{A}_{1}^{*}\right\|_{\mathrm{F}}^{2} + \left\|\mathbf{A}_{2}^{(j)} - \mathbf{A}_{2}^{*}\right\|_{\mathrm{F}}^{2} \leq 50\phi^{4/3}\mathrm{dist}_{(j)}^{2}$, and hence,

$$
\left\langle \mathbf{Z} - \mathbf{Z}^{*}, \widetilde{\mathbf{Z}}\widetilde{\mathbf{Z}}^{\top}\mathbf{Z} \right\rangle \geq \frac{1}{2}\left(\left\|\mathbf{A}_{1}^{(j)}\right\|_{\mathrm{F}}^{2} - \left\|\mathbf{A}_{2}^{(j)}\right\|_{\mathrm{F}}^{2}\right)^{2} - 25\phi^{4/3}\mathrm{dist}_{(j)}^{2}\left|\left\|\mathbf{A}_{1}^{(j)}\right\|_{\mathrm{F}}^{2} - \left\|\mathbf{A}_{2}^{(j)}\right\|_{\mathrm{F}}^{2}\right|.
$$

Combining the two pieces, we have the lower bound of G terms:

$$
\begin{aligned}
&\sum_{i=1}^{2}\left(G_{\mathbf{D}_i} + G_{\mathbf{R}_i} + G_{\mathbf{P}_i} + G_{\mathbf{C}_i}\right) \\
\geq&\frac{1}{2}\left(\left\|\mathbf{A}_{1}^{(j)}\right\|_{\mathrm{F}}^{2} - \left\|\mathbf{A}_{2}^{(j)}\right\|_{\mathrm{F}}^{2}\right)^{2} - 45\phi^{4/3}\mathrm{dist}_{(j)}^{2}\left|\left\|\mathbf{A}_{1}^{(j)}\right\|_{\mathrm{F}}^{2} - \left\|\mathbf{A}_{2}^{(j)}\right\|_{\mathrm{F}}^{2}\right| \\
\geq&\frac{1}{4}\left(\left\|\mathbf{A}_{1}^{(j)}\right\|_{\mathrm{F}}^{2} - \left\|\mathbf{A}_{2}^{(j)}\right\|_{\mathrm{F}}^{2}\right)^{2} - 2025\phi^{8/3}\mathrm{dist}_{(j)}^{4} \\
\geq&\frac{1}{4}\left(\left\|\mathbf{A}_{1}^{(j)}\right\|_{\mathrm{F}}^{2} - \left\|\mathbf{A}_{2}^{(j)}\right\|_{\mathrm{F}}^{2}\right)^{2} - \frac{2025C_D}{\kappa^{2}}\phi^{10/3}\mathrm{dist}_{(j)}^{2}.
\end{aligned}
$$

Step 3.3. (Lower bound for $\sum_{i=1}^{2}\left(T_{\mathbf{R}_i} + T_{\mathbf{P}_i} + T_{\mathbf{C}_i}\right)$)

Similarly to Wang et al. (2023), it can be verified that

$$
\begin{aligned}
&T_{\mathbf{R}_1}+T_{\mathbf{P}_1}+T_{\mathbf{C}_1}+T_{\mathbf{R}_2}+T_{\mathbf{P}_2}+T_{\mathbf{C}_2}\\
=&\left\langle\left[\mathbf{C}_1^{(j)}\ \mathbf{R}_1^{(j)}\right]-\left[\mathbf{C}_1^*\ \mathbf{R}_1^*\right]\mathbf{O}_{1,u}^{(j)},\left[\mathbf{C}_1^{(j)}\ \mathbf{R}_1^{(j)}\right]\left(\left[\mathbf{C}_1^{(j)}\ \mathbf{R}_1^{(j)}\right]^\top\left[\mathbf{C}_1^{(j)}\ \mathbf{R}_1^{(j)}\right]-b^2\mathbf{I}_{r_1}\right)\right\rangle\\
&+\left\langle\left[\mathbf{C}_1^{(j)}\ \mathbf{P}_1^{(j)}\right]-\left[\mathbf{C}_1^*\ \mathbf{P}_1^*\right]\mathbf{O}_{1,v}^{(j)},\left[\mathbf{C}_1^{(j)}\ \mathbf{P}_1^{(j)}\right]\left(\left[\mathbf{C}_1^{(j)}\ \mathbf{P}_1^{(j)}\right]^\top\left[\mathbf{C}_1^{(j)}\ \mathbf{P}_1^{(j)}\right]-b^2\mathbf{I}_{r_1}\right)\right\rangle\\
&+\left\langle\left[\mathbf{C}_2^{(j)}\ \mathbf{R}_2^{(j)}\right]-\left[\mathbf{C}_2^*\ \mathbf{R}_2^*\right]\mathbf{O}_{2,u}^{(j)},\left[\mathbf{C}_2^{(j)}\ \mathbf{R}_2^{(j)}\right]\left(\left[\mathbf{C}_2^{(j)}\ \mathbf{R}_2^{(j)}\right]^\top\left[\mathbf{C}_2^{(j)}\ \mathbf{R}_2^{(j)}\right]-b^2\mathbf{I}_{r_2}\right)\right\rangle\\
&+\left\langle\left[\mathbf{C}_2^{(j)}\ \mathbf{P}_2^{(j)}\right]-\left[\mathbf{C}_2^*\ \mathbf{P}_2^*\right]\mathbf{O}_{2,v}^{(j)},\left[\mathbf{C}_2^{(j)}\ \mathbf{P}_2^{(j)}\right]\left(\left[\mathbf{C}_2^{(j)}\ \mathbf{P}_2^{(j)}\right]^\top\left[\mathbf{C}_2^{(j)}\ \mathbf{P}_2^{(j)}\right]-b^2\mathbf{I}_{r_2}\right)\right\rangle.
\end{aligned}
\tag{S3.16}
$$

Denote $\mathbf{U}_1^{(j)}=[\mathbf{C}_1^{(j)}\ \mathbf{R}_1^{(j)}],\mathbf{V}_1^{(j)}=[\mathbf{C}_1^{(j)}\ \mathbf{P}_1^{(j)}],\mathbf{U}_1^*=[\mathbf{C}_1^*\ \mathbf{R}_1^*],\mathbf{V}_1^*=[\mathbf{C}_1^*\ \mathbf{P}_1^*]$. Recall that $\mathbf{U}_1^{*\top}\mathbf{U}_1^*=b^2\mathbf{I}_{r_1}$ and $\mathbf{V}_1^{*\top}\mathbf{V}_1^*=b^2\mathbf{I}_{r_1}$, for the first term we have

$$
\begin{aligned}
&\left\langle\mathbf{U}_1^{(j)}-\mathbf{U}_1^*\mathbf{O}_{1,u}^{(j)},\mathbf{U}_1^{(j)}\left(\mathbf{U}_1^{(j)\top}\mathbf{U}_1^{(j)}-b^2\mathbf{I}_{r_1}\right)\right\rangle\\
=&\frac{1}{2}\left\langle\mathbf{U}_1^{(j)\top}\mathbf{U}_1^{(j)}-\mathbf{U}_1^{*\top}\mathbf{U}_1^*,\mathbf{U}_1^{(j)\top}\mathbf{U}_1^{(j)}-b^2\mathbf{I}_{r_1}\right\rangle+\frac{1}{2}\left\langle\mathbf{U}_1^{(j)\top}(\mathbf{U}_1^{(j)}-\mathbf{U}_1^*\mathbf{O}_{1,u}^{(j)}),\mathbf{U}_1^{(j)\top}\mathbf{U}_1^{(j)}-b^2\mathbf{I}_{r_1}\right\rangle\\
&+\frac{1}{2}\left\langle\mathbf{U}_1^{*\top}\mathbf{U}_1^*-\mathbf{U}_1^{(j)\top}\mathbf{U}_1^*\mathbf{O}_{1,u}^{(j)},\mathbf{U}_1^{(j)\top}\mathbf{U}_1^{(j)}-b^2\mathbf{I}_{r_1}\right\rangle.
\end{aligned}
$$

Since $\mathbf{U}_1^{(j)\top}\mathbf{U}_1^{(j)}-b^2\mathbf{I}_{r_1}$ is symmetric, we have

$$
\begin{aligned}
&\left\langle\mathbf{U}_1^{*\top}\mathbf{U}_1^*-\mathbf{U}_1^{(j)\top}\mathbf{U}_1^*\mathbf{O}_{1,u}^{(j)},\mathbf{U}_1^{(j)\top}\mathbf{U}_1^{(j)}-b^2\mathbf{I}_{r_1}\right\rangle\\
=&\left\langle\mathbf{U}_1^{*\top}\mathbf{U}_1^*-\mathbf{O}_{1,u}^{(j)\top}\mathbf{U}_1^{*\top}\mathbf{U}_1^{(j)},\mathbf{U}_1^{(j)\top}\mathbf{U}_1^{(j)}-b^2\mathbf{I}_{r_1}\right\rangle\\
=&\left\langle\mathbf{O}_{1,u}^{(j)\top}\mathbf{U}_1^{*\top}(\mathbf{U}_1^*\mathbf{O}_{1,u}^{(j)}-\mathbf{U}_1^{(j)}),\mathbf{U}_1^{(j)\top}\mathbf{U}_1^{(j)}-b^2\mathbf{I}_{r_1}\right\rangle.
\end{aligned}
$$

Therefore, with condition (S3.3),

$$
\begin{aligned}
&\left\langle\mathbf{U}_1^{(j)}-\mathbf{U}_1^*\mathbf{O}_{1,u}^{(j)},\mathbf{U}_1^{(j)}\left(\mathbf{U}_1^{(j)\top}\mathbf{U}_1^{(j)}-b^2\mathbf{I}_{r_1}\right)\right\rangle\\
=&\frac{1}{2}\left\|\mathbf{U}_1^{(j)\top}\mathbf{U}_1^{(j)}-b^2\mathbf{I}_{r_1}\right\|_{\mathrm{F}}^2+\frac{1}{2}\left\langle(\mathbf{U}_1^{(j)}-\mathbf{U}_1^*\mathbf{O}_{1,u}^{(j)})^\top(\mathbf{U}_1^{(j)}-\mathbf{U}_1^*\mathbf{O}_{1,u}^{(j)}),\mathbf{U}_1^{(j)\top}\mathbf{U}_1^{(j)}-b^2\mathbf{I}_{r_1}\right\rangle\\
\geq&\frac{1}{2}\left\|\mathbf{U}_1^{(j)\top}\mathbf{U}_1^{(j)}-b^2\mathbf{I}_{r_1}\right\|_{\mathrm{F}}^2-\frac{1}{2}\mathrm{dist}_{(j)}^2\left\|\mathbf{U}_1^{(j)\top}\mathbf{U}_1^{(j)}-b^2\mathbf{I}_{r_1}\right\|_{\mathrm{F}}\\
\geq&\frac{1}{4}\left\|\mathbf{U}_1^{(j)\top}\mathbf{U}_1^{(j)}-b^2\mathbf{I}_{r_1}\right\|_{\mathrm{F}}^2-\frac{1}{4}\mathrm{dist}_{(j)}^4\\
\geq&\frac{1}{4}\left\|\mathbf{U}_1^{(j)\top}\mathbf{U}_1^{(j)}-b^2\mathbf{I}_{r_1}\right\|_{\mathrm{F}}^2-\frac{C_D}{4}\phi^{2/3}\mathrm{dist}_{(j)}^2.
\end{aligned}
$$

Adding the lower bounds of the four terms in (S3.16) together, we have

$$
\begin{aligned}
& T_{\mathbf{R}_1}+T_{\mathbf{P}_1}+T_{\mathbf{C}_1}+T_{\mathbf{R}_2}+T_{\mathbf{P}_2}+T_{\mathbf{C}_2} \\
\geq & \frac{1}{4} \sum_{i=1}^2\left(\left\|\left[\mathbf{C}_i^{(j)}\ \mathbf{R}_i^{(j)}\right]^{\top}\left[\mathbf{C}_i^{(j)}\ \mathbf{R}_i^{(j)}\right]-b^2 \mathbf{I}_{r_i}\right\|_{\mathrm{F}}^2+\left\|\left[\mathbf{C}_i^{(j)}\ \mathbf{P}_i^{(j)}\right]^{\top}\left[\mathbf{C}_i^{(j)}\ \mathbf{P}_i^{(j)}\right]-b^2 \mathbf{I}_{r_i}\right\|_{\mathrm{F}}^2\right) \\
& -C_D \phi^{2/3} \operatorname{dist}_{(j)}^2 .
\end{aligned}
$$

Step 3.4. (Upper bound for $\sum_{i=1}^2\left(Q_{\mathbf{D}_i, 2}+Q_{\mathbf{R}_i, 2}+Q_{\mathbf{P}_i, 2}+Q_{\mathbf{C}_i, 2}\right)$)

Following the definitions and plugging in $b=\phi^{1/3}$, we have

$$
\begin{aligned}
& \sum_{i=1}^2\left(Q_{\mathbf{D}_i, 2}+Q_{\mathbf{R}_i, 2}+Q_{\mathbf{P}_i, 2}+Q_{\mathbf{C}_i, 2}\right) \\
= & 240 \phi^{10/3}\left(\xi^2\left(r_1, r_2, d_1, d_2\right)+\left\|\nabla \widetilde{\mathcal{L}}\left(\mathbf{A}^{(j)}\right)-\nabla \widetilde{\mathcal{L}}\left(\mathbf{A}^*\right)\right\|_{\mathrm{F}}^2\right) \\
& +\lambda_2^2 \phi^{2/3} \sum_{i=1}^2\left(18\left\|\mathbf{C}_i^{(j)\top} \mathbf{R}_i^{(j)}\right\|_{\mathrm{F}}^2+18\left\|\mathbf{C}_i^{(j)\top} \mathbf{P}_i^{(j)}\right\|_{\mathrm{F}}^2+40\left\|\mathbf{C}_i^{(j)\top} \mathbf{C}_i^{(j)}-b^2 \mathbf{I}_{d_i}\right\|_{\mathrm{F}}^2\right. \\
& \left.+8\left\|\mathbf{R}_i^{(j)\top} \mathbf{R}_i^{(j)}-b^2 \mathbf{I}_{r_i-d_i}\right\|_{\mathrm{F}}^2+8\left\|\mathbf{P}_i^{(j)\top} \mathbf{P}_i^{(j)}-b^2 \mathbf{I}_{r_i-d_i}\right\|_{\mathrm{F}}^2\right) \\
& +120 \lambda_1^2 \phi^{10/3}\left(\left\|\mathbf{A}_1^{(j)}\right\|_{\mathrm{F}}^2-\left\|\mathbf{A}_2^{(j)}\right\|_{\mathrm{F}}^2\right)^2 \\
\leq & 240 \phi^{10/3}\left(\xi^2\left(r_1, r_2, d_1, d_2\right)+\left\|\nabla \widetilde{\mathcal{L}}\left(\mathbf{A}^{(j)}\right)-\nabla \widetilde{\mathcal{L}}\left(\mathbf{A}^*\right)\right\|_{\mathrm{F}}^2\right) \\
& +20 \lambda_2^2 \phi^{2/3} \sum_{i=1}^2\left(\left\|\left[\mathbf{C}_i^{(j)}\ \mathbf{R}_i^{(j)}\right]^{\top}\left[\mathbf{C}_i^{(j)}\ \mathbf{R}_i^{(j)}\right]-b^2 \mathbf{I}_{r_i}\right\|_{\mathrm{F}}^2+\left\|\left[\mathbf{C}_i^{(j)}\ \mathbf{P}_i^{(j)}\right]^{\top}\left[\mathbf{C}_i^{(j)}\ \mathbf{P}_i^{(j)}\right]-b^2 \mathbf{I}_{r_i}\right\|_{\mathrm{F}}^2\right) \\
& +120 \lambda_1^2 \phi^{10/3}\left(\left\|\mathbf{A}_1^{(j)}\right\|_{\mathrm{F}}^2-\left\|\mathbf{A}_2^{(j)}\right\|_{\mathrm{F}}^2\right)^2 .
\end{aligned}
$$

Step 3.5. (Lower bound for $\left\|\mathbf{A}_2^{(j)} \otimes \mathbf{A}_1^{(j)}-\mathbf{A}_2^* \otimes \mathbf{A}_1^*\right\|_{\mathrm{F}}^2$)

So far, we have derived bounds for all parts in (S3.11). Combining these pieces, we have

$$
\begin{aligned}
&\mathrm{dist}^2_{(j+1)} - \mathrm{dist}^2_{(j)} \\
&\leq -\alpha\eta \left\| \mathbf{A}_2^{(j)} \otimes \mathbf{A}_1^{(j)} - \mathbf{A}_2^* \otimes \mathbf{A}_1^* \right\|_{\mathrm{F}}^2 \\
&-2\eta \left(-\frac{C_H \alpha \phi^{10/3}}{\kappa^2} - 36c\phi^{10/3} - C_D \lambda_2 \phi^{2/3} - \frac{2025\lambda_1 C_D}{\kappa^2}\phi^{10/3} \right) \mathrm{dist}^2_{(j)} \\
&+ \left(240\eta^2\phi^{10/3} - \frac{\eta}{2\beta} \right) \left\| \nabla \widetilde{\mathcal{L}}(\mathbf{A}^{(j)}) - \nabla \widetilde{\mathcal{L}}(\mathbf{A}^*) \right\|_{\mathrm{F}}^2 \\
&+ \left(240\eta^2\phi^{10/3} + \frac{2\eta}{c} \right) \xi^2(r_1, r_2, d_1, d_2) \\
&+ \left(20\lambda_2^2\eta^2\phi^{2/3} - \frac{\lambda_2\eta}{2} \right) \\
&\times \sum_{i=1}^{2} \left(\left\| \left[\mathbf{C}_i^{(j)} \ \mathbf{R}_i^{(j)} \right]^\top \left[\mathbf{C}_i^{(j)} \ \mathbf{R}_i^{(j)} \right] - b^2 \mathbf{I}_{r_i} \right\|_{\mathrm{F}}^2 + \left\| \left[\mathbf{C}_i^{(j)} \ \mathbf{P}_i^{(j)} \right]^\top \left[\mathbf{C}_i^{(j)} \ \mathbf{P}_i^{(j)} \right] - b^2 \mathbf{I}_{r_i} \right\|_{\mathrm{F}}^2 \right) \\
&+ \left(120\lambda_1^2\eta^2\phi^{10/3} - \frac{1}{2}\lambda_1\eta \right) \left(\left\| \mathbf{A}_1^{(j)} \right\|_{\mathrm{F}}^2 - \left\| \mathbf{A}_2^{(j)} \right\|_{\mathrm{F}}^2 \right)^2 .
\end{aligned}
\tag{S3.17}
$$

Next, we construct a lower bound of $\left\| \mathbf{A}_2^{(j)} \otimes \mathbf{A}_1^{(j)} - \mathbf{A}_2^* \otimes \mathbf{A}_1^* \right\|_{\mathrm{F}}^2$ related to $\mathrm{dist}^2_{(j)}$. Viewing $\mathrm{vec}(\mathbf{A}_2)$ and $\mathrm{vec}(\mathbf{A}_1)$ as the factors in $\mathrm{vec}(\mathbf{A}_2)\mathrm{vec}(\mathbf{A}_1)^\top$, our first regularizer

$$
\left(\|\mathbf{A}_1\|_{\mathrm{F}}^2 - \|\mathbf{A}_2\|_{\mathrm{F}}^2 \right)^2 = \left(\mathrm{vec}(\mathbf{A}_1)^\top \mathrm{vec}(\mathbf{A}_1) - \mathrm{vec}(\mathbf{A}_2)^\top \mathrm{vec}(\mathbf{A}_2) \right)^2
$$

can be regarded as a special case of Wang et al. (2017), as well as a generalization of Tu et al. (2016). Define the distance of two vectors up to a sign switch as

$$
\mathrm{dist}^2(\mathrm{vec}(\mathbf{A}_i^{(j)}), \mathrm{vec}(\mathbf{A}_i^*)) = \min_{s=\pm 1} \left\| \mathrm{vec}(\mathbf{A}_i^{(j)}) - \mathrm{vec}(\mathbf{A}_i^*) s \right\|_{\mathrm{F}}^2, \quad i = 1, 2.
$$

For more details of this type of distance, see Cai and Zhang (2018). By Lemma S.3, since C_D is small,

$$
\left\| \mathbf{A}_1^{(j)} - \mathbf{A}_1^* \right\|_{\mathrm{F}}^2 + \left\| \mathbf{A}_2^{(j)} - \mathbf{A}_2^* \right\|_{\mathrm{F}}^2 \leq 30\phi^{4/3} D \leq 30 C_D \phi^2 \leq \phi^2, \quad i = 1, 2.
$$

Then, we have $\left\| \mathbf{A}_i^{(j)} - \mathbf{A}_i^* \right\|_{\mathrm{F}} \leq \|\mathbf{A}_i^*\|_{\mathrm{F}}$. Hence, we know that the angle between $\mathrm{vec}(\mathbf{A}_i^{(j)})$

and $\mathrm{vec}(\mathbf{A}_i^*)$ is an acute angle. Thus,

$$\mathrm{dist}^2(\mathrm{vec}(\mathbf{A}_i^{(j)}), \mathrm{vec}(\mathbf{A}_i^*)) = \left\| \mathrm{vec}(\mathbf{A}_i^{(j)}) - \mathrm{vec}(\mathbf{A}_i^*) \right\|_{\mathrm{F}}^2, \quad i = 1, 2.$$

Meanwhile, applying the permutation operator $\mathcal{P}$, the Frobenius norm remains unchanged

$$\left\| \mathbf{A}_2^{(j)} \otimes \mathbf{A}_1^{(j)} - \mathbf{A}_2^* \otimes \mathbf{A}_1^* \right\|_{\mathrm{F}}^2 = \left\| \mathrm{vec}(\mathbf{A}_2^{(j)})\mathrm{vec}(\mathbf{A}_1^{(j)})^\top - \mathrm{vec}(\mathbf{A}_2^*)\mathrm{vec}(\mathbf{A}_1^*)^\top \right\|_{\mathrm{F}}^2.$$

Recalling the notations defined in Step 3.2 and applying Lemma S.4, we have

$$\begin{aligned}
&\left\| \mathbf{A}_1^{(j)} - \mathbf{A}_1^* \right\|_{\mathrm{F}}^2 + \left\| \mathbf{A}_2^{(j)} - \mathbf{A}_2^* \right\|_{\mathrm{F}}^2 \\
=&\mathrm{dist}^2(\mathbf{Z}, \mathbf{Z}^*) \\
\leq&\frac{1}{4(\sqrt{2}-1)\phi^2} \left\| \mathbf{Z}\mathbf{Z}^\top - \mathbf{Z}^*\mathbf{Z}^{*\top} \right\|_{\mathrm{F}}^2 \\
=&\frac{1}{4(\sqrt{2}-1)\phi^2} \left(2 \left\| \mathbf{A}_2^{(j)} \otimes \mathbf{A}_1^{(j)} - \mathbf{A}_2^* \otimes \mathbf{A}_1^* \right\|_{\mathrm{F}}^2 \right. \\
&+ \left\| \mathrm{vec}(\mathbf{A}_1^{(j)})\mathrm{vec}(\mathbf{A}_1^{(j)})^\top - \mathrm{vec}(\mathbf{A}_1^*)\mathrm{vec}(\mathbf{A}_1^*)^\top \right\|_{\mathrm{F}}^2 \\
&\left. + \left\| \mathrm{vec}(\mathbf{A}_2^{(j)})\mathrm{vec}(\mathbf{A}_2^{(j)})^\top - \mathrm{vec}(\mathbf{A}_2^*)\mathrm{vec}(\mathbf{A}_2^*)^\top \right\|_{\mathrm{F}}^2 \right).
\end{aligned} \tag{S3.18}$$

To erase the last two terms, we note that

$$\begin{aligned}
&\left(\left\| \mathbf{A}_1^{(j)} \right\|_{\mathrm{F}}^2 - \left\| \mathbf{A}_2^{(j)} \right\|_{\mathrm{F}}^2 \right)^2 \\
=&\left\| \widetilde{\mathbf{Z}}^\top \mathbf{Z} \right\|_{\mathrm{F}}^2 \\
=&\left\langle \mathbf{Z}\mathbf{Z}^\top, \widetilde{\mathbf{Z}}\widetilde{\mathbf{Z}}^\top \right\rangle \\
=&\left\langle \mathbf{Z}\mathbf{Z}^\top - \mathbf{Z}^*\mathbf{Z}^{*\top}, \widetilde{\mathbf{Z}}\widetilde{\mathbf{Z}}^\top - \widetilde{\mathbf{Z}}^*\widetilde{\mathbf{Z}}^{*\top} \right\rangle + \left\langle \mathbf{Z}\mathbf{Z}^\top, \widetilde{\mathbf{Z}}^*\widetilde{\mathbf{Z}}^{*\top} \right\rangle + \left\langle \mathbf{Z}^*\mathbf{Z}^{*\top}, \widetilde{\mathbf{Z}}\widetilde{\mathbf{Z}}^\top \right\rangle - \left\langle \mathbf{Z}^*\mathbf{Z}^{*\top}, \widetilde{\mathbf{Z}}^*\widetilde{\mathbf{Z}}^{*\top} \right\rangle \\
=&\left\langle \mathbf{Z}\mathbf{Z}^\top - \mathbf{Z}^*\mathbf{Z}^{*\top}, \widetilde{\mathbf{Z}}\widetilde{\mathbf{Z}}^\top - \widetilde{\mathbf{Z}}^*\widetilde{\mathbf{Z}}^{*\top} \right\rangle + 2 \left\| \mathbf{Z}^\top \widetilde{\mathbf{Z}}^* \right\|_{\mathrm{F}}^2 - \underbrace{\left\| \mathbf{Z}^{*\top} \widetilde{\mathbf{Z}}^* \right\|_{\mathrm{F}}^2}_{=0} \\
\geq&\left\langle \mathbf{Z}\mathbf{Z}^\top - \mathbf{Z}^*\mathbf{Z}^{*\top}, \widetilde{\mathbf{Z}}\widetilde{\mathbf{Z}}^\top - \widetilde{\mathbf{Z}}^*\widetilde{\mathbf{Z}}^{*\top} \right\rangle \\
=&\left\| \mathrm{vec}(\mathbf{A}_1^{(j)})\mathrm{vec}(\mathbf{A}_1^{(j)})^\top - \mathrm{vec}(\mathbf{A}_1^*)\mathrm{vec}(\mathbf{A}_1^*)^\top \right\|_{\mathrm{F}}^2 + \left\| \mathrm{vec}(\mathbf{A}_2^{(j)})\mathrm{vec}(\mathbf{A}_2^{(j)})^\top - \mathrm{vec}(\mathbf{A}_2^*)\mathrm{vec}(\mathbf{A}_2^*)^\top \right\|_{\mathrm{F}}^2 \\
&-2 \left\| \mathbf{A}_2^{(j)} \otimes \mathbf{A}_1^{(j)} - \mathbf{A}_2^* \otimes \mathbf{A}_1^* \right\|_{\mathrm{F}}^2.
\end{aligned} \tag{S3.19}$$

Then, combine (S3.18) and (S3.19), we have

$$\left\|\mathbf{A}_1^{(j)}-\mathbf{A}_1^*\right\|_{\mathrm{F}}^2+\left\|\mathbf{A}_2^{(j)}-\mathbf{A}_2^*\right\|_{\mathrm{F}}^2 \leq \frac{4}{\phi^2}\left\|\mathbf{A}_2^{(j)}\otimes\mathbf{A}_1^{(j)}-\mathbf{A}_2^*\otimes\mathbf{A}_1^*\right\|_{\mathrm{F}}^2+\frac{1}{\phi^2}\left(\left\|\mathbf{A}_1^{(j)}\right\|_{\mathrm{F}}^2-\left\|\mathbf{A}_2^{(j)}\right\|_{\mathrm{F}}^2\right)^2. \tag{S3.20}$$

Next, by Lemma S.3 and $c_b \leq 0.01$, we have

$$\begin{aligned}\mathrm{dist}_{(j)}^2 \leq\ & 100\phi^{-4/3}\kappa^2\left(\left\|\mathbf{A}_1^{(j)}-\mathbf{A}_1^*\right\|_{\mathrm{F}}^2+\left\|\mathbf{A}_2^{(j)}-\mathbf{A}_2^*\right\|_{\mathrm{F}}^2\right)\\ &+24\phi^{-2/3}\sum_{i=1}^{2}\left(\left\|\left[\mathbf{C}_i^{(j)}\ \mathbf{R}_i^{(j)}\right]^\top\left[\mathbf{C}_i^{(j)}\ \mathbf{R}_i^{(j)}\right]-b^2\mathbf{I}_{r_i}\right\|_{\mathrm{F}}^2+\left\|\left[\mathbf{C}_i^{(j)}\ \mathbf{P}_i^{(j)}\right]^\top\left[\mathbf{C}_i^{(j)}\ \mathbf{P}_i^{(j)}\right]-b^2\mathbf{I}_{r_i}\right\|_{\mathrm{F}}^2\right).\end{aligned}$$

Then, we obtain a lower bound:

$$\begin{aligned}&\left\|\mathbf{A}_2^{(j)}\otimes\mathbf{A}_1^{(j)}-\mathbf{A}_2^*\otimes\mathbf{A}_1^*\right\|_{\mathrm{F}}^2\\ \geq&\frac{\phi^{10/3}}{400\kappa^2}\mathrm{dist}_{(j)}^2-\frac{1}{4}\left(\left\|\mathbf{A}_1^{(j)}\right\|_{\mathrm{F}}^2-\left\|\mathbf{A}_2^{(j)}\right\|_{\mathrm{F}}^2\right)^2\\ -&\frac{3\phi^{8/3}}{50\kappa^2}\sum_{i=1}^{2}\left(\left\|\left[\mathbf{C}_i^{(j)}\ \mathbf{R}_i^{(j)}\right]^\top\left[\mathbf{C}_i^{(j)}\ \mathbf{R}_i^{(j)}\right]-b^2\mathbf{I}_{r_i}\right\|_{\mathrm{F}}^2+\left\|\left[\mathbf{C}_i^{(j)}\ \mathbf{P}_i^{(j)}\right]^\top\left[\mathbf{C}_i^{(j)}\ \mathbf{P}_i^{(j)}\right]-b^2\mathbf{I}_{r_i}\right\|_{\mathrm{F}}^2\right).\end{aligned} \tag{S3.21}$$

Finally, plugging the lower bound (S3.21) into (S3.17), we have the following upper bound:

$$\begin{aligned}\mathrm{dist}_{(j+1)}^2 \leq&\left(1-2\eta\left(\frac{\alpha\phi^{10/3}}{800\kappa^2}-\frac{C_h^2C_D\alpha\phi^{10/3}}{\kappa^2}-36c\phi^{10/3}-C_D\lambda_2\phi^{2/3}-\frac{2025\lambda_1C_D}{\kappa^2}\phi^{10/3}\right)\right)\mathrm{dist}_{(j)}^2\\ &+\eta\left(\eta 240\phi^{10/3}+\frac{2}{c}\right)\xi^2(r_1,r_2,d_1,d_2)\\ &+\eta\left(240\eta\phi^{10/3}-\frac{1}{2\beta}\right)\left\|\nabla\widetilde{\mathcal{L}}(\mathbf{A}^{(j)})-\nabla\widetilde{\mathcal{L}}(\mathbf{A}^*)\right\|_{\mathrm{F}}^2\\ &+\frac{1}{2}\eta\left(40\lambda_2^2\eta\phi^{2/3}+\frac{3\alpha\phi^{8/3}}{25\kappa^2}-\lambda_2\right)\\ &\quad\times\sum_{i=1}^{2}\left(\left\|\left[\mathbf{C}_i^{(j)}\ \mathbf{R}_i^{(j)}\right]^\top\left[\mathbf{C}_i^{(j)}\ \mathbf{R}_i^{(j)}\right]-b^2\mathbf{I}_{r_i}\right\|_{\mathrm{F}}^2+\left\|\left[\mathbf{C}_i^{(j)}\ \mathbf{P}_i^{(j)}\right]^\top\left[\mathbf{C}_i^{(j)}\ \mathbf{P}_i^{(j)}\right]-b^2\mathbf{I}_{r_i}\right\|_{\mathrm{F}}^2\right)\\ &+\frac{1}{4}\eta\left(\alpha+480\lambda_1^2\eta\phi^{10/3}-2\lambda_1\right)\left(\left\|\mathbf{A}_1^{(j)}\right\|_{\mathrm{F}}^2-\left\|\mathbf{A}_2^{(j)}\right\|_{\mathrm{F}}^2\right)^2.\end{aligned} \tag{S3.22}$$

Step 4. (**Convergence analysis of** $\mathrm{dist}^2_{(j+1)}$)

For the tuning parameters $\eta, \lambda_2, \lambda_1$ and c, let

$$\eta = \frac{\eta_0}{\beta\phi^{10/3}}, \quad \lambda_2 = \frac{\alpha\phi^{8/3}}{\kappa^2}, \quad \lambda_1 = \alpha, \quad \text{and} \quad c = \frac{C_D\alpha}{\kappa^2}.$$

We see that with $\eta_0 \leq 1/480$ and C_D being small enough, the error is reduced by iteration.

For the coefficient of the third term of (S3.22),

$$240\eta\phi^{10/3} - \frac{1}{2\beta} \leq \frac{240}{480\beta} - \frac{1}{2\beta} \leq 0.$$

For the fourth term of (S3.22),

$$\begin{aligned}
40\lambda_2^2\eta\phi^{2/3} + \frac{3\alpha\phi^{8/3}}{25\kappa^2} - \lambda_2 \leq & \frac{40\alpha^2\phi^{8/3}}{480\kappa^4\beta} + \frac{3\alpha\phi^{8/3}}{25\kappa^2} - \frac{\alpha\phi^{8/3}}{\kappa^2} \\
\leq & \frac{\alpha\phi^{8/3}}{\kappa^2}\left(\frac{1}{12} + \frac{3}{25} - 1\right) \\
\leq & 0.
\end{aligned}$$

For the fifth term of (S3.22),

$$\alpha + 480\lambda_1^2\eta\phi^{10/3} - 2\lambda_1 \leq \alpha + \frac{\alpha^2}{\beta} - 2\alpha \leq (1 + 1 - 2)\alpha \leq 0.$$

From now on, for the sake of simplicity, C will denote a constant whose exact value may change in different contexts. For the second term of (S3.22),

$$\eta\left(\eta 240\phi^{10/3} + \frac{2}{c}\right) \leq \frac{\eta_0}{\beta\phi^{10/3}}\left(\frac{240}{480\beta} + \frac{2\kappa^2}{C_D\alpha}\right) \leq \frac{C\eta_0\kappa^2}{\alpha\beta\phi^{10/3}}.$$

For the first term, i.e., the coefficient of $\mathrm{dist}^2_{(j)}$,

$$\begin{aligned}
& \frac{\alpha\phi^{10/3}}{800\kappa^2} - \frac{C_h^2 C_D\alpha\phi^{10/3}}{\kappa^2} - 36c\phi^{10/3} - C_D\lambda_2\phi^{2/3} - \frac{2025\lambda_1 C_D}{\kappa^2}\phi^{10/3} \\
= & \frac{\alpha\phi^{10/3}}{\kappa^2}\left(\frac{1}{800} - \left(C_h^2 + 2062\right)C_D\right) \\
:= & \frac{\alpha\phi^{10/3}}{\kappa^2} \times \frac{C_0}{2}.
\end{aligned}$$

Therefore, we can derive the following recursive relationship:

$$\mathrm{dist}^2_{(j+1)} \leq (1 - C_0\eta_0\alpha\beta^{-1}\kappa^{-2})\mathrm{dist}^2_{(j)} + C\eta_0\kappa^2\alpha^{-1}\beta^{-1}\phi^{-10/3}\xi^2(r_1, r_2, d_1, d_2). \tag{S3.23}$$

When C_D is small enough, C_0 is a positive constant and $1 - C_0\eta_0\alpha\beta^{-1}\kappa^{-2}$ is a constant smaller than 1. Hence, the computational error can be reduced through the proposed gradient descent algorithm.

Step 5. (**Verification of the conditions**)

In this step, we verify that the conditions (S3.4) and (S3.3) in the convergence analysis hold recursively.

For $j = 0$, since initialization condition $\mathrm{dist}^2_{(0)} \leq C_D\alpha\beta^{-1}\kappa^{-2}\phi^{2/3}$ holds, we have

$$\begin{aligned}
\left\|[\mathbf{C}_i^{(0)}\ \mathbf{R}_i^{(0)}]\right\|_{\mathrm{op}} &\leq \left\|[\mathbf{C}_i^{(0)}\ \mathbf{R}_i^{(0)}] - [\mathbf{C}_i^*\ \mathbf{R}_i^*]\mathbf{O}_{i,u}^{(0)}\right\|_{\mathrm{op}} + \left\|[\mathbf{C}_i^*\ \mathbf{R}_i^*]\mathbf{O}_{i,u}^{(0)}\right\|_{\mathrm{op}} \\
&\leq \mathrm{dist}_{(0)} + \phi^{1/3} \\
&\leq (1 + c_b)b, \\
\left\|[\mathbf{C}_i^{(0)}\ \mathbf{P}_i^{(0)}]\right\|_{\mathrm{op}} &\leq \left\|[\mathbf{C}_i^{(0)}\ \mathbf{P}_i^{(0)}] - [\mathbf{C}_i^*\ \mathbf{P}_i^*]\mathbf{O}_{i,v}^{(0)}\right\|_{\mathrm{op}} + \left\|[\mathbf{C}_i^*\ \mathbf{P}_i^*]\mathbf{O}_{i,v}^{(0)}\right\|_{\mathrm{op}} \\
&\leq \mathrm{dist}_{(0)} + \phi^{1/3} \\
&\leq (1 + c_b)b,
\end{aligned}$$

and

$$\begin{aligned}
\left\|\mathbf{D}_i^{(0)}\right\|_{\mathrm{F}} &\leq \left\|\mathbf{D}_i^{(0)} - \mathbf{O}_{i,u}^{(0)\top}\mathbf{D}_i^*\mathbf{O}_{i,v}^{(0)}\right\|_{\mathrm{F}} + \left\|\mathbf{O}_{i,u}^{(0)\top}\mathbf{D}_i^*\mathbf{O}_{i,v}^{(0)}\right\|_{\mathrm{F}} \\
&\leq \sqrt{\mathrm{dist}^2_{(0)}} + \phi^{1/3} \\
&\leq \frac{(1 + c_b)\phi}{b^2}, \quad \text{for } i = 1, 2.
\end{aligned}$$

Then, suppose (S3.3) and (S3.4) hold at step j, for $j + 1$, we have

$$\begin{aligned}
\mathrm{dist}^2_{(j+1)} &\leq (1 - C_0\eta_0\alpha\beta^{-1}\kappa^{-2})\mathrm{dist}^2_{(j)} + C\eta_0\kappa^2\alpha^{-1}\beta^{-1}\phi^{-10/3}\xi^2(r_1, r_2, d_1, d_2) \\
&\leq \frac{C_D\alpha\phi^{2/3}}{\beta\kappa^2} - \eta_0\phi^{2/3}\alpha^2\beta^{-1}\kappa^2\left(\frac{C_0C_D}{\beta\kappa^6} - \frac{C\xi^2(r_1, r_2, d_1, d_2)}{\alpha^3\phi^4}\right).
\end{aligned}$$

Since $\phi^4 \geq C\beta\kappa^6\xi^2\alpha^{-3}$ for some universally big constant, we can verify that

$$\frac{C_2}{\beta\kappa^6} - \frac{C_1\xi^2(r_1, r_2, d_1, d_2)}{\alpha^3\phi^4} \geq 0,$$

then

$$\mathrm{dist}_{(j+1)}^2 \leq \frac{C_D \alpha \phi^{2/3}}{\beta \kappa^2}.$$

By the same argument as $j = 0$, we can verify that condition (S3.4) holds. Then the induction is completed.

Step 6. (**Construction of the upper bounds**)

In the preceding steps, we proved that (S3.3), (S3.4), and the recursive relationship (S3.23) hold for any $j \geq 0$. By iterating this relationship and summing the resulting geometric series, noting that $\sum_{j=0}^{\infty}(1 - C_0\eta_0\alpha\beta^{-1}\kappa^{-2})^j = C_0^{-1}\eta_0^{-1}\alpha^{-1}\beta\kappa^2$, we obtain the upper bound of $\mathrm{dist}_{(j)}^2$:

$$\mathrm{dist}_{(j)}^2 \leq (1 - C_0\eta_0\alpha\beta^{-1}\kappa^{-2})^j \mathrm{dist}_{(0)}^2 + C\kappa^4\alpha^{-2}\phi^{-10/3}\xi^2(r_1, r_2, d_1, d_2),$$

For the error bound of $\mathbf{A}_1$ and $\mathbf{A}_2$, by Lemma S.3, we have

$$\begin{aligned}
&\left\|\mathbf{A}_1^{(j)} - \mathbf{A}_1^*\right\|_{\mathrm{F}}^2 + \left\|\mathbf{A}_2^{(j)} - \mathbf{A}_2^*\right\|_{\mathrm{F}}^2 \\
\leq& C\phi^{4/3}\mathrm{dist}_{(j)}^2 \\
\leq& C\phi^{4/3}(1 - C_0\eta_0\alpha\beta^{-1}\kappa^{-2})^j \mathrm{dist}_{(0)}^2 + C\kappa^4\alpha^{-2}\phi^{-2}\xi^2(r_1, r_2, d_1, d_2) \\
\leq& C\kappa^2(1 - C_0\eta_0\alpha\beta^{-1}\kappa^{-2})^j \left(\left\|\mathbf{A}_1^{(0)} - \mathbf{A}_1^*\right\|_{\mathrm{F}}^2 + \left\|\mathbf{A}_2^{(0)} - \mathbf{A}_2^*\right\|_{\mathrm{F}}^2\right) \\
&+ C\kappa^4\alpha^{-2}\phi^{-2}\xi^2(r_1, r_2, d_1, d_2).
\end{aligned}$$

Also, for the error bound of $\left\|\mathbf{A}_2^{(j)} \otimes \mathbf{A}_1^{(j)} - \mathbf{A}_2^* \otimes \mathbf{A}_1^*\right\|_{\mathrm{F}}^2$, we have

$$
\begin{aligned}
&\left\|\mathbf{A}_2^{(j)} \otimes \mathbf{A}_1^{(j)} - \mathbf{A}_2^* \otimes \mathbf{A}_1^*\right\|_{\mathrm{F}}^2 \\
\leq& 2\left\|\left(\mathbf{A}_2^{(j)} - \mathbf{A}_2^*\right) \otimes \mathbf{A}_1^{(j)}\right\|_{\mathrm{F}}^2 + 2\left\|\mathbf{A}_2^* \otimes \left(\mathbf{A}_1^{(j)} - \mathbf{A}_1^*\right)\right\|_{\mathrm{F}}^2 \\
\leq& 8\phi^2 \left(\left\|\mathbf{A}_1^{(j)} - \mathbf{A}_1^*\right\|_{\mathrm{F}}^2 + \left\|\mathbf{A}_2^{(j)} - \mathbf{A}_2^*\right\|_{\mathrm{F}}^2\right) \\
\leq& C\phi^2\kappa^2(1 - C_0\eta_0\alpha\beta^{-1}\kappa^{-2})^j \left(\left\|\mathbf{A}_1^{(0)} - \mathbf{A}_1^*\right\|_{\mathrm{F}}^2 + \left\|\mathbf{A}_2^{(0)} - \mathbf{A}_2^*\right\|_{\mathrm{F}}^2\right) \\
&\quad + C\kappa^4\alpha^{-2}\xi^2(r_1, r_2, d_1, d_2) \\
\leq& C\kappa^2(1 - C_0\eta_0\alpha\beta^{-1}\kappa^{-2})^j \left(\left\|\mathbf{A}_2^{(0)} \otimes \mathbf{A}_1^{(0)} - \mathbf{A}_2^* \otimes \mathbf{A}_1^*\right\|_{\mathrm{F}}^2\right) \\
&\quad + C\kappa^4\alpha^{-2}\xi^2(r_1, r_2, d_1, d_2).
\end{aligned}
$$

The last inequality is from (S3.20) with the fact that $\left\|\mathbf{A}_1^{(0)}\right\|_{\mathrm{F}} = \left\|\mathbf{A}_2^{(0)}\right\|_{\mathrm{F}}$.

S3.2 Auxiliary Lemmas

The first lemma states that when the Frobenius norm of estimated parameter matrices are close to their true values, the Frobenius norm of high-order perturbation terms can be controlled by the running error in (S3.1).

Lemma S.1. *Define the following matrices*

$$
\mathbf{A}_1^* = [\mathbf{C}_1^* \ \mathbf{R}_1^*]\mathbf{D}_1^*[\mathbf{C}_1^* \ \mathbf{P}_1^*]^\top, \mathbf{A}_2^* = [\mathbf{C}_2^* \ \mathbf{R}_2^*]\mathbf{D}_2^*[\mathbf{C}_2^* \ \mathbf{P}_2^*]^\top,
$$

$$
\mathbf{A}_1 = [\mathbf{C}_1 \ \mathbf{R}_1]\mathbf{D}_1[\mathbf{C}_1 \ \mathbf{P}_1]^\top, \ \textit{and}\ \mathbf{A}_2 = [\mathbf{C}_2 \ \mathbf{R}_2]\mathbf{D}_2[\mathbf{C}_2 \ \mathbf{P}_2]^\top,
$$

where $\mathbf{D}_i^*,\ \mathbf{D}_i \in \mathbb{R}^{r_i \times r_i},\ \mathbf{C}_i^*, \mathbf{C}_i \in \mathbb{R}^{p_i \times d_i},$ *and* $\mathbf{R}_i, \mathbf{R}_i^*, \mathbf{P}_i, \mathbf{P}_i^* \in \mathbb{R}^{p_i \times (r_i - d_i)}$, *for* $i = 1, 2$. *Meanwhile,* $\|\mathbf{A}_1^*\|_{\mathrm{F}} = \|\mathbf{A}_2^*\|_{\mathrm{F}} = \phi$ *and* $[\mathbf{C}_i^* \ \mathbf{R}_i^*]^\top [\mathbf{C}_i^* \ \mathbf{R}_i^*] = [\mathbf{C}_i^* \ \mathbf{P}_i^*]^\top [\mathbf{C}_i^* \ \mathbf{P}_i^*] = b^2\mathbf{I}_{r_i}$ *for some balance scalar* b. *Suppose that there exists a constant* c, *such that*

$$
\|[\mathbf{C}_i \ \mathbf{R}_i]\|_{\mathrm{op}} \leq (1+c)b, \quad \|[\mathbf{C}_i \ \mathbf{P}_i]\|_{\mathrm{op}} \leq (1+c)b, \textit{and}\ \|\mathbf{D}_i\|_{\mathrm{F}} \leq \frac{(1+c)\phi}{b^2}, \quad \textit{for}\ i = 1, 2. \tag{S3.24}
$$

For $i = 1, 2$, let

$$\mathcal{E}_{i,u} = [\mathbf{C}_i^* \ \mathbf{R}_i^*]\mathbf{O}_{i,u} - [\mathbf{C}_i \ \mathbf{R}_i],$$

$$\mathcal{E}_{i,v} = [\mathbf{C}_i^* \ \mathbf{P}_i^*]\mathbf{O}_{i,v} - [\mathbf{C}_i \ \mathbf{P}_i],$$

$$\mathcal{E}_{i,D} = \mathbf{O}_{i,u}^\top \mathbf{D}_i^* \mathbf{O}_{i,v} - \mathbf{D}_i,$$

where $\mathbf{O}_{i,c} \in \mathbb{O}^{d_i \times d_i}, \mathbf{O}_{i,r}, \mathbf{O}_{i,p} \in \mathbb{O}^{(r_i - d_i)\times(r_i - d_i)}, \mathbf{O}_{i,u} = \operatorname{diag}(\mathbf{O}_{i,c}, \mathbf{O}_{i,r}), \mathbf{O}_{i,v} = \operatorname{diag}(\mathbf{O}_{i,c}, \mathbf{O}_{i,p})$.

Then, let

$$\begin{aligned}\mathbf{H}_1 := &- \mathbf{A}_2 \otimes \left([\mathbf{C}_1 \ \mathbf{R}_1]\mathbf{D}_1\mathcal{E}_{1,v}^\top + [\mathbf{C}_1 \ \mathbf{R}_1]\mathcal{E}_{1,D}[\mathbf{C}_1 \ \mathbf{P}_1]^\top + \mathcal{E}_{1,u}\mathbf{D}_1[\mathbf{C}_1 \ \mathbf{P}_1]^\top\right) \\ &- \left([\mathbf{C}_2 \ \mathbf{R}_2]\mathbf{D}_2\mathcal{E}_{2,v}^\top + [\mathbf{C}_2 \ \mathbf{R}_2]\mathcal{E}_{2,D}[\mathbf{C}_2 \ \mathbf{P}_2]^\top + \mathcal{E}_{2,u}\mathbf{D}_2[\mathbf{C}_2 \ \mathbf{P}_2]^\top\right) \otimes \mathbf{A}_1,\end{aligned}$$

which contains the first-order perturbation terms, and

$$\mathbf{H} := \mathbf{A}_2^* \otimes \mathbf{A}_1^* - \mathbf{A}_2 \otimes \mathbf{A}_1 + \mathbf{H}_1,$$

which represents second- and higher-order terms perturbed from $\mathbf{A}_2 \otimes \mathbf{A}_1$.

Meanwhile, define the distance as

$$\begin{aligned}D := &\min_{\substack{\mathbf{O}'_{i,c} \in \mathbb{O}^{d_i \times d_i} \\ \mathbf{O}'_{i,r}, \mathbf{O}'_{i,p} \in \mathbb{O}^{(r_i - d_i)\times(r_i-d_i)} \\ i=1,2}} \sum_{i=1,2} \Big\{ \|\mathbf{C}_i - \mathbf{C}_i^*\mathbf{O}'_{i,c}\|_{\mathrm{F}}^2 + \|\mathbf{R}_i - \mathbf{R}_i^*\mathbf{O}'_{i,r}\|_{\mathrm{F}}^2 + \|\mathbf{P}_i - \mathbf{P}_i^*\mathbf{O}'_{i,p}\|_{\mathrm{F}}^2 \\ &\qquad + \|\mathbf{D}_i - \operatorname{diag}(\mathbf{O}'_{i,c}, \mathbf{O}'_{i,r})^\top \mathbf{D}_i^* \operatorname{diag}(\mathbf{O}'_{i,c}, \mathbf{O}'_{i,p})\|_{\mathrm{F}}^2 \Big\} \\ := &\sum_{i=1,2} \Big\{ \|\mathbf{C}_i - \mathbf{C}_i^*\mathbf{O}_{i,c}\|_{\mathrm{F}}^2 + \|\mathbf{R}_i - \mathbf{R}_i^*\mathbf{O}_{i,r}\|_{\mathrm{F}}^2 + \|\mathbf{P}_i - \mathbf{P}_i^*\mathbf{O}_{i,p}\|_{\mathrm{F}}^2 + \|\mathbf{D}_i - \mathbf{O}_{i,u}^\top \mathbf{D}_i^* \mathbf{O}_{i,v}\|_{\mathrm{F}}^2 \Big\}.\end{aligned}$$

Assume that $D \le C_D \phi^{2/3}$ for some constant C_D. If $b \asymp \phi^{1/3}$, there exists a constant C_h, such that

$$\|\mathbf{H}\|_{\mathrm{F}} \le C_h \phi^{4/3} D.$$

Proof of Lemma S.1. We start by decomposing the perturbed matrix $\mathbf{A}_2^* \otimes \mathbf{A}_1^*$. Using notations defined above, we split the perturbed matrix to 64 terms and classify them into three types: zeroth-order perturbation, first-order perturbation, and high-order perturbation. Then we control the Frobenius norm of high-order perturbation, $\|\mathbf{H}\|_{\mathrm{F}}$.

$$
\begin{aligned}
&\mathbf{A}_2^* \otimes \mathbf{A}_1^* \\
=&[\mathbf{C}_2^* \ \mathbf{R}_2^*]\mathbf{D}_2^*[\mathbf{C}_2^* \ \mathbf{P}_2^*]^\top \otimes [\mathbf{C}_1^* \ \mathbf{R}_1^*]\mathbf{D}_1^*[\mathbf{C}_1^* \ \mathbf{P}_1^*]^\top \\
=&\left([\mathbf{C}_2 \ \mathbf{R}_2] + \mathcal{E}_{2,u}\right)\left(\mathbf{D}_2 + \mathcal{E}_{2,D}\right)\left([\mathbf{C}_2 \ \mathbf{P}_2] + \mathcal{E}_{2,v}\right)^\top \otimes \left([\mathbf{C}_1 \ \mathbf{R}_1] + \mathcal{E}_{1,u}\right)\left(\mathbf{D}_1 + \mathcal{E}_{1,D}\right)\left([\mathbf{C}_1 \ \mathbf{P}_1] + \mathcal{E}_{1,v}\right)^\top \\
=&\begin{pmatrix} [\mathbf{C}_2 \ \mathbf{R}_2]\mathbf{D}_2[\mathbf{C}_2 \ \mathbf{P}_2]^\top + [\mathbf{C}_2 \ \mathbf{R}_2]\mathbf{D}_2\mathcal{E}_{2,v}^\top + [\mathbf{C}_2 \ \mathbf{R}_2]\mathcal{E}_{2,D}[\mathbf{C}_2 \ \mathbf{P}_2]^\top + [\mathbf{C}_2 \ \mathbf{R}_2]\mathcal{E}_{2,D}\mathcal{E}_{2,v}^\top \\ +\mathcal{E}_{2,u}\mathbf{D}_2[\mathbf{C}_2 \ \mathbf{P}_2]^\top + \mathcal{E}_{2,u}\mathbf{D}_2\mathcal{E}_{2,v}^\top + \mathcal{E}_{2,u}\mathcal{E}_{2,D}[\mathbf{C}_2 \ \mathbf{P}_2]^\top + \mathcal{E}_{2,u}\mathcal{E}_{2,D}\mathcal{E}_{2,v}^\top \end{pmatrix} \\
&\otimes \begin{pmatrix} [\mathbf{C}_1 \ \mathbf{R}_1]\mathbf{D}_1[\mathbf{C}_1 \ \mathbf{P}_1]^\top + [\mathbf{C}_1 \ \mathbf{R}_1]\mathbf{D}_1\mathcal{E}_{1,v}^\top + [\mathbf{C}_1 \ \mathbf{R}_1]\mathcal{E}_{1,D}[\mathbf{C}_1 \ \mathbf{P}_1]^\top + [\mathbf{C}_1 \ \mathbf{R}_1]\mathcal{E}_{1,D}\mathcal{E}_{1,v}^\top \\ +\mathcal{E}_{1,u}\mathbf{D}_1[\mathbf{C}_1 \ \mathbf{P}_1]^\top + \mathcal{E}_{1,u}\mathbf{D}_1\mathcal{E}_{1,v}^\top + \mathcal{E}_{1,u}\mathcal{E}_{1,D}[\mathbf{C}_1 \ \mathbf{P}_1]^\top + \mathcal{E}_{1,u}\mathcal{E}_{1,D}\mathcal{E}_{1,v}^\top \end{pmatrix} \\
=&\mathbf{A}_2 \otimes \mathbf{A}_1 + \mathbf{A}_2 \otimes \left([\mathbf{C}_1 \ \mathbf{R}_1]\mathbf{D}_1\mathcal{E}_{1,v}^\top + [\mathbf{C}_1 \ \mathbf{R}_1]\mathcal{E}_{1,D}[\mathbf{C}_1 \ \mathbf{P}_1]^\top + \mathcal{E}_{1,u}\mathbf{D}_1[\mathbf{C}_1 \ \mathbf{P}_1]^\top\right) \\
&+ \left([\mathbf{C}_2 \ \mathbf{R}_2]\mathbf{D}_2\mathcal{E}_{2,v}^\top + [\mathbf{C}_2 \ \mathbf{R}_2]\mathcal{E}_{2,D}[\mathbf{C}_2 \ \mathbf{P}_2]^\top + \mathcal{E}_{2,u}\mathbf{D}_2[\mathbf{C}_2 \ \mathbf{P}_2]^\top\right) \otimes \mathbf{A}_1 \\
&+ (57 \text{ terms containing 2 or more } \mathcal{E}\text{s}) \\
=&\mathbf{A}_2 \otimes \mathbf{A}_1 - \mathbf{H}_1 + (57 \text{ terms containing 2 or more } \mathcal{E}\text{s}).
\end{aligned}
$$

Therefore, $\mathbf{H}$ is the summation of 57 terms of higher order perturbation. Next, we upper bound $\|\mathbf{H}\|_{\mathrm{F}}$ by upper bounding every piece of $\mathbf{H}$.

$$
\|\mathbf{H}\|_{\mathrm{F}} \leq \sum_{i=1}^{57} \|\text{the } i\text{th term of } \mathbf{H}\|_{\mathrm{F}}.
$$

It can be easily verified that for every $\mathcal{E}$ defined above (we just ignore the subscripts), $\|\mathcal{E}\|_{\mathrm{F}}^2 \leq D$. Writing out every one of the 57 terms, we find that the Frobenius norm of each term can be upper bounded by one of the values in $\{CD^3,\ CbD^{5/2},\ C\phi b^{-2}D^{5/2},\ Cb^2D^2,\ C\phi b^{-1}D^2,\ C\phi^2 b^{-4}D^2,\ Cb^2D^2,\ C\phi^2 b^{-3}D^{3/2},\ C\phi D^{3/2},\ Cb^3D^{3/2},\ C\phi bD,\ C\phi^2 b^{-2}D,\ Cb^4D\}$. Using $b = \phi^{1/3}$ and $D \leq C_D\phi^{2/3}$, it is clear that all those values are less than or equal to

$C\phi^{4/3}D$, where C is another constant. Finally, adding together the 57 upper bounds gives

$$\|\mathbf{H}\|_{\mathrm{F}} \leq C_h \phi^{4/3} D.$$

□

The following lemma is adapted from Lemma A.2 in Wang et al. (2023), using the Frobenius norm $\|\mathbf{B}^*\|_{\mathrm{F}}$ instead of the original spectral norm, as presented below. The proof follows an analogous argument and is omitted here for brevity. It establishes the approximate equivalence between the estimation error of components and that of the aggregated matrix $\mathbf{B}$.

Lemma S.2. *(Wang et al., 2023) Suppose that* $\mathbf{B}^* = [\mathbf{C}^*\ \mathbf{R}^*]\,\mathbf{D}^*\,[\mathbf{C}^*\ \mathbf{P}^*]^\top$, $[\mathbf{C}^*\ \mathbf{R}^*]^\top [\mathbf{C}^*\ \mathbf{R}^*] = b^2\mathbf{I}_r$, $[\mathbf{C}^*\ \mathbf{P}^*]^\top [\mathbf{C}^*\ \mathbf{P}^*] = b^2\mathbf{I}_r$, $\phi = \|\mathbf{B}^*\|_{\mathrm{F}}$, *and* $\sigma_r = \sigma_r(\mathbf{B}^*)$. *Let* $\mathbf{B} = [\mathbf{C}\ \mathbf{R}]\mathbf{D}[\mathbf{C}\ \mathbf{P}]^\top$ *with* $\|[\mathbf{C}\ \mathbf{R}]\|_{\mathrm{op}} \leq (1+c)\,b$, $\|[\mathbf{C}\ \mathbf{P}]\|_{\mathrm{op}} \leq (1+c)\,b$, *and* $\|\mathbf{D}\|_{\mathrm{op}} \leq (1+c)\,\phi/b^2$ *for some constant* $c > 0$. *Define*

$$E := \min_{\substack{\mathbf{O}_c \in \mathbb{O}^{d\times d}\\ \mathbf{O}_r, \mathbf{O}_p \in \mathbb{O}^{(r-d)\times(r-d)}}} \Big(\|[\mathbf{C}\ \mathbf{R}] - [\mathbf{C}^*\ \mathbf{R}^*]\operatorname{diag}(\mathbf{O}_c, \mathbf{O}_r)\|_{\mathrm{F}}^2 + \|\,[\mathbf{C}\ \mathbf{P}]$$
$$- [\mathbf{C}^*\ \mathbf{P}^*]\operatorname{diag}(\mathbf{O}_c, \mathbf{O}_p)\,\|_{\mathrm{F}}^2 + \|\,\mathbf{D} - \operatorname{diag}(\mathbf{O}_c, \mathbf{O}_r)^\top \mathbf{D}^* \operatorname{diag}(\mathbf{O}_c, \mathbf{O}_p)\,\|_{\mathrm{F}}^2\Big).$$

Then, we have

$$E \leq \left(4b^{-4} + \frac{8b^2}{\sigma_r^2} C_b\right)\|\mathbf{B} - \mathbf{B}^*\|_{\mathrm{F}}^2$$
$$+ 2b^{-2} C_b \left(\left\|[\mathbf{C}\ \mathbf{R}]^\top [\mathbf{C}\ \mathbf{R}] - b^2\mathbf{I}_r\right\|_{\mathrm{F}}^2 + \left\|[\mathbf{C}\ \mathbf{P}]^\top [\mathbf{C}\ \mathbf{P}] - b^2\mathbf{I}_r\right\|_{\mathrm{F}}^2\right)$$

and

$$\|\mathbf{B} - \mathbf{B}^*\|_{\mathrm{F}}^2 \leq 3b^4\left[1 + 4\phi^2 b^{-6}(1+c)^4\right] E,$$

where $C_b = 1 + 4\phi^2 b^{-6}\left((1+c)^4 + (1+c)^2(2+c)^2/2\right)$.

Applying Lemma S.2 on $\mathbf{A}_1$ and $\mathbf{A}_2$, we have the following equivalence relationship between the combined distance D defined in Lemma S.1 and the squared errors of $\mathbf{A}_1$ and $\mathbf{A}_2$.

Lemma S.3. *Consider* $\mathbf{A}_1^*, \mathbf{A}_1, \mathbf{A}_2^*, \mathbf{A}_2$ *and* D *defined as in Lemma S.1, and* (S3.24) *holds. Define* $\underline{\sigma} := \min(\sigma_{1,r_1}, \sigma_{2,r_2})$. *Then, we have*

$$D \leq \left(4b^{-4} + \frac{8b^2}{\underline{\sigma}^2} C_b\right) \left(\|\mathbf{A}_1 - \mathbf{A}_1^*\|_{\mathrm{F}}^2 + \|\mathbf{A}_2 - \mathbf{A}_2^*\|_{\mathrm{F}}^2\right)$$
$$+ 2C_b b^{-2} \sum_{i=1}^{2} \left(\left\| [\mathbf{C}_i \ \mathbf{R}_i]^\top [\mathbf{C}_i \ \mathbf{R}_i] - b^2 \mathbf{I}_{r_i} \right\|_{\mathrm{F}}^2 + \left\| [\mathbf{C}_i \ \mathbf{P}_i]^\top [\mathbf{C}_i \ \mathbf{P}_i] - b^2 \mathbf{I}_{r_i} \right\|_{\mathrm{F}}^2 \right).$$

In addition,

$$\|\mathbf{A}_1 - \mathbf{A}_1^*\|_{\mathrm{F}}^2 + \|\mathbf{A}_2 - \mathbf{A}_2^*\|_{\mathrm{F}}^2 \leq 6b^4 D + 24\phi^2 b^{-2}(1+c)^4 D.$$

Proof of Lemma S.3. The conditions specified in (S3.24) satisfy the premises of Lemma S.2. Applying Lemma S.2 to the pairs $(\mathbf{A}_1, \mathbf{A}_1^*)$ and $(\mathbf{A}_2, \mathbf{A}_2^*)$ separately yields the desired bounds. Specifically, summing the resulting inequalities for $i = 1$ and $i = 2$ establishes the first inequality. The second inequality follows directly by applying the reverse bound from Lemma S.2 and summing over $i = 1, 2$. This concludes the proof. □

The last lemma is useful in analyzing the upper bound of the distance between our estimates and their true values.

Lemma S.4. *(Lemma 5.4., Tu et al. (2016)) For any* $\mathbf{U}, \mathbf{X} \in \mathbb{R}^{p \times r}$, *where* $p \geq r$ *and* $\mathbf{X}$ *is full-rank, define*

$$\operatorname{dist}(\mathbf{U}, \mathbf{X})^2 = \min_{\mathbf{O} \in \mathbb{O}^{r \times r}} \|\mathbf{U} - \mathbf{X}\mathbf{O}\|_{\mathrm{F}}^2,$$

we have

$$\operatorname{dist}(\mathbf{U}, \mathbf{X})^2 \leq \frac{1}{2(\sqrt{2}-1)\sigma_r^2(\mathbf{X})} \left\|\mathbf{U}\mathbf{U}^\top - \mathbf{X}\mathbf{X}^\top\right\|_{\mathrm{F}}^2.$$

S4 Statistical Convergence Analysis

In this appendix, we present the stochastic analysis underlying our statistical theory, including verification of the RSC/RSS condition and bounding of the deviation term ξ and initialization errors. The conclusions are derived under assumptions presented in Section 4 of the main article. Notations are inherited from Appendix S3.

S4.1 Proof of Theorem 2

We first verify the conditions needed in Theorem 1 for the local convergence. Under Assumptions 1 and 2, by Proposition 1, we have the conclusion that with probability at least $1-C\exp\{-C(p_1r_1+p_2r_2+4r_1^2r_2^2)\}$, the RSC/RSS condition holds with $\alpha=\alpha_{\text{RSC}}$ and $\beta=\beta_{\text{RSS}}$. Meanwhile, by Proposition 2, the initialization error satisfies $\text{dist}_{(0)}^2\lesssim\kappa^{-2}\phi^{2/3}\alpha_{\text{RSC}}\beta_{\text{RSS}}^{-1}$ with probability at least $1-C\exp(-Cp_1p_2)$. In addition, by Proposition S.2, the condition $\xi^2\lesssim\kappa^{-6}\phi^4\alpha_{\text{RSC}}^3\beta_{\text{RSS}}^{-1}$ is satisfied with high probability.

Therefore, along with other conditions of Theorem 2, the conditions of Theorem 1 are satisfied, implying that for $j\geq 1$,

$$
\begin{aligned}
&\left\|\mathbf{A}_1^{(j)}-\mathbf{A}_1^*\right\|_{\text{F}}^2+\left\|\mathbf{A}_2^{(j)}-\mathbf{A}_2^*\right\|_{\text{F}}^2\\
\lesssim&\kappa^2(1-C_0\eta_0\alpha_{\text{RSC}}\beta_{\text{RSS}}^{-1}\kappa^{-2})^j\left(\left\|\mathbf{A}_1^{(0)}-\mathbf{A}_1^*\right\|_{\text{F}}^2+\left\|\mathbf{A}_2^{(0)}-\mathbf{A}_2^*\right\|_{\text{F}}^2\right)+\kappa^4\phi^{-2}\alpha_{\text{RSC}}^{-2}\xi^2(r_1,r_2,d_1,d_2)\\
\lesssim&\kappa^2(1-C_0\eta_0\alpha_{\text{RSC}}\beta_{\text{RSS}}^{-1}\kappa^{-2})^j\phi^{4/3}\text{dist}_{(0)}^2+\kappa^4\phi^{-2}\alpha_{\text{RSC}}^{-2}\xi^2(r_1,r_2,d_1,d_2)
\end{aligned}
$$

For the initial error term, we apply Proposition 2 again to see that with probability at least $1-C\exp(-Cp_1p_2)$,

$$\text{dist}_{(0)}^2\lesssim\phi^{2/3}\underline{\sigma}^{-4}g_{\min}^{-4}\alpha_{\text{RSC}}^{-2}\tau^2M_1^2\frac{p_1p_2r_1r_2}{T}.$$

In addition, by Proposition S.1, the statistical error is bounded as

$$\xi(r_1,r_2,d_1,d_2)\lesssim\tau M_1\sqrt{\frac{\text{df}_{\text{MARCF}}}{T}},$$

with probability at least $1 - C\exp\left(-C(p_1r_1 + p_2r_2)\right)$.

To ensure the upper bound of the statistical error term dominates the computational error term after j iterations, we need

$$\kappa^2(1 - C_0\eta_0\alpha_{\mathrm{RSC}}\beta_{\mathrm{RSS}}^{-1}\kappa^{-2})^j\phi^2\underline{\sigma}^{-4}g_{\min}^{-4}\alpha_{\mathrm{RSC}}^{-2}\tau^2M_1^2\frac{p_1p_2r_1r_2}{T} \lesssim \kappa^4\alpha_{\mathrm{RSC}}^{-2}\phi^{-2}\tau^2M_1^2\frac{\mathrm{df}_{\mathrm{MARCF}}}{T}.$$

It implies that when

$$J \gtrsim \frac{\log\left(g_{\min}^4\kappa^{-2}\mathrm{df}_{\mathrm{MARCF}}/(p_1p_2r_1r_2)\right)}{\log(1 - C_0\eta_0\alpha_{\mathrm{RSC}}\beta_{\mathrm{RSS}}^{-1}\kappa^{-2})},$$

the computational error is absorbed, and then

$$\left\|\mathbf{A}_1^{(J)} - \mathbf{A}_1^*\right\|_{\mathrm{F}}^2 + \left\|\mathbf{A}_2^{(J)} - \mathbf{A}_2^*\right\|_{\mathrm{F}}^2 \lesssim \kappa^4\phi^{-2}\alpha_{\mathrm{RSC}}^{-2}\tau^2M_1^2\frac{\mathrm{df}_{\mathrm{MARCF}}}{T}.$$

By similar analysis on $\left\|\mathbf{A}_2^{(J)} \otimes \mathbf{A}_1^{(J)} - \mathbf{A}_2^* \otimes \mathbf{A}_1^*\right\|_{\mathrm{F}}^2$, we also have

$$\left\|\mathbf{A}_2^{(J)} \otimes \mathbf{A}_1^{(J)} - \mathbf{A}_2^* \otimes \mathbf{A}_1^*\right\|_{\mathrm{F}}^2 \lesssim \kappa^4\alpha_{\mathrm{RSC}}^{-2}\tau^2M_1^2\frac{\mathrm{df}_{\mathrm{MARCF}}}{T}.$$

□

S4.2 Verification of RSC/RSS Condition

In this appendix, we prove Proposition 1. We first recall some notations. Let $\mathbf{A}^* = \mathbf{A}_2^* \otimes \mathbf{A}_1^*$ denote the true transition matrix in the vectorized MARCF model $\mathbf{y}_t = \mathbf{A}^*\mathbf{y}_{t-1} + \mathbf{e}_t$, where $\mathbf{y}_t := \mathrm{vec}(\mathbf{Y}_t)$ and $\mathbf{e}_t := \mathrm{vec}(\mathbf{E}_t)$. The RSC and RSS properties are defined with respect to an arbitrary candidate parameter matrix $\mathbf{A} = \mathbf{A}_2 \otimes \mathbf{A}_1$ satisfying the MARCF structural restrictions. Throughout this proof, unstarred $\mathbf{A}$, $\mathbf{A}_1$, and $\mathbf{A}_2$ refer to such candidate matrices. Then, the least-squares loss function with respect to $\mathbf{A}$ can be written as

$$\widetilde{\mathcal{L}}(\mathbf{A}) = \frac{1}{2T}\sum_{t=1}^{T}\|\mathbf{y}_t - \mathbf{A}\mathbf{y}_{t-1}\|_2^2.$$

In addition, recall that the RSC and RSS coefficients defined in Proposition 1 are $\alpha_{\mathrm{RSC}} = \lambda_{\min}\left(\boldsymbol{\Sigma}_{\mathbf{e}}\right)/(2\mu_{\max}(\mathcal{A}))$ and $\beta_{\mathrm{RSS}} = 3\lambda_{\max}\left(\boldsymbol{\Sigma}_{\mathbf{e}}\right)/(2\mu_{\min}(\mathcal{A}))$, and the spectral quantity used for the sample complexity is $M_2 = \lambda_{\min}(\mathbf{\Sigma_e})\,\mu_{\min}(\mathcal{A})\,/\,(\lambda_{\max}(\mathbf{\Sigma_e})\,\mu_{\max}(\mathcal{A}))$. In the following, we give the proof of Proposition 1.

Proof of Proposition 1. Straightforward algebraic manipulation yields

$$\widetilde{\mathcal{L}}(\mathbf{A}) - \widetilde{\mathcal{L}}(\mathbf{A}^*) - \left\langle \nabla\widetilde{\mathcal{L}}(\mathbf{A}^*), \mathbf{A}-\mathbf{A}^* \right\rangle = \frac{1}{2T}\sum_{t=0}^{T-1} \|(\mathbf{A}-\mathbf{A}^*)\mathbf{y}_t\|_2^2 .$$

Hence, it suffices to prove that

$$\frac{\alpha_{\mathrm{RSC}}}{2}\|\mathbf{A}-\mathbf{A}^*\|_{\mathrm{F}}^2 \le \frac{1}{2T}\sum_{t=0}^{T-1}\|(\mathbf{A}-\mathbf{A}^*)\mathbf{y}_t\|_{\mathrm{F}}^2 \le \frac{\beta_{\mathrm{RSS}}}{2}\|\mathbf{A}-\mathbf{A}^*\|_{\mathrm{F}}^2$$

with high probability.

Based on the moving average representation of VAR(1), we can rewrite $\mathbf{y}_t$ as a VMA(∞) process,

$$\mathbf{y}_t = \mathbf{e}_t + \mathbf{A}^*\mathbf{e}_{t-1} + (\mathbf{A}^*)^2\mathbf{e}_{t-2} + (\mathbf{A}^*)^3\mathbf{e}_{t-3} + \cdots$$

Let $\mathbf{z} = \left(\mathbf{y}_{T-1}^\top, \mathbf{y}_{T-2}^\top, \dots, \mathbf{y}_0^\top\right)^\top$, $\mathbf{e} = \left(\mathbf{e}_{T-1}^\top, \mathbf{e}_{T-2}^\top, \dots, \mathbf{e}_0^\top, \dots\right)^\top$. Now $\mathbf{z} = \widetilde{\mathbf{A}}\mathbf{e}$, where $\widetilde{\mathbf{A}}$ is a matrix with Tp_1p_2 rows and ∞ columns, defined as

$$\widetilde{\mathbf{A}} = \begin{bmatrix} \mathbf{I}_{p_1p_2} & \mathbf{A}^* & (\mathbf{A}^*)^2 & (\mathbf{A}^*)^3 & \dots \\ \mathbf{O} & \mathbf{I}_{p_1p_2} & \mathbf{A}^* & (\mathbf{A}^*)^2 & \dots \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{O} & \mathbf{O} & \mathbf{O} & \mathbf{I}_{p_1p_2} & \dots \end{bmatrix} = \begin{bmatrix} \mathbf{I}_{p_1p_2} & \mathbf{A}_2^*\otimes\mathbf{A}_1^* & (\mathbf{A}_2^*\otimes\mathbf{A}_1^*)^2 & (\mathbf{A}_2^*\otimes\mathbf{A}_1^*)^3 & \dots \\ \mathbf{O} & \mathbf{I}_{p_1p_2} & \mathbf{A}_2^*\otimes\mathbf{A}_1^* & (\mathbf{A}_2^*\otimes\mathbf{A}_1^*)^2 & \dots \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{O} & \mathbf{O} & \mathbf{O} & \mathbf{I}_{p_1p_2} & \dots \end{bmatrix}.$$

Let $\boldsymbol{\zeta} = \left(\boldsymbol{\zeta}_{T-1}^\top, \boldsymbol{\zeta}_{T-2}^\top, \dots, \boldsymbol{\zeta}_0^\top, \dots\right)^\top$. By assumption on noise we know that $\mathbf{e} = \widetilde{\mathbf{\Sigma}}_{\mathbf{e}}^{1/2}\boldsymbol{\zeta}$ where

$$\widetilde{\mathbf{\Sigma}}_{\mathbf{e}}^{1/2} = \begin{pmatrix} \mathbf{\Sigma}_{\mathbf{e}}^{1/2} & \mathbf{O} & \mathbf{O} & \cdots \\ \mathbf{O} & \mathbf{\Sigma}_{\mathbf{e}}^{1/2} & \mathbf{O} & \cdots \\ \mathbf{O} & \mathbf{O} & \mathbf{\Sigma}_{\mathbf{e}}^{1/2} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

Let $\boldsymbol{\Delta} = \mathbf{A} - \mathbf{A}^*$. It suffices to show that for $\|\boldsymbol{\Delta}\|_{\mathrm{F}} = 1$, there exist $0 < \alpha \leq \beta$, such that

$$\frac{\alpha}{2} \leq \frac{1}{2T} \sum_{t=0}^{T-1} \|\boldsymbol{\Delta}\mathbf{y}_t\|_2^2 \leq \frac{\beta}{2}.$$

We define the set of unit-norm differences of low-rank Kronecker products as:

$$\mathcal{S}(r_1, r_2, p_1, p_2) := \Big\{\mathbf{M} : \mathbf{M} = \mathbf{A}_2 \otimes \mathbf{A}_1 - \mathbf{A}_2' \otimes \mathbf{A}_1', \\ \mathbf{A}_i, \mathbf{A}_i' \in \mathbb{R}^{p_i \times p_i},\ \mathrm{rank}(\mathbf{A}_i),\ \mathrm{rank}(\mathbf{A}_i') \leq r_i,\ \|\mathbf{M}\|_{\mathrm{F}} = 1,\ i = 1, 2\Big\}.$$

For brevity, we denote this set as $\mathcal{S}$ when the dimensions and ranks are clear from the context. By construction, $\boldsymbol{\Delta} \in \mathcal{S}$. Denoting $R_T(\boldsymbol{\Delta}) := \sum_{t=0}^{T-1} \|\boldsymbol{\Delta}\mathbf{y}_t\|_2^2$, we have

$$\begin{aligned} R_T(\boldsymbol{\Delta}) &= \sum_{t=0}^{T-1} \mathbf{y}_t^\top \boldsymbol{\Delta}^\top \boldsymbol{\Delta} \mathbf{y}_t \\ &= \left(\mathbf{y}_{T-1}^\top, \mathbf{y}_{T-2}^\top, \dots, \mathbf{y}_0^\top\right) \begin{pmatrix} \boldsymbol{\Delta}^\top\boldsymbol{\Delta} & & \\ & \ddots & \\ & & \boldsymbol{\Delta}^\top\boldsymbol{\Delta} \end{pmatrix} \begin{pmatrix} \mathbf{y}_{T-1} \\ \mathbf{y}_{T-2} \\ \vdots \\ \mathbf{y}_0 \end{pmatrix} \qquad \text{(S4.1)} \\ &= \mathbf{z}^\top \left(\mathbf{I}_T \otimes \boldsymbol{\Delta}^\top \boldsymbol{\Delta}\right) \mathbf{z} \\ &= \mathbf{e}^\top \widetilde{\mathbf{A}}^\top \left(\mathbf{I}_T \otimes \boldsymbol{\Delta}^\top \boldsymbol{\Delta}\right) \widetilde{\mathbf{A}} \mathbf{e} \\ &= \boldsymbol{\zeta}^\top \widetilde{\boldsymbol{\Sigma}}_{\mathbf{e}}^{1/2} \widetilde{\mathbf{A}}^\top \left(\mathbf{I}_T \otimes \boldsymbol{\Delta}^\top \boldsymbol{\Delta}\right) \widetilde{\mathbf{A}} \widetilde{\boldsymbol{\Sigma}}_{\mathbf{e}}^{1/2} \boldsymbol{\zeta} \\ &:= \boldsymbol{\zeta}^\top \boldsymbol{\Sigma}_{\boldsymbol{\Delta}} \boldsymbol{\zeta}. \end{aligned}$$

Note that $R_T(\boldsymbol{\Delta}) \geq \mathbb{E}R_T(\boldsymbol{\Delta}) - \sup_{\boldsymbol{\Delta} \in \mathcal{S}} |R_T(\boldsymbol{\Delta}) - \mathbb{E}R_T(\boldsymbol{\Delta})|$, we will derive a lower bound for $\mathbb{E}R_T(\boldsymbol{\Delta})$ and an upper bound for $\sup_{\boldsymbol{\Delta} \in \mathcal{S}} |R_T(\boldsymbol{\Delta}) - \mathbb{E}R_T(\boldsymbol{\Delta})|$ to complete the proof of RSC part.

For $\mathbb{E}R_T(\boldsymbol{\Delta})$, by (S4.1) and properties of Frobenius norm,

$$\begin{aligned}\mathbb{E}R_T(\boldsymbol{\Delta}) &= \mathbb{E}\left[\boldsymbol{\zeta}^\top \widetilde{\boldsymbol{\Sigma}}_{\mathbf{e}}^{1/2} \widetilde{\mathbf{A}}^\top \left(\mathbf{I}_T \otimes \boldsymbol{\Delta}^\top \boldsymbol{\Delta}\right) \widetilde{\mathbf{A}} \widetilde{\boldsymbol{\Sigma}}_{\mathbf{e}}^{1/2} \boldsymbol{\zeta}\right] \\ &= \operatorname{tr}\left(\widetilde{\boldsymbol{\Sigma}}_{\mathbf{e}}^{1/2} \widetilde{\mathbf{A}}^\top \left(\mathbf{I}_T \otimes \boldsymbol{\Delta}^\top \boldsymbol{\Delta}\right) \widetilde{\mathbf{A}} \widetilde{\boldsymbol{\Sigma}}_{\mathbf{e}}^{1/2}\right) \\ &= \left\|(\mathbf{I}_T \otimes \boldsymbol{\Delta}) \widetilde{\mathbf{A}} \widetilde{\boldsymbol{\Sigma}}_{\mathbf{e}}^{1/2}\right\|_{\mathrm{F}}^2 \\ &\geq T\lambda_{\min}(\widetilde{\mathbf{A}}\widetilde{\mathbf{A}}^\top)\lambda_{\min}(\boldsymbol{\Sigma}_{\mathbf{e}}).\end{aligned}$$

For $|R_T(\boldsymbol{\Delta}) - \mathbb{E}R_T(\boldsymbol{\Delta})|$, noting that

$$\|\boldsymbol{\Sigma}_{\boldsymbol{\Delta}}\|_{\mathrm{F}}^2 \leq \|\mathbf{I}_T \otimes \boldsymbol{\Delta}\|_{\mathrm{F}}^2 \|\mathbf{I}_T \otimes \boldsymbol{\Delta}\|_{\mathrm{op}}^2 \|\widetilde{\boldsymbol{\Sigma}}_{\mathbf{e}}^{1/2}\|_{\mathrm{op}}^4 \|\widetilde{\mathbf{A}}\|_{\mathrm{op}}^4 \leq T\lambda_{\max}^2\left(\boldsymbol{\Sigma}_{\mathbf{e}}\right)\lambda_{\max}^2\left(\widetilde{\mathbf{A}}\widetilde{\mathbf{A}}^\top\right)$$

and

$$\|\boldsymbol{\Sigma}_{\boldsymbol{\Delta}}\|_{\mathrm{op}} \leq \|\mathbf{I}_T \otimes \boldsymbol{\Delta}\|_{\mathrm{op}}^2 \|\widetilde{\boldsymbol{\Sigma}}_{\mathbf{e}}^{1/2}\|_{\mathrm{op}}^2 \|\widetilde{\mathbf{A}}\|_{\mathrm{op}}^2 \leq \lambda_{\max}\left(\boldsymbol{\Sigma}_{\mathbf{e}}\right)\lambda_{\max}\left(\widetilde{\mathbf{A}}\widetilde{\mathbf{A}}^\top\right),$$

by Hanson-Wright inequality (Lemma S.5), we have that for any $x \geq 0$,

$$\begin{aligned}&\mathbb{P}\left(|R_T(\boldsymbol{\Delta}) - \mathbb{E}R_T(\boldsymbol{\Delta})| \geq \tau^2 x\right) \\ \leq& 2\exp\left\{-C\min\left(\frac{x}{\lambda_{\max}\left(\boldsymbol{\Sigma}_{\mathbf{e}}\right)\lambda_{\max}\left(\widetilde{\mathbf{A}}\widetilde{\mathbf{A}}^\top\right)}, \frac{x^2}{T\lambda_{\max}^2\left(\boldsymbol{\Sigma}_{\mathbf{e}}\right)\lambda_{\max}^2\left(\widetilde{\mathbf{A}}\widetilde{\mathbf{A}}^\top\right)}\right)\right\},\end{aligned} \tag{S4.2}$$

where C is a constant.

For any $\varepsilon \in (0,1)$, consider an ε-net of $\mathcal{S}$ with respect to the Frobenius norm, denoted by $\overline{\mathcal{S}}$. Then for any $\boldsymbol{\Delta} \in \mathcal{S}$, there exists $\overline{\boldsymbol{\Delta}} \in \overline{\mathcal{S}}$, such that $\left\|\boldsymbol{\Delta} - \overline{\boldsymbol{\Delta}}\right\|_{\mathrm{F}} \leq \varepsilon$. By Lemma S.6, we have

$$|\overline{\mathcal{S}}| \leq \left(\frac{15}{\varepsilon}\right)^{4p_1r_1+4p_2r_2+16r_1^2r_2^2}.$$

Therefore,

$$
\begin{aligned}
&|R_T(\boldsymbol{\Delta}) - \mathbb{E}R_T(\boldsymbol{\Delta})| \\
=&\left|\sum_{t=0}^{T-1}\left(\left\|\mathbf{y}_t^\top \otimes \mathbf{I}_{p_1p_2}\mathrm{vec}(\boldsymbol{\Delta})\right\|_{\mathrm{F}}^2 - \mathbb{E}\left\|\mathbf{y}_t^\top \otimes \mathbf{I}_{p_1p_2}\mathrm{vec}(\boldsymbol{\Delta})\right\|_{\mathrm{F}}^2\right)\right| \\
=&\left|\mathrm{vec}(\boldsymbol{\Delta})^\top \sum_{t=0}^{T-1}\left(\mathbf{y}_t\mathbf{y}_t^\top \otimes \mathbf{I}_{p_1p_2} - \mathbb{E}\mathbf{y}_t\mathbf{y}_t^\top \otimes \mathbf{I}_{p_1p_2}\right)\mathrm{vec}(\boldsymbol{\Delta})\right| \\
:=&\left|\mathrm{vec}(\boldsymbol{\Delta})^\top \mathbf{M}(\mathbf{y}_t)\mathrm{vec}(\boldsymbol{\Delta})\right| \\
\leq&\left|\mathrm{vec}(\overline{\boldsymbol{\Delta}})^\top \mathbf{M}(\mathbf{y}_t)\mathrm{vec}(\overline{\boldsymbol{\Delta}})\right| + 2\left|\mathrm{vec}(\overline{\boldsymbol{\Delta}})^\top \mathbf{M}(\mathbf{y}_t)(\mathrm{vec}(\boldsymbol{\Delta}) - \mathrm{vec}(\overline{\boldsymbol{\Delta}}))\right| \\
&+\left|(\mathrm{vec}(\boldsymbol{\Delta})^\top - \mathrm{vec}(\overline{\boldsymbol{\Delta}})^\top)\mathbf{M}(\mathbf{y}_t)(\mathrm{vec}(\boldsymbol{\Delta}) - \mathrm{vec}(\overline{\boldsymbol{\Delta}}))\right| \\
\leq&\max_{\overline{\boldsymbol{\Delta}}\in\overline{\mathcal{S}}}|R_T(\overline{\boldsymbol{\Delta}}) - \mathbb{E}R_T(\overline{\boldsymbol{\Delta}})| + (2\varepsilon + \varepsilon^2)\sup_{\boldsymbol{\Delta}\in\mathcal{S}}|R_T(\boldsymbol{\Delta}) - \mathbb{E}R_T(\boldsymbol{\Delta})|.
\end{aligned}
$$

When $\varepsilon < 1/3$, we have

$$
\sup_{\boldsymbol{\Delta}\in\mathcal{S}}|R_T(\boldsymbol{\Delta}) - \mathbb{E}R_T(\boldsymbol{\Delta})| \leq (1-3\varepsilon)^{-1}\max_{\overline{\boldsymbol{\Delta}}\in\overline{\mathcal{S}}}|R_T(\overline{\boldsymbol{\Delta}}) - \mathbb{E}R_T(\overline{\boldsymbol{\Delta}})|.
$$

Letting $\varepsilon = 0.1$, by union bounds and (S4.2),

$$
\begin{aligned}
&\mathbb{P}\left(\sup_{\boldsymbol{\Delta}\in\mathcal{S}}|R_T(\boldsymbol{\Delta}) - \mathbb{E}R_T(\boldsymbol{\Delta})| \geq \tau^2 x\right) \\
\leq&\mathbb{P}\left(\max_{\overline{\boldsymbol{\Delta}}\in\overline{\mathcal{S}}}|R_T(\overline{\boldsymbol{\Delta}}) - \mathbb{E}R_T(\overline{\boldsymbol{\Delta}})| \geq (1-3\varepsilon)\tau^2 x\right) \\
\leq&\sum_{\overline{\boldsymbol{\Delta}}\in\overline{\mathcal{S}}}\mathbb{P}\left(|R_T(\overline{\boldsymbol{\Delta}}) - \mathbb{E}R_T(\overline{\boldsymbol{\Delta}})| \geq \frac{1}{2}\tau^2 x\right) \\
\leq&2\times 150^{4p_1r_1+4p_2r_2+16r_1^2r_2^2}\exp\left\{-C\min\left(\frac{x}{\lambda_{\max}(\boldsymbol{\Sigma}_{\mathbf{e}})\lambda_{\max}\left(\widetilde{\mathbf{A}}\widetilde{\mathbf{A}}^\top\right)}, \frac{x^2}{T\lambda_{\max}^2(\boldsymbol{\Sigma}_{\mathbf{e}})\lambda_{\max}^2\left(\widetilde{\mathbf{A}}\widetilde{\mathbf{A}}^\top\right)}\right)\right\} \\
\leq&2\exp\left\{21(p_1r_1+p_2r_2+4r_1^2r_2^2) - C\min\left(\frac{x}{\lambda_{\max}(\boldsymbol{\Sigma}_{\mathbf{e}})\lambda_{\max}\left(\widetilde{\mathbf{A}}\widetilde{\mathbf{A}}^\top\right)}, \frac{x^2}{T\lambda_{\max}^2(\boldsymbol{\Sigma}_{\mathbf{e}})\lambda_{\max}^2\left(\widetilde{\mathbf{A}}\widetilde{\mathbf{A}}^\top\right)}\right)\right\}.
\end{aligned}
$$

Here we take $x = T\lambda_{\min}(\boldsymbol{\Sigma}_{\mathbf{e}})\lambda_{\min}\left(\widetilde{\mathbf{A}}\widetilde{\mathbf{A}}^\top\right)/(2\tau^2)$ and define

$$
M_2 := \lambda_{\min}(\boldsymbol{\Sigma}_{\mathbf{e}})\lambda_{\min}\left(\widetilde{\mathbf{A}}\widetilde{\mathbf{A}}^\top\right)/\lambda_{\max}(\boldsymbol{\Sigma}_{\mathbf{e}})\lambda_{\max}\left(\widetilde{\mathbf{A}}\widetilde{\mathbf{A}}^\top\right) \leq 1,
$$

then

$$\begin{aligned}
&\mathbb{P}\left(\sup_{\boldsymbol{\Delta}\in\mathcal{S}}|R_T(\boldsymbol{\Delta})-\mathbb{E}R_T(\boldsymbol{\Delta})|\geq\frac{T}{2}\lambda_{\min}\left(\boldsymbol{\Sigma}_{\mathbf{e}}\right)\lambda_{\min}\left(\widetilde{\mathbf{A}}\widetilde{\mathbf{A}}^\top\right)\right)\\
\leq&2\exp\left\{21(p_1r_1+p_2r_2+4r_1^2r_2^2)-C\min\left(\frac{M_2}{2\tau^2},\frac{M_2^2}{4\tau^4}\right)T\right\} \qquad \text{(S4.3)}\\
\leq&2\exp\left\{21(p_1r_1+p_2r_2+4r_1^2r_2^2)-CM_2^2\min\left(\tau^{-2},\tau^{-4}\right)T\right\}.
\end{aligned}$$

Hence, when $T\gtrsim(p_1r_1+p_2r_2+4r_1^2r_2^2)M_2^{-2}\max(\tau^2,\tau^4)$, we have that with probability at least $1-2\exp\{-C(p_1r_1+p_2r_2+4r_1^2r_2^2)\}$,

$$R_T(\boldsymbol{\Delta})\geq\frac{T}{2}\lambda_{\min}\left(\boldsymbol{\Sigma}_{\mathbf{e}}\right)\lambda_{\min}\left(\widetilde{\mathbf{A}}\widetilde{\mathbf{A}}^\top\right).$$

Then the RSC coefficient $\alpha_{\text{RSC}}=\lambda_{\min}\left(\boldsymbol{\Sigma}_{\mathbf{e}}\right)\lambda_{\min}\left(\widetilde{\mathbf{A}}\widetilde{\mathbf{A}}^\top\right)/2$.

As for RSS part, similarly, we have $R_T(\boldsymbol{\Delta})\leq\mathbb{E}R_T(\boldsymbol{\Delta})+\sup_{\boldsymbol{\Delta}\in\mathcal{S}}|R_T(\boldsymbol{\Delta})-\mathbb{E}R_T(\boldsymbol{\Delta})|$ and $\mathbb{E}R_T(\boldsymbol{\Delta})\leq T\lambda_{\max}(\widetilde{\mathbf{A}}\widetilde{\mathbf{A}}^\top)\lambda_{\max}(\boldsymbol{\Sigma}_{\mathbf{e}})$. Since the event

$$\left\{\sup_{\boldsymbol{\Delta}\in\mathcal{S}}|R_T(\boldsymbol{\Delta})-\mathbb{E}R_T(\boldsymbol{\Delta})|\leq T\lambda_{\min}\left(\boldsymbol{\Sigma}_{\mathbf{e}}\right)\lambda_{\min}\left(\widetilde{\mathbf{A}}\widetilde{\mathbf{A}}^\top\right)/2\right\}$$

implies

$$\left\{\sup_{\boldsymbol{\Delta}\in\mathcal{S}}|R_T(\boldsymbol{\Delta})-\mathbb{E}R_T(\boldsymbol{\Delta})|\leq T\lambda_{\max}\left(\boldsymbol{\Sigma}_{\mathbf{e}}\right)\lambda_{\max}\left(\widetilde{\mathbf{A}}\widetilde{\mathbf{A}}^\top\right)/2\right\},$$

we have when the event above occurs,

$$R_T(\boldsymbol{\Delta})\leq\frac{3T}{2}\lambda_{\max}\left(\boldsymbol{\Sigma}_{\mathbf{e}}\right)\lambda_{\max}\left(\widetilde{\mathbf{A}}\widetilde{\mathbf{A}}^\top\right). \qquad \text{(S4.4)}$$

Thus $\beta_{\text{RSS}}:=3\lambda_{\max}\left(\boldsymbol{\Sigma}_{\mathbf{e}}\right)\lambda_{\max}\left(\widetilde{\mathbf{A}}\widetilde{\mathbf{A}}^\top\right)/2$.

Finally, since $\widetilde{\mathbf{A}}$ is related to VMA(∞) process, by the spectral measure of ARMA process discussed in Basu and Michailidis (2015) we replace $\lambda_{\max}\left(\widetilde{\mathbf{A}}\widetilde{\mathbf{A}}^\top\right)$ and $\lambda_{\min}\left(\widetilde{\mathbf{A}}\widetilde{\mathbf{A}}^\top\right)$ with $1/\mu_{\min}(\mathcal{A})$ and $1/\mu_{\max}(\mathcal{A})$, respectively. $\square$

S4.3 Property of Deviation Bound

In this appendix, we derive an upper bound for ξ defined in Definition 2, which demonstrates the benefit of dimension reduction conferred by the MARCF model.

Proposition S.1. *Under Assumptions 1 and 2, if* $T \gtrsim (p_1r_1+p_2r_2+4r_1^2r_2^2)M_2^{-2}\max(\tau^2,\tau^4)$, *then with probability at least* $1-2\exp(-C(p_1r_1+p_2r_2))$,

$$\xi(r_1,r_2,d_1,d_2) \lesssim \tau M_1\sqrt{\frac{\mathrm{df}_{\mathrm{MARCF}}}{T}},$$

where $M_1 = \lambda_{\max}(\boldsymbol{\Sigma}_{\mathbf{e}})/\mu_{\min}^{1/2}(\mathcal{A})$, $\mathrm{df}_{\mathrm{MARCF}} = p_1(2r_1-d_1)+p_2(2r_2-d_2)+r_1^2+r_2^2$, *and* $M_2 = \lambda_{\min}(\boldsymbol{\Sigma}_{\mathbf{e}})\mu_{\min}(\mathcal{A})/(\lambda_{\max}(\boldsymbol{\Sigma}_{\mathbf{e}})\mu_{\max}(\mathcal{A}))$.

Proof of Proposition S.1. We first define the parameter spaces for $\mathbf{A}_1$ and $\mathbf{A}_2$ in the MARCF model. Specifically, we consider the following set of matrices with unit Frobenius norm and common column and row subspaces, denoted by $\mathcal{W}(r,d;p)$:

$$\begin{aligned}\mathcal{W}(r,d;p) := \big\{&\mathbf{W}\in\mathbb{R}^{p\times p}:\mathbf{W}=[\mathbf{C}\ \mathbf{R}]\mathbf{D}[\mathbf{C}\ \mathbf{P}]^\top, \mathbf{C}\in\mathbb{O}^{p\times d}, \mathbf{R},\mathbf{P}\in\mathbb{O}^{p\times(r-d)},\\ &\mathbf{C}^\top\mathbf{R}=\mathbf{C}^\top\mathbf{P}=\mathbf{0},\ \text{and}\ \|\mathbf{W}\|_{\mathrm{F}}=1\big\}.\end{aligned}$$

Let $\overline{\mathcal{W}}(r,d;p)$ be the $\varepsilon/2$-net of $\mathcal{W}(r,d,p)$ with respect to the Frobenius norm. By Lemma S.8, $\overline{\mathcal{W}}(r,d;p)\subset\mathcal{W}(r,d;p)$, and

$$|\overline{\mathcal{W}}(r,d;p)| \le \left(\frac{48}{\varepsilon}\right)^{p(2r-d)+r^2}.$$

Define

$$\mathcal{V}(r_1,r_2,d_1,d_2) := \{\mathbf{W}_2\otimes\mathbf{W}_1:\mathbf{W}_2\in\mathcal{W}(r_2,d_2;p_2),\mathbf{W}_1\in\mathcal{W}(r_1,d_1;p_1)\},$$

then $\overline{\mathcal{V}}(r_1,r_2,d_1,d_2) := \big\{\overline{\mathbf{W}}_2\otimes\overline{\mathbf{W}}_1:\overline{\mathbf{W}}_2\in\overline{\mathcal{W}}(r_2,d_2;p_2),\overline{\mathbf{W}}_1\in\overline{\mathcal{W}}(r_1,d_1;p_1)\big\}$ is an ε-net of $\mathcal{V}(r_1,r_2,d_1,d_2)$, which can be directly verified. Meanwhile,

$$|\overline{\mathcal{V}}(r_1,r_2,d_1,d_2)| \le \left(\frac{48}{\varepsilon}\right)^{2(p_1r_1+p_2r_2)-p_1d_1-p_2d_2+r_1^2+r_2^2}. \tag{S4.5}$$

In addition, note that $\nabla\mathcal{L}(\mathbf{A}^*) = -(\sum_{t=1}^T \mathbf{e}_t\mathbf{y}_{t-1}^\top)/T$, we have

$$\xi(r_1,r_2,d_1,d_2) = \sup_{\mathbf{A}\in\mathcal{V}(r_1,r_2,d_1,d_2)} \left\langle \frac{1}{T}\sum_{t=1}^T \mathbf{e}_t\mathbf{y}_{t-1}^\top, \mathbf{A}\right\rangle. \tag{S4.6}$$

In the following, we start from establishing an upper bound for ξ when there is no common space, i.e., $d_1 = d_2 = 0$. In this case, elements of $\mathbf{W}_1$ and $\mathbf{W}_2$ are merely low rank matrices.

For every $\mathbf{A} = \mathbf{W}_2\otimes\mathbf{W}_1 \in \mathcal{V}(r_1,r_2,0,0)$, let $\overline{\mathbf{A}} = \overline{\mathbf{W}}_2\otimes\overline{\mathbf{W}}_1 \in \overline{\mathcal{V}}(r_1,r_2,0,0)$ be its covering matrix. By splitting SVD with common space (Lemma S.7), we know that $\mathbf{W}_2-\overline{\mathbf{W}}_2$ can be decomposed as $\mathbf{W}_2-\overline{\mathbf{W}}_2 = \mathbf{\Delta}_{2,1}+\mathbf{\Delta}_{2,2}$, where $\mathbf{\Delta}_{2,1},\mathbf{\Delta}_{2,2}$ are both rank-r_2 and $\langle\mathbf{\Delta}_{2,1},\mathbf{\Delta}_{2,2}\rangle = 0$. For $\mathbf{W}_1-\overline{\mathbf{W}}_1$, there exist $\mathbf{\Delta}_{1,1}$ and $\mathbf{\Delta}_{1,2}$ following the same property. Then with Cauchy's inequality and $\|\mathbf{\Delta}_{i,1}+\mathbf{\Delta}_{i,2}\|_{\mathrm{F}}^2 = \|\mathbf{\Delta}_{i,1}\|_{\mathrm{F}}^2+\|\mathbf{\Delta}_{i,2}\|_{\mathrm{F}}^2$ we know that $\|\mathbf{\Delta}_{i,1}\|_{\mathrm{F}}+\|\mathbf{\Delta}_{i,2}\|_{\mathrm{F}} \leq \sqrt{2}\,\|\mathbf{\Delta}_{i,1}+\mathbf{\Delta}_{i,2}\|_{\mathrm{F}} \leq \sqrt{2}\varepsilon$, for $i=1,2$. Since $\mathbf{\Delta}_{i,j}/\|\mathbf{\Delta}_{i,j}\|_{\mathrm{F}} \in \mathcal{W}(r_i,0;p_i)$ we have

$$\begin{aligned}
&\left\langle \frac{1}{T}\sum_{t=1}^T \mathbf{e}_t\mathbf{y}_{t-1}^\top, \mathbf{A}\right\rangle \\
=&\left\langle \frac{1}{T}\sum_{t=1}^T \mathbf{e}_t\mathbf{y}_{t-1}^\top, \overline{\mathbf{A}}\right\rangle \\
&+\left\langle \frac{1}{T}\sum_{t=1}^T \mathbf{e}_t\mathbf{y}_{t-1}^\top, (\mathbf{W}_2-\overline{\mathbf{W}}_2)\otimes\mathbf{W}_1\right\rangle + \left\langle \frac{1}{T}\sum_{t=1}^T \mathbf{e}_t\mathbf{y}_{t-1}^\top, \overline{\mathbf{W}}_2\otimes(\mathbf{W}_1-\overline{\mathbf{W}}_1)\right\rangle \\
=&\left\langle \frac{1}{T}\sum_{t=1}^T \mathbf{e}_t\mathbf{y}_{t-1}^\top, \overline{\mathbf{A}}\right\rangle \\
&+\left\langle \frac{1}{T}\sum_{t=1}^T \mathbf{e}_t\mathbf{y}_{t-1}^\top, \frac{\mathbf{\Delta}_{2,1}}{\|\mathbf{\Delta}_{2,1}\|_{\mathrm{F}}}\otimes\mathbf{W}_1\right\rangle \|\mathbf{\Delta}_{2,1}\|_{\mathrm{F}} + \left\langle \frac{1}{T}\sum_{t=1}^T \mathbf{e}_t\mathbf{y}_{t-1}^\top, \frac{\mathbf{\Delta}_{2,2}}{\|\mathbf{\Delta}_{2,2}\|_{\mathrm{F}}}\otimes\mathbf{W}_1\right\rangle \|\mathbf{\Delta}_{2,2}\|_{\mathrm{F}} \\
&+\left\langle \frac{1}{T}\sum_{t=1}^T \mathbf{e}_t\mathbf{y}_{t-1}^\top, \overline{\mathbf{W}}_2\otimes\frac{\mathbf{\Delta}_{1,1}}{\|\mathbf{\Delta}_{1,1}\|_{\mathrm{F}}}\right\rangle \|\mathbf{\Delta}_{1,1}\|_{\mathrm{F}} + \left\langle \frac{1}{T}\sum_{t=1}^T \mathbf{e}_t\mathbf{y}_{t-1}^\top, \overline{\mathbf{W}}_2\otimes\frac{\mathbf{\Delta}_{1,2}}{\|\mathbf{\Delta}_{1,2}\|_{\mathrm{F}}}\right\rangle \|\mathbf{\Delta}_{1,2}\|_{\mathrm{F}}
\end{aligned}$$

$$
\begin{aligned}
&\leq \max_{\overline{A}\in\overline{\mathcal{V}}(r_1,r_2,0,0)} \left\langle \frac{1}{T}\sum_{t=1}^{T}\mathbf{e}_t\mathbf{y}_{t-1}^\top, \overline{\mathbf{A}} \right\rangle + \xi(r_1,r_2,0,0)\left(\|\boldsymbol{\Delta}_{2,1}\|_{\mathrm{F}} + \|\boldsymbol{\Delta}_{2,2}\|_{\mathrm{F}} + \|\boldsymbol{\Delta}_{1,1}\|_{\mathrm{F}} + \|\boldsymbol{\Delta}_{1,2}\|_{\mathrm{F}}\right) \\
&\leq \max_{\overline{A}\in\overline{\mathcal{V}}(r_1,r_2,0,0)} \left\langle \frac{1}{T}\sum_{t=1}^{T}\mathbf{e}_t\mathbf{y}_{t-1}^\top, \overline{\mathbf{A}} \right\rangle + 2\sqrt{2}\varepsilon\xi(r_1,r_2,0,0).
\end{aligned}
$$

Hence,

$$
\xi(r_1,r_2,0,0) \leq (1-2\sqrt{2}\varepsilon)^{-1} \max_{\overline{A}\in\overline{\mathcal{V}}(r_1,r_2,0,0)} \left\langle \frac{1}{T}\sum_{t=1}^{T}\mathbf{e}_t\mathbf{y}_{t-1}^\top, \overline{\mathbf{A}} \right\rangle.
$$

Define $R_T(\mathbf{M}) := \sum_{t=0}^{T-1}\|\mathbf{M}\mathbf{y}_t\|_{\mathrm{F}}^2$ and $S_T(\mathbf{M}) := \sum_{t=1}^{T}\langle \mathbf{e}_t, \mathbf{M}\mathbf{y}_{t-1}\rangle$ for any $p_1p_2 \times p_1p_2$ real matrix $\mathbf{M}$ with unit Frobenius norm. By the proof of LemmaS5 in Wang et al. (2024), for any z_1 and $z_2 \geq 0$,

$$
\mathbb{P}\left[\{S_T(\mathbf{M}) \geq z_1\} \cap \{R_T(\mathbf{M}) \leq z_2\}\right] \leq \exp\left(-\frac{z_1^2}{2\tau^2\lambda_{\max}(\boldsymbol{\Sigma}_{\mathbf{e}})\, z_2}\right).
$$

With the derivation of (S4.3) and (S4.4), when $T \gtrsim (p_1r_1+p_2r_2+4r_1^2r_2^2)M_2^{-2}\max(\tau^2,\tau^4)$, using proper constants C here, we have that

$$
\mathbb{P}\left(R_T(\boldsymbol{\Delta}) \geq \frac{3T}{2}\lambda_{\max}(\boldsymbol{\Sigma}_{\mathbf{e}})\,\lambda_{\max}\left(\widetilde{\mathbf{A}}\widetilde{\mathbf{A}}^\top\right)\right) \leq 2\exp\{-15(p_1r_1+p_2r_2+4r_1^2r_2^2)\}.
$$

Then, using the pieces above, we have for any $x \geq 0$,

$$
\begin{aligned}
&\mathbb{P}\left(\xi(r_1,r_2,0,0) \geq x\right) \\
\leq&\mathbb{P}\left(\max_{\overline{A}\in\overline{\mathcal{V}}(r_1,r_2,0,0)} \left\langle \frac{1}{T}\sum_{t=1}^{T}\mathbf{e}_t\mathbf{y}_{t-1}^\top, \overline{\mathbf{A}} \right\rangle \geq (1-2\sqrt{2}\varepsilon)x\right) \\
\leq&\sum_{\overline{A}\in\overline{\mathcal{V}}(r_1,r_2,0,0)} \mathbb{P}\left(S_T(\overline{\mathbf{A}}) \geq (1-2\sqrt{2}\varepsilon)Tx\right) \\
\leq&\sum_{\overline{A}\in\overline{\mathcal{V}}(r_1,r_2,0,0)} \mathbb{P}\left(\left\{S_T(\overline{\mathbf{A}}) \geq (1-2\sqrt{2}\varepsilon)Tx\right\}\bigcap\left\{R_T(\overline{\mathbf{A}}) \leq \frac{3T}{2}\lambda_{\max}(\boldsymbol{\Sigma}_{\mathbf{e}})\,\lambda_{\max}\left(\widetilde{\mathbf{A}}\widetilde{\mathbf{A}}^\top\right)\right\}\right) \\
&+\sum_{\overline{A}\in\overline{\mathcal{V}}(r_1,r_2,0,0)} \mathbb{P}\left(R_T(\overline{\mathbf{A}}) \geq \frac{3T}{2}\lambda_{\max}(\boldsymbol{\Sigma}_{\mathbf{e}})\,\lambda_{\max}\left(\widetilde{\mathbf{A}}\widetilde{\mathbf{A}}^\top\right)\right) \\
\leq&|\overline{\mathcal{V}}(r_1,r_2,0,0)|\left(\exp\left\{-\frac{(1-2\sqrt{2}\varepsilon)^2Tx^2}{3\tau^2\lambda_{\max}^2(\boldsymbol{\Sigma}_{\mathbf{e}})\,\lambda_{\max}\left(\widetilde{\mathbf{A}}\widetilde{\mathbf{A}}^\top\right)}\right\} + \exp\left\{-15(p_1r_1+p_2r_2+4r_1^2r_2^2)\right\}\right).
\end{aligned}
$$

Here we take $\varepsilon = 0.1$ and $x = 4\sqrt{3}\tau\lambda_{\max}(\mathbf{\Sigma_e})\lambda_{\max}^{1/2}\left(\widetilde{\mathbf{A}}\widetilde{\mathbf{A}}^\top\right)\sqrt{(2p_1r_1+2p_2r_2+r_1^2+r_2^2)/T}$, then

$$
\begin{aligned}
&\mathbb{P}\left(\xi(r_1,r_2,0,0)\geq x\right)\\
\leq&\left(\frac{48}{0.1}\right)^{2(p_1r_1+p_2r_2)+r_1^2+r_2^2}\\
&\times\left(\exp\left\{-\frac{Tx^2}{6\tau^2\lambda_{\max}^2(\mathbf{\Sigma_e})\lambda_{\max}\left(\widetilde{\mathbf{A}}\widetilde{\mathbf{A}}^\top\right)}\right\}+\exp\left\{-15(p_1r_1+p_2r_2+4r_1^2r_2^2)\right\}\right)\\
\leq&\exp\left\{(7-8)(2p_1r_1+2p_2r_2+r_1^2+r_2^2)\right\}\\
&+2\exp\left\{7(2p_1r_1+2p_2r_2+r_1^2+r_2^2)-15(p_1r_1+p_2r_2+4r_1^2r_2^2)\right\}\\
\leq&C\exp\{-C(p_1r_1+p_2r_2)\}.
\end{aligned}
$$

Define $\mathrm{df}_{\mathrm{RRMAR}}=2p_1r_1+2p_2r_2+r_1^2+r_2^2$ and $M_1=\lambda_{\max}(\mathbf{\Sigma_e})/\mu_{\min}^{1/2}(\mathcal{A})=\lambda_{\max}(\mathbf{\Sigma_e})\lambda_{\max}^{1/2}\left(\widetilde{\mathbf{A}}\widetilde{\mathbf{A}}^\top\right)$. Therefore, when $T\gtrsim(p_1r_1+p_2r_2+4r_1^2r_2^2)M_2^{-2}\max(\tau^2,\tau^4)$, we have that with probability at least $1-C\exp\{-C(p_1r_1+p_2r_2)\}$,

$$
\xi(r_1,r_2,0,0)\lesssim\tau M_1\sqrt{\frac{\mathrm{df}_{\mathrm{RRMAR}}}{T}}.
$$

Next, we construct the upper bound of ξ when common spaces exist. By Lemma S.7, now $\mathbf{W}_2-\overline{\mathbf{W}}_2$ can be decomposed as $\mathbf{W}_2-\overline{\mathbf{W}}_2=\mathbf{\Delta}_{2,1}+\mathbf{\Delta}_{2,2}+\mathbf{\Delta}_{2,3}+\mathbf{\Delta}_{2,4}$, where $\mathbf{\Delta}_{2,1},\mathbf{\Delta}_{2,2}$ are both rank-r_2 with common dimension d_2. $\mathbf{\Delta}_{2,3},\mathbf{\Delta}_{2,4}$ are rank-r_2. Moreover, $\langle\mathbf{\Delta}_{2,j},\mathbf{\Delta}_{2,k}\rangle=0$ for any $j,k=1,2,3,4, j\neq k$. Then with Cauchy's inequality and $\left\|\sum_{s=1}^4\mathbf{\Delta}_{i,s}\right\|_{\mathrm{F}}^2=\sum_{s=1}^4\|\mathbf{\Delta}_{i,s}\|_{\mathrm{F}}^2$ we know that $\sum_{s=1}^4\|\mathbf{\Delta}_{i,s}\|_{\mathrm{F}}\leq 2\left\|\sum_{s=1}^4\mathbf{\Delta}_{i,s}\right\|_{\mathrm{F}}\leq 2\varepsilon, i=1,2$. Similarly, we can decompose each side of the Kronecker product into four parts. The first two parts contain common space while the last two do not. Thus, $\xi(r_1,r_2,d_1,d_2)$ can be upper bounded by the covering of $\mathcal{V}(r_1, r_2, d_1, d_2)$ and $\xi(r_1, r_2, 0, 0)$ as follows:

$$\begin{aligned}
&\left\langle \frac{1}{T}\sum_{t=1}^{T} \mathbf{e}_t \mathbf{y}_{t-1}^\top, \mathbf{A} \right\rangle \\
=&\left\langle \frac{1}{T}\sum_{t=1}^{T} \mathbf{e}_t \mathbf{y}_{t-1}^\top, \overline{\mathbf{A}} \right\rangle \\
&+ \sum_{s=1}^{4} \left\langle \frac{1}{T}\sum_{t=1}^{T} \mathbf{e}_t \mathbf{y}_{t-1}^\top, \frac{\mathbf{\Delta}_{2,s}}{\|\mathbf{\Delta}_{2,s}\|_{\mathrm{F}}} \otimes \mathbf{W}_1 \right\rangle \|\mathbf{\Delta}_{2,s}\|_{\mathrm{F}} \\
&+ \sum_{s=1}^{4} \left\langle \frac{1}{T}\sum_{t=1}^{T} \mathbf{e}_t \mathbf{y}_{t-1}^\top, \overline{\mathbf{W}}_2 \otimes \frac{\mathbf{\Delta}_{1,s}}{\|\mathbf{\Delta}_{1,s}\|_{\mathrm{F}}} \right\rangle \|\mathbf{\Delta}_{1,s}\|_{\mathrm{F}} \\
\leq& \max_{\overline{A} \in \overline{\mathcal{V}}(r_1,r_2,d_1,d_2)} \left\langle \frac{1}{T}\sum_{t=1}^{T} \mathbf{e}_t \mathbf{y}_{t-1}^\top, \overline{\mathbf{A}} \right\rangle + 4\varepsilon\xi(r_1, r_2, 0, 0) + 4\varepsilon\xi(r_1, r_2, d_1, d_2).
\end{aligned}$$

Therefore,

$$\xi(r_1, r_2, d_1, d_2) \leq (1-4\varepsilon)^{-1} \left(\max_{\overline{A} \in \overline{\mathcal{V}}(r_1,r_2,d_1,d_2)} \left\langle \frac{1}{T}\sum_{t=1}^{T} \mathbf{e}_t \mathbf{y}_{t-1}^\top, \overline{\mathbf{A}} \right\rangle + 4\varepsilon\xi(r_1, r_2, 0, 0) \right).$$

Similar to the derivation of $\xi(r_1, r_2, 0, 0)$, we have

$$\begin{aligned}
&\mathbb{P}\left(\max_{\overline{A} \in \overline{\mathcal{V}}(r_1,r_2,d_1,d_2)} \left\langle \frac{1}{T}\sum_{t=1}^{T} \mathbf{e}_t \mathbf{y}_{t-1}^\top, \overline{\mathbf{A}} \right\rangle \geq x \right) \\
\leq& |\overline{\mathcal{V}}(r_1, r_2, d_1, d_2)| \left(\exp\left\{ -\frac{Tx^2}{3\tau^2 \lambda_{\max}^2(\mathbf{\Sigma}_{\mathbf{e}}) \lambda_{\max}\left(\widetilde{\mathbf{A}}\widetilde{\mathbf{A}}^\top\right)} \right\} + \exp\left\{-C(p_1 r_1 + p_2 r_2 + 4r_1^2 r_2^2)\right\} \right).
\end{aligned}$$

Define

$$\mathrm{df}_{\mathrm{MARCF}} := p_1(2r_1 - d_1) + p_2(2r_2 - d_2) + r_1^2 + r_2^2.$$

Taking $\varepsilon = 0.1$ and $x = C\tau\lambda_{\max}(\mathbf{\Sigma}_{\mathbf{e}}) \lambda_{\max}^{1/2}\left(\widetilde{\mathbf{A}}\widetilde{\mathbf{A}}^\top\right) \sqrt{\mathrm{df}_{\mathrm{MARCF}}/T}$ and with (S4.5), similarly we have

$$\mathbb{P}\left(\max_{\overline{A} \in \overline{\mathcal{V}}(r_1,r_2,d_1,d_2)} \left\langle \frac{1}{T}\sum_{t=1}^{T} \mathbf{e}_t \mathbf{y}_{t-1}^\top, \overline{\mathbf{A}} \right\rangle \gtrsim \tau M_1 \sqrt{\mathrm{df}_{\mathrm{MARCF}}/T} \right) \leq \exp\left\{-C(p_1 r_1 + p_2 r_2)\right\}$$

Combining the upper bound of $\xi(r_1, r_2, 0, 0)$, we have

$$\xi(r_1, r_2, d_1, d_2) \lesssim \tau M_1 \left(\sqrt{\frac{\mathrm{df}_{\mathrm{MARCF}}}{T}} + \sqrt{\frac{\mathrm{df}_{\mathrm{RRMAR}}}{T}} \right),$$

with probability at least $1-C\exp\left(-C(p_1r_1+p_2r_2)\right)$. Moreover, since $\mathrm{df}_{\mathrm{RRMAR}} \leq 2\mathrm{df}_{\mathrm{MARCF}}$,

$$\xi(r_1,r_2,d_1,d_2) \lesssim \tau M_1\sqrt{\frac{\mathrm{df}_{\mathrm{MARCF}}}{T}}.$$

$\square$

With this upper bound, we verify the condition on the upper bound of ξ in Theorem 1.

Proposition S.2. *Under Assumptions 1 and 2, if* $T \gtrsim (p_1r_1+p_2r_2+4r_1^2r_2^2)M_2^{-2}\max\left(\tau^2,\tau^4\right)$ *as well as* $T \gtrsim \kappa^2\underline{\sigma}^{-4}\alpha_{\mathrm{RSC}}^{-3}\beta_{\mathrm{RSS}}\tau^2M_1^2(p_1r_1+p_2r_2)$*, then we have that, with probability at least* $1-C\exp\left(-C(p_1r_1+p_2r_2)\right)$*,*

$$\xi^2 \lesssim \frac{\phi^4\alpha_{\mathrm{RSC}}^3}{\kappa^6\beta_{\mathrm{RSS}}}.$$

Proof of Proposition S.2. By Proposition S.1, we know that with probability at least $1-C\exp\left(-C(p_1r_1+p_2r_2)\right)$,

$$\xi(r_1,r_2,d_1,d_2) \lesssim \tau M_1\sqrt{\frac{\mathrm{df}_{\mathrm{MARCF}}}{T}}.$$

Then, when $T \gtrsim \kappa^2\underline{\sigma}^{-4}\alpha_{\mathrm{RSC}}^{-3}\beta_{\mathrm{RSS}}\tau^2M_1^2(p_1r_1+p_2r_2) \gtrsim \kappa^2\underline{\sigma}^{-4}\alpha_{\mathrm{RSC}}^{-3}\beta_{\mathrm{RSS}}\tau^2M_1^2\mathrm{df}_{\mathrm{MARCF}}$, we have

$$\xi^2 \lesssim \frac{\tau^2M_1^2\mathrm{df}_{\mathrm{MARCF}}}{\kappa^2\underline{\sigma}^{-4}\alpha_{\mathrm{RSC}}^{-3}\beta_{\mathrm{RSS}}\tau^2M_1^2\mathrm{df}_{\mathrm{MARCF}}} = \frac{\phi^4\alpha_{\mathrm{RSC}}^3}{\kappa^6\beta_{\mathrm{RSS}}}.$$

$\square$

S4.4 Properties of Initialization

In this appendix, we establish the theoretical guarantee for the initialization procedure in Section 3.2.

Proof of Proposition 2. Letting $\boldsymbol{\Delta} = \widehat{\mathbf{A}}^{\mathrm{VAR}} - \mathbf{A}^*$, by the optimality of $\widehat{\mathbf{A}}^{\mathrm{VAR}}$, we have

$$\widetilde{\mathcal{L}}(\mathbf{A}^*+\boldsymbol{\Delta}) + \lambda_0\|\mathbf{A}^*+\boldsymbol{\Delta}\|_* \leq \widetilde{\mathcal{L}}(\mathbf{A}^*) + \lambda_0\|\mathbf{A}^*\|_*.$$

Expanding the quadratic loss gives

$$\widetilde{\mathcal{L}}(\mathbf{A}^* + \mathbf{\Delta}) - \widetilde{\mathcal{L}}(\mathbf{A}^*) = -\left\langle \frac{1}{T}\sum_{t=1}^{T} \mathbf{e}_t \mathbf{y}_{t-1}^\top, \mathbf{\Delta} \right\rangle + \frac{1}{2T}\sum_{t=0}^{T-1} \|\mathbf{\Delta}\mathbf{y}_t\|_2^2.$$

Noting that $\text{rank}(\mathbf{A}^*) = r_1 r_2$, we let the compact SVD of $\mathbf{A}^*$ be $\mathbf{A}^* = \mathbf{U}^*\mathbf{S}^*\mathbf{V}^{*\top}$, where $\mathbf{U}^*, \mathbf{V}^* \in \mathbb{O}^{p_1p_2 \times r_1r_2}$. By Holder's inequality and the triangle inequality, we have

$$\begin{aligned} \frac{1}{2T}\sum_{t=0}^{T-1} \|\mathbf{\Delta}\mathbf{y}_t\|_2^2 &\leq \left\langle \frac{1}{T}\sum_{t=1}^{T} \mathbf{e}_t \mathbf{y}_{t-1}^\top, \mathbf{\Delta} \right\rangle + \lambda_0 \left(\|\mathbf{A}^*\|_* - \|\mathbf{A}^* + \mathbf{\Delta}\|_*\right) \\ &\leq \left\| \frac{1}{T}\sum_{t=1}^{T} \mathbf{e}_t \mathbf{y}_{t-1}^\top \right\|_{\text{op}} \|\mathbf{\Delta}\|_* + \lambda_0 \left(\|\mathbf{A}^*\|_* - \|\mathbf{A}^* + \mathbf{\Delta}\|_*\right). \end{aligned} \tag{S4.7}$$

The first part of proof follows the literature on generic nuclear-norm penalized matrix estimation (Negahban and Wainwright, 2011). Denote $\mathbb{T}_{\mathbf{A}^*}$ to be the tangent space at $\mathbf{A}^*$:

$$\mathbb{T}_{\mathbf{A}^*} = \left\{ \mathbf{U}^*\mathbf{B}_1^\top + \mathbf{B}_2\mathbf{V}^{*\top} : \mathbf{B}_1, \mathbf{B}_2 \in \mathbb{R}^{p_1p_2 \times r_1r_2} \right\}.$$

Thus, the orthogonal complement of $\mathbb{T}_{\mathbf{A}^*}$ is defined as

$$\mathbb{T}_{\mathbf{A}^*}^\perp = \left\{ \mathbf{M} \in \mathbb{R}^{p_1p_2 \times p_1p_2} : \mathbf{U}^{*\top}\mathbf{M} = \mathbf{0},\, \mathbf{M}\mathbf{V}^* = \mathbf{0} \right\}.$$

Define $\mathbf{\Delta}''$ to be the projection of $\mathbf{\Delta}$ onto $\mathbb{T}_{\mathbf{A}^*}^\perp$ and $\mathbf{\Delta}' := \mathbf{\Delta} - \mathbf{\Delta}''$. We have

$$\|\mathbf{\Delta} + \mathbf{A}^*\|_* = \|\mathbf{\Delta}' + \mathbf{\Delta}'' + \mathbf{A}^*\|_* \geq \|\mathbf{\Delta}'' + \mathbf{A}^*\|_* - \|\mathbf{\Delta}'\|_* = \|\mathbf{\Delta}''\|_* + \|\mathbf{A}^*\|_* - \|\mathbf{\Delta}'\|_*. \tag{S4.8}$$

The last equality holds because $\mathbf{\Delta}''$ is orthogonal to $\mathbf{A}^*$.

Since $T \gtrsim M_2^{-2} \max(\tau^2, \tau^4)\, p_1p_2$, by Proposition S.3, and choosing a proper constant for λ_0, where $\lambda_0 \asymp \tau M_1\sqrt{p_1p_2/T}$, we have $\left\|\sum_{t=1}^T \mathbf{e}_t\mathbf{y}_{t-1}^\top/T\right\|_{\text{op}} \leq \lambda_0/2$. Therefore, for (S4.7), plugging in (S4.8), we have

$$\frac{1}{2T}\sum_{t=0}^{T-1} \|\mathbf{\Delta}\mathbf{y}_t\|_2^2 \leq \frac{\lambda_0}{2}\|\mathbf{\Delta}' + \mathbf{\Delta}''\|_* + \lambda_0(-\|\mathbf{\Delta}''\|_* + \|\mathbf{\Delta}'\|_*) \leq \frac{3\lambda_0}{2}\|\mathbf{\Delta}'\|_*.$$

By Lemma 1 of Negahban and Wainwright (2011), we know that $\text{rank}(\mathbf{\Delta}') \leq 2r_1r_2$. Therefore,

$$\frac{1}{2T}\sum_{t=0}^{T-1} \|\mathbf{\Delta}\mathbf{y}_t\|_2^2 \leq \frac{3\lambda_0}{2}\|\mathbf{\Delta}'\|_* \leq \frac{3\lambda_0}{2}\sqrt{2r_1r_2}\,\|\mathbf{\Delta}'\|_{\text{F}} \leq \frac{3\lambda_0}{2}\sqrt{2r_1r_2}\,\|\mathbf{\Delta}\|_{\text{F}}.$$

Combining (a) of Proposition S.3 and $\lambda_0 \asymp \tau M_1\sqrt{p_1p_2/T}$ with the above inequality, we have

$$\left\|\widehat{\mathbf{A}}^{\mathrm{VAR}} - \mathbf{A}^*\right\|_{\mathrm{F}} \lesssim \alpha_{\mathrm{RSC}}^{-1}\tau M_1\sqrt{\frac{p_1p_2r_1r_2}{T}}.$$

with probability at least $1 - C\exp(-Cp_1p_2)$.

Next, we propagate the error through the nearest Kronecker product approximation. As in Chen et al. (2021), utilizing the re-arrangement operator $\mathcal{P}$,

$$\left\|\mathbf{A}_2 \otimes \mathbf{A}_1 - \widehat{\mathbf{A}}^{\mathrm{VAR}}\right\|_{\mathrm{F}} = \left\|\mathrm{vec}(\mathbf{A}_2)\mathrm{vec}(\mathbf{A}_1)^\top - \mathcal{P}\left[\widehat{\mathbf{A}}^{\mathrm{VAR}}\right]\right\|_{\mathrm{F}}.$$

Define the largest singular value of $\mathcal{P}[\widehat{\mathbf{A}}^{\mathrm{VAR}}]$ as $\widetilde{\sigma}_1$, and the corresponding left and right singular vectors as $\widetilde{\mathbf{u}}$ and $\widetilde{\mathbf{v}}$. By the best rank-1 approximation property of the SVD (the Eckart–Young–Mirsky theorem), the explicit solution to (3.12) is

$$\widehat{\mathbf{A}}_1^{\mathrm{KP}} = \mathrm{mat}_{p_1\times p_1}\left(\widetilde{\sigma}_1^{1/2}\widetilde{\mathbf{v}}\right), \quad \text{and} \quad \widehat{\mathbf{A}}_2^{\mathrm{KP}} = \mathrm{mat}_{p_2\times p_2}\left(\widetilde{\sigma}_1^{1/2}\widetilde{\mathbf{u}}\right),$$

where $\mathrm{mat}_{m\times n}(\cdot)$ is the matricization operator from a vector to a (m,n)-dimensional matrix. Because $\widehat{\mathbf{A}}_2^{\mathrm{KP}} \otimes \widehat{\mathbf{A}}_1^{\mathrm{KP}}$ is the nearest Kronecker product approximation to $\widehat{\mathbf{A}}^{\mathrm{VAR}}$, and $\mathbf{A}^* = \mathbf{A}_2^* \otimes \mathbf{A}_1^*$ has the same Kronecker product structure, we have

$$\left\|\widehat{\mathbf{A}}_2^{\mathrm{KP}} \otimes \widehat{\mathbf{A}}_1^{\mathrm{KP}} - \mathbf{A}^*\right\|_{\mathrm{F}} \le \left\|\widehat{\mathbf{A}}_2^{\mathrm{KP}} \otimes \widehat{\mathbf{A}}_1^{\mathrm{KP}} - \widehat{\mathbf{A}}^{\mathrm{VAR}}\right\|_{\mathrm{F}} + \left\|\widehat{\mathbf{A}}^{\mathrm{VAR}} - \mathbf{A}^*\right\|_{\mathrm{F}} \le 2\left\|\widehat{\mathbf{A}}^{\mathrm{VAR}} - \mathbf{A}^*\right\|_{\mathrm{F}}.$$

For $T \gtrsim \phi^{-4}\alpha_{\mathrm{RSC}}^{-2}\tau^2M_1^2p_1p_2r_1r_2$ with a suitable constant, we have

$$\left\|\widehat{\mathbf{A}}_2^{\mathrm{KP}} \otimes \widehat{\mathbf{A}}_1^{\mathrm{KP}} - \mathbf{A}^*\right\|_{\mathrm{F}}^2 \le \|\mathbf{A}^*\|_{\mathrm{F}}^2,$$

which means that the angle between $\mathrm{vec}(\widehat{\mathbf{A}}_2^{\mathrm{KP}} \otimes \widehat{\mathbf{A}}_1^{\mathrm{KP}})$ and $\mathrm{vec}(\mathbf{A}^*)$ is an acute angle. Therefore, by an argument analogous to the derivation of (S3.20), we can separate the estimation error as

$$\begin{aligned}\left\|\widehat{\mathbf{A}}_2^{\mathrm{KP}}-\mathbf{A}_2^*\right\|_{\mathrm{F}}^2+\left\|\widehat{\mathbf{A}}_1^{\mathrm{KP}}-\mathbf{A}_1^*\right\|_{\mathrm{F}}^2 \leq& \frac{4}{\phi^2}\left\|\widehat{\mathbf{A}}_2^{\mathrm{KP}}\otimes\widehat{\mathbf{A}}_1^{\mathrm{KP}}-\mathbf{A}_2^*\otimes\mathbf{A}_1^*\right\|_{\mathrm{F}}^2\\ \leq& \frac{16}{\phi^2}\left\|\widehat{\mathbf{A}}^{\mathrm{VAR}}-\mathbf{A}^*\right\|_{\mathrm{F}}^2\\ \lesssim& \phi^{-2}\alpha_{\mathrm{RSC}}^{-2}\tau^2M_1^2\frac{p_1p_2r_1r_2}{T}.\end{aligned}$$

Next, we account for the rank-truncation step applied after the NKP approximation. Again, since $\widehat{\mathbf{A}}_i^{\mathrm{RR}}$ denotes the best rank-r_i approximation of $\widehat{\mathbf{A}}_i^{\mathrm{KP}}$ in Frobenius norm obtained by truncated SVD and $\mathrm{rank}(\mathbf{A}_i^*)=r_i$, the triangle inequality and the Eckart–Young–Mirsky theorem give

$$\left\|\widehat{\mathbf{A}}_i^{\mathrm{RR}}-\mathbf{A}_i^*\right\|_{\mathrm{F}}\leq 2\left\|\widehat{\mathbf{A}}_i^{\mathrm{KP}}-\mathbf{A}_i^*\right\|_{\mathrm{F}},\quad \text{for } i=1,2.$$

Consequently,

$$\left\|\widehat{\mathbf{A}}_1^{\mathrm{RR}}-\mathbf{A}_1^*\right\|_{\mathrm{F}}^2+\left\|\widehat{\mathbf{A}}_2^{\mathrm{RR}}-\mathbf{A}_2^*\right\|_{\mathrm{F}}^2\lesssim\phi^{-2}\alpha_{\mathrm{RSC}}^{-2}\tau^2M_1^2\frac{p_1p_2r_1r_2}{T}.$$

Finally, with Assumption 3, applying a matrix perturbation analysis analogous to Lemma B.4 in Wang et al. (2023), we bound the error of the final decomposition step as

$$\mathrm{dist}_{(0)}^2\lesssim\phi^{8/3}\underline{\sigma}^{-4}g_{\min}^{-4}\left(\left\|\widehat{\mathbf{A}}_1^{\mathrm{RR}}-\mathbf{A}_1^*\right\|_{\mathrm{F}}^2+\left\|\widehat{\mathbf{A}}_2^{\mathrm{RR}}-\mathbf{A}_2^*\right\|_{\mathrm{F}}^2\right).$$

Combining the above two upper bounds, we have the initialization error bound as

$$\mathrm{dist}_{(0)}^2\lesssim\phi^{2/3}\underline{\sigma}^{-4}g_{\min}^{-4}\alpha_{\mathrm{RSC}}^{-2}\tau^2M_1^2\frac{p_1p_2r_1r_2}{T}.$$

□

The following proposition proves two auxiliary bounds used in the proof of Proposition 2. The first bound corresponds to the RSC property of the VAR loss with respect to arbitrary test matrices, and the second one corresponds to the statistical error of the convex nuclear-norm penalized VAR estimator, which is also the key to verify the choice of λ_0.

Proposition S.3. *Under Assumptions 1 and 2, if* $T \gtrsim M_2^{-2} \max(\tau^2, \tau^4) p_1 p_2$*, we have that with probability at least* $1 - C\exp(-Cp_1p_2)$*,*

(a) for any matrix $\mathbf{A} \in \mathbb{R}^{p_1p_2 \times p_1p_2}$*,*

$$\frac{1}{T}\sum_{t=0}^{T-1} \|(\mathbf{A}-\mathbf{A}^*)\mathbf{y}_t\|_2^2 \geq \alpha_{\mathrm{RSC}} \|(\mathbf{A}-\mathbf{A}^*)\|_{\mathrm{F}}^2,$$

and (b)

$$\left\| \frac{1}{T}\sum_{t=1}^{T} \mathbf{e}_t \mathbf{y}_{t-1}^\top \right\|_{\mathrm{op}} \lesssim \tau M_1 \sqrt{\frac{p_1p_2}{T}}.$$

Proof of Proposition S.3. Noting that

$$\frac{1}{T}\sum_{t=0}^{T-1} \|\boldsymbol{\Delta}\mathbf{y}_t\|_2^2 = \frac{1}{T}\sum_{t=0}^{T-1} \mathrm{tr}\left(\boldsymbol{\Delta}\mathbf{y}_t\mathbf{y}_t^\top\boldsymbol{\Delta}^\top\right) \geq \lambda_{\min}\left\{\widehat{\boldsymbol{\Gamma}}_y(0)\right\} \|\boldsymbol{\Delta}\|_{\mathrm{F}}^2,$$

where $\widehat{\boldsymbol{\Gamma}}_y(0) = \sum_{t=1}^{T} \mathbf{y}_{t-1}\mathbf{y}_{t-1}^\top / T$. Defining $\boldsymbol{\Gamma}_y(0) = \mathbb{E}\left(\mathbf{y}_t\mathbf{y}_t^\top\right)$, since Weyl's inequality gives

$$\lambda_{\min}\left\{\widehat{\boldsymbol{\Gamma}}_y(0)\right\} \geq \lambda_{\min}\left\{\boldsymbol{\Gamma}_y(0)\right\} - \left\|\widehat{\boldsymbol{\Gamma}}_y(0) - \boldsymbol{\Gamma}_y(0)\right\|_{\mathrm{op}},$$

we first give a concentration bound for $\|\widehat{\boldsymbol{\Gamma}}_y(0) - \boldsymbol{\Gamma}_y(0)\|_{\mathrm{op}}$. The proof is analogous to the proof of Proposition 1, but we give necessary details for completeness.

Recall the VMA(∞) representation and the related notations defined in Appendix S4.2, we know that $\mathbf{z} = (\mathbf{y}_{T-1}^\top, \ldots, \mathbf{y}_0^\top)^\top = \widetilde{\mathbf{A}}\widetilde{\boldsymbol{\Sigma}}_{\mathbf{e}}^{1/2}\boldsymbol{\zeta}$. Define $\mathbb{S}^{p_1p_2}$ to be the unit sphere in $\mathbb{R}^{p_1p_2}$ with respect to Euclidean norm. For any fixed $\mathbf{v} \in \mathbb{S}^{p_1p_2}$, defining $Q_T(\mathbf{v}) = T \cdot \mathbf{v}^\top\widehat{\boldsymbol{\Gamma}}_y(0)\mathbf{v}$, we have

$$\begin{aligned} Q_T(\mathbf{v}) = T \cdot \mathbf{v}^\top\widehat{\boldsymbol{\Gamma}}_y(0)\mathbf{v} = & \sum_{t=0}^{T-1} \mathbf{y}_t^\top\mathbf{v}\mathbf{v}^\top\mathbf{y}_t \\ = & \mathbf{z}^\top(\mathbf{I}_T \otimes \mathbf{v}\mathbf{v}^\top)\mathbf{z} \\ = & \boldsymbol{\zeta}^\top\widetilde{\boldsymbol{\Sigma}}_{\mathbf{e}}^{1/2}(\widetilde{\mathbf{A}})^\top(\mathbf{I}_T \otimes \mathbf{v}\mathbf{v}^\top)\widetilde{\mathbf{A}}\widetilde{\boldsymbol{\Sigma}}_{\mathbf{e}}^{1/2}\boldsymbol{\zeta} \end{aligned}$$

Here, $Q_T(\mathbf{v})$ is the analogue of $R_T(\boldsymbol{\Delta})$ in the proof of Proposition 1. Thus, by an analogous argument, using the Hanson-Wright inequality, we have that for any $x \geq 0$,

$$\begin{aligned}&\mathbb{P}\left(\left|Q_T(\mathbf{v})-\mathbb{E}Q_T(\mathbf{v})\right| \geq \tau^2 x\right)\\ \leq &2\exp\left\{-C\min\left(\frac{x}{\lambda_{\max}\left(\boldsymbol{\Sigma}_{\mathbf{e}}\right)\lambda_{\max}\left(\widetilde{\mathbf{A}}\widetilde{\mathbf{A}}^\top\right)}, \frac{x^2}{T\lambda_{\max}^2\left(\boldsymbol{\Sigma}_{\mathbf{e}}\right)\lambda_{\max}^2\left(\widetilde{\mathbf{A}}\widetilde{\mathbf{A}}^\top\right)}\right)\right\}.\end{aligned}$$

Next, by the discretization tool in the proof of Proposition 1, we apply the union bound over a ϵ-net $\mathcal{N}$ of $\mathbb{S}^{p_1p_2}$, which has cardinality at most $(3/\varepsilon)^{p_1p_2}$, to get

$$\begin{aligned}&\mathbb{P}\left(\sup_{\mathbf{v}\in\mathbb{S}^{p_1p_2}}\left|Q_T(\mathbf{v})-\mathbb{E}Q_T(\mathbf{v})\right| \geq \frac{T}{2}\lambda_{\min}\left(\boldsymbol{\Sigma}_{\mathbf{e}}\right)\lambda_{\min}\left(\widetilde{\mathbf{A}}\widetilde{\mathbf{A}}^\top\right)\right)\\ \leq &C_3\exp\left\{C_1p_1p_2 - C_2M_2^2\min\left(\tau^{-2},\tau^{-4}\right)T\right\}.\end{aligned}$$

Hence, when $T \gtrsim p_1p_2M_2^{-2}\max\left(\tau^2,\tau^4\right)$, we have that with probability at least $1-C\exp\left\{-Cp_1p_2\right\}$,

$$\left\|\widehat{\boldsymbol{\Gamma}}_y(0)-\boldsymbol{\Gamma}_y(0)\right\|_{\text{op}} = \frac{1}{T}\sup_{\mathbf{v}\in\mathbb{S}^{p_1p_2}}\left|Q_T(\mathbf{v})-\mathbb{E}Q_T(\mathbf{v})\right| \leq \frac{1}{2}\lambda_{\min}\left(\boldsymbol{\Sigma}_{\mathbf{e}}\right)\lambda_{\min}\left(\widetilde{\mathbf{A}}\widetilde{\mathbf{A}}^\top\right) = \alpha_{\text{RSC}}.$$

In addition, by the spectral approach as in Basu and Michailidis (2015), $\lambda_{\min}\left\{\boldsymbol{\Gamma}_y(0)\right\} \geq 2\alpha_{\text{RSC}}$. Finally, combining the two bounds, we have

$$\frac{1}{T}\sum_{t=0}^{T-1}\left\|\boldsymbol{\Delta}\mathbf{y}_t\right\|_2^2 \geq \alpha_{\text{RSC}}\left\|\boldsymbol{\Delta}\right\|_{\text{F}}^2.$$

Next, we give the upper bound for $\left\|\sum_{t=1}^{T}\mathbf{e}_t\mathbf{y}_{t-1}^\top/T\right\|_{\text{op}}$. Using the variational characterization of the operator norm, we have

$$\left\|\sum_{t=1}^{T}\frac{1}{T}\mathbf{e}_t\mathbf{y}_{t-1}^\top\right\|_{\text{op}} = \sup_{\mathbf{u},\mathbf{v}\in\mathbb{S}^{p_1p_2}}\left|\mathbf{u}^\top\left(\sum_{t=1}^{T}\frac{1}{T}\mathbf{e}_t\mathbf{y}_{t-1}^\top\right)\mathbf{v}\right| = \sup_{\mathbf{u},\mathbf{v}\in\mathbb{S}^{p_1p_2}}\left\langle\sum_{t=1}^{T}\frac{1}{T}\mathbf{e}_t\mathbf{y}_{t-1}^\top, \mathbf{u}\mathbf{v}^\top\right\rangle. \tag{S4.9}$$

Note that $\mathbf{u}$ and $\mathbf{v}$ can be viewed as two rank-1 matrices with unit Frobenius norm, and $\mathbf{u}\mathbf{v}^\top = \mathbf{u}\otimes\mathbf{v}^\top$. The quantity on the right hand side of (S4.9) can be viewed as a rectangular rank-1 analogue of $\xi(1,1,0,0)$ used in (S4.6) in the proof of Proposition S.1, with $\mathbf{W}_1 = \mathbf{u}\in\mathbb{R}^{p_1p_2\times 1}$ and $\mathbf{W}_2 = \mathbf{v}^\top\in\mathbb{R}^{1\times p_1p_2}$. Therefore, we use the same discretization and union bound argument from the proof of the upper bound of $\xi(r_1,r_2,0,0)$ in Appendix S4.3 to obtain that when $T \gtrsim p_1p_2M_2^{-2}\max(\tau^2, \tau^4)$,

$$\left\| \frac{1}{T}\sum_{t=1}^{T} \mathbf{e}_t\mathbf{y}_{t-1}^\top \right\|_{\mathrm{op}} \lesssim \tau M_1 \sqrt{\frac{p_1p_2}{T}}$$

holds with probability at least $1 - C\exp(-C(p_1p_2))$. □

Therefore, when T is large, the initialization condition for local convergence in Theorem 1 holds with high probability.

S4.5 Auxiliary Lemmas

The first lemma is Hanson–Wright inequality and can be found in high-dimensional statistics monograph (Wainwright, 2019).

Lemma S.5. *(Wainwright, 2019) Given random variables $\{X_i\}_{i=1}^n$ and a positive semidefinite matrix $\mathbf{Q} \in \mathcal{S}_+^{n\times n}$, consider the random quadratic form*

$$Z = \sum_{i=1}^{n}\sum_{j=1}^{n} \mathbf{Q}_{ij}X_iX_j.$$

If the random variables $\{X_i\}_{i=1}^n$ are i.i.d. with mean zero, unit variance, and σ-sub-Gaussian, then there are universal constants (c_1, c_2) such that

$$\mathbb{P}[|Z - \mathbb{E}Z| \geq \sigma t] \leq 2\exp\left\{ -\min\left(\frac{c_1 t}{\|\mathbf{Q}\|_2}, \frac{c_2 t^2}{\|\mathbf{Q}\|_{\mathrm{F}}^2} \right) \right\},$$

where $\|\mathbf{Q}\|_2$ and $\|\mathbf{Q}\|_{\mathrm{F}}$ denote the operator and Frobenius norms, respectively.

The second lemma establishes the covering number of $\mathcal{S}(r_1, r_2, p_1, p_2)$ introduced in the proof of Proposition 1.

Lemma S.6. *Define the set of unit-norm differences of low-rank Kronecker products as:*

$$\mathcal{S}(r_1, r_2, p_1, p_2) := \Big\{ \mathbf{M} : \mathbf{M} = \mathbf{A}_2 \otimes \mathbf{A}_1 - \mathbf{A}_2' \otimes \mathbf{A}_1', \quad \mathbf{A}_i, \mathbf{A}_i' \in \mathbb{R}^{p_i\times p_i},\ \mathrm{rank}(\mathbf{A}_i),\ \mathrm{rank}(\mathbf{A}_i') \leq r_i,\ \|\mathbf{M}\|_{\mathrm{F}} = 1,\ i = 1, 2 \Big\}.$$

Let $\overline{\mathcal{S}}(r_1, r_2, p_1, p_2)$ be an ε-net of $\mathcal{S}(r_1, r_2, p_1, p_2)$ with respect to the Frobenius norm, where $\varepsilon \in (0, 1)$. Then,

$$\left|\overline{\mathcal{S}}(r_1, r_2, p_1, p_2)\right| \leq \left(\frac{15}{\varepsilon}\right)^{4p_1r_1+4p_2r_2+16r_1^2r_2^2}.$$

Proof of Lemma S.6. For brevity, given (r_1, r_2, p_1, p_2), we use $\mathcal{S}$ and $\overline{\mathcal{S}}$ to represent $\mathcal{S}(r_1, r_2, p_1, p_2)$ and $\overline{\mathcal{S}}(r_1, r_2, p_1, p_2)$, respectively. Since the proof utilizes tensor operations as techniques, we give some basic notations and tensor algebra relevant to the proof for clarity. We use 4-dimensional tensors to illustrate the concepts.

Our definitions align with Kolda and Bader (2009). For any tensor $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{d_1\times d_2\times d_3\times d_4}$, denote its (i_1, i_2, i_3, i_4)-th entry as $\mathcal{X}_{i_1,i_2,i_3,i_4}$. The Frobenius norm is defined by

$$\|\boldsymbol{\mathcal{X}}\|_{\mathrm{F}} = \sqrt{\sum_{i_1=1}^{d_1}\sum_{i_2=1}^{d_2}\sum_{i_3=1}^{d_3}\sum_{i_4=1}^{d_4}\mathcal{X}_{i_1,i_2,i_3,i_4}^2}.$$

The mode-k matricization of a tensor $\boldsymbol{\mathcal{X}}$, denoted by $\boldsymbol{\mathcal{X}}_{(k)}$, is a $d_k \times (d_1d_2d_3d_4/d_k)$ matrix. Specifically, $\mathcal{X}_{i_1,i_2,i_3,i_4}$ maps to matrix element (i_k, j), where

$$j = 1 + \sum_{\substack{n=1\\n\neq k}}^{4}(i_n - 1)\,J_n \quad \text{with} \quad J_n = \prod_{\substack{m=1\\m\neq k}}^{n-1} d_m.$$

For a matrix $\mathbf{U} \in \mathbb{R}^{p_1\times d_1}$, the 1-mode product, denoted by $\boldsymbol{\mathcal{X}} \times_1 \mathbf{U}$, is of size $p_1 \times d_2 \times d_3 \times d_4$. It is defined by

$$(\boldsymbol{\mathcal{X}} \times_1 \mathbf{U})_{i',j,k,l} = \sum_{i=1}^{d_1}\mathbf{U}_{i',i}\mathcal{X}_{i,j,k,l},$$

The 2,3 and 4-mode products are defined analogously. Let $\boldsymbol{\mathcal{Y}} = \boldsymbol{\mathcal{X}} \times_1 \mathbf{U}_1 \times_2 \mathbf{U}_2 \times_3 \mathbf{U}_3 \times_4 \mathbf{U}_4$ with $\mathbf{U}_k \in \mathbb{R}^{p_k\times d_k}$, its (i', j', k', l')-th entry can be written as

$$\mathcal{Y}_{i',j',k',l'} = \sum_{i=1}^{d_1}\sum_{j=1}^{d_2}\sum_{k=1}^{d_3}\sum_{l=1}^{d_4}\mathcal{X}_{i,j,k,l}(\mathbf{U}_1)_{i',i}(\mathbf{U}_2)_{j',j}(\mathbf{U}_3)_{k',k}(\mathbf{U}_4)_{l',l}. \tag{S4.10}$$

An important property is that

$$\boldsymbol{\mathcal{Y}}_{(k)} = \mathbf{U}_k\boldsymbol{\mathcal{X}}_{(k)}\left(\mathbf{U}_4 \otimes \cdots \otimes \mathbf{U}_{(k+1)} \otimes \mathbf{U}_{(k-1)} \otimes \cdots \otimes \mathbf{U}_1\right)^{\top}. \tag{S4.11}$$

Next, we introduce the concept of Tucker rank. Defining $r_k = \text{rank}(\boldsymbol{\mathcal{X}}_{(k)})$ for $k = 1, 2, 3, 4$, the Tucker rank of $\boldsymbol{\mathcal{X}}$ is given by (r_1, r_2, r_3, r_4). If $\boldsymbol{\mathcal{X}}$ is low-Tucker-rank, i.e., $r_k \leq \bar{r}_k$ for some $\bar{r}_k$ and $k = 1, 2, 3, 4$, it can be decomposed as

$$\boldsymbol{\mathcal{X}} = \boldsymbol{\mathcal{G}} \times_1 \mathbf{U}_1 \times_2 \mathbf{U}_2 \times_3 \mathbf{U}_3 \times_4 \mathbf{U}_4,$$

where $\mathbf{U}_k \in \mathbb{R}^{d_k \times \bar{r}_k}$ has orthonormal columns and $\boldsymbol{\mathcal{G}} \in \mathbb{R}^{\bar{r}_1 \times \bar{r}_2 \times \bar{r}_3 \times \bar{r}_4}$. The decomposition can be obtained by HOSVD, see De Lathauwer et al. (2000) for more details.

Now we start our proof. We first validate that Kronecker products can be viewed as 4-dimensional tensors. For any matrix $\mathbf{M} \in \mathbb{R}^{p_2 p_1 \times p_2 p_1}$, define the rearrangement operator $\mathcal{T} : \mathbb{R}^{p_2 p_1 \times p_2 p_1} \to \mathbb{R}^{p_2 \times p_1 \times p_2 \times p_1}$. The operation is defined element-wise as follows: if $\mathbf{M} = \mathbf{M}_2 \otimes \mathbf{M}_1$ where $\mathbf{M}_i \in \mathbb{R}^{p_i \times p_i}$ for $i = 1, 2$, then

$$[\mathcal{T}(\mathbf{M}_2 \otimes \mathbf{M}_1)]_{i_1, j_1, i_2, j_2} = (\mathbf{M}_2)_{i_1, i_2} (\mathbf{M}_1)_{j_1, j_2}, \text{ for all } i_1, i_2 = 1 \ldots, p_2, \text{ and } j_1, j_2 = 1, \ldots, p_1.$$

Since $\mathcal{T}$ is a bijection, the inverse operator $\mathcal{T}^{-1}$ is well-defined. Furthermore, $\mathcal{T}$ and $\mathcal{T}^{-1}$ are linear isometries with respect to the Frobenius norm. Therefore, we find the ε-net of $\mathcal{S}$ by applying $\mathcal{T}^{-1}$ on the ε-net of $\mathcal{T}(\mathcal{S}) := \{\mathcal{T}(\mathbf{M}) : \mathbf{M} \in \mathcal{S}\}$. Denote $\overline{\mathcal{T}(\mathcal{S})}$ to be the ε-net of $\mathcal{T}(\mathcal{S})$ with respect to the tensor Frobenius norm, we have $|\overline{\mathcal{S}}| = |\overline{\mathcal{T}(\mathcal{S})}|$.

It remains to characterize the structure of $\mathcal{T}(\mathcal{S})$ and bound the covering number of it. For any $\mathbf{M} \in \mathcal{S}$, it is decomposed as $\mathbf{M} = \mathbf{A}_2 \otimes \mathbf{A}_1 - \mathbf{A}'_2 \otimes \mathbf{A}'_1$, where $\mathbf{A}_i, \mathbf{A}'_i \in \mathbb{R}^{p_i \times p_i}$ with $\text{rank}(\mathbf{A}_i)$ and $\text{rank}(\mathbf{A}'_i) \leq r_i$. Define their SVD components as $\mathbf{A}_i = \mathbf{U}_i \mathbf{S}_i \mathbf{V}_i^\top$ and $\mathbf{A}'_i = \mathbf{U}'_i \mathbf{S}'_i \mathbf{V}_i'^\top$, where $\mathbf{U}_i$, $\mathbf{U}'_i$, $\mathbf{V}_i$, and $\mathbf{V}'_i \in \mathbb{O}^{p_i \times r_i}$, for $i = 1, 2$. Denote σ_k and γ_k to be the k-th largest singular value of $\mathbf{A}_2$ and $\mathbf{A}_1$, respectively. Applying $\mathcal{T}$ on $\mathbf{A}_2 \otimes \mathbf{A}_1$, the

(i_1, j_1, i_2, j_2)-th entry is

$$
\begin{aligned}
&[\mathcal{T}(\mathbf{A}_2 \otimes \mathbf{A}_1)]_{i_1,j_1,i_2,j_2} \\
=& \left(\sum_{k=1}^{r_2} \sigma_k (\mathbf{U}_2)_{i_1,k} (\mathbf{V}_2)_{i_2,k}\right) \left(\sum_{l=1}^{r_1} \gamma_l (\mathbf{U}_1)_{j_1,l} (\mathbf{V}_1)_{j_2,l}\right) \\
=& \sum_{k=1}^{r_2} \sum_{l=1}^{r_1} \sigma_k \gamma_l (\mathbf{U}_2)_{i_1,k} (\mathbf{V}_2)_{i_2,k} (\mathbf{U}_1)_{j_1,l} (\mathbf{V}_1)_{j_2,l} \\
=& \sum_{k=1}^{r_2} \sum_{l=1}^{r_1} \sum_{m=1}^{r_2} \sum_{n=1}^{r_1} \sigma_k \gamma_l 1_{\{m=k\}} 1_{\{n=l\}} (\mathbf{U}_2)_{i_1,k} (\mathbf{V}_2)_{i_2,m} (\mathbf{U}_1)_{j_1,l} (\mathbf{V}_1)_{j_2,n},
\end{aligned}
\tag{S4.12}
$$

where $1_{\{\cdot\}}$ is the indicator function. It corresponds to the form of mode product given in (S4.10), which implies that

$$\mathcal{T}(\mathbf{A}_2 \otimes \mathbf{A}_1) = \boldsymbol{\mathcal{C}} \times_1 \mathbf{U}_2 \times_2 \mathbf{U}_1 \times_3 \mathbf{V}_2 \times_4 \mathbf{V}_1,$$

where $\boldsymbol{\mathcal{C}} \in \mathbb{R}^{r_2 \times r_1 \times r_2 \times r_1}$ is a low-dimensional core tensor being defined by the last line of (S4.12). Then, by matricization with (S4.11) and checking the ranks of four modes, we know that $\mathcal{T}(\mathbf{A}_2 \otimes \mathbf{A}_1)$ is low-Tucker-rank with ranks of four modes being at most (r_2, r_1, r_2, r_1). Similar argument holds analogously for $\mathcal{T}(\mathbf{A}_2' \otimes \mathbf{A}_1')$. Therefore, by the property of matrix ranks, the Tucker rank of $\mathcal{T}(\mathbf{M}) = \mathcal{T}(\mathbf{A}_2 \otimes \mathbf{A}_1) - \mathcal{T}(\mathbf{A}_2' \otimes \mathbf{A}_1')$ are at most $(2r_2, 2r_1, 2r_2, 2r_1)$. Applying Lemma S.9, we have that

$$|\overline{\mathcal{S}}| = |\overline{\mathcal{T}(\mathcal{S})}| \leq \left(\frac{15}{\varepsilon}\right)^{4p_1 r_1 + 4p_2 r_2 + 16 r_1^2 r_2^2}.$$

□

The next lemma gives some results on splitting matrices with common dimensions. Its goal is to show that we can decompose the sum of two matrices equipped with common dimensions as the sum of four matrices. Each of them is perpendicular to the others, two of them are low-rank without common dimensions and the other two are equipped with the same common dimensions.

Lemma S.7. *Suppose that there are two $p \times p$ rank-r matrices $\mathbf{W}_1$ and $\mathbf{W}_2$. Then,*

1. *There exist two $p \times p$ rank-r matrices $\widetilde{\mathbf{W}}_1, \widetilde{\mathbf{W}}_2$, such that $\mathbf{W}_1 + \mathbf{W}_2 = \widetilde{\mathbf{W}}_1 + \widetilde{\mathbf{W}}_2$ and $\left\langle \widetilde{\mathbf{W}}_1, \widetilde{\mathbf{W}}_2 \right\rangle = 0$.*

2. *Suppose that now $\mathbf{W}_1$ and $\mathbf{W}_2$ both have common dimension d. That is, Define*

$$\mathcal{W}'(r, d; p) = \big\{\mathbf{W} \in \mathbb{R}^{p\times p} : \mathbf{W} = [\mathbf{C}\ \mathbf{R}]\mathbf{D}[\mathbf{C}\ \mathbf{P}]^\top, \mathbf{C} \in \mathbb{O}^{p\times d}, \mathbf{R}, \mathbf{P} \in \mathbb{O}^{p\times(r-d)},$$
$$\mathbf{C}^\top\mathbf{R} = \mathbf{C}^\top\mathbf{P} = \mathbf{0}\big\}$$

to be the model space of MARCF model with arbitrary scale, $\mathbf{W}_1, \mathbf{W}_2 \in \mathcal{W}'(r, d; p)$. Then there exist $\widetilde{\mathbf{W}}_1, \widetilde{\mathbf{W}}_2 \in \mathcal{W}'(r, d; p)$ and $\widetilde{\mathbf{W}}_3, \widetilde{\mathbf{W}}_4 \in \mathcal{W}'(r, 0; p)$, such that $\mathbf{W}_1 + \mathbf{W}_2 = \widetilde{\mathbf{W}}_1 + \widetilde{\mathbf{W}}_2 + \widetilde{\mathbf{W}}_3 + \widetilde{\mathbf{W}}_4$ and $\left\langle \widetilde{\mathbf{W}}_j, \widetilde{\mathbf{W}}_k \right\rangle = 0$ for every $j, k = 1, 2, 3, 4, j \neq k$.

Proof of Lemma S.7. For the first part of the lemma, by SVD decomposition we know that there exist $p \times r$ matrices $\mathbf{U}_1, \mathbf{U}_2, \mathbf{V}_1, \mathbf{V}_2$ with mutually orthogonal columns, such that $\mathbf{U}_1^\top\mathbf{U}_1 = \mathbf{I}_r, \mathbf{W}_1 = \mathbf{U}_1\mathbf{V}_1^\top$ and $\mathbf{W}_2 = \mathbf{U}_2\mathbf{V}_2^\top$. Let $\widetilde{\mathbf{W}}_1 = \mathbf{U}_1(\mathbf{V}_1^\top + \mathbf{U}_1^\top\mathbf{U}_2\mathbf{V}_2^\top)$ and $\widetilde{\mathbf{W}}_2 = (\mathbf{I} - \mathbf{U}_1\mathbf{U}_1^\top)\mathbf{U}_2\mathbf{V}_2^\top$, then $\widetilde{\mathbf{W}}_1$ and $\widetilde{\mathbf{W}}_2$ satisfy the conditions.

For the second part, decompose and write together $\mathbf{W}_1 + \mathbf{W}_2$ gives

$$\mathbf{W}_1 + \mathbf{W}_2 = [\mathbf{C}_1\ \mathbf{R}_1\ \mathbf{C}_2\ \mathbf{R}_2] \begin{bmatrix} \mathbf{D}_1 & \mathbf{O} \\ \mathbf{O} & \mathbf{D}_2 \end{bmatrix} [\mathbf{C}_1\ \mathbf{P}_1\ \mathbf{C}_2\ \mathbf{P}_2]^\top.$$

By applying Gram-Schmidt orthogonalization procedure (or equivalently, QR decomposition), we can obtain two orthogonal bases,

$$[\mathbf{C}_1\ \widetilde{\mathbf{R}}_1\ \widetilde{\mathbf{C}}_2\ \widetilde{\mathbf{R}}_2] \in \mathbb{O}^{p\times 2r}, \quad [\mathbf{C}_1\ \widetilde{\mathbf{P}}_1\ \widetilde{\mathbf{C}}_2\ \widetilde{\mathbf{P}}_2] \in \mathbb{O}^{p\times 2r}.$$

Then

$$\begin{aligned}
\mathbf{W}_1+\mathbf{W}_2 =&[\mathbf{C}_1\ \widetilde{\mathbf{R}}_1\ \widetilde{\mathbf{C}}_2\ \widetilde{\mathbf{R}}_2]\begin{bmatrix}\widetilde{\mathbf{D}}_1 & \widetilde{\mathbf{D}}_3\\ \widetilde{\mathbf{D}}_4 & \widetilde{\mathbf{D}}_2\end{bmatrix}[\mathbf{C}_1\ \widetilde{\mathbf{P}}_1\ \widetilde{\mathbf{C}}_2\ \widetilde{\mathbf{P}}_2]^\top\\
=&[\mathbf{C}_1\ \widetilde{\mathbf{R}}_1]\widetilde{\mathbf{D}}_1[\mathbf{C}_1\ \widetilde{\mathbf{P}}_1]^\top+[\widetilde{\mathbf{C}}_2\ \widetilde{\mathbf{R}}_2]\widetilde{\mathbf{D}}_2[\widetilde{\mathbf{C}}_2\ \widetilde{\mathbf{P}}_2]^\top\\
&+[\mathbf{C}_1\ \widetilde{\mathbf{R}}_1]\widetilde{\mathbf{D}}_3[\widetilde{\mathbf{C}}_2\ \widetilde{\mathbf{P}}_2]^\top+[\widetilde{\mathbf{C}}_2\ \widetilde{\mathbf{R}}_2]\widetilde{\mathbf{D}}_4[\mathbf{C}_1\ \widetilde{\mathbf{P}}_1]^\top\\
:=&\widetilde{\mathbf{W}}_1+\widetilde{\mathbf{W}}_2+\widetilde{\mathbf{W}}_3+\widetilde{\mathbf{W}}_4.
\end{aligned}$$

It can be directly verified that $\left\langle\widetilde{\mathbf{W}}_j,\widetilde{\mathbf{W}}_k\right\rangle=0$ for every $j,k=1,2,3,4, j\neq k$. $\square$

The next lemma gives the covering number of low rank matrices with common column and row spaces.

Lemma S.8. *(Lemma B.6, Wang et al. (2023)) Define*

$$\mathcal{W}(r,d;p)=\big\{\mathbf{W}\in\mathbb{R}^{p\times p}:\mathbf{W}=[\mathbf{C}\ \mathbf{R}]\mathbf{D}[\mathbf{C}\ \mathbf{P}]^\top,\mathbf{C}\in\mathbb{O}^{p\times d},\mathbf{R},\mathbf{P}\in\mathbb{O}^{p\times(r-d)},\\ \mathbf{C}^\top\mathbf{R}=\mathbf{C}^\top\mathbf{P}=\mathbf{0},\ \text{and}\ \|\mathbf{W}\|_{\mathrm{F}}=1\big\}.$$

Let $\overline{\mathcal{W}}(r,d;p)$ be an ϵ-net of $\mathcal{W}(r,d;p)$, where $\epsilon\in(0,1]$. Then

$$|\overline{\mathcal{W}}(r,d;p)|\leq\left(\frac{24}{\epsilon}\right)^{p(2r-d)+r^2}.$$

Finally, the following lemma establishes the covering number bound for the set of low-Tucker-rank tensors.

Lemma S.9. *(Lemma 2, Rauhut et al. (2017)) Define*

$$\mathcal{H}=\left\{\boldsymbol{\mathcal{X}}\in\mathbb{R}^{n_1\times n_2\times\cdots\times n_N}:\mathrm{rank}(\boldsymbol{\mathcal{X}}_{(k)})\leq r_k,k=1,2,...,N,\|\boldsymbol{\mathcal{X}}\|_{\mathrm{F}}=1\right\}.$$

For any $\varepsilon\in(0,1)$, the covering number of $\mathcal{H}$ with respect to the Frobenius norm satisfies:

$$\mathcal{N}\left(\mathcal{H},\|\cdot\|_{\mathrm{F}},\varepsilon\right)\leq(3(N+1)/\varepsilon)^{r_1r_2\cdots r_N+\sum_{k=1}^N n_kr_k}.$$

S5 Consistency of Rank Selection

Proof of Theorem 3: We focus on the consistency for selecting r_1, as the result for r_2 can be developed analogously. Under Assumptions 1 and 2, with $T \gtrsim \{M_2^{-2}\max(\tau^2,\tau^4) + \phi^{-4}\alpha_{\text{RSC}}^{-2}\tau^2 M_1^2 r_1 r_2\}p_1 p_2$, Appendix S4.4 implies the first-stage consistency of the VAR-based estimator. Since $\bar{r}_i > r_i = \text{rank}(\mathbf{A}_i^*)$, the Eckart–Young–Mirsky theorem gives the same truncation bound for $\widehat{\mathbf{A}}_i^{\text{RR}}(\bar{r}_i)$ from the proof of Proposition 2. Hence,

$$\left\|\widehat{\mathbf{A}}_1^{\text{RR}}(\bar{r}_1) - \mathbf{A}_1^*\right\|_{\text{F}}^2 + \left\|\widehat{\mathbf{A}}_2^{\text{RR}}(\bar{r}_2) - \mathbf{A}_2^*\right\|_{\text{F}}^2 \lesssim \phi^{-2}\alpha_{\text{RSC}}^{-2}\tau^2 M_1^2 \frac{p_1 p_2 r_1 r_2}{T}.$$

with probability approaching 1 as $T \to \infty$ and $p_1, p_2 \to \infty$. Since $\widehat{\mathbf{A}}_1^{\text{RR}} - \mathbf{A}_1^*$ is rank-$(\bar{r}_1 + r_1)$, by the fact that L_∞ norm is smaller than L_2 norm and Mirsky's singular value inequality Mirsky (1960),

$$\begin{aligned}
\max_{1\le j\le \bar{r}_1+r_1} |\sigma_j(\widehat{\mathbf{A}}_1^{\text{RR}}(\bar{r}_1)) - \sigma_j(\mathbf{A}_1^*)|^2 &\le \sum_{j=1}^{\bar{r}_1+r_1}\left(\sigma_j(\widehat{\mathbf{A}}_1^{\text{RR}}(\bar{r}_1)) - \sigma_j(\mathbf{A}_1^*)\right)^2 \\
&\le \sum_{j=1}^{\bar{r}_1+r_1}\sigma_j^2(\widehat{\mathbf{A}}_1^{\text{RR}}(\bar{r}_1) - \mathbf{A}_1^*) \\
&= \left\|\widehat{\mathbf{A}}_1^{\text{RR}}(\bar{r}_1) - \mathbf{A}_1^*\right\|_{\text{F}}^2 \\
&\lesssim \phi^{-2}\alpha_{\text{RSC}}^{-2}\tau^2 M_1^2 \frac{p_1 p_2 r_1 r_2}{T}.
\end{aligned}$$

Then, $\forall j = 1, 2, ..., \bar{r}_1$,

$$|\sigma_j(\widehat{\mathbf{A}}_1^{\text{RR}}(\bar{r}_1)) - \sigma_j(\mathbf{A}_1^*)| = O(\phi^{-1}\alpha_{\text{RSC}}^{-1}\tau M_1\sqrt{p_1 p_2 r_1 r_2/T}).$$

Next, we show that as $T, p_1, p_2 \to \infty$, the ratio $(\sigma_{j+1}(\widehat{\mathbf{A}}_1^{\text{RR}}(\bar{r}_1)) + s(p_1,p_2,T))/(\sigma_j(\widehat{\mathbf{A}}_1^{\text{RR}}(\bar{r}_1)) + s(p_1,p_2,T))$ achieves its minimum at $j = r_1$. For $j > r_1$, $\sigma_j(\mathbf{A}_1^*) = 0$ and

$$\sigma_j(\widehat{\mathbf{A}}_1^{\text{RR}}(\bar{r}_1)) = O(\phi^{-1}\alpha_{\text{RSC}}^{-1}\tau M_1\sqrt{p_1 p_2 r_1 r_2/T}) = o(s(p_1,p_2,T)).$$

Therefore, we have

$$\frac{\sigma_{j+1}(\widehat{\mathbf{A}}_1^{\text{RR}}(\bar{r}_1)) + s(p_1,p_2,T)}{\sigma_j(\widehat{\mathbf{A}}_1^{\text{RR}}(\bar{r}_1)) + s(p_1,p_2,T)} \to 1.$$

For $j < r_1$,

$$\begin{aligned}
&\lim_{\substack{T\to\infty \\ p_1,p_2\to\infty}} \frac{\sigma_{j+1}(\widehat{\mathbf{A}}_1^{\mathrm{RR}}(\bar{r}_1)) + s(p_1,p_2,T)}{\sigma_j(\widehat{\mathbf{A}}_1^{\mathrm{RR}}(\bar{r}_1)) + s(p_1,p_2,T)} \\
=&\lim_{\substack{T\to\infty \\ p_1,p_2\to\infty}} \frac{\sigma_{j+1}(\mathbf{A}_1^*) + o(s(p_1,p_2,T)) + s(p_1,p_2,T)}{\sigma_j(\mathbf{A}_1^*) + o(s(p_1,p_2,T)) + s(p_1,p_2,T)} \\
=&\lim_{p_1,p_2\to\infty} \frac{\sigma_{j+1}(\mathbf{A}_1^*)}{\sigma_j(\mathbf{A}_1^*)} \leq 1.
\end{aligned}$$

For $j = r_1$,

$$\begin{aligned}
\frac{\sigma_{j+1}(\widehat{\mathbf{A}}_1^{\mathrm{RR}}(\bar{r}_1)) + s(p_1,p_2,T)}{\sigma_j(\widehat{\mathbf{A}}_1^{\mathrm{RR}}(\bar{r}_1)) + s(p_1,p_2,T)} =& \frac{o(s(p_1,p_2,T)) + s(p_1,p_2,T)}{\sigma_{r_1}(\mathbf{A}_1^*) + o(s(p_1,p_2,T)) + s(p_1,p_2,T)} \\
\to& \frac{s(p_1,p_2,T)}{\sigma_{r_1}(\mathbf{A}_1^*)} \\
=& o\left(\min_{1\leq j\leq r_1-1} \frac{\sigma_{j+1}(\mathbf{A}_1^*)}{\sigma_j(\mathbf{A}_1^*)}\right).
\end{aligned}$$

Therefore, when $T, p_1, p_2 \to \infty$, the ratio will finally achieve its minimum at $j = r_1$, with a probability approaching one.

Bibliography

Basu, S. and Michailidis, G. (2015). Regularized estimation in sparse high-dimensional time series models. *The Annals of Statistics*, 43(4):1535 – 1567.

Cai, T. T. and Zhang, A. (2018). Rate-optimal perturbation bounds for singular subspaces with applications to high-dimensional statistics. *The Annals of Statistics*, 46(1):60 – 89.

Chen, R., Xiao, H., and Yang, D. (2021). Autoregressive models for matrix-valued time series. *Journal of Econometrics*, 222(1, Part B):539–560.

De Lathauwer, L., De Moor, B., and Vandewalle, J. (2000). A multilinear singular value decomposition. *SIAM Journal on Matrix Analysis and Applications*, 21(4):1253–1278.

Han, R., Willett, R., and Zhang, A. R. (2022). An optimal statistical and computational framework for generalized tensor estimation. *The Annals of Statistics*, 50(1):1 – 29.

Kolda, T. G. and Bader, B. W. (2009). Tensor decompositions and applications. *SIAM Review*, 51(3):455–500.

Mirsky, L. (1960). Symmetric gauge functions and unitarily invariant norms. *The Quarterly Journal of Mathematics*, 11(1):50–59.

Negahban, S. and Wainwright, M. J. (2011). Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics*, 39(2):1069 – 1097.

Nesterov, Y. (2004). *Introductory Lectures on Convex Optimization: A Basic Course*. Springer New York, NY, 1 edition.

Rauhut, H., Schneider, R., and Stojanac, Ž. (2017). Low rank tensor recovery via iterative hard thresholding. *Linear Algebra and its Applications*, 523:220–262.

Tu, S., Boczar, R., Simchowitz, M., Soltanolkotabi, M., and Recht, B. (2016). Low-rank solutions of linear matrix equations via Procrustes Flow. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48, pages 964–973.

Wainwright, M. J. (2019). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*, pages 21–57. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Wang, D., Zhang, X., Li, G., and Tsay, R. (2023). High-dimensional vector autoregression with common response and predictor factors. *arXiv preprint arXiv:2203.15170*.

Wang, D., Zheng, Y., and Li, G. (2024). High-dimensional low-rank tensor autoregressive time series modeling. *Journal of Econometrics*, 238(1):105544.

Wang, L., Zhang, X., and Gu, Q. (2017). A unified computational and statistical framework for nonconvex low-rank matrix estimation. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54, pages 981–990.