Sparse Tucker Decomposition and Graph Regularization for High-Dimensional Time Series Forecasting

Sijia Xia[1], Michael K. Ng[2], and Xiongjun Zhang[1]

[1]*Central China Normal University and* [2]*Hong Kong Baptist University*

Supplementary Material

This supplementary file is organized into six sections. Section S1 provides proofs of auxiliary lemmas for Theorem 1. Section S2 presents the proof of Theorem 1. Section S3 presents solutions to the subproblems in the PALM algorithm, where the computational complexity of PALM is also given. Section S4 presents the proof of Theorem 2. Section S5 reports additional numerical experiments on synthetic data, traffic flow datasets, comprehensive climate dataset, and macroeconomic dataset. Section S6 gives more discussions about the STDGR models, including parameters analysis, choice of lag order, different initialization, different choices of Tucker rank, construction of graphs, ablation analysis, and the extensions to uniform noise.

S1 Auxiliary Lemmas

For any two tensors $\mathcal{X} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ and $\mathcal{Y} \in \mathbb{R}^{n_4 \times n_5 \times n_6}$, their tensor outer product, denoted by $\mathcal{X} \circ \mathcal{Y} \in \mathbb{R}^{n_1 \times n_2 \times n_3 \times n_4 \times n_5 \times n_6}$, is defined as $(\mathcal{X} \circ \mathcal{Y})_{i_1 i_2 i_3 i_4 i_5 i_6} = \mathcal{X}_{i_1 i_2 i_3} \mathcal{Y}_{i_4 i_5 i_6}$, where $1 \leq i_j \leq n_j, j = 1, \ldots, 6$. For any matrix

$\mathbf{A} \in \mathbb{C}^{n_1 \times n_2}, \mathbf{B} \in \mathbb{C}^{n_3 \times n_4}$, their Kronecker product (Golub and Van Loan, 2013) is defined as

$$\mathbf{A} \otimes \mathbf{B} = \begin{pmatrix} a_{11}\mathbf{B} & a_{12}\mathbf{B} & \cdots & a_{1n_2}\mathbf{B} \\ a_{21}\mathbf{B} & a_{22}\mathbf{B} & \cdots & a_{2n_2}\mathbf{B} \\ \vdots & \vdots & & \vdots \\ a_{n_1 1}\mathbf{B} & a_{n_1 2}\mathbf{B} & \cdots & a_{n_1 n_2}\mathbf{B} \end{pmatrix}.$$

Proposition 1. *For any matrices* $\mathbf{A} \in \mathbb{C}^{n_1 \times n_2}$ *and* $\mathbf{B} \in \mathbb{C}^{n_3 \times n_4}$*, the following assertions hold:*

(1) $(\mathbf{A} \otimes \mathbf{B})' = \mathbf{A}' \otimes \mathbf{B}'$ *(Golub and Van Loan, 2013).*

(2) Given matrices $\mathbf{D} \in \mathbb{C}^{n_2 \times n_5}, \mathbf{E} \in \mathbb{C}^{n_4 \times n_6}$, $(\mathbf{A} \otimes \mathbf{B})(\mathbf{D} \otimes \mathbf{E}) = (\mathbf{AD}) \otimes (\mathbf{BE})$ *(Golub and Van Loan, 2013).*

(3) $\|\mathbf{A} \otimes \mathbf{B}\| = \|\mathbf{A}\|\|\mathbf{B}\|$ *(Gueridi and Kittaneh, 2022).*

(4) $\|\mathbf{A} \otimes \mathbf{B}\|_F = \|\mathbf{A}\|_F \|\mathbf{B}\|_F$ *(Gueridi and Kittaneh, 2022).*

Proposition 2. *For any matrix* $\mathbf{A} \in \mathbb{C}^{n_1 \times n_2}$ *with* $\mathbf{A}'\mathbf{A} = \mathbf{I}_{n_2}$, $\mathbf{B} \in \mathbb{C}^{n_3 \times n_4}$ *with* $\mathbf{B}'\mathbf{B} = \mathbf{I}_{n_4}$*, it holds that*

$$\|\mathbf{A} \otimes \mathbf{B}\| \leq 1.$$

Proof. For any matrix $\mathbf{A}$ satisfying $\mathbf{A}'\mathbf{A} = \mathbf{I}_{n_2}$, it follows from Theorem 3.9 of Stewart and Sun (1990) that $\|\mathbf{A}\| \leq 1$. Taking this together with Proposition 1(3) yields

$$\|\mathbf{A} \otimes \mathbf{B}\| = \|\mathbf{A}\|\|\mathbf{B}\| \leq 1,$$

which completes the proof. □

Proposition 3. *For any vector* $\mathbf{a} \in \mathbb{R}^{n_1}$ *and matrix* $\mathbf{X} \in \mathbb{R}^{n_2 \times n_3}$*, one has*

$$(\mathbf{a} \circ \mathbf{X})_{(2)} = \mathbf{X} \otimes \mathbf{a}', \ (\mathbf{a} \circ \mathbf{X})_{(3)} = \mathbf{X}' \otimes \mathbf{a}'.$$

Proof. For vector $\mathbf{a} = (a_1, \ldots, a_{n_1})' \in \mathbb{R}^{n_1}$ and matrix $\mathbf{X} \in \mathbb{R}^{n_2 \times n_3}$ with the (j, l)-th element x_{jl} $(1 \leq j \leq n_2, 1 \leq l \leq n_3)$, the outer product of $\mathbf{a}$ and $\mathbf{X}$, denoted by $\mathcal{T} = \mathbf{a} \circ \mathbf{X}$, is a third-order tensor with size $n_1 \times n_2 \times n_3$, and its each entry is given by

$$\mathcal{T}_{ijl} = a_i x_{jl}.$$

For the mode-2 unfolding of $\mathcal{T}$, denoted $\mathcal{T}_{(2)} \in \mathbb{R}^{n_2 \times n_1 n_3}$, the corresponding element-wise representation is given by

$$(\mathcal{T}_{(2)})_{j,(l-1)n_3+i} = \mathcal{T}_{ijl} = a_i x_{jl}. \tag{S1.1}$$

Let $\mathbf{B} = \mathbf{X} \otimes \mathbf{a}' \in \mathbb{R}^{n_2 \times n_1 n_3}$. Then the $(j, (l-1)n_3 + i)$-th element of $\mathbf{B}$, denoted by $\mathbf{B}_{j,(l-1)n_3+i}$, is given by

$$\mathbf{B}_{j,(l-1)n_3+i} = x_{jl} a_i,$$

which combined with (S1.1) yields that $\mathcal{T}_{(2)} = (\mathbf{a} \circ \mathbf{X})_{(2)} = \mathbf{X} \otimes \mathbf{a}'$. Similarly, we can deduce that $(\mathbf{a} \circ \mathbf{X})_{(3)} = \mathbf{X}' \otimes \mathbf{a}'$. □

Proposition 4. *For any vector* $\mathbf{a} \in \mathbb{R}^{n_1}$ *and matrix* $\mathbf{X} \in \mathbb{R}^{n_2 \times n_3}$*, one has*

$$\|\mathbf{a} \circ \mathbf{X}\|_F = \|\mathbf{a}\|_2 \|\mathbf{X}\|_F.$$

Proof. Combining Proposition 3 with Proposition 1, we have

$$\|\mathbf{a} \circ \mathbf{X}\|_F = \|(\mathbf{a} \circ \mathbf{X})_{(2)}\|_F = \|\mathbf{X} \otimes \mathbf{a}'\|_F = \|\mathbf{X}\|_F \|\mathbf{a}'\|_2 = \|\mathbf{X}\|_F \|\mathbf{a}\|_2.$$

The proof is completed. □

Consider the weighted and undirected graph $G_i = (V_i, E_i, \widetilde{\mathbf{W}}_i), i = 1, 2, 3$, where V_i is the finite set of vertices based on the row of $\mathbf{A}_i$ and E_i the finite set of edges, and $\widetilde{\mathbf{W}}_i$ denotes the weighted adjacency matrix. Assume that the graph G_i is connected. Then we have the following inequality with respect to the graph regularization term.

Lemma 1. *Denote $\Delta_{\mathbf{A}_i} := \widehat{\mathbf{A}}_i - \mathbf{A}_i$. Suppose that Assumption 5 holds and the graph G_i is connected. Then the matrix $\mathbf{A}_i \in \mathbb{R}^{m \times r_i}, i = 1, 2, \mathbf{A}_3 \in \mathbb{R}^{p \times r_3}$ satisfy the following inequality:*

$$\operatorname{tr}\big(\Delta'_{\mathbf{A}_i} \mathbf{L}_i \Delta_{\mathbf{A}_i}\big) \geq \frac{\lambda_2(\mathbf{L}_i)}{2} \left\| \widehat{\mathbf{A}}_i - \mathbf{A}_i \right\|_F^2.$$

where $\lambda_2(\mathbf{L}_i) > 0$ denotes the second smallest eigenvalue of $\mathbf{L}_i, i = 1, 2, 3$.

Proof. Let the eigenvalue decomposition of $\mathbf{L}_1 \in \mathbb{R}^{m \times m}$ be $\mathbf{L}_1 = \mathbf{Q}\Lambda\mathbf{Q}'$. Define the matrix $\mathbf{M} := \mathbf{Q}'\Delta_{\mathbf{A}_1} \in \mathbb{R}^{m \times r_1}$. Then the l-th column of $\mathbf{M}$ is given by $\mathbf{M}_{:,l} = \mathbf{Q}'(\Delta_{\mathbf{A}_1})_{:,l}$, where $(\Delta_{\mathbf{A}_1})_{:,l}$ denotes the l-th column of $\Delta_{\mathbf{A}_1}$. Note that $\mathbf{L}_1$ is a symmetric semi-positive definite matrix with real and non-negative eigenvalues. Let $\lambda_j(\mathbf{L}_1)$ be the j-th smallest eigenvalue of $\mathbf{L}_1$

with $\lambda_1(\mathbf{L}_1) \leq \lambda_2(\mathbf{L}_1) \leq \cdots \leq \lambda_m(\mathbf{L}_1)$. It can be easily demonstrated that

$$\begin{aligned} \operatorname{tr}\left(\Delta_{\mathbf{A}_1}' \mathbf{L}_1 \Delta_{\mathbf{A}_1}\right) &= \sum_{l=1}^{r_1} (\Delta_{\mathbf{A}_1})_{:,l}' \mathbf{L}_1 (\Delta_{\mathbf{A}_1})_{:,l} \\ &= \sum_{l=1}^{r_1} (\Delta_{\mathbf{A}_1})_{:,l}' \mathbf{Q} \Lambda \mathbf{Q}' (\Delta_{\mathbf{A}_1})_{:,l} = \sum_{l=1}^{r_1} \sum_{j=1}^{m} \lambda_j(\mathbf{L}_1) \mathbf{M}_{jl}^2, \end{aligned} \tag{S1.2}$$

where $\mathbf{M}_{jl}$ denotes the (j,l)-th entry of $\mathbf{M}$. By Section 3.1 of Yang et al. (2019) and Theorem 5 of Hirn (2021), we can conclude that $\lambda_1(\mathbf{L}_1) = 0$ and $\lambda_2(\mathbf{L}_1) > 0$. Then, we have

$$\begin{aligned} \sum_{j=1}^{m} \lambda_j(\mathbf{L}_1) \mathbf{M}_{jl}^2 = \sum_{j=2}^{m} \lambda_j(\mathbf{L}_1) \mathbf{M}_{jl}^2 &\geq \sum_{j=2}^{m} \lambda_2(\mathbf{L}_1) \mathbf{M}_{jl}^2 \\ &= \sum_{j=1}^{m} \lambda_2(\mathbf{L}_1) \mathbf{M}_{jl}^2 - \lambda_2(\mathbf{L}_1) \mathbf{M}_{1l}^2 \\ &= \lambda_2(\mathbf{L}_1) \left\| (\Delta_{\mathbf{A}_1})_{:,l} \right\|_2^2 - \lambda_2(\mathbf{L}_1) \mathbf{M}_{1l}^2. \end{aligned} \tag{S1.3}$$

Substituting (S1.3) into (S1.2) yields

$$\begin{aligned} \operatorname{tr}\left(\Delta_{\mathbf{A}_1}' \mathbf{L}_1 \Delta_{\mathbf{A}_1}\right) &\geq \sum_{l=1}^{r_1} \left(\lambda_2(\mathbf{L}_1) \left\| (\Delta_{\mathbf{A}_1})_{:,l} \right\|_2^2 - \lambda_2(\mathbf{L}_1) \mathbf{M}_{1l}^2 \right) \\ &= \lambda_2(\mathbf{L}_1) \|\Delta_{\mathbf{A}_1}\|_F^2 - \lambda_2(\mathbf{L}_1) \left\| (\mathbf{Q}_{:,1})' \Delta_{\mathbf{A}_1} \right\|_2^2, \end{aligned} \tag{S1.4}$$

where the last equality follows from the fact $\mathbf{M} = \mathbf{Q}' \Delta_{\mathbf{A}_1}$. Note that the first column of $\mathbf{Q}$ satisfies $\mathbf{Q}_{:,1} = (1/\sqrt{m}, 1/\sqrt{m}, \ldots, 1/\sqrt{m})'$ (Hirn, 2021). Consequently, we get that $\|(\mathbf{Q}_{:,1})' \Delta_{\mathbf{A}_1}\|_2^2 = \sum_{l=1}^{r_1} \frac{1}{m} \left(\sum_{j=1}^{m} (\Delta_{\mathbf{A}_1})_{jl} \right)^2$. Combining $\|\Delta_{\mathbf{A}_1}\|_F^2 = \sum_{l=1}^{r_1} \sum_{j=1}^{m} (\Delta_{\mathbf{A}_1})_{jl}^2$ with (S1.4) yields

$$\operatorname{tr}\left(\Delta_{\mathbf{A}_1}' \mathbf{L}_1 \Delta_{\mathbf{A}_1}\right) \geq \lambda_2(\mathbf{L}_1) \sum_{l=1}^{r_1} \sum_{j=1}^{m} (\Delta_{\mathbf{A}_1})_{jl}^2 - \lambda_2(\mathbf{L}_1) \sum_{l=1}^{r_1} \frac{1}{m} \Big(\sum_{j=1}^{m} (\Delta_{\mathbf{A}_1})_{jl} \Big)^2.$$

By Assumption 5, we can conclude that $\sum_{l=1}^{r_1}\frac{1}{m}(\sum_{j=1}^{m}(\Delta_{\mathbf{A}_1})_{jl})^2 \leq \frac{1}{2}\sum_{l=1}^{r_1}\sum_{j=1}^{m}(\Delta_{\mathbf{A}_1})_{jl}^2$. Hence, we obtain

$$\begin{aligned}\operatorname{tr}\left(\Delta_{\mathbf{A}_1}'\mathbf{L}_1\Delta_{\mathbf{A}_1}\right) &\geq \lambda_2(\mathbf{L}_1)\sum_{l=1}^{r_1}\sum_{j=1}^{m}(\Delta_{\mathbf{A}_1})_{jl}^2 - \frac{\lambda_2(\mathbf{L}_1)}{2}\sum_{l=1}^{r_1}\sum_{j=1}^{m}(\Delta_{\mathbf{A}_1})_{jl}^2 \\ &= \frac{\lambda_2(\mathbf{L}_1)}{2}\|\Delta_{\mathbf{A}_1}\|_F^2 = \frac{\lambda_2(\mathbf{L}_1)}{2}\|\widehat{\mathbf{A}}_1 - \mathbf{A}_1\|_F^2.\end{aligned}$$

By Assumption 5, using similar arguments leads to

$$\begin{aligned}\operatorname{tr}(\Delta_{\mathbf{A}_2}'\mathbf{L}_2\Delta_{\mathbf{A}_2}) &\geq \frac{\lambda_2(\mathbf{L}_2)}{2}\|\widehat{\mathbf{A}}_2 - \mathbf{A}_2\|_F^2, \\ \operatorname{tr}(\Delta_{\mathbf{A}_3}'\mathbf{L}_3\Delta_{\mathbf{A}_3}) &\geq \frac{\lambda_2(\mathbf{L}_3)}{2}\|\widehat{\mathbf{A}}_3 - \mathbf{A}_3\|_F^2,\end{aligned} \tag{S1.5}$$

where $\lambda_2(\mathbf{L}_2), \lambda_2(\mathbf{L}_3) > 0$ represent the second smallest eigenvalues of $\mathbf{L}_2$ and $\mathbf{L}_3$, respectively. This completes the proof. □

Lemma 2. *Let $\widehat{\Delta}$ be defined in (S2.42), $\mathbf{Z}$ and $\mathbf{e}$ be defined in (S2.33). Suppose that Assumptions 1 and 2 hold, and the sample size satisfies $T \geq a_1 \log(m^2 p)$. Then the following inequality holds with probability at least $1 - 6\exp(-a\log(m^2p))$:*

$$\frac{1}{T}\langle \mathbf{Z}\widehat{\Delta}, \mathbf{e}\rangle \leq 2\pi\vartheta a_2\sqrt{\log(m^2p)/T}\|\widehat{\Delta}\|_1,$$

where $\vartheta := \lambda_{\max}(\mathbf{\Sigma}_{\boldsymbol{\varepsilon}})(1 + \frac{\mu_{\max}(\mathbf{W})}{\mu_{\min}(\mathbf{W})})$ and the constants $a > 0, a_1, a_2 \geq 1$.

Proof. By Hölder's inequality, we have that

$$\frac{1}{T}\langle \mathbf{Z}\widehat{\Delta}, \mathbf{e}\rangle = \frac{1}{T}\langle \mathbf{Z}'\mathbf{e}, \widehat{\Delta}\rangle \leq \frac{1}{T}\|\mathbf{Z}'\mathbf{e}\|_\infty\|\widehat{\Delta}\|_1. \tag{S1.6}$$

By a simple calculation, we can deduce

$$\|\mathbf{Z}'\mathbf{e}\|_\infty = \|\mathbf{X}'\mathbf{E}\|_\infty = \max_{1\le i\le mp, 1\le j\le m} |\mathbf{e}_i'\mathbf{X}'\mathbf{E}\mathbf{e}_j|,$$

where $\mathbf{e}_i$ denotes the standard basis vector with the i-th entry being 1 and other entries being 0. According to Proposition 2.4 of Basu and Michailidis (2015) with Assumptions 1 and 2, for any vectors $\mathbf{u}$ and $\mathbf{v}$ with $\|\mathbf{u}\|_2 \le 1, \|\mathbf{v}\|_2 \le 1$ and any $\eta > 0$, there exists a constant $a_4 > 0$ such that

$$\begin{aligned}&\mathbb{P}\left[|\mathbf{u}'\left(\mathbf{X}'\mathbf{E}/T\right)\mathbf{v}| > 2\pi\left(\lambda_{\max}\left(\mathbf{\Sigma}_{\boldsymbol{\varepsilon}}\right)\left(1+\frac{\mu_{\max}(\mathbf{W})}{\mu_{\min}(\mathbf{W})}\right)\right)\eta\right]\\ \le\ & 6\exp(-a_4 T\min(\eta,\eta^2)).\end{aligned}$$

Denote $\vartheta := \lambda_{\max}\left(\mathbf{\Sigma}_{\boldsymbol{\varepsilon}}\right)[1+\mu_{\max}(\mathbf{W})/\mu_{\min}(\mathbf{W})]$. By using a union bound, we get that

$$\begin{aligned}&\mathbb{P}\left[\max_{1\le i\le mp, 1\le j\le m} |\mathbf{e}_i'\mathbf{X}'\mathbf{E}\mathbf{e}_j/T| > 2\pi\vartheta\eta\right]\\ \le\ & \sum_{1\le i\le mp, 1\le j\le m} \mathbb{P}\left[|\mathbf{e}_i'\mathbf{X}'\mathbf{E}\mathbf{e}_j/T| > 2\pi\vartheta\eta\right] \qquad \text{(S1.7)}\\ \le\ & 6m^2p\exp(-a_4 T\min(\eta,\eta^2)).\end{aligned}$$

Suppose that $a_4 > 1$, and let $\eta = \sqrt{(\log(m^2p))/T}$, where $T \ge \log(m^2p)$. Then, it is straightforward to obtain that $\min(\eta,\eta^2) = (\log(m^2p))/T$. As a consequence, we get

$$\begin{aligned}6m^2p\exp(-a_4 T\min(\eta,\eta^2)) &= 6\exp(\log(m^2p))\exp(-a_4\log(m^2p))\\ &= 6\exp(-(a_4-1)\log(m^2p)).\end{aligned} \qquad \text{(S1.8)}$$

Suppose that $a_4 \in (0,1]$ and $T \geq \frac{2}{a_4}\log(m^2p)$. Let $\eta = \sqrt{\frac{2}{a_4}}\sqrt{\frac{\log(m^2p)}{T}}$, where $\eta \leq 1$. Then, we have

$$\begin{aligned} 6m^2p\exp(-a_4T\min(\eta,\eta^2)) &= 6m^2p\exp(-2\frac{a_4}{2}T\min(\eta,\eta^2)) \\ &= 6m^2p\exp(-2T\frac{a_4}{2}\eta^2) \\ &= 6m^2p\exp(-2\log\left(m^2p\right)) \\ &= 6\exp(-\log\left(m^2p\right)). \end{aligned} \tag{S1.9}$$

By combining (S1.8), (S1.9), and the condition $T \geq a_1\log(m^2p)$, the following inequality holds:

$$\mathbb{P}\left[\max_{1\leq i\leq mp, 1\leq j\leq m}\left|\mathbf{e}_i'\mathbf{X}'\mathbf{E}\mathbf{e}_j/T\right| > 2\pi\vartheta a_2\sqrt{\log\left(m^2p\right)/T}\right] \leq 6\exp(-a\log(m^2p)),$$

where $a_1, a_2 \geq 1$ and $a > 0$. This taken together with (S1.6) indicates that

$$\frac{1}{T}\langle\mathbf{Z}\widehat{\Delta},\mathbf{e}\rangle \leq \frac{1}{T}\|\mathbf{Z}'\mathbf{e}\|_\infty\|\widehat{\Delta}\|_1 \leq 2\pi\vartheta a_2\sqrt{\log(m^2p)/T}\|\widehat{\Delta}\|_1$$

with probability at least $1 - 6\exp(-a\log(m^2p))$. □

Lemma 3. *Let $\widehat{\Delta}$ and $\mathbf{X}$ be defined in (S2.42) and (S2.32), respectively. Suppose that Assumptions 1, 2 and 3 hold, and the sample size $T \geq \frac{8\varsigma^2}{a_3}\bar{p}\min(\log(mp), \log(21emp/\bar{p}))$, where $\varsigma := \frac{\lambda_{\min}(\mathbf{\Sigma_\varepsilon})/\mu_{\max}(\mathbf{W})}{\lambda_{\max}(\mathbf{\Sigma_\varepsilon})/\mu_{\min}(\mathbf{W})}$, $a_3 > 0$ is a constant and $\bar{p}$ depends on τ. Then*

$$T^{-1}\left\|(\mathbf{I}_m \otimes \mathbf{X})\,\widehat{\Delta}\right\|_2^2 \geq \varpi\|\widehat{\Delta}\|_2^2/2$$

holds with probability at least $1-2\exp(-\bar{p}\min(\log mp, \log(21emp/\bar{p})))$*, where* $\varpi := \lambda_{\min}\left(\mathbf{\Sigma}_{\boldsymbol{\varepsilon}}\right)/\mu_{\max}(\mathbf{W})$.

Proof. For $\widehat{\Delta}$ in (S2.42), we partition it into m parts, namely $\widehat{\Delta} = (\widehat{\Delta}_1', \ldots, \widehat{\Delta}_m')'$, where $\widehat{\Delta}_i \in \mathbb{R}^{mp}$. Then we get

$$\left\|\left(\mathbf{I}_m \otimes \mathbf{X}\right)\widehat{\Delta}\right\|_2^2 = \sum_{i=1}^{m}\left\|\mathbf{X}\widehat{\Delta}_i\right\|_2^2.$$

Denote $\widehat{\mathbf{\Gamma}} = \mathbf{X}'\mathbf{X}/T$ and $\mathbf{\Gamma} = \mathbb{E}\widehat{\mathbf{\Gamma}}$. It can be seen that

$$\begin{aligned}
\widehat{\Delta}'\left(\mathbf{I}_m \otimes \mathbf{\Gamma}\right)\widehat{\Delta} = \widehat{\Delta}'\big(\mathbf{I}_m \otimes \mathbb{E}\widehat{\mathbf{\Gamma}}\big)\widehat{\Delta} &= \widehat{\Delta}'\Big(\mathbf{I}_m \otimes \mathbb{E}\big(\mathbf{X}'\mathbf{X}/T\big)\Big)\widehat{\Delta} \\
&= \widehat{\Delta}'\mathbb{E}\Big(\mathbf{I}_m \otimes \big(\mathbf{X}'\mathbf{X}/T\big)\Big)\widehat{\Delta} \\
&= \mathbb{E}\Big(\widehat{\Delta}'\Big(\mathbf{I}_m \otimes \big(\mathbf{X}'\mathbf{X}/T\big)\Big)\widehat{\Delta}\Big) \qquad \text{(S1.10)} \\
&= T^{-1}\mathbb{E}\Big(\widehat{\Delta}'\Big(\mathbf{I}_m \otimes \big(\mathbf{X}'\mathbf{X}\big)\Big)\widehat{\Delta}\Big) \\
&= T^{-1}\mathbb{E}\left(\|(\mathbf{I}_m \otimes \mathbf{X})\widehat{\Delta}\|_2^2\right).
\end{aligned}$$

Additionally,

$$\begin{aligned}
T^{-1}\|\left(\mathbf{I}_m \otimes \mathbf{X}\right)\widehat{\Delta}\|_2^2 &= \widehat{\Delta}'(\mathbf{I}_m \otimes \widehat{\mathbf{\Gamma}})\widehat{\Delta} \\
&= \widehat{\Delta}'\left(\mathbf{I}_m \otimes \mathbf{\Gamma}\right)\widehat{\Delta} + \widehat{\Delta}'\left(\mathbf{I}_m \otimes (\widehat{\mathbf{\Gamma}} - \mathbf{\Gamma})\right)\widehat{\Delta} \\
&= \widehat{\Delta}'\left(\mathbf{I}_m \otimes \mathbf{\Gamma}\right)\widehat{\Delta} + \sum_{i=1}^{m}\widehat{\Delta}_i'(\widehat{\mathbf{\Gamma}} - \mathbf{\Gamma})\widehat{\Delta}_i \\
&= T^{-1}\mathbb{E}\left(\|(\mathbf{I}_m \otimes \mathbf{X})\widehat{\Delta}\|_2^2\right) + \sum_{i=1}^{m}\widehat{\Delta}_i'(\widehat{\mathbf{\Gamma}} - \mathbf{\Gamma})\widehat{\Delta}_i,
\end{aligned} \qquad \text{(S1.11)}$$

where the last equality follows from (S1.10). Specifically, it follows from the proof of Proposition 4.2 in Basu and Michailidis (2015) and Assumptions

1, 2 that

$$\lambda_{\min}(\mathbf{\Gamma}) \geq \lambda_{\min}\left(\mathbf{\Sigma}_{\boldsymbol{\varepsilon}}\right)/\mu_{\max}(\mathbf{W}). \tag{S1.12}$$

Then, we have

$$\begin{aligned} T^{-1}\mathbb{E}\left(\|(\mathbf{I}_m \otimes \mathbf{X})\widehat{\Delta}\|_2^2\right) = \widehat{\Delta}'\left(\mathbf{I}_m \otimes \mathbf{\Gamma}\right)\widehat{\Delta} &\geq \lambda_{\min}(\mathbf{\Gamma})\|\widehat{\Delta}\|_2^2 \\ &\geq \lambda_{\min}\left(\mathbf{\Sigma}_{\boldsymbol{\varepsilon}}\right)/\mu_{\max}(\mathbf{W})\|\widehat{\Delta}\|_2^2 \\ &= \varpi\|\widehat{\Delta}\|_2^2, \end{aligned} \tag{S1.13}$$

where the first inequality follows from Theorem 4.2.2 of Horn and Johnson (2013) and $\lambda_{\min}(\mathbf{\Gamma}) = \lambda_{\min}(\mathbf{I}_m \otimes \mathbf{\Gamma})$, the equality follows from $\varpi := \frac{\lambda_{\min}(\mathbf{\Sigma}_{\boldsymbol{\varepsilon}})}{\mu_{\max}(\mathbf{W})}$. By Lemma 6, for any $i = 1, 2, \ldots, m$, we have $\|\widehat{\Delta}_i\|_0 \leq q_i mp$, where $q_i \in [0, 1]$ is related to τ. Let $\bar{p} = \max\{q_1 mp, q_2 mp, \ldots, q_m mp\}$. We define $\mathcal{K}(\bar{p}) := \{\boldsymbol{v} \in \mathbb{R}^{mp} : \|\boldsymbol{v}\|_0 \leq \bar{p}, \|\boldsymbol{v}\|_2 \leq 1\}$ as the set of $\bar{p}$-sparse vectors. Note that

$$\begin{aligned} & T^{-1}\|\left(\mathbf{I}_m \otimes \mathbf{X}\right)\widehat{\Delta}\|_2^2 \\ \geq\ & T^{-1}\mathbb{E}\left(\|(\mathbf{I}_m \otimes \mathbf{X})\widehat{\Delta}\|_2^2\right) - \sup_{\widehat{\Delta}_i \in \mathcal{K}(\bar{p})}\left|T^{-1}\|\left(\mathbf{I}_m \otimes \mathbf{X}\right)\widehat{\Delta}\|_2^2 - T^{-1}\mathbb{E}\left(\|(\mathbf{I}_m \otimes \mathbf{X})\widehat{\Delta}\|_2^2\right)\right| \\ =\ & T^{-1}\mathbb{E}\left(\|(\mathbf{I}_m \otimes \mathbf{X})\widehat{\Delta}\|_2^2\right) - \sup_{\widehat{\Delta}_i \in \mathcal{K}(\bar{p})}\left|\sum_{i=1}^{m}\widehat{\Delta}_i'(\widehat{\mathbf{\Gamma}} - \mathbf{\Gamma})\widehat{\Delta}_i\right| \\ \geq\ & \varpi\|\widehat{\Delta}\|_2^2 - \sup_{\widehat{\Delta}_i \in \mathcal{K}(\bar{p})}\left|\sum_{i=1}^{m}\widehat{\Delta}_i'(\widehat{\mathbf{\Gamma}} - \mathbf{\Gamma})\widehat{\Delta}_i\right|, \end{aligned}$$

where the equality comes from (S1.11) and the second inequality comes

from (S1.13). Then, we have

$$T^{-1}\|\left(\mathbf{I}_m \otimes \mathbf{X}\right)\widehat{\Delta}\|_2^2 - \frac{\varpi}{2}\|\widehat{\Delta}\|_2^2 \geq \frac{\varpi}{2}\|\widehat{\Delta}\|_2^2 - \sup_{\widehat{\Delta}_i \in \mathcal{K}(\bar{p})}\left|\sum_{i=1}^{m}\widehat{\Delta}_i'(\widehat{\mathbf{\Gamma}} - \mathbf{\Gamma})\widehat{\Delta}_i\right|,$$

which readily implies that

$$\begin{aligned}
&\mathbb{P}\left[T^{-1}\|\left(\mathbf{I}_m \otimes \mathbf{X}\right)\widehat{\Delta}\|_2^2 - \frac{\varpi}{2}\|\widehat{\Delta}\|_2^2 \geq 0\right] \\
&\geq \mathbb{P}\left[\frac{\varpi}{2}\|\widehat{\Delta}\|_2^2 - \sup_{\widehat{\Delta}_i \in \mathcal{K}(\bar{p})}\left|\sum_{i=1}^{m}\widehat{\Delta}_i'(\widehat{\mathbf{\Gamma}} - \mathbf{\Gamma})\widehat{\Delta}_i\right| \geq 0\right] \\
&= \mathbb{P}\left[\frac{\varpi}{2}\sum_{i=1}^{m}\|\widehat{\Delta}_i\|_2^2 - \sup_{\widehat{\Delta}_i \in \mathcal{K}(\bar{p})}\left|\sum_{i=1}^{m}\widehat{\Delta}_i'(\widehat{\mathbf{\Gamma}} - \mathbf{\Gamma})\widehat{\Delta}_i\right| \geq 0\right].
\end{aligned} \tag{S1.14}$$

For stable VAR(p) processes, as stated in Section 4 of Basu and Michailidis (2015), the spectral density function $f_{\boldsymbol{X}}(\theta)$ is defined as $f_{\boldsymbol{X}}(\theta) := \frac{1}{2\pi}\left(\mathbf{W}^{-1}\left(e^{-i\theta}\right)\right)\mathbf{\Sigma}_{\boldsymbol{\varepsilon}}\left(\mathbf{W}^{-1}\left(e^{-i\theta}\right)\right)'$ for $\theta \in [-\pi, \pi]$. We then define $\mathcal{M}\left(f_{\boldsymbol{X}}\right) := \operatorname{ess\,sup}_{\theta\in[-\pi,\pi]}\lambda_{\max}\left(f_{\boldsymbol{X}}(\theta)\right)$ (Basu and Michailidis, 2015). According to Proposition 2.4 of Basu and Michailidis (2015), for any $\eta > 0$, $\|\mathbf{u}\|_2 \leq 1$, there exists a constant $a_3 > 0$ such that

$$\mathbb{P}\left[\left|\mathbf{u}'(\widehat{\mathbf{\Gamma}} - \mathbf{\Gamma})\mathbf{u}\right| > 2\pi\mathcal{M}\left(f_{\boldsymbol{X}}\right)\eta\right] \leq 2\exp\left(-a_3 T\min(\eta, \eta^2)\right). \tag{S1.15}$$

Let $\eta := \frac{\lambda_{\min}(\mathbf{\Sigma}_{\boldsymbol{\varepsilon}})/\mu_{\max}(\mathbf{W})}{4\pi\mathcal{M}(f_{\boldsymbol{X}})}$. Therefore, we get that

$$\mathbb{P}\left[\left|\mathbf{u}'(\widehat{\mathbf{\Gamma}} - \mathbf{\Gamma})\mathbf{u}\right| > \frac{\varpi}{2}\right] \leq 2\exp\left(-a_3 T\min(\eta, \eta^2)\right). \tag{S1.16}$$

By Proposition 2.3 of Basu and Michailidis (2015), we have $2\pi\mathcal{M}\left(f_{\boldsymbol{X}}\right) \geq \lambda_{\min}(\mathbf{\Gamma})$. This combined with (S1.12) yields $2\pi\mathcal{M}\left(f_{\boldsymbol{X}}\right) \geq \lambda_{\min}(\mathbf{\Gamma}) \geq \frac{\lambda_{\min}(\mathbf{\Sigma}_{\boldsymbol{\varepsilon}})}{\mu_{\max}(\mathbf{W})}$.

As a consequence, $\eta \leq 1/2$. It can be shown that (S1.14) together with (S1.15) implies

$$
\begin{aligned}
&\mathbb{P}\left[T^{-1}\left\|(\mathbf{I}_m \otimes \mathbf{X})\,\widehat{\Delta}\right\|_2^2 \geq \frac{\varpi}{2}\|\widehat{\Delta}\|_2^2\right] \\
&\geq \mathbb{P}\left[\sup_{\widehat{\Delta}_i \in \mathcal{K}(\bar{p})}\left|\sum_{i=1}^{m}\widehat{\Delta}_i'(\widehat{\mathbf{\Gamma}}-\mathbf{\Gamma})\widehat{\Delta}_i\right| \leq \frac{\varpi}{2}\sum_{i=1}^{m}\|\widehat{\Delta}_i\|_2^2\right] \\
&\geq \mathbb{P}\left[\sup_{\widehat{\Delta}_i \in \mathcal{K}(\bar{p})}\left|\widehat{\Delta}_i'(\widehat{\mathbf{\Gamma}}-\mathbf{\Gamma})\widehat{\Delta}_i\right| \leq \frac{\varpi}{2}\|\widehat{\Delta}_i\|_2^2\right] \\
&= \mathbb{P}\left[\sup_{\widehat{\Delta}_i \in \mathcal{K}(\bar{p})}\left|\frac{\widehat{\Delta}_i'}{\|\widehat{\Delta}_i\|_2}(\widehat{\mathbf{\Gamma}}-\mathbf{\Gamma})\frac{\widehat{\Delta}_i}{\|\widehat{\Delta}_i\|_2}\right| \leq \frac{\varpi}{2}\right] \\
&\geq 1 - 2\exp\left(-a_3 T\frac{\varpi^2}{(4\pi\mathcal{M}\left(f_{\boldsymbol{X}}\right))^2} + \bar{p}\min(\log(mp), \log(21emp/\bar{p}))\right) \\
&\geq 1 - 2\exp\left(-a_3 T\frac{\varpi^2}{4(\lambda_{\max}\left(\mathbf{\Sigma}_{\boldsymbol{\epsilon}}\right)/\mu_{\min}(\mathbf{W}))^2} + \bar{p}\min(\log(mp), \log(21emp/\bar{p}))\right),
\end{aligned}
\tag{S1.17}
$$

where the third inequality follows from Lemma F.2 of Basu and Michailidis (2015) with $\eta = \frac{\lambda_{\min}(\mathbf{\Sigma}_{\boldsymbol{\epsilon}})/\mu_{\max}(\mathbf{W})}{4\pi\mathcal{M}(f_{\boldsymbol{X}})} = \frac{\varpi}{4\pi\mathcal{M}(f_{\boldsymbol{X}})}$ and the last inequality follows from $2\pi\mathcal{M}\left(f_{\boldsymbol{X}}\right) \leq \lambda_{\max}\left(\mathbf{\Sigma}_{\boldsymbol{\epsilon}}\right)/\mu_{\min}(\mathbf{W})$ (Basu and Michailidis, 2015). Suppose that $T \geq \frac{8(\lambda_{\max}(\mathbf{\Sigma}_{\boldsymbol{\epsilon}})/\mu_{\min}(\mathbf{W}))^2}{a_3\varpi^2}\bar{p}\min\left(\log(mp), \log(21emp/\bar{p})\right)$, then

$$
\begin{aligned}
&1 - 2\exp\left(-a_3 T\frac{\varpi^2}{4(\lambda_{\max}\left(\mathbf{\Sigma}_{\boldsymbol{\epsilon}}\right)/\mu_{\min}(\mathbf{W}))^2} + \bar{p}\min(\log(mp), \log(21emp/\bar{p}))\right) \\
&\geq 1 - 2\exp\left(-\bar{p}\min(\log(mp), \log(21emp/\bar{p}))\right).
\end{aligned}
$$

This combined with (S1.17) yields

$$
\mathbb{P}\left[T^{-1}\left\|(\mathbf{I}_m \otimes \mathbf{X})\,\widehat{\Delta}\right\|_2^2 \geq \frac{\varpi}{2}\|\widehat{\Delta}\|_2^2\right] \geq 1 - 2\exp(-\bar{p}\min(\log mp, \log(21emp/\bar{p}))).
$$

This completes the proof. □

Lemma 4. *Let $\widehat{\Delta}$ be defined in (S2.42). Then*

$$\|\widehat{\Delta}\|_2^2 \le 2\|\operatorname{vec}(\widehat{\mathcal{G}}'_{(1)}) - \operatorname{vec}(\mathcal{G}'_{(1)})\|_2^2 + 6c^2 r_1 r_2 r_3 \Big(\|\widehat{\mathbf{A}}_3 - \mathbf{A}_3\|_F^2 + \|\widehat{\mathbf{A}}_2 - \mathbf{A}_2\|_F^2 + \|\widehat{\mathbf{A}}_1 - \mathbf{A}_1\|_F^2 \Big),$$

where $\widehat{\mathcal{G}}, \mathcal{G} \in \mathbb{R}^{r_1 \times r_2 \times r_3}$ with $\|\widehat{\mathcal{G}}\|_\infty \le c, \|\mathcal{G}\|_\infty \le c$, c is defined in (2.4), $\widehat{\mathbf{A}}_i, \mathbf{A}_i \in \mathbb{R}^{n_i \times r_i}$, and $\widehat{\mathbf{A}}_i'\widehat{\mathbf{A}}_i = \mathbf{I}_{r_i}, \mathbf{A}_i'\mathbf{A}_i = \mathbf{I}_{r_i}, i = 1, 2, 3.$

Proof. From the definitions of $\widehat{\mathbf{A}}, \mathbf{A}, \widehat{\mathcal{G}}, \mathcal{G}$, it can be easily seen that

$$\begin{aligned}
0 \le \|\widehat{\Delta}\|_2 &= \Big\| (\widehat{\mathbf{A}}_1 \otimes \widehat{\mathbf{A}}_3 \otimes \widehat{\mathbf{A}}_2) \operatorname{vec}(\widehat{\mathcal{G}}'_{(1)}) - (\mathbf{A}_1 \otimes \mathbf{A}_3 \otimes \mathbf{A}_2) \operatorname{vec}(\mathcal{G}'_{(1)}) \Big\|_2 \\
&= \Big\| (\widehat{\mathbf{A}}_1 \otimes \widehat{\mathbf{A}}_3 \otimes \widehat{\mathbf{A}}_2) \operatorname{vec}(\widehat{\mathcal{G}}'_{(1)}) - (\mathbf{A}_1 \otimes \mathbf{A}_3 \otimes \mathbf{A}_2) \operatorname{vec}(\widehat{\mathcal{G}}'_{(1)}) \\
&\quad + (\mathbf{A}_1 \otimes \mathbf{A}_3 \otimes \mathbf{A}_2) \operatorname{vec}(\widehat{\mathcal{G}}'_{(1)}) - (\mathbf{A}_1 \otimes \mathbf{A}_3 \otimes \mathbf{A}_2) \operatorname{vec}(\mathcal{G}'_{(1)}) \Big\|_2 \\
&\le \Big\| (\widehat{\mathbf{A}}_1 \otimes \widehat{\mathbf{A}}_3 \otimes \widehat{\mathbf{A}}_2) \operatorname{vec}(\widehat{\mathcal{G}}'_{(1)}) - (\mathbf{A}_1 \otimes \mathbf{A}_3 \otimes \mathbf{A}_2) \operatorname{vec}(\widehat{\mathcal{G}}'_{(1)}) \Big\|_2 \\
&\quad + \Big\| (\mathbf{A}_1 \otimes \mathbf{A}_3 \otimes \mathbf{A}_2) \operatorname{vec}(\widehat{\mathcal{G}}'_{(1)}) - (\mathbf{A}_1 \otimes \mathbf{A}_3 \otimes \mathbf{A}_2) \operatorname{vec}(\mathcal{G}'_{(1)}) \Big\|_2 .
\end{aligned}$$

Then, we deduce

$$\begin{aligned}
\|\widehat{\Delta}\|_2^2 &\le \Big(\Big\| (\widehat{\mathbf{A}}_1 \otimes \widehat{\mathbf{A}}_3 \otimes \widehat{\mathbf{A}}_2) \operatorname{vec}(\widehat{\mathcal{G}}'_{(1)}) - (\mathbf{A}_1 \otimes \mathbf{A}_3 \otimes \mathbf{A}_2) \operatorname{vec}\operatorname{vec}(\mathcal{G}'_{(1)}) \Big\|_2 \\
&\quad + \Big\| (\mathbf{A}_1 \otimes \mathbf{A}_3 \otimes \mathbf{A}_2) \operatorname{vec}(\widehat{\mathcal{G}}'_{(1)}) - (\mathbf{A}_1 \otimes \mathbf{A}_3 \otimes \mathbf{A}_2) \operatorname{vec}(\mathcal{G}'_{(1)}) \Big\|_2 \Big)^2 \\
&\le 2 \Big\| (\widehat{\mathbf{A}}_1 \otimes \widehat{\mathbf{A}}_3 \otimes \widehat{\mathbf{A}}_2) \operatorname{vec}(\widehat{\mathcal{G}}'_{(1)}) - (\mathbf{A}_1 \otimes \mathbf{A}_3 \otimes \mathbf{A}_2) \operatorname{vec}(\widehat{\mathcal{G}}'_{(1)}) \Big\|_2^2 \\
&\quad + 2 \Big\| (\mathbf{A}_1 \otimes \mathbf{A}_3 \otimes \mathbf{A}_2) \operatorname{vec}(\widehat{\mathcal{G}}'_{(1)}) - (\mathbf{A}_1 \otimes \mathbf{A}_3 \otimes \mathbf{A}_2) \operatorname{vec}(\mathcal{G}'_{(1)}) \Big\|_2^2 ,
\end{aligned} \tag{S1.18}$$

where the last inequality follows from the fact that $\frac{a+b}{2} \le \sqrt{\frac{a^2+b^2}{2}}$ for any

$a, b \geq 0$. Note that

$$
\begin{aligned}
&\left\|(\mathbf{A}_1 \otimes \mathbf{A}_3 \otimes \mathbf{A}_2)\operatorname{vec}(\widehat{\mathcal{G}}'_{(1)}) - (\mathbf{A}_1 \otimes \mathbf{A}_3 \otimes \mathbf{A}_2)\operatorname{vec}(\mathcal{G}'_{(1)})\right\|_2 \\
=&\left\|(\mathbf{A}_1 \otimes \mathbf{A}_3 \otimes \mathbf{A}_2)(\operatorname{vec}(\widehat{\mathcal{G}}'_{(1)}) - \operatorname{vec}(\mathcal{G}'_{(1)}))\right\|_2 \\
\leq&\|(\mathbf{A}_1 \otimes \mathbf{A}_3 \otimes \mathbf{A}_2)\| \left\|\operatorname{vec}(\widehat{\mathcal{G}}'_{(1)}) - \operatorname{vec}(\mathcal{G}'_{(1)})\right\|_2 \\
=&\|\mathbf{A}_3\|\,\|\mathbf{A}_2\|\,\|\mathbf{A}_1\| \left\|\operatorname{vec}(\widehat{\mathcal{G}}'_{(1)}) - \operatorname{vec}(\mathcal{G}'_{(1)})\right\|_2 \\
\leq&\left\|\operatorname{vec}(\widehat{\mathcal{G}}'_{(1)}) - \operatorname{vec}(\mathcal{G}'_{(1)})\right\|_2,
\end{aligned}
\tag{S1.19}
$$

where the first inequality follows from Lemma 2.1 of Chen et al. (2021), the second equality follows from Proposition 1 and the second inequality follows from Proposition 2.

On the other hand, observe that

$$
\begin{aligned}
&\left\|(\widehat{\mathbf{A}}_1 \otimes \widehat{\mathbf{A}}_3 \otimes \widehat{\mathbf{A}}_2)\operatorname{vec}(\widehat{\mathcal{G}}'_{(1)}) - (\mathbf{A}_1 \otimes \mathbf{A}_3 \otimes \mathbf{A}_2)\operatorname{vec}(\widehat{\mathcal{G}}'_{(1)})\right\|_2 \\
=&\Big\|(\widehat{\mathbf{A}}_1 \otimes \widehat{\mathbf{A}}_3 \otimes \widehat{\mathbf{A}}_2)\operatorname{vec}(\widehat{\mathcal{G}}'_{(1)}) - (\widehat{\mathbf{A}}_1 \otimes \widehat{\mathbf{A}}_3 \otimes \mathbf{A}_2)\operatorname{vec}(\widehat{\mathcal{G}}'_{(1)}) \\
&+ (\widehat{\mathbf{A}}_1 \otimes \widehat{\mathbf{A}}_3 \otimes \mathbf{A}_2)\operatorname{vec}(\widehat{\mathcal{G}}'_{(1)}) - (\widehat{\mathbf{A}}_1 \otimes \mathbf{A}_3 \otimes \mathbf{A}_2)\operatorname{vec}(\widehat{\mathcal{G}}'_{(1)}) \\
&+ (\widehat{\mathbf{A}}_1 \otimes \mathbf{A}_3 \otimes \mathbf{A}_2)\operatorname{vec}(\widehat{\mathcal{G}}'_{(1)}) - (\mathbf{A}_1 \otimes \mathbf{A}_3 \otimes \mathbf{A}_2)\operatorname{vec}(\widehat{\mathcal{G}}'_{(1)})\Big\|_2 \\
\leq&\left\|(\widehat{\mathbf{A}}_1 \otimes \widehat{\mathbf{A}}_3 \otimes \widehat{\mathbf{A}}_2)\operatorname{vec}(\widehat{\mathcal{G}}'_{(1)}) - (\widehat{\mathbf{A}}_1 \otimes \widehat{\mathbf{A}}_3 \otimes \mathbf{A}_2)\operatorname{vec}(\widehat{\mathcal{G}}'_{(1)})\right\|_2 \\
&+ \left\|(\widehat{\mathbf{A}}_1 \otimes \widehat{\mathbf{A}}_3 \otimes \mathbf{A}_2)\operatorname{vec}(\widehat{\mathcal{G}}'_{(1)}) - (\widehat{\mathbf{A}}_1 \otimes \mathbf{A}_3 \otimes \mathbf{A}_2)\operatorname{vec}(\widehat{\mathcal{G}}'_{(1)})\right\|_2 \\
&+ \left\|(\widehat{\mathbf{A}}_1 \otimes \mathbf{A}_3 \otimes \mathbf{A}_2)\operatorname{vec}(\widehat{\mathcal{G}}'_{(1)}) - (\mathbf{A}_1 \otimes \mathbf{A}_3 \otimes \mathbf{A}_2)\operatorname{vec}(\widehat{\mathcal{G}}'_{(1)})\right\|_2.
\end{aligned}
\tag{S1.20}
$$

For the first term on the right-hand side of the above inequality, it can be

seen that

$$
\begin{aligned}
&\left\|(\widehat{\mathbf{A}}_1 \otimes \widehat{\mathbf{A}}_3 \otimes \widehat{\mathbf{A}}_2)\operatorname{vec}(\widehat{\mathcal{G}}'_{(1)}) - (\widehat{\mathbf{A}}_1 \otimes \widehat{\mathbf{A}}_3 \otimes \mathbf{A}_2)\operatorname{vec}(\widehat{\mathcal{G}}'_{(1)})\right\|_2 \\
&= \left\|(\widehat{\mathbf{A}}_1 \otimes \widehat{\mathbf{A}}_3 \otimes (\widehat{\mathbf{A}}_2 - \mathbf{A}_2))\operatorname{vec}(\widehat{\mathcal{G}}'_{(1)})\right\|_2 \\
&\leq \left\|\widehat{\mathbf{A}}_1 \otimes \widehat{\mathbf{A}}_3 \otimes (\widehat{\mathbf{A}}_2 - \mathbf{A}_2)\right\| \, \|\operatorname{vec}(\widehat{\mathcal{G}}'_{(1)})\|_2 \\
&= \|\widehat{\mathbf{A}}_1\|\|\widehat{\mathbf{A}}_3\|\|\widehat{\mathbf{A}}_2 - \mathbf{A}_2\|\|\operatorname{vec}(\widehat{\mathcal{G}}'_{(1)})\|_2 \\
&\leq \|\widehat{\mathbf{A}}_2 - \mathbf{A}_2\|\|\operatorname{vec}(\widehat{\mathcal{G}}'_{(1)})\|_2 \\
&= \|\widehat{\mathbf{A}}_2 - \mathbf{A}_2\|\|\widehat{\mathcal{G}}\|_F \\
&\leq c\sqrt{r_1 r_2 r_3}\|\widehat{\mathbf{A}}_2 - \mathbf{A}_2\| \\
&\leq c\sqrt{r_1 r_2 r_3}\|\widehat{\mathbf{A}}_2 - \mathbf{A}_2\|_F,
\end{aligned}
\tag{S1.21}
$$

where the first inequality comes from Lemma 2.1 of Chen et al. (2021), the second equality comes from Proposition 1, the second inequality comes from Proposition 2, the third inequality comes from $\|\widehat{\mathcal{G}}\|_F \leq \sqrt{r_1 r_2 r_3}\|\widehat{\mathcal{G}}\|_\infty \leq \sqrt{r_1 r_2 r_3}c$, and the last inequality comes from $\|\mathbf{A}\| \leq \|\mathbf{A}\|_F$ for any matrix $\mathbf{A}$. Similarly,

$$
\begin{aligned}
&\left\|(\widehat{\mathbf{A}}_1 \otimes \widehat{\mathbf{A}}_3 \otimes \mathbf{A}_2)\operatorname{vec}(\widehat{\mathcal{G}}'_{(1)}) - (\widehat{\mathbf{A}}_1 \otimes \mathbf{A}_3 \otimes \mathbf{A}_2)\operatorname{vec}(\widehat{\mathcal{G}}'_{(1)})\right\|_2 \\
&\leq c\sqrt{r_1 r_2 r_3}\|\widehat{\mathbf{A}}_3 - \mathbf{A}_3\|_F, \\
&\left\|(\widehat{\mathbf{A}}_1 \otimes \mathbf{A}_3 \otimes \mathbf{A}_2)\operatorname{vec}(\widehat{\mathcal{G}}'_{(1)}) - (\mathbf{A}_1 \otimes \mathbf{A}_3 \otimes \mathbf{A}_2)\operatorname{vec}(\widehat{\mathcal{G}}'_{(1)})\right\|_2 \\
&\leq c\sqrt{r_1 r_2 r_3}\|\widehat{\mathbf{A}}_1 - \mathbf{A}_1\|_F.
\end{aligned}
\tag{S1.22}
$$

Substituting (S1.21), (S1.22) into (S1.20), we obtain

$$\begin{aligned}0 &\leq \left\|(\widehat{\mathbf{A}}_1 \otimes \widehat{\mathbf{A}}_3 \otimes \widehat{\mathbf{A}}_2)\operatorname{vec}(\widehat{\mathcal{G}}'_{(1)}) - (\mathbf{A}_1 \otimes \mathbf{A}_3 \otimes \mathbf{A}_2)\operatorname{vec}(\widehat{\mathcal{G}}'_{(1)})\right\|_2 \\ &\leq c\sqrt{r_1 r_2 r_3}(\|\widehat{\mathbf{A}}_3 - \mathbf{A}_3\|_F + \|\widehat{\mathbf{A}}_2 - \mathbf{A}_2\|_F + \|\widehat{\mathbf{A}}_1 - \mathbf{A}_1\|_F).\end{aligned}$$

Further, we have

$$\begin{aligned}&\left\|(\widehat{\mathbf{A}}_1 \otimes \widehat{\mathbf{A}}_3 \otimes \widehat{\mathbf{A}}_2)\operatorname{vec}(\widehat{\mathcal{G}}'_{(1)}) - (\mathbf{A}_1 \otimes \mathbf{A}_3 \otimes \mathbf{A}_2)\operatorname{vec}(\widehat{\mathcal{G}}'_{(1)})\right\|_2^2 \\ &\leq c^2 r_1 r_2 r_3 \left(\|\widehat{\mathbf{A}}_3 - \mathbf{A}_3\|_F + \|\widehat{\mathbf{A}}_2 - \mathbf{A}_2\|_F + \|\widehat{\mathbf{A}}_1 - \mathbf{A}_1\|_F\right)^2 \\ &= c^2 r_1 r_2 r_3 \Big(\|\widehat{\mathbf{A}}_3 - \mathbf{A}_3\|_F^2 + \|\widehat{\mathbf{A}}_2 - \mathbf{A}_2\|_F^2 + \|\widehat{\mathbf{A}}_1 - \mathbf{A}_1\|_F^2 + 2\|\widehat{\mathbf{A}}_3 - \mathbf{A}_3\|_F \\ &\quad \cdot \|\widehat{\mathbf{A}}_2 - \mathbf{A}_2\|_F + 2\|\widehat{\mathbf{A}}_3 - \mathbf{A}_3\|_F\|\widehat{\mathbf{A}}_1 - \mathbf{A}_1\|_F + 2\|\widehat{\mathbf{A}}_1 - \mathbf{A}_1\|_F\|\widehat{\mathbf{A}}_2 - \mathbf{A}_2\|_F\Big).\end{aligned} \tag{S1.23}$$

For any non-negative numbers a, b, c, it is known that $a^2 + b^2 + c^2 \geq ab + ac + bc$. Therefore, we can conclude that

$$\begin{aligned}&\|\widehat{\mathbf{A}}_3 - \mathbf{A}_3\|_F\|\widehat{\mathbf{A}}_2 - \mathbf{A}_2\|_F + \|\widehat{\mathbf{A}}_3 - \mathbf{A}_3\|_F\|\widehat{\mathbf{A}}_1 - \mathbf{A}_1\|_F \\ &+ \|\widehat{\mathbf{A}}_1 - \mathbf{A}_1\|_F\|\widehat{\mathbf{A}}_2 - \mathbf{A}_2\|_F \leq \|\widehat{\mathbf{A}}_3 - \mathbf{A}_3\|_F^2 + \|\widehat{\mathbf{A}}_2 - \mathbf{A}_2\|_F^2 + \|\widehat{\mathbf{A}}_1 - \mathbf{A}_1\|_F^2,\end{aligned}$$

which taken collectively with (S1.23) gives

$$\begin{aligned}&\left\|(\widehat{\mathbf{A}}_3 \otimes \widehat{\mathbf{A}}_2 \otimes \widehat{\mathbf{A}}_1)\operatorname{vec}(\widehat{\mathcal{G}}'_{(1)}) - (\mathbf{A}_3 \otimes \mathbf{A}_2 \otimes \mathbf{A}_1)\operatorname{vec}(\widehat{\mathcal{G}}'_{(1)})\right\|_2^2 \\ &\leq c^2 r_1 r_2 r_3 \left(3\|\widehat{\mathbf{A}}_3 - \mathbf{A}_3\|_F^2 + 3\|\widehat{\mathbf{A}}_2 - \mathbf{A}_2\|_F^2 + 3\|\widehat{\mathbf{A}}_1 - \mathbf{A}_1\|_F^2\right).\end{aligned} \tag{S1.24}$$

Substituting (S1.24), (S1.19) into (S1.18) yields

$$\begin{aligned}\|\widehat{\Delta}\|_2^2 \leq 2\|\operatorname{vec}(\widehat{\mathcal{G}}'_{(1)}) - \operatorname{vec}(\mathcal{G}'_{(1)})\|_2^2 + 6c^2 r_1 r_2 r_3 \Big(&\|\widehat{\mathbf{A}}_3 - \mathbf{A}_3\|_F^2 \\ &+ \|\widehat{\mathbf{A}}_2 - \mathbf{A}_2\|_F^2 + \|\widehat{\mathbf{A}}_1 - \mathbf{A}_1\|_F^2\Big).\end{aligned}$$

This completes the proof. □

Lemma 5. *Let* $\mathcal{W} = [\![\mathcal{G}; \mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3]\!]$ *and* $\widetilde{\mathcal{W}} = [\![\widetilde{\mathcal{G}}; \widetilde{\mathbf{A}}_1, \widetilde{\mathbf{A}}_2, \widetilde{\mathbf{A}}_3]\!]$ *be the HOSVD of* $\mathcal{W} \in \mathbb{R}^{m\times m\times p}$ *and* $\widetilde{\mathcal{W}} \in \mathbb{R}^{m\times m\times p}$ *with the same multilinear ranks* (r_1, r_2, r_3)*, respectively. Suppose that Assumption 4 holds for* $\mathcal{W}$*. Furthermore, assume that* $\|\widetilde{\mathcal{G}}\|_\infty \le c$ *and* $\|\mathcal{G}\|_\infty \le c$*. Then, we have*

$$\|\widetilde{\mathcal{G}} - \mathcal{G}\|_F \le b\|\widetilde{\mathcal{W}} - \mathcal{W}\|_F,$$

and

$$\|\widetilde{\mathbf{A}}_i - \mathbf{A}_i\|_F \le 2^{\frac{3}{2}} (\eta_i/\rho) \|\widetilde{\mathcal{W}} - \mathcal{W}\|_F,$$

where $\eta_i = \left(3\sigma_1(\mathcal{W}_{(i)}) + c\sqrt{r_1 r_2 r_3}\right) \sqrt{\sum_{j=1}^{r_i} \frac{1}{\sigma_j^4(\mathcal{W}_{(i)})}}$*,* $b = 1 + \frac{2^{\frac{3}{2}}\sigma_1(\mathcal{W}_{(1)}) \sum_{i=1}^3 \eta_i}{\rho}$ *and* ρ *is defined in Assumption 4.*

Proof. By Theorem 3 of Yu et al. (2015), under Assumption 4, for the j-th singular vector of $\mathcal{W}_{(i)}$,

$$\begin{aligned}\left\|(\widetilde{\mathbf{A}}_i)_{:,j} - (\mathbf{A}_i)_{:,j}\right\|_2 &\le \frac{2^{\frac{3}{2}}\left(2\sigma_1(\mathcal{W}_{(i)}) + \|\widetilde{\mathcal{W}}_{(i)} - \mathcal{W}_{(i)}\|\right)\|\widetilde{\mathcal{W}}_{(i)} - \mathcal{W}_{(i)}\|_F}{\min\left[\sigma_{j-1}^2(\mathcal{W}_{(i)}) - \sigma_j^2(\mathcal{W}_{(i)}), \sigma_j^2(\mathcal{W}_{(i)}) - \sigma_{j+1}^2(\mathcal{W}_{(i)})\right]} \\ &\le \frac{2^{\frac{3}{2}}\left(2\sigma_1(\mathcal{W}_{(i)}) + \|\widetilde{\mathcal{W}}_{(i)} - \mathcal{W}_{(i)}\|\right)\|\widetilde{\mathcal{W}}_{(i)} - \mathcal{W}_{(i)}\|_F}{\rho\sigma_j^2(\mathcal{W}_{(i)})},\end{aligned} \tag{S1.25}$$

where $(\mathbf{A}_i)_{:,j}$ denotes the j-th left singular vector of the singular value decomposition of $\mathcal{W}_{(i)}$, $(\widetilde{\mathbf{A}}_i)_{:,j}$ denotes the j-th left singular vector of the singular value decomposition of $\widetilde{\mathcal{W}}_{(i)}$ and $\sigma_1(\mathcal{W}_{(i)}) = \|\mathcal{W}_{(i)}\|$. Additionally, let the SVD of $\widetilde{\mathcal{G}}_{(1)}$ be $\widetilde{\mathcal{G}}_{(1)} = \bar{\mathbf{U}}_1\bar{\mathbf{S}}_1\bar{\mathbf{V}}_1'$, where $\bar{\mathbf{U}}_1 \in \mathbb{R}^{r_1\times r_1}$ and $\bar{\mathbf{V}}_1 \in$

$\mathbb{R}^{r_2r_3\times r_2r_3}$ are orthogonal matrices, $\bar{\mathbf{S}}_1 \in \mathbb{R}^{r_1\times r_2r_3}$ is a diagonal matrix. It follows from Proposition 1 that

$$\begin{aligned}(\widetilde{\mathbf{A}}_3 \otimes \widetilde{\mathbf{A}}_2)'(\widetilde{\mathbf{A}}_3 \otimes \widetilde{\mathbf{A}}_2) = (\widetilde{\mathbf{A}}_3' \otimes \widetilde{\mathbf{A}}_2')(\widetilde{\mathbf{A}}_3 \otimes \widetilde{\mathbf{A}}_2) &= (\widetilde{\mathbf{A}}_3'\widetilde{\mathbf{A}}_3) \otimes (\widetilde{\mathbf{A}}_2'\widetilde{\mathbf{A}}_2) \\ &= \mathbf{I}_{r_3} \otimes \mathbf{I}_{r_2} = \mathbf{I}_{r_2r_3},\end{aligned}$$

which implies that $\widetilde{\mathbf{A}}_3 \otimes \widetilde{\mathbf{A}}_2$ has orthogonal columns. Further, note that

$$\begin{aligned}\widetilde{\mathcal{W}}_{(1)} = \widetilde{\mathbf{A}}_1\widetilde{\mathcal{G}}_{(1)}(\widetilde{\mathbf{A}}_3 \otimes \widetilde{\mathbf{A}}_2)' = \widetilde{\mathbf{A}}_1\bar{\mathbf{U}}_1\bar{\mathbf{S}}_1\bar{\mathbf{V}}_1'(\widetilde{\mathbf{A}}_3 \otimes \widetilde{\mathbf{A}}_2)' \\ = (\widetilde{\mathbf{A}}_1\bar{\mathbf{U}}_1)\bar{\mathbf{S}}_1((\widetilde{\mathbf{A}}_3 \otimes \widetilde{\mathbf{A}}_2)\bar{\mathbf{V}}_1)'.\end{aligned} \tag{S1.26}$$

Here, $\widetilde{\mathbf{A}}_1\bar{\mathbf{U}}_1$ and $(\widetilde{\mathbf{A}}_3 \otimes \widetilde{\mathbf{A}}_2)\bar{\mathbf{V}}_1$ have orthogonal columns, and (S1.26) is the SVD of $\widetilde{\mathcal{W}}_{(1)}$. This consequently implies that $\|\widetilde{\mathcal{W}}_{(1)}\| = \|\widetilde{\mathcal{G}}_{(1)}\|$. Similarly, we can obtain that $\|\widetilde{\mathcal{W}}_{(2)}\| = \|\widetilde{\mathcal{G}}_{(2)}\|$ and $\|\widetilde{\mathcal{W}}_{(3)}\| = \|\widetilde{\mathcal{G}}_{(3)}\|$. Then, for any $i = 1, 2, 3$, we have

$$\|\widetilde{\mathcal{W}}_{(i)} - \mathcal{W}_{(i)}\| \le \|\widetilde{\mathcal{W}}_{(i)}\| + \|\mathcal{W}_{(i)}\| = \|\widetilde{\mathcal{G}}_{(i)}\| + \sigma_1(\mathcal{W}_{(i)}) \le c\sqrt{r_1r_2r_3} + \sigma_1(\mathcal{W}_{(i)}),$$

where the last inequality follows from $\|\widetilde{\mathcal{G}}_{(i)}\| \le \|\widetilde{\mathcal{G}}_{(i)}\|_F = \|\widetilde{\mathcal{G}}\|_F \le c\sqrt{r_1r_2r_3}$. Combining the above inequality with (S1.25) immediately yields

$$\left\|(\widetilde{\mathbf{A}}_i)_{:,j} - (\mathbf{A}_i)_{:,j}\right\|_2 \le \frac{2^{\frac{3}{2}}\left(3\sigma_1(\mathcal{W}_{(i)}) + c\sqrt{r_1r_2r_3}\right)\|\widetilde{\mathcal{W}}_{(i)} - \mathcal{W}_{(i)}\|_F}{\rho\sigma_j^2(\mathcal{W}_{(i)})}.$$

Consequently,

$$\begin{aligned}\left\|\widetilde{\mathbf{A}}_i - \mathbf{A}_i\right\|_F^2 &= \sum_{j=1}^{r_i}\left\|(\widetilde{\mathbf{A}}_i)_{:,j} - (\mathbf{A}_i)_{:,j}\right\|_2^2 \\ &\le \sum_{j=1}^{r_i}\frac{2^3\left(3\sigma_1\left(\mathcal{W}_{(i)}\right) + c\sqrt{r_1r_2r_3}\right)^2\|\widetilde{\mathcal{W}}_{(i)} - \mathcal{W}_{(i)}\|_F^2}{\rho^2\sigma_j^4\left(\mathcal{W}_{(i)}\right)}.\end{aligned}$$

Then, we can further demonstrate that

$$\left\|\widetilde{\mathbf{A}}_i - \mathbf{A}_i\right\|_F \leq 2^{\frac{3}{2}} \left(\eta_i/\rho\right) \|\widetilde{\mathcal{W}} - \mathcal{W}\|_F, \tag{S1.27}$$

where $\eta_i = \left(3\sigma_1\left(\mathcal{W}_{(i)}\right) + c\sqrt{r_1 r_2 r_3}\right) \sqrt{\sum_{j=1}^{r_i} \frac{1}{\sigma_j^4\left(\mathcal{W}_{(i)}\right)}}$.

On the other hand, we make the observation that

$$\begin{aligned}
&\|\widetilde{\mathcal{G}} - \mathcal{G}\|_F \\
=&\left\|\widetilde{\mathcal{W}} \times_1 \widetilde{\mathbf{A}}_1' \times_2 \widetilde{\mathbf{A}}_2' \times_3 \widetilde{\mathbf{A}}_3' - \mathcal{W} \times_1 \mathbf{A}_1' \times_2 \mathbf{A}_2' \times_3 \mathbf{A}_3'\right\|_F \\
=&\left\|\widetilde{\mathcal{W}} \times_1 \widetilde{\mathbf{A}}_1' \times_2 \widetilde{\mathbf{A}}_2' \times_3 \widetilde{\mathbf{A}}_3' - \mathcal{W} \times_1 \widetilde{\mathbf{A}}_1' \times_2 \widetilde{\mathbf{A}}_2' \times_3 \widetilde{\mathbf{A}}_3'\right. \\
&+ \mathcal{W} \times_1 \widetilde{\mathbf{A}}_1' \times_2 \widetilde{\mathbf{A}}_2' \times_3 \widetilde{\mathbf{A}}_3' - \mathcal{W} \times_1 \mathbf{A}_1' \times_2 \widetilde{\mathbf{A}}_2' \times_3 \widetilde{\mathbf{A}}_3' \\
&+ \mathcal{W} \times_1 \mathbf{A}_1' \times_2 \widetilde{\mathbf{A}}_2' \times_3 \widetilde{\mathbf{A}}_3' - \mathcal{W} \times_1 \mathbf{A}_1' \times_2 \mathbf{A}_2' \times_3 \widetilde{\mathbf{A}}_3' \\
&\left.+ \mathcal{W} \times_1 \mathbf{A}_1' \times_2 \mathbf{A}_2' \times_3 \widetilde{\mathbf{A}}_3' - \mathcal{W} \times_1 \mathbf{A}_1' \times_2 \mathbf{A}_2' \times_3 \mathbf{A}_3'\right\|_F \\
\leq&\left\|\widetilde{\mathcal{W}} \times_1 \widetilde{\mathbf{A}}_1' \times_2 \widetilde{\mathbf{A}}_2' \times_3 \widetilde{\mathbf{A}}_3' - \mathcal{W} \times_1 \widetilde{\mathbf{A}}_1' \times_2 \widetilde{\mathbf{A}}_2' \times_3 \widetilde{\mathbf{A}}_3'\right\|_F \\
&+ \left\|\mathcal{W} \times_1 \widetilde{\mathbf{A}}_1' \times_2 \widetilde{\mathbf{A}}_2' \times_3 \widetilde{\mathbf{A}}_3' - \mathcal{W} \times_1 \mathbf{A}_1' \times_2 \widetilde{\mathbf{A}}_2' \times_3 \widetilde{\mathbf{A}}_3'\right\|_F \\
&+ \left\|\mathcal{W} \times_1 \mathbf{A}_1' \times_2 \widetilde{\mathbf{A}}_2' \times_3 \widetilde{\mathbf{A}}_3' - \mathcal{W} \times_1 \mathbf{A}_1' \times_2 \mathbf{A}_2' \times_3 \widetilde{\mathbf{A}}_3'\right\|_F \\
&+ \left\|\mathcal{W} \times_1 \mathbf{A}_1' \times_2 \mathbf{A}_2' \times_3 \widetilde{\mathbf{A}}_3' - \mathcal{W} \times_1 \mathbf{A}_1' \times_2 \mathbf{A}_2' \times_3 \mathbf{A}_3'\right\|_F \\
\leq&\|\widetilde{\mathcal{W}} - \mathcal{W}\|_F + \left\|\mathcal{W} \times_1 \widetilde{\mathbf{A}}_1' \times_2 \widetilde{\mathbf{A}}_2' \times_3 \widetilde{\mathbf{A}}_3' - \mathcal{W} \times_1 \mathbf{A}_1' \times_2 \widetilde{\mathbf{A}}_2' \times_3 \widetilde{\mathbf{A}}_3'\right\|_F \\
&+ \left\|\mathcal{W} \times_1 \mathbf{A}_1' \times_2 \widetilde{\mathbf{A}}_2' \times_3 \widetilde{\mathbf{A}}_3' - \mathcal{W} \times_1 \mathbf{A}_1' \times_2 \mathbf{A}_2' \times_3 \widetilde{\mathbf{A}}_3'\right\|_F \\
&+ \left\|\mathcal{W} \times_1 \mathbf{A}_1' \times_2 \mathbf{A}_2' \times_3 \widetilde{\mathbf{A}}_3' - \mathcal{W} \times_1 \mathbf{A}_1' \times_2 \mathbf{A}_2' \times_3 \mathbf{A}_3'\right\|_F.
\end{aligned} \tag{S1.28}$$

Similar to (S1.21), we have

$$\begin{aligned}
&\left\|\mathcal{W}\times_1\widetilde{\mathbf{A}}_1'\times_2\widetilde{\mathbf{A}}_2'\times_3\widetilde{\mathbf{A}}_3'-\mathcal{W}\times_1\mathbf{A}_1'\times_2\widetilde{\mathbf{A}}_2'\times_3\widetilde{\mathbf{A}}_3'\right\|_F\\
=&\left\|\widetilde{\mathbf{A}}_1'\mathcal{W}_{(1)}(\widetilde{\mathbf{A}}_3\otimes\widetilde{\mathbf{A}}_2)-\mathbf{A}_1'\mathcal{W}_{(1)}(\widetilde{\mathbf{A}}_3\otimes\widetilde{\mathbf{A}}_2)\right\|_F\\
=&\left\|(\widetilde{\mathbf{A}}_1'-\mathbf{A}_1')\mathcal{W}_{(1)}(\widetilde{\mathbf{A}}_3\otimes\widetilde{\mathbf{A}}_2)\right\|_F\\
\le&\|\widetilde{\mathbf{A}}_1'-\mathbf{A}_1'\|_F\|\mathcal{W}_{(1)}(\widetilde{\mathbf{A}}_3\otimes\widetilde{\mathbf{A}}_2)\|\\
\le&\|\widetilde{\mathbf{A}}_1'-\mathbf{A}_1'\|_F\|\mathcal{W}_{(1)}\|\|\widetilde{\mathbf{A}}_3\otimes\widetilde{\mathbf{A}}_2\|\\
\le&\|\widetilde{\mathbf{A}}_1'-\mathbf{A}_1'\|_F\|\mathcal{W}_{(1)}\|\\
=&\sigma_1(\mathcal{W}_{(1)})\|\widetilde{\mathbf{A}}_1'-\mathbf{A}_1'\|_F.
\end{aligned}$$

The above inequality combined with (S1.27) yields

$$\begin{aligned}
&\left\|\mathcal{W}\times_1\widetilde{\mathbf{A}}_1'\times_2\widetilde{\mathbf{A}}_2'\times_3\widetilde{\mathbf{A}}_3'-\mathcal{W}\times_1\mathbf{A}_1'\times_2\widetilde{\mathbf{A}}_2'\times_3\widetilde{\mathbf{A}}_3'\right\|_F\\
\le&\frac{2^{\frac{3}{2}}\eta_1\sigma_1(\mathcal{W}_{(1)})}{\rho}\|\widetilde{\mathcal{W}}-\mathcal{W}\|_F.
\end{aligned}\tag{S1.29}$$

Similarly,

$$\begin{aligned}
&\left\|\mathcal{W}\times_1\mathbf{A}_1'\times_2\widetilde{\mathbf{A}}_2'\times_3\widetilde{\mathbf{A}}_3'-\mathcal{W}\times_1\mathbf{A}_1'\times_2\mathbf{A}_2'\times_3\widetilde{\mathbf{A}}_3'\right\|_F\\
\le&\frac{2^{\frac{3}{2}}\eta_2\sigma_1(\mathcal{W}_{(1)})}{\rho}\|\widetilde{\mathcal{W}}-\mathcal{W}\|_F,\\
&\left\|\mathcal{W}\times_1\mathbf{A}_1'\times_2\mathbf{A}_2'\times_3\widetilde{\mathbf{A}}_3'-\mathcal{W}\times_1\mathbf{A}_1'\times_2\mathbf{A}_2'\times_3\mathbf{A}_3'\right\|_F\\
\le&\frac{2^{\frac{3}{2}}\eta_3\sigma_1(\mathcal{W}_{(1)})}{\rho}\|\widetilde{\mathcal{W}}-\mathcal{W}\|_F.
\end{aligned}\tag{S1.30}$$

Substituting (S1.29), (S1.30) into (S1.28) then gives

$$\|\widetilde{\mathcal{G}}-\mathcal{G}\|_F\le\|\widetilde{\mathcal{W}}-\mathcal{W}\|_F+\sum_{i=1}^{3}\frac{2^{\frac{3}{2}}\eta_i\sigma_1(\mathcal{W}_{(1)})}{\rho}\|\widetilde{\mathcal{W}}-\mathcal{W}\|_F=b\|\widetilde{\mathcal{W}}-\mathcal{W}\|_F,$$

where $b:=1+\frac{2^{\frac{3}{2}}\sigma_1(\mathcal{W}_{(1)})\sum_{i=1}^{3}\eta_i}{\rho}>0$. □

Lemma 6. *Suppose that Assumption 3 holds. Then* $\|\mathbf{Ag}\|_0 \leq s_1 m^2 p$*, where* $\mathbf{A}$ *and* $\mathbf{g}$ *are defined in (S2.33), and* $s_1 \in [0,1]$ *is the sparsity threshold related to* τ*.*

Proof. Recall that $\mathbf{g} = \mathrm{vec}(\mathcal{G}'_{(1)}) \in \mathbb{R}^{r_1 r_2 r_3}$. The ratio of zero entries in a vector or matrix is denoted as $\bar{z}(\cdot)$ (Wang et al., 2024). It follows from the fact that $\mathcal{G} \in \mathbb{R}^{r_1 \times r_2 \times r_3}$ belongs to the set $\mathcal{U}$ that the sparsity of the vector $\mathbf{g}$ is equal to τ, i.e., $\bar{z}(\mathbf{g}) = \tau$. The matrix $\mathbf{G} \in \mathbb{R}^{r_1 r_2 r_3 \times m^2 p}$ is constructed such that $\mathbf{G}_{:,j} = \mathbf{g}$ for each $j = 1, 2, \ldots, m^2 p$, where $\mathbf{G}_{:,j}$ denotes the j-th column of $\mathbf{G}$. The sparsity of each row of $\mathbf{G}$ is denoted as $\bar{z}(\mathbf{G}_{i,:})$, where $\mathbf{G}_{i,:}$ represents the i-th row of $\mathbf{G}$ and $\bar{z}(\mathbf{G}_{i,:}) \in \{0,1\}$ for $i = 1, 2, \ldots, r_1 r_2 r_3$. Similarly, the sparsity of the i-th column of $\mathbf{A} \in \mathbb{R}^{m^2 p \times r_1 r_2 r_3}$ is defined as $\bar{z}(\mathbf{A}_{:,i})$. By Proposition 1 in Wang et al. (2024), we deduce that

$$\bar{z}(\mathbf{AG}) \geq \max\big\{0, 1 + \sum_{i=1}^{r_1 r_2 r_3} \left(\bar{z}(\mathbf{G}_{i,:}) + \bar{z}\left(\mathbf{A}_{:,i}\right) - \bar{z}(\mathbf{G}_{i,:})\bar{z}\left(\mathbf{A}_{:,i}\right)\right) - r_1 r_2 r_3\big\} := z_1. \tag{S1.31}$$

It can be easily verified that $\bar{z}(\mathbf{Ag}) = \bar{z}(\mathbf{AG}) \geq z_1$. Consequently, we have $\|\mathbf{Ag}\|_0 \leq (1 - z_1) m^2 p := s_1 m^2 p$. Note that $\sum_{i=1}^{r_1 r_2 r_3} \bar{z}(\mathbf{G}_{i,:})/(r_1 r_2 r_3) = \tau$, which implies that s_1 depends on τ. □

Throughout this paper, we assume that the core tensor in model (2.4) also satisfies the constraint $\mathcal{U} := \{\mathcal{G} : \|\mathcal{G}\|_0 \leq s\}$. By using similar tech-

niques, it follows from the fact that $\widehat{\mathcal{G}} \in \mathbb{R}^{r_1 \times r_2 \times r_3}$ belongs to the set $\mathcal{U}$ that $\|\widehat{\mathbf{A}}\widehat{\mathbf{g}}\|_0 \leq s_2 m^2 p$, where $s_2 \in [0,1]$ related to τ.

S2 Proof of Theorem 1

The VAR model in (1.1) can be expressed equivalently as

$$\underbrace{\begin{bmatrix} \mathbf{y}_1' \\ \mathbf{y}_2' \\ \vdots \\ \mathbf{y}_T' \end{bmatrix}}_{\mathbf{Y}} = \underbrace{\begin{bmatrix} \mathbf{y}_0' & \mathbf{y}_{-1}' & \cdots & \mathbf{y}_{-p+1}' \\ \mathbf{y}_1' & \mathbf{y}_0' & \cdots & \mathbf{y}_{-p+2}' \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{y}_{T-1}' & \mathbf{y}_{T-2}' & \cdots & \mathbf{y}_{T-p}' \end{bmatrix}}_{\mathbf{X}} \underbrace{\begin{bmatrix} \mathbf{W}_1' \\ \mathbf{W}_2' \\ \vdots \\ \mathbf{W}_p' \end{bmatrix}}_{\mathcal{W}_{(1)}'} + \underbrace{\begin{bmatrix} \boldsymbol{\varepsilon}_1' \\ \boldsymbol{\varepsilon}_2' \\ \vdots \\ \boldsymbol{\varepsilon}_T' \end{bmatrix}}_{\mathbf{E}}. \tag{S2.32}$$

Notice that $\mathcal{W} = \mathcal{G} \times_1 \mathbf{A}_1 \times_2 \mathbf{A}_2 \times_3 \mathbf{A}_3$ represents the HOSVD of $\mathcal{W}$ and $\mathcal{W}_{(1)} = \mathbf{A}_1 \mathcal{G}_{(1)} (\mathbf{A}_3 \otimes \mathbf{A}_2)'$ (Kolda and Bader, 2009). Based on the formulation in (S2.32), we can rewrite (1.1) as $\mathbf{Y} = \mathbf{X}(\mathbf{A}_3 \otimes \mathbf{A}_2)\mathcal{G}_{(1)}' \mathbf{A}_1' + \mathbf{E}$. Which is equivalent to

$$\begin{aligned} \mathbf{y} := \operatorname{vec}(\mathbf{Y}) &= \operatorname{vec}\left(\mathbf{X}(\mathbf{A}_3 \otimes \mathbf{A}_2)\mathcal{G}_{(1)}' \mathbf{A}_1'\right) + \operatorname{vec}(\mathbf{E}) \\ &= \operatorname{vec}\left(\mathbf{X}(\mathbf{A}_3 \otimes \mathbf{A}_2)\mathcal{G}_{(1)}' \mathbf{A}_1' \mathbf{I}_m\right) + \operatorname{vec}(\mathbf{E}) \\ &= (\mathbf{I}_m' \otimes \mathbf{X}) \operatorname{vec}\left((\mathbf{A}_3 \otimes \mathbf{A}_2)\mathcal{G}_{(1)}' \mathbf{A}_1'\right) + \operatorname{vec}(\mathbf{E}) \\ &= (\mathbf{I}_m \otimes \mathbf{X}) \left(\mathbf{A}_1 \otimes (\mathbf{A}_3 \otimes \mathbf{A}_2)\right) \operatorname{vec}(\mathcal{G}_{(1)}') + \operatorname{vec}(\mathbf{E}) \\ &= \underbrace{(\mathbf{I}_m \otimes \mathbf{X})}_{\mathbf{Z}} \underbrace{(\mathbf{A}_1 \otimes \mathbf{A}_3 \otimes \mathbf{A}_2)}_{\mathbf{A}} \underbrace{\operatorname{vec}(\mathcal{G}_{(1)}')}_{\mathbf{g}} + \underbrace{\operatorname{vec}(\mathbf{E})}_{\mathbf{e}}, \end{aligned} \tag{S2.33}$$

where the third and fourth equalities follow from Equation 12.3.9 of Golub and Van Loan (2013). Note that $\widehat{\mathcal{W}} = \widehat{\mathcal{G}} \times_1 \widehat{\mathbf{A}}_1 \times_2 \widehat{\mathbf{A}}_2 \times_3 \widehat{\mathbf{A}}_3$. Denote

$$\widehat{\mathbf{A}} = \widehat{\mathbf{A}}_1 \otimes \widehat{\mathbf{A}}_3 \otimes \widehat{\mathbf{A}}_2, \quad \widehat{\mathbf{g}} = \mathrm{vec}(\widehat{\mathcal{G}}'_{(1)}). \tag{S2.34}$$

Then, it can be easily verified that

$$\begin{aligned} \frac{1}{2T}\sum_{t=1}^{T}\left\|\mathbf{y}_t - (\mathcal{G} \times_1 \mathbf{A}_1 \times_2 \mathbf{A}_2 \times_3 \mathbf{A}_3)_{(1)}\mathbf{x}_t\right\|_2^2 &= \frac{1}{2T}\|\mathbf{y} - \mathbf{Z}\mathbf{A}\mathbf{g}\|_2^2, \\ \frac{1}{2T}\sum_{t=1}^{T}\left\|\mathbf{y}_t - (\widehat{\mathcal{G}} \times_1 \widehat{\mathbf{A}}_1 \times_2 \widehat{\mathbf{A}}_2 \times_3 \widehat{\mathbf{A}}_3)_{(1)}\mathbf{x}_t\right\|_2^2 &= \frac{1}{2T}\|\mathbf{y} - \mathbf{Z}\widehat{\mathbf{A}}\widehat{\mathbf{g}}\|_2^2. \end{aligned} \tag{S2.35}$$

Due to the optimality of $\widehat{\mathcal{W}}$ in (2.4), we have

$$\begin{aligned} &\frac{1}{2T}\sum_{t=1}^{T}\left\|\mathbf{y}_t - (\widehat{\mathcal{G}} \times_1 \widehat{\mathbf{A}}_1 \times_2 \widehat{\mathbf{A}}_2 \times_3 \widehat{\mathbf{A}}_3)_{(1)}\mathbf{x}_t\right\|_2^2 + \beta\|\widehat{\mathbf{g}}\|_1 + \sum_{i=1}^{3}\alpha_i \,\mathrm{tr}(\widehat{\mathbf{A}}_i'\mathbf{L}_i\widehat{\mathbf{A}}_i) \\ \le\ &\frac{1}{2T}\sum_{t=1}^{T}\left\|\mathbf{y}_t - (\mathcal{G} \times_1 \mathbf{A}_1 \times_2 \mathbf{A}_2 \times_3 \mathbf{A}_3)_{(1)}\mathbf{x}_t\right\|_2^2 + \beta\|\mathbf{g}\|_1 + \sum_{i=1}^{3}\alpha_i \,\mathrm{tr}(\mathbf{A}_i'\mathbf{L}_i\mathbf{A}_i). \end{aligned}$$

Using (S2.35) further leads to

$$\begin{aligned} &\frac{1}{2T}\|\mathbf{y} - \mathbf{Z}\widehat{\mathbf{A}}\widehat{\mathbf{g}}\|_2^2 + \beta\|\widehat{\mathbf{g}}\|_1 + \sum_{i=1}^{3}\alpha_i\left(\mathrm{tr}(\widehat{\mathbf{A}}_i'\mathbf{L}_i\widehat{\mathbf{A}}_i) - \mathrm{tr}(\mathbf{A}_i'\mathbf{L}_i\mathbf{A}_i)\right) \\ \le\ &\frac{1}{2T}\|\mathbf{y} - \mathbf{Z}\mathbf{A}\mathbf{g}\|_2^2 + \beta\|\mathbf{g}\|_1. \end{aligned} \tag{S2.36}$$

Denote $\Delta_{\mathbf{A}_i} = \widehat{\mathbf{A}}_i - \mathbf{A}_i, i = 1, 2, 3$. Then we get

$$\begin{aligned} &\mathrm{tr}(\widehat{\mathbf{A}}_1'\mathbf{L}_1\widehat{\mathbf{A}}_1) - \mathrm{tr}(\mathbf{A}_1'\mathbf{L}_1\mathbf{A}_1) \\ =\ &\mathrm{tr}((\mathbf{A}_1 + \Delta_{\mathbf{A}_1})'\mathbf{L}_1(\mathbf{A}_1 + \Delta_{\mathbf{A}_1})) - \mathrm{tr}(\mathbf{A}_1'\mathbf{L}_1\mathbf{A}_1) \\ =\ &\mathrm{tr}(\mathbf{A}_1'\mathbf{L}_1\mathbf{A}_1 + \mathbf{A}_1'\mathbf{L}_1\Delta_{\mathbf{A}_1} + \Delta_{\mathbf{A}_1}'\mathbf{L}_1\mathbf{A}_1 + \Delta_{\mathbf{A}_1}'\mathbf{L}_1\Delta_{\mathbf{A}_1}) - \mathrm{tr}(\mathbf{A}_1'\mathbf{L}_1\mathbf{A}_1) \\ =\ &2\,\mathrm{tr}(\mathbf{A}_1'\mathbf{L}_1\Delta_{\mathbf{A}_1}) + \mathrm{tr}(\Delta_{\mathbf{A}_1}'\mathbf{L}_1\Delta_{\mathbf{A}_1}), \end{aligned} \tag{S2.37}$$

where the last equality follows from the fact that $\mathbf{L}_1$ is symmetric. Note that

$$\mathrm{tr}(\mathbf{A}_1'\mathbf{L}_1\Delta_{\mathbf{A}_1}) = \langle \mathbf{L}_1\mathbf{A}_1, \Delta_{\mathbf{A}_1}\rangle \geq -\|\mathbf{L}_1\mathbf{A}_1\|_F\|\Delta_{\mathbf{A}_1}\|_F, \tag{S2.38}$$

where the inequality holds by Cauchy-Schwarz inequality. Combining the above inequality with Lemma 1 and (S2.37) yields

$$\mathrm{tr}(\widehat{\mathbf{A}}_1'\mathbf{L}_1\widehat{\mathbf{A}}_1) - \mathrm{tr}(\mathbf{A}_1'\mathbf{L}_1\mathbf{A}_1) \geq \frac{\lambda_2(\mathbf{L}_1)}{2}\|\widehat{\mathbf{A}}_1 - \mathbf{A}_1\|_F^2 - 2\|\mathbf{L}_1\mathbf{A}_1\|_F\|\Delta_{\mathbf{A}_1}\|_F. \tag{S2.39}$$

By using similar techniques, we can easily obtain

$$\begin{aligned}
\mathrm{tr}(\widehat{\mathbf{A}}_2'\mathbf{L}_2\widehat{\mathbf{A}}_2) - \mathrm{tr}(\mathbf{A}_2'\mathbf{L}_2\mathbf{A}_2) &\geq \frac{\lambda_2(\mathbf{L}_2)}{2}\|\widehat{\mathbf{A}}_2 - \mathbf{A}_2\|_F^2 - 2\|\mathbf{L}_2\mathbf{A}_2\|_F\|\Delta_{\mathbf{A}_2}\|_F, \\
\mathrm{tr}(\widehat{\mathbf{A}}_3'\mathbf{L}_3\widehat{\mathbf{A}}_3) - \mathrm{tr}(\mathbf{A}_3'\mathbf{L}_3\mathbf{A}_3) &\geq \frac{\lambda_2(\mathbf{L}_3)}{2}\|\widehat{\mathbf{A}}_3 - \mathbf{A}_3\|_F^2 - 2\|\mathbf{L}_3\mathbf{A}_3\|_F\|\Delta_{\mathbf{A}_3}\|_F.
\end{aligned} \tag{S2.40}$$

Substituting (S2.39), (S2.40) into (S2.36) yields

$$\begin{aligned}
&\frac{1}{2T}\|\mathbf{y} - \mathbf{Z}\widehat{\mathbf{A}}\widehat{\mathbf{g}}\|_2^2 + \beta\|\widehat{\mathbf{g}}\|_1 + \sum_{i=1}^{3}\alpha_i\left(\frac{\lambda_2(\mathbf{L}_i)}{2}\|\widehat{\mathbf{A}}_i - \mathbf{A}_i\|_F^2 - 2\|\mathbf{L}_i\mathbf{A}_i\|_F\|\Delta_{\mathbf{A}_i}\|_F\right) \\
&\leq \frac{1}{2T}\|\mathbf{y} - \mathbf{Z}\mathbf{A}\mathbf{g}\|_2^2 + \beta\|\mathbf{g}\|_1.
\end{aligned} \tag{S2.41}$$

Denote

$$\widehat{\Delta} := \widehat{\mathbf{A}}\widehat{\mathbf{g}} - \mathbf{A}\mathbf{g}, \tag{S2.42}$$

where $\mathbf{A}, \mathbf{g}$ and $\widehat{\mathbf{A}}, \widehat{\mathbf{g}}$ are defined in (S2.33) and (S2.34), respectively. We

can easily get that

$$\frac{1}{2T}\|\mathbf{y}-\mathbf{Z}\widehat{\mathbf{A}}\widehat{\mathbf{g}}\|_2^2=\frac{1}{2T}\|\mathbf{y}-\mathbf{Z}(\widehat{\Delta}+\mathbf{A}\mathbf{g})\|_2^2=\frac{1}{2T}\|-\mathbf{Z}\widehat{\Delta}+\mathbf{e}\|_2^2$$
$$=\frac{1}{2T}\|\mathbf{Z}\widehat{\Delta}\|_2^2+\frac{1}{2T}\|\mathbf{e}\|_2^2-\frac{1}{T}\langle\mathbf{Z}\widehat{\Delta},\mathbf{e}\rangle,$$

where the second equality follows from (S2.33). Combining the above result with (S2.33) and (S2.41) immediately yields

$$\frac{1}{2T}\|\mathbf{Z}\widehat{\Delta}\|_2^2+\frac{1}{2T}\|\mathbf{e}\|_2^2-\frac{1}{T}\langle\mathbf{Z}\widehat{\Delta},\mathbf{e}\rangle+\beta\|\widehat{\mathbf{g}}\|_1$$
$$+\sum_{i=1}^{3}\alpha_i\left(\frac{\lambda_2(\mathbf{L}_i)}{2}\|\widehat{\mathbf{A}}_i-\mathbf{A}_i\|_F^2-2\|\mathbf{L}_i\mathbf{A}_i\|_F\|\Delta_{\mathbf{A}_i}\|_F\right)\leq\frac{1}{2T}\|\mathbf{e}\|_2^2+\beta\|\mathbf{g}\|_1,$$

which is equivalent to

$$\frac{1}{2T}\|\mathbf{Z}\widehat{\Delta}\|_2^2+\beta\|\widehat{\mathbf{g}}\|_1+\sum_{i=1}^{3}\frac{\alpha_i\lambda_2(\mathbf{L}_i)}{2}\|\widehat{\mathbf{A}}_i-\mathbf{A}_i\|_F^2$$
$$\leq\beta\|\mathbf{g}\|_1+\frac{1}{T}\langle\mathbf{Z}\widehat{\Delta},\mathbf{e}\rangle+\sum_{i=1}^{3}2\alpha_i\|\mathbf{L}_i\mathbf{A}_i\|_F\|\Delta_{\mathbf{A}_i}\|_F.$$

Then, by Lemma 2 and the assumption $\beta\geq 2\pi\vartheta a_2\sqrt{\log(m^2p)/T}$, with probability at least $1-6\exp(-a\log(m^2p))$, the following inequality holds:

$$\begin{aligned}&\frac{1}{2T}\|\mathbf{Z}\widehat{\Delta}\|_2^2+\beta\|\widehat{\mathbf{g}}\|_1+\sum_{i=1}^{3}\frac{\alpha_i\lambda_2(\mathbf{L}_i)}{2}\|\widehat{\mathbf{A}}_i-\mathbf{A}_i\|_F^2\\&\leq\beta\|\mathbf{g}\|_1+\beta\|\widehat{\Delta}\|_1+\sum_{i=1}^{3}2\alpha_i\|\mathbf{L}_i\mathbf{A}_i\|_F\|\Delta_{\mathbf{A}_i}\|_F.\end{aligned}\tag{S2.43}$$

Let $\Delta_{\mathbf{g}}:=\widehat{\mathbf{g}}-\mathbf{g}$. Denote the nonzero index set of $\mathrm{vec}(\mathcal{G}')$ by $\mathbb{S}_g$, and let $\mathbb{S}_{g^c}$ represent its complementary set. Then, we have

$$\|\widehat{\mathbf{g}}\|_1=\left\|\widehat{\mathbf{g}}_{\mathbb{S}_g}\right\|_1+\left\|\widehat{\mathbf{g}}_{\mathbb{S}_{g^c}}\right\|_1=\left\|(\mathbf{g}+\Delta_{\mathbf{g}})_{\mathbb{S}_g}\right\|_1+\left\|\widehat{\mathbf{g}}_{\mathbb{S}_{g^c}}\right\|_1$$
$$\geq\left\|\mathbf{g}_{\mathbb{S}_g}\right\|_1-\left\|(\Delta_{\mathbf{g}})_{\mathbb{S}_g}\right\|_1+\left\|\widehat{\mathbf{g}}_{\mathbb{S}_{g^c}}\right\|_1,$$

which taken collectively with (S2.43) gives

$$\frac{1}{2T}\|\mathbf{Z}\widehat{\Delta}\|_2^2+\beta\left\|\mathbf{g}_{\mathbb{S}_g}\right\|_1-\beta\left\|(\Delta_{\mathbf{g}})_{\mathbb{S}_g}\right\|_1+\beta\left\|\widehat{\mathbf{g}}_{\mathbb{S}_{g^c}}\right\|_1+\sum_{i=1}^{3}\frac{\alpha_i\lambda_2(\mathbf{L}_i)}{2}\|\widehat{\mathbf{A}}_i-\mathbf{A}_i\|_F^2$$
$$\leq\beta\|\mathbf{g}\|_1+\beta\|\widehat{\Delta}\|_1+\sum_{i=1}^{3}2\alpha_i\|\mathbf{L}_i\mathbf{A}_i\|_F\|\Delta_{\mathbf{A}_i}\|_F.$$

The above inequality can be rewritten as

$$\frac{1}{2T}\|\mathbf{Z}\widehat{\Delta}\|_2^2+\beta\left\|\widehat{\mathbf{g}}_{\mathbb{S}_{g^c}}\right\|_1+\sum_{i=1}^{3}\frac{\alpha_i\lambda_2(\mathbf{L}_i)}{2}\|\widehat{\mathbf{A}}_i-\mathbf{A}_i\|_F^2$$
$$\leq\beta\|\widehat{\Delta}\|_1+\sum_{i=1}^{3}2\alpha_i\|\mathbf{L}_i\mathbf{A}_i\|_F\|\Delta_{\mathbf{A}_i}\|_F+\beta\left\|(\Delta_{\mathbf{g}})_{\mathbb{S}_g}\right\|_1.$$

It follows from $\beta\left\|\widehat{\mathbf{g}}_{\mathbb{S}_{g^c}}\right\|_1\geq 0$ that

$$\frac{1}{2T}\|\mathbf{Z}\widehat{\Delta}\|_2^2+\sum_{i=1}^{3}\frac{\alpha_i\lambda_2(\mathbf{L}_i)}{2}\|\widehat{\mathbf{A}}_i-\mathbf{A}_i\|_F^2$$
$$\leq\beta\|\widehat{\Delta}\|_1+\sum_{i=1}^{3}2\alpha_i\|\mathbf{L}_i\mathbf{A}_i\|_F\|\Delta_{\mathbf{A}_i}\|_F+\beta\left\|(\Delta_{\mathbf{g}})_{\mathbb{S}_g}\right\|_1.$$

Note that $T\geq\frac{8\varsigma^2}{a_3}\bar{p}\min(\log(mp),\log(21emp/\bar{p}))$. Lemma 3 implies that the following inequality holds

$$\begin{aligned}&\frac{\varpi}{4}\|\widehat{\Delta}\|_2^2+\sum_{i=1}^{3}\frac{\alpha_i\lambda_2(\mathbf{L}_i)}{2}\|\widehat{\mathbf{A}}_i-\mathbf{A}_i\|_F^2\\&\leq\beta\|\widehat{\Delta}\|_1+\sum_{i=1}^{3}2\alpha_i\|\mathbf{L}_i\mathbf{A}_i\|_F\|\Delta_{\mathbf{A}_i}\|_F+\beta\left\|(\Delta_{\mathbf{g}})_{\mathbb{S}_g}\right\|_1\end{aligned}\tag{S2.44}$$

with probability at least $1-2\exp(-\bar{p}\min(\log mp,\log(21emp/\bar{p})))$, where $\bar{p}$ is related to τ.

Note that $\alpha_i \geq \frac{6c^2}{\lambda_2(\mathbf{L}_i)}\sqrt{\frac{s\log(m^2p)}{T}} = \frac{6c^2 r_1r_2r_3}{\lambda_2(\mathbf{L}_i)}\sqrt{\frac{s\log(m^2p)}{(r_1r_2r_3)^2T}}$. Then, we have

$$\begin{aligned}
&\sum_{i=1}^{3}\alpha_i\frac{\lambda_2(\mathbf{L}_i)}{2}\|\widehat{\mathbf{A}}_i-\mathbf{A}_i\|_F^2+\sqrt{\frac{s\log(m^2p)}{(r_1r_2r_3)^2T}}\|\operatorname{vec}(\widehat{\mathcal{G}}'_{(1)})-\operatorname{vec}(\mathcal{G}'_{(1)})\|_2^2\\
\geq\ & 3c^2r_1r_2r_3\sqrt{\frac{s\log(m^2p)}{(r_1r_2r_3)^2T}}\left(\|\widehat{\mathbf{A}}_3-\mathbf{A}_3\|_F^2+\|\widehat{\mathbf{A}}_2-\mathbf{A}_2\|_F^2+\|\widehat{\mathbf{A}}_1-\mathbf{A}_1\|_F^2\right)\\
&+\sqrt{\frac{s\log(m^2p)}{(r_1r_2r_3)^2T}}\|\operatorname{vec}(\widehat{\mathcal{G}}'_{(1)})-\operatorname{vec}(\mathcal{G}'_{(1)})\|_2^2\\
\geq\ & \frac{1}{2}\sqrt{\frac{s\log(m^2p)}{(r_1r_2r_3)^2T}}\|\widehat{\Delta}\|_2^2,
\end{aligned}$$

where the last inequality follows from Lemma 4. This combined with (S2.44) leads to

$$\begin{aligned}
\frac{\varpi}{4}\|\widehat{\Delta}\|_2^2+\frac{1}{2}\sqrt{\frac{s\log(m^2p)}{(r_1r_2r_3)^2T}}\|\widehat{\Delta}\|_2^2 \leq \beta\|\widehat{\Delta}\|_1+\sum_{i=1}^{3}2\alpha_i\|\mathbf{L}_i\mathbf{A}_i\|_F\|\Delta_{\mathbf{A}_i}\|_F+\beta\left\|(\Delta_{\mathbf{g}})_{\mathbb{S}_g}\right\|_1\\
+\sqrt{\frac{s\log(m^2p)}{(r_1r_2r_3)^2T}}\|\operatorname{vec}(\widehat{\mathcal{G}}'_{(1)})-\operatorname{vec}(\mathcal{G}'_{(1)})\|_2^2
\end{aligned} \tag{S2.45}$$

with probability at least $1-2\exp(-\bar{p}\min(\log mp, \log(21emp/\bar{p})))$.

Recall that $\tau = 1-s/(r_1r_2r_3) \in [0,1]$. Since both $\mathcal{G}$ and $\widehat{\mathcal{G}}$ belong to the set $\mathcal{U}$, it follows from Lemma 6 that $\|\mathbf{Ag}\|_0 \leq s_1m^2p$ and $\|\widehat{\mathbf{A}}\widehat{\mathbf{g}}\|_0 \leq s_2m^2p$, where $s_1, s_2 \in [0,1]$ are the sparsity thresholds related to τ. Then, we have

$$\begin{aligned}
\|\widehat{\Delta}\|_0 = \|\widehat{\mathbf{A}}\widehat{\mathbf{g}}-\mathbf{Ag}\|_0 &\leq \|\widehat{\mathbf{A}}\widehat{\mathbf{g}}\|_0+\|\mathbf{Ag}\|_0\\
&\leq s_2m^2p+s_1m^2p := s_{\text{aux}}.
\end{aligned} \tag{S2.46}$$

As a consequence, we get

$$\|\widehat{\Delta}\|_1 \leq \sqrt{s_{\text{aux}}}\|\widehat{\Delta}\|_2. \tag{S2.47}$$

In addition,

$$\left\|(\Delta_{\mathbf{g}})_{\mathbb{S}_g}\right\|_1 \leq \sqrt{s}\|\Delta_{\mathbf{g}}\|_2 \leq b\sqrt{s}\|\widehat{\Delta}\|_2, \tag{S2.48}$$

where the last inequality follows from Lemma 5. By Hölder's inequality, we have

$$\begin{aligned}
\|\operatorname{vec}(\widehat{\mathcal{G}}'_{(1)}) - \operatorname{vec}(\mathcal{G}'_{(1)})\|_2^2 &\leq \|\operatorname{vec}(\widehat{\mathcal{G}}'_{(1)}) - \operatorname{vec}(\mathcal{G}'_{(1)})\|_\infty \|\operatorname{vec}(\widehat{\mathcal{G}}'_{(1)}) - \operatorname{vec}(\mathcal{G}'_{(1)})\|_1 \\
&\leq 2c\|\operatorname{vec}(\widehat{\mathcal{G}}'_{(1)}) - \operatorname{vec}(\mathcal{G}'_{(1)})\|_1 \\
&= 2c\|\Delta_{\mathbf{g}}\|_1 \\
&\leq 2c\sqrt{2s}\,\|\Delta_{\mathbf{g}}\|_2 \\
&\leq 2bc\sqrt{2s}\|\widehat{\Delta}\|_2,
\end{aligned} \tag{S2.49}$$

where the third inequality follows from the fact that both $\mathcal{G}$ and $\widehat{\mathcal{G}}$ belong to the set $\mathcal{U}$ and the last inequality follows from Lemma 5. Substituting (S2.47), (S2.48) and (S2.49) into (S2.45) then gives

$$\begin{aligned}
&\left(\frac{\varpi}{4} + \frac{1}{2}\sqrt{\frac{s\log(m^2p)}{(r_1r_2r_3)^2T}}\right)\|\widehat{\Delta}\|_2^2 \\
&\leq \beta\sqrt{s_{\text{aux}}}\|\widehat{\Delta}\|_2 + \sum_{i=1}^{3} 2\alpha_i\|\mathbf{L}_i\mathbf{A}_i\|_F\|\Delta_{\mathbf{A}_i}\|_F + b\beta\sqrt{s}\|\widehat{\Delta}\|_2 + 2\sqrt{2}bc\sqrt{\frac{s^2\log(m^2p)}{(r_1r_2r_3)^2T}}\|\widehat{\Delta}\|_2 \\
&\leq \beta\sqrt{s_{\text{aux}}}\|\widehat{\Delta}\|_2 + \sum_{i=1}^{3} 2\alpha_i\|\mathbf{L}_i\mathbf{A}_i\|_F\|\Delta_{\mathbf{A}_i}\|_F + b\beta\sqrt{s}\|\widehat{\Delta}\|_2 + 2\sqrt{2}bc\sqrt{\frac{\log(m^2p)}{T}}\|\widehat{\Delta}\|_2
\end{aligned} \tag{S2.50}$$

with probability at least $1-2\exp(-\bar{p}\min(\log mp, \log(21emp/\bar{p})))$, where the last inequality follows from $s \leq r_1 r_2 r_3$. Note that

$$2\alpha_i \|\mathbf{L}_i \mathbf{A}_i\|_F \|\Delta_{\mathbf{A}_i}\|_F \leq 2\alpha_i \|\mathbf{L}_i\|_F \|\mathbf{A}_i\| \|\Delta_{\mathbf{A}_i}\|_F \leq 2\alpha_i l_i \|\Delta_{\mathbf{A}_i}\|_F,$$

where $l_i := \|\mathbf{L}_i\|_F$. Using Lemma 5 further leads to

$$2\alpha_i l_i \|\Delta_{\mathbf{A}_i}\|_F \leq 2\alpha_i l_i \frac{2^{\frac{3}{2}}\eta_i}{\rho} \|\widehat{\Delta}\|_2, i = 1, 2, 3.$$

Taken the above inequality together with (S2.50) demonstrates that

$$\begin{aligned}
&\left(\frac{\varpi}{4} + \frac{1}{2}\sqrt{\frac{s\log(m^2 p)}{(r_1 r_2 r_3)^2 T}}\right) \|\widehat{\Delta}\|_2^2 \\
&\leq \beta\sqrt{s_{\text{aux}}} \|\widehat{\Delta}\|_2 + \sum_{i=1}^{3} \alpha_i l_i \frac{2^{\frac{5}{2}}\eta_i}{\rho} \|\widehat{\Delta}\|_2 + b\beta\sqrt{s} \|\widehat{\Delta}\|_2 + 2\sqrt{2}bc\sqrt{\frac{\log(m^2 p)}{T}} \|\widehat{\Delta}\|_2
\end{aligned}$$

with probability at least $1-2\exp(-\bar{p}\min(\log mp, \log(21emp/\bar{p})))$. Consequently, we obtain

$$\begin{aligned}
\|\widehat{\Delta}\|_2 &\leq \frac{4\beta\sqrt{s_{\text{aux}}} + \sum_{i=1}^{3} 2^{\frac{9}{2}} \alpha_i l_i \frac{\eta_i}{\rho} + 4b\beta\sqrt{s} + 8\sqrt{2}bc\sqrt{(\log(m^2 p))/T}}{\varpi + 2\sqrt{(s\log(m^2 p))/((r_1 r_2 r_3)^2 T)}} \\
&= \frac{4\beta(\sqrt{s_{\text{aux}}} + b\sqrt{s}) + \hat{a}\kappa + c_0 \sum_{i=1}^{3} \alpha_i l_i \eta_i}{\varpi + 2\kappa\sqrt{s/(r_1 r_2 r_3)^2}}
\end{aligned} \tag{S2.51}$$

with probability at least $1-2\exp(-\bar{p}\min(\log mp, \log(21emp/\bar{p})))$, where $\hat{a} = 8\sqrt{2}bc$, $\kappa = \sqrt{(\log(m^2 p))/T}$ and $c_0 = \frac{2^{\frac{9}{2}}}{\rho}$.

Furthermore, with probability at least $1-2\exp(-\bar{p}\min(\log mp, \log(\frac{21emp}{\bar{p}})))$,

$$
\begin{aligned}
& T^{-1}\sum_{t=1}^{T}\|\widehat{\mathcal{W}}_{(1)}\mathbf{x}_t - \mathcal{W}_{(1)}\mathbf{x}_t\|_2^2 \\
&= T^{-1}\|\mathbf{X}\widehat{\mathcal{W}}_{(1)}' - \mathbf{X}\mathcal{W}_{(1)}'\|_F^2 \\
&= T^{-1}\|\operatorname{vec}(\mathbf{X}\widehat{\mathcal{W}}_{(1)}') - \operatorname{vec}(\mathbf{X}\mathcal{W}_{(1)}')\|_2^2 \\
&= T^{-1}\|\left(\mathbf{I}_m \otimes \mathbf{X}\right)\widehat{\mathbf{A}}\widehat{\mathbf{g}} - \left(\mathbf{I}_m \otimes \mathbf{X}\right)\mathbf{A}\mathbf{g}\|_2^2 \\
&= T^{-1}\|\left(\mathbf{I}_m \otimes \mathbf{X}\right)\widehat{\Delta}\|_2^2 \\
&= \widehat{\Delta}'\left(\mathbf{I}_m \otimes (\mathbf{X}'\mathbf{X}/T)\right)\widehat{\Delta} \\
&\leq \lambda_{\max}(\mathbf{X}'\mathbf{X}/T)\|\widehat{\Delta}\|_2^2 \\
&\leq \lambda_{\max}\left(\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}\right)/\mu_{\min}(\mathbf{W})\|\widehat{\Delta}\|_2^2 \\
&\leq \frac{\lambda_{\max}\left(\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}\right)\left(4\beta(\sqrt{s_{\text{aux}}} + b\sqrt{s}) + \hat{a}\kappa + c_0\sum_{i=1}^{3}\alpha_i l_i \eta_i\right)^2}{\mu_{\min}(\mathbf{W})(\varpi + 2\kappa\sqrt{s/(r_1 r_2 r_3)^2})^2},
\end{aligned}
$$

where the first equality follows from (S2.32), the third equality follows from (S2.33), the fourth equality follows from $\widehat{\Delta} = \widehat{\mathbf{A}}\widehat{\mathbf{g}} - \mathbf{A}\mathbf{g}$, the first inequality follows from Theorem 4.2.2 of Horn and Johnson (2013) and $\lambda_{\max}\left(\mathbf{I}_m \otimes (\mathbf{X}'\mathbf{X}/T)\right) = \lambda_{\max}((\mathbf{X}'\mathbf{X}/T))$, the second inequality follows from the equation (6) of Masini et al. (2022) and the third inequality follows from (S2.51). This completes the proof.

S3 PALM Algorithm

We recall the definition of subgradient of a general function, which is given as follows.

Definition 1. Rockafellar and Wets (1998) Consider a function $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ and any $\mathbf{x} \in \mathrm{dom}(f)$. For a vector $\mathbf{v} \in \mathbb{R}^n$, one says

(a) $\mathbf{v}$ is a regular subgradient of f at $\mathbf{x}$, written as $\hat{\partial} f(\mathbf{x})$, if

$$\liminf_{\substack{\mathbf{y}\neq\mathbf{x}\\ \mathbf{y}\to\mathbf{x}}} \frac{1}{\|\mathbf{x}-\mathbf{y}\|_2}\left[f(\mathbf{y}) - f(\mathbf{x}) - \langle \mathbf{v}, \mathbf{y}-\mathbf{x}\rangle\right] \geq 0.$$

(b) $\mathbf{v}$ is a subgradient of f at $\mathbf{x}$, written as $\partial f(\mathbf{x})$, if there exist sequences $\mathbf{x}^k \to \mathbf{x}$, $f(\mathbf{x}^k) \to f(\mathbf{x})$, $\mathbf{v}^k \in \hat{\partial} f(\mathbf{x}^k)$ with $\mathbf{v}^k \to \mathbf{v}$.

Before proceeding, for simplicity, we denote

$$\begin{aligned} F(\mathcal{G}, \mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3, \mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3) &:= \frac{1}{2T}\sum_{t=1}^{T}\left\|\mathbf{y}_t - (\mathcal{G}\times_1 \mathbf{A}_1 \times_2 \mathbf{A}_2 \times_3 \mathbf{A}_3)_{(1)}\mathbf{x}_t\right\|_2^2 \\ &+ \beta\|\mathcal{G}\|_1 + \sum_{i=1}^{3}\alpha_i \,\mathrm{tr}\left(\mathbf{U}_i'\mathbf{L}_i\mathbf{U}_i\right) + \sum_{i=1}^{3}\frac{\gamma_i}{2}\|\mathbf{U}_i - \mathbf{A}_i\|_F^2 + \sum_{i=1}^{3}\delta_{\mathfrak{B}_i}(\mathbf{A}_i) + \delta_{\mathcal{D}}(\mathcal{G}) \end{aligned} \tag{S3.52}$$

and

$$Q([\![\mathcal{G}; \mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3]\!]) := \frac{1}{2T}\sum_{t=1}^{T}\left\|\mathbf{y}_t - (\mathcal{G}\times_1 \mathbf{A}_1 \times_2 \mathbf{A}_2 \times_3 \mathbf{A}_3)_{(1)}\mathbf{x}_t\right\|_2^2. \tag{S3.53}$$

Then, by the definition of Ψ, we get that

$$\Psi(\mathcal{G}, \mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3, \mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3) = Q([\![\mathcal{G}; \mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3]\!]) + \sum_{i=1}^{3} \frac{\gamma_i}{2} \|\mathbf{U}_i - \mathbf{A}_i\|_F^2 \tag{S3.54}$$

and

$$\begin{aligned} F(\mathcal{G}, \mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3, \mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3) = \Psi(\mathcal{G}, \mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3, \mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3) + \beta\|\mathcal{G}\|_1 \\ + \sum_{i=1}^{3} \alpha_i \operatorname{tr}\left(\mathbf{U}_i' \mathbf{L}_i \mathbf{U}_i\right) + \sum_{i=1}^{3} \delta_{\mathfrak{B}_i}(\mathbf{A}_i) + \delta_{\mathcal{D}}(\mathcal{G}). \end{aligned}$$

It follows from Section D.2 of Wang et al. (2022) that the gradient of the loss function Q with respect to the coefficient tensor $\mathcal{W}$ is given by

$$\nabla Q([\![\mathcal{G}; \mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3]\!]) = \frac{1}{T} \sum_{t=1}^{T} \left(\mathcal{W}_{(1)} \mathbf{x}_t - \mathbf{y}_t\right) \circ \mathbf{X}_t = \frac{1}{T} \sum_{t=1}^{T} (-\boldsymbol{\varepsilon}_t) \circ \mathbf{X}_t, \tag{S3.55}$$

where $\mathbf{X}_t = (\mathbf{y}_{t-1}, \mathbf{y}_{t-2}, \ldots, \mathbf{y}_{t-p}) \in \mathbb{R}^{m \times p}$.

It follows from Lemma 2.1 of Han et al. (2022) that the partial gradients of Q are given as follows:

$$\begin{cases} \nabla_{\mathcal{G}} Q([\![\mathcal{G}; \mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3]\!]) = \nabla Q([\![\mathcal{G}; \mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3]\!]) \times_1 \mathbf{A}_1' \times_2 \mathbf{A}_2' \times_3 \mathbf{A}_3', \\ \nabla_{\mathbf{A}_1} Q([\![\mathcal{G}; \mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3]\!]) = (\nabla Q([\![\mathcal{G}; \mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3]\!]))_{(1)} (\mathbf{A}_3 \otimes \mathbf{A}_2) \mathcal{G}_{(1)}', \\ \nabla_{\mathbf{A}_2} Q([\![\mathcal{G}; \mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3]\!]) = (\nabla Q([\![\mathcal{G}; \mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3]\!]))_{(2)} (\mathbf{A}_1 \otimes \mathbf{A}_3) \mathcal{G}_{(2)}', \\ \nabla_{\mathbf{A}_3} Q([\![\mathcal{G}; \mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3]\!]) = (\nabla Q([\![\mathcal{G}; \mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3]\!]))_{(3)} (\mathbf{A}_2 \otimes \mathbf{A}_1) \mathcal{G}_{(3)}'. \end{cases} \tag{S3.56}$$

Next, we can deduce the gradient of Ψ (defined in (S3.54)) with respect to

$\mathcal{G}, \mathbf{A}_i, \mathbf{U}_i$ as follows:

$$\begin{cases} \nabla_{\mathcal{G}}\Psi(\mathcal{G}, \mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3, \mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3) = \nabla_{\mathcal{G}}Q([\![\mathcal{G}; \mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3]\!]), \\ \nabla_{\mathbf{A}_1}\Psi(\mathcal{G}, \mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3, \mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3) = \nabla_{\mathbf{A}_1}Q([\![\mathcal{G}; \mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3]\!]) - \gamma_1(\mathbf{U}_1 - \mathbf{A}_1), \\ \nabla_{\mathbf{A}_2}\Psi(\mathcal{G}, \mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3, \mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3) = \nabla_{\mathbf{A}_2}Q([\![\mathcal{G}; \mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3]\!]) - \gamma_2(\mathbf{U}_2 - \mathbf{A}_2), \\ \nabla_{\mathbf{A}_3}\Psi(\mathcal{G}, \mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3, \mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3) = \nabla_{\mathbf{A}_3}Q([\![\mathcal{G}; \mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3]\!]) - \gamma_3(\mathbf{U}_3 - \mathbf{A}_3), \\ \nabla_{\mathbf{U}_1}\Psi(\mathcal{G}, \mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3, \mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3) = \gamma_1(\mathbf{U}_1 - \mathbf{A}_1), \\ \nabla_{\mathbf{U}_2}\Psi(\mathcal{G}, \mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3, \mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3) = \gamma_2(\mathbf{U}_2 - \mathbf{A}_2), \\ \nabla_{\mathbf{U}_3}\Psi(\mathcal{G}, \mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3, \mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3) = \gamma_3(\mathbf{U}_3 - \mathbf{A}_3). \end{cases} \tag{S3.57}$$

We now provide detailed solutions for $\mathcal{G}$, $\mathbf{A}_i$, and $\mathbf{U}_i$ for $i = 1, 2, 3$.

The subproblem about $\mathcal{G}$ in (3.13) can be expressed as

$$\begin{aligned} &\mathcal{G}^{k+1} \\ =\ &\underset{\|\mathcal{G}\|_\infty \le c}{\arg\min}\, \beta\|\mathcal{G}\|_1 + \left\langle \nabla_{\mathcal{G}}\Psi(\mathcal{G}^k, \mathbf{A}_1^k, \mathbf{A}_2^k, \mathbf{A}_3^k, \mathbf{U}_1^k, \mathbf{U}_2^k, \mathbf{U}_3^k), \mathcal{G} - \mathcal{G}^k \right\rangle + \frac{\rho_1}{2}\|\mathcal{G} - \mathcal{G}^k\|_F^2 \\ =\ &\underset{\|\mathcal{G}\|_\infty \le c}{\arg\min}\, \beta\|\mathcal{G}\|_1 + \frac{\rho_1}{2}\left\| \mathcal{G} - \mathcal{G}^k + \frac{1}{\rho_1}\nabla_{\mathcal{G}}\Psi(\mathcal{G}^k, \mathbf{A}_1^k, \mathbf{A}_2^k, \mathbf{A}_3^k, \mathbf{U}_1^k, \mathbf{U}_2^k, \mathbf{U}_3^k) \right\|_F^2 \\ =\ &\underset{\|\mathcal{G}\|_\infty \le c}{\arg\min}\, \beta\|\mathcal{G}\|_1 + \frac{\rho_1}{2}\|\mathcal{G} - \mathcal{L}\|_F^2, \end{aligned}$$

where $\mathcal{L} = \mathcal{G}^k - \frac{1}{\rho_1}\nabla_{\mathcal{G}}\Psi(\mathcal{G}^k, \mathbf{A}_1^k, \mathbf{A}_2^k, \mathbf{A}_3^k, \mathbf{U}_1^k, \mathbf{U}_2^k, \mathbf{U}_3^k)$. By Example 6.34 of Beck (2017), simple calculations show that the closed form solution with

respect to $\mathcal{G}$ can be given by

$$\mathcal{G}_{ijl}^{k+1} = \begin{cases} \operatorname{sign}(\mathcal{L}_{ijl}) \max\{|\mathcal{L}_{ijl}| - \beta/\rho_1, 0\}, & \text{if } |\mathcal{L}_{ijl}| \leq c + \beta/\rho_1, \\ \operatorname{sign}(\mathcal{L}_{ijl})\, c, & \text{if } |\mathcal{L}_{ijl}| > c + \beta/\rho_1, \end{cases} \tag{S3.58}$$

where $\operatorname{sign}(\cdot)$ denotes the signum function.

The subproblem with respect to $\mathbf{A}_1$ can be equivalently expressed as

$$\begin{aligned} & \mathbf{A}_1^{k+1} \\ = & \arg\min_{\mathbf{A}_1} \delta_{\mathfrak{B}_1}(\mathbf{A}_1) + \left\langle \nabla_{\mathbf{A}_1} \Psi(\mathcal{G}^{k+1}, \mathbf{A}_1^k, \mathbf{A}_2^k, \mathbf{A}_3^k, \mathbf{U}_1^k, \mathbf{U}_2^k, \mathbf{U}_3^k), \mathbf{A}_1 - \mathbf{A}_1^k \right\rangle \\ & + \frac{\rho_2}{2} \|\mathbf{A}_1 - \mathbf{A}_1^k\|_F^2 \\ = & \arg\min_{\mathbf{A}_1} \delta_{\mathfrak{B}_1}(\mathbf{A}_1) + \frac{\rho_2}{2} \left\| \mathbf{A}_1 - (\mathbf{A}_1^k - \frac{1}{\rho_2} \nabla_{\mathbf{A}_1} \Psi(\mathcal{G}^{k+1}, \mathbf{A}_1^k, \mathbf{A}_2^k, \mathbf{A}_3^k, \mathbf{U}_1^k, \mathbf{U}_2^k, \mathbf{U}_3^k)) \right\|_F^2. \end{aligned}$$

It follows from Theorem 4 of Zou et al. (2006) that the optimal solution of the above problem is easily computed as

$$\mathbf{A}_1^{k+1} = \bar{\mathbf{U}}_1 \bar{\mathbf{V}}_1'. \tag{S3.59}$$

Here $\mathbf{A}_1^k - \frac{1}{\rho_2} \nabla_{\mathbf{A}_1} \Psi(\mathcal{G}^{k+1}, \mathbf{A}_1^k, \mathbf{A}_2^k, \mathbf{A}_3^k, \mathbf{U}_1^k, \mathbf{U}_2^k, \mathbf{U}_3^k) = \bar{\mathbf{U}}_1 \Sigma_1 \bar{\mathbf{V}}_1'$.

Similar techniques are employed to solve the subproblems with respect to $\mathbf{A}_2$ and $\mathbf{A}_3$. By performing SVDs on the two matrices $\mathbf{A}_2^k - \frac{1}{\rho_3} \nabla_{\mathbf{A}_2} \Psi(\mathcal{G}^{k+1}, \mathbf{A}_1^{k+1}, \mathbf{A}_2^k, \mathbf{A}_3^k, \mathbf{U}_1^k, \mathbf{U}_2^k, \mathbf{U}_3^k)$ and $\mathbf{A}_3^k - \frac{1}{\rho_4} \nabla_{\mathbf{A}_3} \Psi(\mathcal{G}^{k+1}, \mathbf{A}_1^{k+1}, \mathbf{A}_2^{k+1}, \mathbf{A}_3^k, \mathbf{U}_1^k, \mathbf{U}_2^k, \mathbf{U}_3^k)$, i.e.,

$$\mathbf{A}_2^k - \frac{1}{\rho_3} \nabla_{\mathbf{A}_2} \Psi(\mathcal{G}^{k+1}, \mathbf{A}_1^{k+1}, \mathbf{A}_2^k, \mathbf{A}_3^k, \mathbf{U}_1^k, \mathbf{U}_2^k, \mathbf{U}_3^k) = \bar{\mathbf{U}}_2 \Sigma_2 \bar{\mathbf{V}}_2'$$

and

$$\mathbf{A}_3^k - \frac{1}{\rho_4}\nabla_{\mathbf{A}_3}\Psi(\mathcal{G}^{k+1}, \mathbf{A}_1^{k+1}, \mathbf{A}_2^{k+1}, \mathbf{A}_3^k, \mathbf{U}_1^k, \mathbf{U}_2^k, \mathbf{U}_3^k) = \bar{\mathbf{U}}_3\Sigma_3\bar{\mathbf{V}}_3',$$

we get that the closed form solutions with respect to $\mathbf{A}_2$ and $\mathbf{A}_3$ in (3.13) are given by

$$\mathbf{A}_2^{k+1} = \bar{\mathbf{U}}_2\bar{\mathbf{V}}_2', \quad \mathbf{A}_3^{k+1} = \bar{\mathbf{U}}_3\bar{\mathbf{V}}_3'. \tag{S3.60}$$

The subproblem about $\mathbf{U}_1$ in (3.13) can be rewritten equivalently as

$$\begin{aligned}\mathbf{U}_1^{k+1} = & \arg\min_{\mathbf{U}_1} \alpha_1 \operatorname{tr}(\mathbf{U}_1'\mathbf{L}_1\mathbf{U}_1) \\ & + \frac{\rho_5}{2}\left\|\mathbf{U}_1 - \mathbf{U}_1^k + \frac{1}{\rho_5}\nabla_{\mathbf{U}_1}\Psi(\mathcal{G}^{k+1}, \mathbf{A}_1^{k+1}, \mathbf{A}_2^{k+1}, \mathbf{A}_3^{k+1}, \mathbf{U}_1^k, \mathbf{U}_2^k, \mathbf{U}_3^k)\right\|_F^2.\end{aligned} \tag{S3.61}$$

By the optimality condition of (S3.61), the closed form solution with respect to $\mathbf{U}_1$ can be given by

$$\mathbf{U}_1^{k+1} = (2\alpha_1\mathbf{L}_1 + \rho_5\mathbf{I}_m)^{-1}(\rho_5\mathbf{U}_1^k - \nabla_{\mathbf{U}_1}\Psi(\mathcal{G}^{k+1}, \mathbf{A}_1^{k+1}, \mathbf{A}_2^{k+1}, \mathbf{A}_3^{k+1}, \mathbf{U}_1^k, \mathbf{U}_2^k, \mathbf{U}_3^k)). \tag{S3.62}$$

Similar to the calculation of $\mathbf{U}_1^{k+1}$, the optimal solution with respect to $\mathbf{U}_2$ and $\mathbf{U}_3$ are given explicitly by

$$\begin{aligned}&\mathbf{U}_2^{k+1} \\ =&(2\alpha_2\mathbf{L}_2 + \rho_6\mathbf{I}_m)^{-1}(\rho_6\mathbf{U}_2^k - \nabla_{\mathbf{U}_2}\Psi(\mathcal{G}^{k+1}, \mathbf{A}_1^{k+1}, \mathbf{A}_2^{k+1}, \mathbf{A}_3^{k+1}, \mathbf{U}_1^{k+1}, \mathbf{U}_2^k, \mathbf{U}_3^k)),\end{aligned} \tag{S3.63}$$

$$\mathbf{U}_3^{k+1}$$

$$=(2\alpha_3\mathbf{L}_3+\rho_7\mathbf{I}_p)^{-1}(\rho_7\mathbf{U}_3^k-\nabla_{\mathbf{U}_3}\Psi(\mathcal{G}^{k+1},\mathbf{A}_1^{k+1},\mathbf{A}_2^{k+1},\mathbf{A}_3^{k+1},\mathbf{U}_1^{k+1},\mathbf{U}_2^{k+1},\mathbf{U}_3^k)). \tag{S3.64}$$

Now we summarize the PALM algorithm for solving problem (3.12) in Algorithm 1.

Algorithm 1 A PALM Algorithm for Solving Problem (3.12)

1: **Input:** $\{\mathbf{y}_t\},\mathcal{G}^0,\mathbf{A}_1^0,\mathbf{A}_2^0,\mathbf{A}_3^0,\mathbf{U}_1^0,\mathbf{U}_2^0,\mathbf{U}_3^0,$ parameters $\beta,\alpha_1,\alpha_2,\alpha_3,\gamma_1,\gamma_2,\gamma_3,\rho_i,i=1,\ldots,7.$

2: **repeat**

3: **Step 1**. Compute $\mathcal{G}^{k+1}$ by (S3.58).

4: **Step 2**. Compute $\mathbf{A}_i^{k+1}$ by (S3.59) and (S3.60).

5: **Step 3**. Compute $\mathbf{U}_i^{k+1}$ by (S3.62), (S3.63), and (S3.64).

6: **until** A stopping condition is satisfied.

7: **Output:** $\mathcal{G}^{k+1}\times_1\mathbf{A}_1^{k+1}\times_2\mathbf{A}_2^{k+1}\times_3\mathbf{A}_3^{k+1}$.

The computational complexity of Algorithm 1 is given as follows. The main computational complexity of $\mathcal{G}^{k+1}$ is that of computing $\nabla_{\mathcal{G}}\Psi$, which is on the order of $O(m^2pT)$. The computational cost of $\mathbf{A}_1^{k+1}$ is $O(m^2pT+r_1m^2p+r_1r_2r_3mp)$. And the computational costs of $\mathbf{A}_2^{k+1}$ and $\mathbf{A}_3^{k+1}$ are $O(m^2pT+r_2m^2p+r_1r_2r_3mp)$ and $O(m^2pT+r_3m^2p+r_1r_2r_3m^2)$, respectively. The computational complexities of $\mathbf{U}_1^{k+1}$ and $\mathbf{U}_2^{k+1}$ are both $O(m^3)$, and the

computational complexity of $\mathbf{U}_3^{k+1}$ is $O(p^3)$. Therefore, the computational complexity of Algorithm 1 is $O((r_1+r_2+r_3+T)m^2p + r_1r_2r_3m(m+p) + m^3+p^3)$ at each iteration.

S4 Proof of Theorem 2

S4.1 Some lemmas used for the proof of Theorem 2

Lemma 7. *Let $f(\mathcal{X}) : \mathbb{R}^{n_1\times n_2\times n_3} \to \mathbb{R}$ be a differentiable function, and $\bar{f}(\mathcal{X}_{(1)}) : \mathbb{R}^{n_1\times n_2n_3} \to \mathbb{R}$ be a differentiable function with respect to the mode-1 unfolding of $\mathcal{X}$, where $f(\mathcal{X}) = \bar{f}(\mathcal{X}_{(1)})$ for any $\mathcal{X}$. Then, one has*

$$(\nabla f(\mathcal{X}))_{(1)} = \nabla \bar{f}(\mathcal{X}_{(1)}).$$

Proof. For any third-order tensor $\mathcal{X} \in \mathbb{R}^{n_1\times n_2\times n_3}$, we get that the (i,n)-th element of $\mathcal{X}_{(1)}$ is given by

$$(\mathcal{X}_{(1)})_{i,n} = \mathcal{X}_{ijl}, \quad \text{where } n = (l-1)n_2 + j, \quad i \in [n_1], j \in [n_2], l \in [n_3].$$

Note that $(\nabla f(\mathcal{X}))_{ijl} = \frac{\partial f}{\partial \mathcal{X}_{ijl}}$, and $((\nabla f(\mathcal{X}))_{(1)})_{i,n} = (\nabla f(\mathcal{X}))_{ijl}$, where $n = (l-1)n_2+j$. For the gradient of $\bar{f}$ with respect to $\mathcal{X}_{(1)}$, denoted as $\nabla \bar{f}(\mathcal{X}_{(1)}) \in \mathbb{R}^{n_1\times n_2n_3}$, one has

$$(\nabla \bar{f}(\mathcal{X}_{(1)}))_{i,n} = \frac{\partial \bar{f}}{\partial (\mathcal{X}_{(1)})_{i,n}}.$$

Since $(\mathcal{X}_{(1)})_{i,n} = \mathcal{X}_{ijl}$ with $n = (l-1)n_2 + j$ and $f(\mathcal{X}) = \bar{f}(\mathcal{X}_{(1)})$, we get

$$\frac{\partial \bar{f}}{\partial (\mathcal{X}_{(1)})_{i,n}} = \frac{\partial f}{\partial \mathcal{X}_{ijl}} = (\nabla f(\mathcal{X}))_{ijl}.$$

Therefore, we obtain that $(\nabla f(\mathcal{X}))_{(1)} = \nabla \bar{f}(\mathcal{X}_{(1)})$. □

Remark 1. Let $\bar{Q}(\mathcal{W}_{(1)}) := \frac{1}{2T}\sum_{t=1}^{T} \left\|\mathbf{y}_t - \mathcal{W}_{(1)}\mathbf{x}_t\right\|_2^2$ be a function with respect to $\mathcal{W}_{(1)}$. Denote the gradient of $\bar{Q}$ with respect to $\mathcal{W}_{(1)}$ by $\nabla\bar{Q}(\mathcal{W}_{(1)})$. By Lemma 7 and (S3.53), we have

$$\nabla\bar{Q}(\mathcal{W}_{(1)}) = \frac{1}{T}\sum_{t=1}^{T}\left(\mathcal{W}_{(1)}\mathbf{x}_t - \mathbf{y}_t\right)\mathbf{x}_t' = (\nabla Q([\![\mathcal{G}; \mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3]\!]))_{(1)}, \quad \text{(S4.65)}$$

where $\nabla Q([\![\mathcal{G}; \mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3]\!])$ is defined in (S3.55).

Lemma 8. *Suppose that $\mathbf{A}_i \in \mathbb{R}^{n_i \times r_i}$ with $\mathbf{A}_i \in \mathfrak{B}_i$, where $n_1 = n_2 = m, n_3 = p$. Consequently, $\nabla_{\mathcal{G}}\Psi(\mathcal{G}, \mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3, \mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3)$ is Lipschitz continuous with Lipschitz constant $L_1 := \frac{1}{T}\sum_{t=1}^{T}\|\mathbf{x}_t\|_2^2$, that is, there exists a constant $L_1 > 0$ such that for any $\mathcal{G}^1, \mathcal{G}^2 \in \mathbb{R}^{r_1 \times r_2 \times r_3}$, the following inequality holds:*

$$\left\|\nabla_{\mathcal{G}}\Psi(\mathcal{G}^1, \mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3, \mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3) - \nabla_{\mathcal{G}}\Psi(\mathcal{G}^2, \mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3, \mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3)\right\|_F$$
$$\leq L_1 \left\|\mathcal{G}^1 - \mathcal{G}^2\right\|_F.$$

Proof. Suppose that $\mathcal{G}^1, \mathcal{G}^2 \in \mathbb{R}^{r_1 \times r_2 \times r_3}$, $\mathbf{A}_i \in \mathfrak{B}_i, i = 1, 2, 3$, denote $\mathcal{W}^i = \mathcal{G}^i \times_1 \mathbf{A}_1 \times_2 \mathbf{A}_2 \times_3 \mathbf{A}_3, i = 1, 2$. Combining with (S3.55) and (S3.56) together,

we obtain

$$
\begin{aligned}
&\left\|\nabla_{\mathcal{G}} Q\left(\llbracket \mathcal{G}^1; \mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3 \rrbracket\right) - \nabla_{\mathcal{G}} Q\left(\llbracket \mathcal{G}^2; \mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3 \rrbracket\right)\right\|_F \\
=&\left\| \left(\frac{1}{T} \sum_{t=1}^{T} (\mathcal{W}^1_{(1)} \mathbf{x}_t - \mathbf{y}_t) \circ \mathbf{X}_t \right) \times \mathbf{A}_1' \times \mathbf{A}_2' \times \mathbf{A}_3' \right. \\
&\left. - \left(\frac{1}{T} \sum_{t=1}^{T} (\mathcal{W}^2_{(1)} \mathbf{x}_t - \mathbf{y}_t) \circ \mathbf{X}_t \right) \times \mathbf{A}_1' \times \mathbf{A}_2' \times \mathbf{A}_3' \right\|_F \\
=&\left\| \left(\frac{1}{T} \sum_{t=1}^{T} \left((\mathcal{W}^1_{(1)} \mathbf{x}_t - \mathbf{y}_t) \circ \mathbf{X}_t - (\mathcal{W}^2_{(1)} \mathbf{x}_t - \mathbf{y}_t) \circ \mathbf{X}_t \right) \right) \times \mathbf{A}_1' \times \mathbf{A}_2' \times \mathbf{A}_3' \right\|_F \\
=&\left\| \frac{1}{T} \sum_{t=1}^{T} (\mathcal{W}^1_{(1)} \mathbf{x}_t - \mathbf{y}_t) \circ \mathbf{X}_t - \frac{1}{T} \sum_{t=1}^{T} (\mathcal{W}^2_{(1)} \mathbf{x}_t - \mathbf{y}_t) \circ \mathbf{X}_t \right\|_F \\
=&\left\| \left(\frac{1}{T} \sum_{t=1}^{T} (\mathcal{W}^1_{(1)} \mathbf{x}_t - \mathbf{y}_t) \circ \mathbf{X}_t \right)_{(1)} - \left(\frac{1}{T} \sum_{t=1}^{T} (\mathcal{W}^2_{(1)} \mathbf{x}_t - \mathbf{y}_t) \circ \mathbf{X}_t \right)_{(1)} \right\|_F,
\end{aligned}
\tag{S4.66}
$$

where the third equality holds since the Frobenius norm is unitarily invariant. By (S4.65) and (S3.55), we have

$$
\begin{aligned}
&\left\| \left(\frac{1}{T} \sum_{t=1}^{T} (\mathcal{W}^1_{(1)} \mathbf{x}_t - \mathbf{y}_t) \circ \mathbf{X}_t \right)_{(1)} - \left(\frac{1}{T} \sum_{t=1}^{T} (\mathcal{W}^2_{(1)} \mathbf{x}_t - \mathbf{y}_t) \circ \mathbf{X}_t \right)_{(1)} \right\|_F \\
=&\left\| \frac{1}{T} \sum_{t=1}^{T} (\mathcal{W}^1_{(1)} \mathbf{x}_t - \mathbf{y}_t) \mathbf{x}_t' - \frac{1}{T} \sum_{t=1}^{T} (\mathcal{W}^2_{(1)} \mathbf{x}_t - \mathbf{y}_t) \mathbf{x}_t' \right\|_F \\
=&\left\| \frac{1}{T} \sum_{t=1}^{T} (\mathcal{W}^1_{(1)} - \mathcal{W}^2_{(1)}) \mathbf{x}_t \mathbf{x}_t' \right\|_F \\
\leq&\left\| \mathcal{W}^1_{(1)} - \mathcal{W}^2_{(1)} \right\|_F \left\| \frac{1}{T} \sum_{t=1}^{T} \mathbf{x}_t \mathbf{x}_t' \right\|_F \\
\leq&\left\| \mathcal{W}^1_{(1)} - \mathcal{W}^2_{(1)} \right\|_F \frac{1}{T} \sum_{t=1}^{T} \left\| \mathbf{x}_t \mathbf{x}_t' \right\|_F \\
=&\, L_1 \left\| \mathcal{W}^1_{(1)} - \mathcal{W}^2_{(1)} \right\|_F,
\end{aligned}
\tag{S4.67}
$$

where the first inequality follows from the Cauchy-Schwarz inequality and

the last equality follows from $L_1 := \frac{1}{T}\sum_{t=1}^{T}\|\mathbf{x}_t\|_2^2 = \frac{1}{T}\sum_{t=1}^{T}\|\mathbf{x}_t\mathbf{x}_t'\|_F$. Observe that

$$\begin{aligned}\|\mathcal{W}^1_{(1)} - \mathcal{W}^2_{(1)}\|_F &= \|\mathcal{W}^1 - \mathcal{W}^2\|_F \\ &= \|\mathcal{G}^1 \times_1 \mathbf{A}_1 \times_2 \mathbf{A}_2 \times_3 \mathbf{A}_3 - \mathcal{G}^2 \times_1 \mathbf{A}_1 \times_2 \mathbf{A}_2 \times_3 \mathbf{A}_3\|_F \\ &= \|(\mathcal{G}^1 - \mathcal{G}^2) \times_1 \mathbf{A}_1 \times_2 \mathbf{A}_2 \times_3 \mathbf{A}_3\|_F \\ &= \|\mathcal{G}^1 - \mathcal{G}^2\|_F.\end{aligned}$$

Combining the above results and (S4.67), (S4.66), we have

$$\|\nabla_{\mathcal{G}} Q([\![\mathcal{G}^1; \mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3]\!]) - \nabla_{\mathcal{G}} Q([\![\mathcal{G}^2; \mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3]\!])\|_F \le L_1\|\mathcal{G}^1 - \mathcal{G}^2\|_F.$$

From (S3.57), we further obtain

$$\begin{aligned}&\|\nabla_{\mathcal{G}}\Psi(\mathcal{G}^1, \mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3, \mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3) - \nabla_{\mathcal{G}}\Psi(\mathcal{G}^2, \mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3, \mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3)\|_F \\ &\le L_1\|\mathcal{G}^1 - \mathcal{G}^2\|_F.\end{aligned}$$

This completes the proof. □

Lemma 9. *Suppose that* $\mathcal{G} \in \mathbb{R}^{r_1\times r_2\times r_3}$ *with* $\mathcal{G} \in \mathcal{D}$, *and* $\mathbf{A}_i \in \mathbb{R}^{n_i\times r_i}, i = 2, 3$ *with* $\mathbf{A}_i \in \mathfrak{B}_i$, *where* $n_2 = m, n_3 = p$. *Then* $\nabla_{\mathbf{A}_1}\Psi(\mathcal{G}, \mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3, \mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3)$ *is Lipschitz continuous with Lipschitz constant* $L_2 := \nu^2 c_1 + \gamma_1$, *where* $\nu := \sqrt{r_1 r_2 r_3}c$ *and* $c_1 := \frac{1}{T}\sum_{t=1}^{T}\|\mathbf{x}_t\|_2^2$, *that is, there exists a constant* $L_2 > 0$ *such that for any* $\mathbf{A}_1^1, \mathbf{A}_1^2 \in \mathbb{R}^{m\times r_1}$, *the following inequality holds:*

$$\begin{aligned}&\|\nabla_{\mathbf{A}_1}\Psi(\mathcal{G}, \mathbf{A}_1^1, \mathbf{A}_2, \mathbf{A}_3, \mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3) - \nabla_{\mathbf{A}_1}\Psi(\mathcal{G}, \mathbf{A}_1^2, \mathbf{A}_2, \mathbf{A}_3, \mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3)\|_F \\ &\le L_2\|\mathbf{A}_1^1 - \mathbf{A}_1^2\|_F.\end{aligned}$$

Proof. Suppose that $\mathcal{G} \in \mathcal{D}$, $\mathbf{A}_i \in \mathfrak{B}_i, i = 2, 3$. By (S3.55) and (S3.56), for any $\mathbf{A}_1^1, \mathbf{A}_1^2 \in \mathbb{R}^{m \times r_1}$, we have

$$
\begin{aligned}
&\left\|\nabla_{\mathbf{A}_1} Q([\![\mathcal{G}; \mathbf{A}_1^1, \mathbf{A}_2, \mathbf{A}_3]\!]) - \nabla_{\mathbf{A}_1} Q([\![\mathcal{G}; \mathbf{A}_1^2, \mathbf{A}_2, \mathbf{A}_3]\!])\right\|_F \\
=&\left\| \left(\frac{1}{T} \sum_{t=1}^{T} \left(\mathbf{A}_1^1 \mathcal{G}_{(1)} \left(\mathbf{A}_3 \otimes \mathbf{A}_2 \right)' \mathbf{x}_t - \mathbf{y}_t \right) \circ \mathbf{X}_t \right)_{(1)} \left(\mathbf{A}_3 \otimes \mathbf{A}_2 \right) \mathcal{G}_{(1)}' \right. \\
&\left. - \left(\frac{1}{T} \sum_{t=1}^{T} \left(\mathbf{A}_1^2 \mathcal{G}_{(1)} \left(\mathbf{A}_3 \otimes \mathbf{A}_2 \right)' \mathbf{x}_t - \mathbf{y}_t \right) \circ \mathbf{X}_t \right)_{(1)} \left(\mathbf{A}_3 \otimes \mathbf{A}_2 \right) \mathcal{G}_{(1)}' \right\|_F \\
\leq&\left\| \left(\frac{1}{T} \sum_{t=1}^{T} \left(\mathbf{A}_1^1 \mathcal{G}_{(1)} \left(\mathbf{A}_3 \otimes \mathbf{A}_2 \right)' \mathbf{x}_t - \mathbf{y}_t \right) \circ \mathbf{X}_t \right)_{(1)} \right. \\
&\left. - \left(\frac{1}{T} \sum_{t=1}^{T} \left(\mathbf{A}_1^2 \mathcal{G}_{(1)} \left(\mathbf{A}_3 \otimes \mathbf{A}_2 \right)' \mathbf{x}_t - \mathbf{y}_t \right) \circ \mathbf{X}_t \right)_{(1)} \right\|_F \left\| \left(\mathbf{A}_3 \otimes \mathbf{A}_2 \right) \mathcal{G}_{(1)}' \right\|_F \\
=&\left\| \frac{1}{T} \sum_{t=1}^{T} \left(\mathbf{A}_1^1 \mathcal{G}_{(1)} \left(\mathbf{A}_3 \otimes \mathbf{A}_2 \right)' \mathbf{x}_t - \mathbf{y}_t \right) \circ \mathbf{X}_t \right. \\
&\left. - \frac{1}{T} \sum_{t=1}^{T} \left(\mathbf{A}_1^2 \mathcal{G}_{(1)} \left(\mathbf{A}_3 \otimes \mathbf{A}_2 \right)' \mathbf{x}_t - \mathbf{y}_t \right) \circ \mathbf{X}_t \right\|_F \left\| \left(\mathbf{A}_3 \otimes \mathbf{A}_2 \right) \mathcal{G}_{(1)}' \right\|_F,
\end{aligned}
\tag{S4.68}
$$

where the first inequality follows from the Cauchy-Schwarz inequality. Proposition 2 together with Lemma 2.1 of Chen et al. (2021) gives

$$
\begin{aligned}
\left\| \left(\mathbf{A}_3 \otimes \mathbf{A}_2 \right) \mathcal{G}_{(1)}' \right\|_F &\leq \left\| \mathbf{A}_3 \otimes \mathbf{A}_2 \right\| \left\| \mathcal{G}_{(1)}' \right\|_F \\
&\leq \left\| \mathcal{G}_{(1)}' \right\|_F = \left\| \mathcal{G}_{(1)} \right\|_F = \left\| \mathcal{G} \right\|_F \leq \nu,
\end{aligned}
\tag{S4.69}
$$

where the last inequality follows from $\|\mathcal{G}\|_F \leq \sqrt{r_1 r_2 r_3} \|\mathcal{G}\|_\infty \leq \sqrt{r_1 r_2 r_3} c :=$

ν for any $\mathcal{G} \in \mathcal{D}$. In addition,

$$
\begin{aligned}
&\Bigg\| \frac{1}{T} \sum_{t=1}^{T} \left(\mathbf{A}_1^1 \mathcal{G}_{(1)} \left(\mathbf{A}_3 \otimes \mathbf{A}_2 \right)' \mathbf{x}_t - \mathbf{y}_t \right) \circ \mathbf{X}_t \\
&\quad - \frac{1}{T} \sum_{t=1}^{T} \left(\mathbf{A}_1^2 \mathcal{G}_{(1)} \left(\mathbf{A}_3 \otimes \mathbf{A}_2 \right)' \mathbf{x}_t - \mathbf{y}_t \right) \circ \mathbf{X}_t \Bigg\|_F \\
= &\left\| \frac{1}{T} \sum_{t=1}^{T} \left(\mathbf{A}_1^1 \mathcal{G}_{(1)} \left(\mathbf{A}_3 \otimes \mathbf{A}_2 \right)' \mathbf{x}_t - \mathbf{A}_1^2 \mathcal{G}_{(1)} \left(\mathbf{A}_3 \otimes \mathbf{A}_2 \right)' \mathbf{x}_t \right) \circ \mathbf{X}_t \right\|_F \\
= &\left\| \frac{1}{T} \sum_{t=1}^{T} \left(\left(\mathbf{A}_1^1 - \mathbf{A}_1^2 \right) \mathcal{G}_{(1)} \left(\mathbf{A}_3 \otimes \mathbf{A}_2 \right)' \mathbf{x}_t \right) \circ \mathbf{X}_t \right\|_F \\
\leq &\frac{1}{T} \sum_{t=1}^{T} \left\| \left(\left(\mathbf{A}_1^1 - \mathbf{A}_1^2 \right) \mathcal{G}_{(1)} \left(\mathbf{A}_3 \otimes \mathbf{A}_2 \right)' \mathbf{x}_t \right) \circ \mathbf{X}_t \right\|_F \\
= &\frac{1}{T} \sum_{t=1}^{T} \left\| \left(\mathbf{A}_1^1 - \mathbf{A}_1^2 \right) \mathcal{G}_{(1)} \left(\mathbf{A}_3 \otimes \mathbf{A}_2 \right)' \mathbf{x}_t \right\|_2 \left\| \mathbf{X}_t \right\|_F \\
\leq &\frac{1}{T} \sum_{t=1}^{T} \left\| \mathbf{A}_1^1 - \mathbf{A}_1^2 \right\|_F \left\| \mathcal{G}_{(1)} \left(\mathbf{A}_3 \otimes \mathbf{A}_2 \right)' \right\|_F \left\| \mathbf{x}_t \right\|_2 \left\| \mathbf{X}_t \right\|_F \\
\leq &\frac{1}{T} \sum_{t=1}^{T} \nu \left\| \mathbf{A}_1^1 - \mathbf{A}_1^2 \right\|_F \left\| \mathbf{x}_t \right\|_2 \left\| \mathbf{X}_t \right\|_F,
\end{aligned}
\tag{S4.70}
$$

where the last equality holds by Proposition 4, the second inequality holds by the Cauchy-Schwarz inequality and the last inequality holds by (S4.69) and the fact that $\left\| \mathcal{G}_{(1)} \left(\mathbf{A}_3 \otimes \mathbf{A}_2 \right)' \right\|_F = \left\| \left(\mathbf{A}_3 \otimes \mathbf{A}_2 \right) \mathcal{G}_{(1)}' \right\|_F$. From the definitions of $\mathbf{x}_t$ and $\mathbf{X}_t$, it is easily seen that

$$
\left\| \mathbf{x}_t \right\|_2 = \left\| \mathbf{X}_t \right\|_F . \tag{S4.71}
$$

Combining (S4.68), (S4.69) and (S4.70), we can easily obtain that

$$\begin{aligned}
&\|\nabla_{\mathbf{A}_1} Q(\llbracket \mathcal{G}; \mathbf{A}_1^1, \mathbf{A}_2, \mathbf{A}_3 \rrbracket) - \nabla_{\mathbf{A}_1} Q(\llbracket \mathcal{G}; \mathbf{A}_1^2, \mathbf{A}_2, \mathbf{A}_3 \rrbracket)\|_F \\
&\leq \nu^2 \frac{1}{T} \sum_{t=1}^{T} \|\mathbf{x}_t\|_2^2 \, \|\mathbf{A}_1^1 - \mathbf{A}_1^2\|_F \\
&= \nu^2 c_1 \|\mathbf{A}_1^1 - \mathbf{A}_1^2\|_F,
\end{aligned}$$

where $c_1 := \frac{1}{T} \sum_{t=1}^{T} \|\mathbf{x}_t\|_2^2$. This taken collectively with (S3.57) yields

$$\begin{aligned}
&\|\nabla_{\mathbf{A}_1} \Psi(\mathcal{G}, \mathbf{A}_1^1, \mathbf{A}_2, \mathbf{A}_3, \mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3) - \nabla_{\mathbf{A}_1} \Psi(\mathcal{G}, \mathbf{A}_1^2, \mathbf{A}_2, \mathbf{A}_3, \mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3)\|_F \\
&= \|\nabla_{\mathbf{A}_1} Q(\llbracket \mathcal{G}; \mathbf{A}_1^1, \mathbf{A}_2, \mathbf{A}_3 \rrbracket) - \nabla_{\mathbf{A}_1} Q(\llbracket \mathcal{G}; \mathbf{A}_1^2, \mathbf{A}_2, \mathbf{A}_3 \rrbracket) + \gamma_1 (\mathbf{A}_1^1 - \mathbf{A}_1^2)\|_F \\
&\leq \|\nabla_{\mathbf{A}_1} Q(\llbracket \mathcal{G}; \mathbf{A}_1^1, \mathbf{A}_2, \mathbf{A}_3 \rrbracket) - \nabla_{\mathbf{A}_1} Q(\llbracket \mathcal{G}; \mathbf{A}_1^2, \mathbf{A}_2, \mathbf{A}_3 \rrbracket)\|_F + \gamma_1 \|\mathbf{A}_1^1 - \mathbf{A}_1^2\|_F \\
&\leq \nu^2 c_1 \|\mathbf{A}_1^1 - \mathbf{A}_1^2\|_F + \gamma_1 \|\mathbf{A}_1^1 - \mathbf{A}_1^2\|_F \\
&= (\nu^2 c_1 + \gamma_1) \|\mathbf{A}_1^1 - \mathbf{A}_1^2\|_F = L_2 \|\mathbf{A}_1^1 - \mathbf{A}_1^2\|_F,
\end{aligned} \tag{S4.72}$$

where $L_2 := \nu^2 c_1 + \gamma_1$. This completes the proof. □

Lemma 10. *Suppose that* $\mathcal{G} \in \mathbb{R}^{r_1 \times r_2 \times r_3}$ *with* $\mathcal{G} \in \mathcal{D}$, $\mathbf{A}_i \in \mathbb{R}^{n_i \times r_i}$ *with* $\mathbf{A}_i \in \mathfrak{B}_i, i = 1, 3$, *where* $n_1 = m, n_3 = p$. *Then* $\nabla_{\mathbf{A}_2} \Psi(\mathcal{G}, \mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3, \mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3)$ *is Lipschitz continuous with Lipschitz constant* $L_3 := \nu^2 c_1 + \gamma_2$, *where* $\nu := \sqrt{r_1 r_2 r_3} c$ *and* $c_1 := \frac{1}{T} \sum_{t=1}^{T} \|\mathbf{x}_t\|_2^2$, *that is, there exists a constant* $L_3 > 0$ *such that for any* $\mathbf{A}_2^1, \mathbf{A}_2^2 \in \mathbb{R}^{m \times r_2}$, *the following inequality holds:*

$$\begin{aligned}
&\left\|\nabla_{\mathbf{A}_2} \Psi(\mathcal{G}, \mathbf{A}_1, \mathbf{A}_2^1, \mathbf{A}_3, \mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3) - \nabla_{\mathbf{A}_2} \Psi(\mathcal{G}, \mathbf{A}_1, \mathbf{A}_2^2, \mathbf{A}_3, \mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3)\right\|_F \\
&\leq L_3 \left\|\mathbf{A}_2^1 - \mathbf{A}_2^2\right\|_F.
\end{aligned}$$

Proof. Similar to the analysis of $\nabla_{\mathbf{A}_1} Q$, we have

$$
\begin{aligned}
&\left\|\nabla_{\mathbf{A}_2} Q([\![\mathcal{G};\mathbf{A}_1,\mathbf{A}_2^1,\mathbf{A}_3]\!]) - \nabla_{\mathbf{A}_2} Q([\![\mathcal{G};\mathbf{A}_1,\mathbf{A}_2^2,\mathbf{A}_3]\!])\right\|_F \\
=& \left\| \left(\frac{1}{T}\sum_{t=1}^{T} \left(\mathbf{A}_1 \mathcal{G}_{(1)} \left(\mathbf{A}_3 \otimes \mathbf{A}_2^1\right)' \mathbf{x}_t - \mathbf{y}_t \right) \circ \mathbf{X}_t \right)_{(2)} \left(\mathbf{A}_1 \otimes \mathbf{A}_3\right) \mathcal{G}_{(2)}' \right. \\
&\left. - \left(\frac{1}{T}\sum_{t=1}^{T} \left(\mathbf{A}_1 \mathcal{G}_{(1)} \left(\mathbf{A}_3 \otimes \mathbf{A}_2^2\right)' \mathbf{x}_t - \mathbf{y}_t \right) \circ \mathbf{X}_t \right)_{(2)} \left(\mathbf{A}_1 \otimes \mathbf{A}_3\right) \mathcal{G}_{(2)}' \right\|_F \\
\leq& \left\| \left(\frac{1}{T}\sum_{t=1}^{T} \left(\mathbf{A}_1 \mathcal{G}_{(1)} \left(\mathbf{A}_3 \otimes \mathbf{A}_2^1\right)' \mathbf{x}_t - \mathbf{y}_t \right) \circ \mathbf{X}_t \right)_{(2)} \right. \\
&\left. - \left(\frac{1}{T}\sum_{t=1}^{T} \left(\mathbf{A}_1 \mathcal{G}_{(1)} \left(\mathbf{A}_3 \otimes \mathbf{A}_2^2\right)' \mathbf{x}_t - \mathbf{y}_t \right) \circ \mathbf{X}_t \right)_{(2)} \right\|_F \nu \\
=& \left\| \frac{1}{T}\sum_{t=1}^{T} \mathbf{X}_t \otimes \left(\mathbf{A}_1 \mathcal{G}_{(1)} \left(\mathbf{A}_3 \otimes \mathbf{A}_2^1\right)' \mathbf{x}_t - \mathbf{y}_t \right)' \right. \\
&\left. - \frac{1}{T}\sum_{t=1}^{T} \mathbf{X}_t \otimes \left(\mathbf{A}_1 \mathcal{G}_{(1)} \left(\mathbf{A}_3 \otimes \mathbf{A}_2^2\right)' \mathbf{x}_t - \mathbf{y}_t \right)' \right\|_F \nu,
\end{aligned}
\tag{S4.73}
$$

where $\nu := \sqrt{r_1 r_2 r_3}c$ and the last equality follows from Proposition 3. Note that

$$
\begin{aligned}
&\left\| \frac{1}{T}\sum_{t=1}^{T} \mathbf{X}_t \otimes \left(\mathbf{A}_1 \mathcal{G}_{(1)} \left(\mathbf{A}_3 \otimes \mathbf{A}_2^1\right)' \mathbf{x}_t - \mathbf{y}_t \right)' \right. \\
&\left. - \frac{1}{T}\sum_{t=1}^{T} \mathbf{X}_t \otimes \left(\mathbf{A}_1 \mathcal{G}_{(1)} \left(\mathbf{A}_3 \otimes \mathbf{A}_2^2\right)' \mathbf{x}_t - \mathbf{y}_t \right)' \right\|_F \\
=& \left\| \frac{1}{T}\sum_{t=1}^{T} \mathbf{X}_t \otimes \left(\mathbf{A}_1 \mathcal{G}_{(1)} \left(\mathbf{A}_3 \otimes \mathbf{A}_2^1\right)' \mathbf{x}_t - \mathbf{A}_1 \mathcal{G}_{(1)} \left(\mathbf{A}_3 \otimes \mathbf{A}_2^2\right)' \mathbf{x}_t \right)' \right\|_F \\
=& \left\| \frac{1}{T}\sum_{t=1}^{T} \mathbf{X}_t \otimes \left(\mathbf{A}_1 \mathcal{G}_{(1)} \left(\mathbf{A}_3 \otimes (\mathbf{A}_2^1 - \mathbf{A}_2^2)\right)' \mathbf{x}_t \right)' \right\|_F \\
\leq& \frac{1}{T}\sum_{t=1}^{T} \left\| \mathbf{X}_t \otimes \left(\mathbf{A}_1 \mathcal{G}_{(1)} \left(\mathbf{A}_3 \otimes (\mathbf{A}_2^1 - \mathbf{A}_2^2)\right)' \mathbf{x}_t \right)' \right\|_F \\
=& \frac{1}{T}\sum_{t=1}^{T} \|\mathbf{X}_t\|_F \left\| \mathbf{A}_1 \mathcal{G}_{(1)} \left(\mathbf{A}_3 \otimes (\mathbf{A}_2^1 - \mathbf{A}_2^2)\right)' \mathbf{x}_t \right\|_2,
\end{aligned}
\tag{S4.74}
$$

where the third equality holds by Proposition 1. It follows from Example 5.6.6 of Horn and Johnson (2013) that

$$\begin{aligned}
&\left\|\mathbf{A}_1\mathcal{G}_{(1)}\left(\mathbf{A}_3\otimes(\mathbf{A}_2^1-\mathbf{A}_2^2)\right)'\mathbf{x}_t\right\|_2\\
\leq\ &\|\mathbf{A}_1\|\left\|\mathcal{G}_{(1)}\left(\mathbf{A}_3\otimes(\mathbf{A}_2^1-\mathbf{A}_2^2)\right)'\mathbf{x}_t\right\|_2\\
\leq\ &\left\|\mathcal{G}_{(1)}\left(\mathbf{A}_3\otimes(\mathbf{A}_2^1-\mathbf{A}_2^2)\right)'\mathbf{x}_t\right\|_2\\
\leq\ &\left\|\mathcal{G}_{(1)}\left(\mathbf{A}_3\otimes(\mathbf{A}_2^1-\mathbf{A}_2^2)\right)'\right\|_F\|\mathbf{x}_t\|_2\\
\leq\ &\left\|\mathcal{G}_{(1)}\right\|_F\left\|\left(\mathbf{A}_3\otimes(\mathbf{A}_2^1-\mathbf{A}_2^2)\right)'\right\|\|\mathbf{x}_t\|_2\\
=\ &\|\mathcal{G}\|_F\|\mathbf{A}_3\|\left\|(\mathbf{A}_2^1-\mathbf{A}_2^2)'\right\|\|\mathbf{x}_t\|_2\\
\leq\ &\nu\|\mathbf{A}_3\|\left\|(\mathbf{A}_2^1-\mathbf{A}_2^2)'\right\|\|\mathbf{x}_t\|_2\leq\ \nu\left\|(\mathbf{A}_2^1-\mathbf{A}_2^2)'\right\|_F\|\mathbf{x}_t\|_2,
\end{aligned}\tag{S4.75}$$

where the second inequality holds by $\mathbf{A}_1\in\mathfrak{B}_1$ with $\|\mathbf{A}_1\|\leq 1$, the third inequality holds by the Cauchy-Schwarz inequality, the fourth inequality holds by Lemma 2.1 of Chen et al. (2021), the fifth inequality holds by $\mathcal{G}\in\mathcal{D}$ with $\|\mathcal{G}\|_F\leq\nu$ and the last inequality holds by $\|\mathbf{A}_3\|\leq 1$ and any matrix $\mathbf{X}$ satisfies $\|\mathbf{X}\|\leq\|\mathbf{X}\|_F$. Combining (S4.75) and (S4.74), we have

$$\begin{aligned}
&\left\|\frac{1}{T}\sum_{t=1}^{T}\mathbf{X}_t\otimes\left(\mathbf{A}_1\mathcal{G}_{(1)}\left(\mathbf{A}_3\otimes\mathbf{A}_2^1\right)'\mathbf{x}_t-\mathbf{y}_t\right)'\right.\\
&\left.-\frac{1}{T}\sum_{t=1}^{T}\mathbf{X}_t\otimes\left(\mathbf{A}_1\mathcal{G}_{(1)}\left(\mathbf{A}_3\otimes\mathbf{A}_2^2\right)'\mathbf{x}_t-\mathbf{y}_t\right)'\right\|_F\\
\leq\ &\frac{1}{T}\sum_{t=1}^{T}\|\mathbf{X}_t\|_F\nu\left\|(\mathbf{A}_2^1-\mathbf{A}_2^2)'\right\|_F\|\mathbf{x}_t\|_2\\
=\ &\frac{1}{T}\sum_{t=1}^{T}\|\mathbf{x}_t\|_2^2\nu\left\|\mathbf{A}_2^1-\mathbf{A}_2^2\right\|_F=\ \nu c_1\left\|\mathbf{A}_2^1-\mathbf{A}_2^2\right\|_F,
\end{aligned}\tag{S4.76}$$

where the first equality follows from (S4.71) and $c_1 := \frac{1}{T}\sum_{t=1}^{T}\|\mathbf{x}_t\|_2^2$. Then together with (S4.73), one can obtain that

$$\|\nabla_{\mathbf{A}_2} Q([\![\mathcal{G}; \mathbf{A}_1, \mathbf{A}_2^1, \mathbf{A}_3]\!]) - \nabla_{\mathbf{A}_2} Q([\![\mathcal{G}; \mathbf{A}_1, \mathbf{A}_2^2, \mathbf{A}_3]\!])\|_F \leq \nu^2 c_1 \|\mathbf{A}_2^1 - \mathbf{A}_2^2\|_F.$$

Similar to (S4.72), we have

$$\|\nabla_{\mathbf{A}_2} \Psi(\mathcal{G}, \mathbf{A}_1, \mathbf{A}_2^1, \mathbf{A}_3, \mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3) - \nabla_{\mathbf{A}_2} \Psi(\mathcal{G}, \mathbf{A}_1, \mathbf{A}_2^2, \mathbf{A}_3, \mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3)\|_F$$
$$\leq L_3 \|\mathbf{A}_2^1 - \mathbf{A}_2^2\|_F,$$

where $L_3 := \nu^2 c_1 + \gamma_2$. The proof is completed. □

Lemma 11. *Suppose that* $\mathcal{G} \in \mathbb{R}^{r_1 \times r_2 \times r_3}$ *with* $\mathcal{G} \in \mathcal{D}$, $\mathbf{A}_i \in \mathbb{R}^{n_i \times r_i}$ *with* $\mathbf{A}_i \in \mathfrak{B}_i, i = 1, 2$, *where* $n_1 = n_2 = m$. *Then* $\nabla_{\mathbf{A}_3} \Psi(\mathcal{G}, \mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3, \mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3)$ *is Lipschitz continuous with Lipschitz constant* $L_4 := \nu^2 c_1 + \gamma_3$, *where* $\nu := \sqrt{r_1 r_2 r_3} c$ *and* $c_1 := \frac{1}{T}\sum_{t=1}^{T}\|\mathbf{x}_t\|_2^2$, *that is, there exists a constant* $L_4 > 0$ *such that for any* $\mathbf{A}_3^1, \mathbf{A}_3^2 \in \mathbb{R}^{p \times r_3}$, *the following inequality holds:*

$$\|\nabla_{\mathbf{A}_3} \Psi(\mathcal{G}, \mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3^1, \mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3) - \nabla_{\mathbf{A}_3} \Psi(\mathcal{G}, \mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3^2, \mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3)\|_F$$
$$\leq L_4 \|\mathbf{A}_3^1 - \mathbf{A}_3^2\|_F.$$

Proof. Similar to the analysis of $\nabla_{\mathbf{A}_1} Q$, by (S3.56), we have

$$
\begin{aligned}
&\left\|\nabla_{\mathbf{A}_3} Q(\llbracket \mathcal{G}; \mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3^1 \rrbracket) - \nabla_{\mathbf{A}_3} Q(\llbracket \mathcal{G}; \mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3^2 \rrbracket)\right\|_F \\
=&\left\| \left(\frac{1}{T} \sum_{t=1}^{T} \left(\mathbf{A}_1 \mathcal{G}_{(1)} \left(\mathbf{A}_3^1 \otimes \mathbf{A}_2 \right)' \mathbf{x}_t - \mathbf{y}_t \right) \circ \mathbf{X}_t \right)_{(3)} (\mathbf{A}_2 \otimes \mathbf{A}_1) \mathcal{G}_{(3)}' \right. \\
&\left. - \left(\frac{1}{T} \sum_{t=1}^{T} \left(\mathbf{A}_1 \mathcal{G}_{(1)} \left(\mathbf{A}_3^2 \otimes \mathbf{A}_2 \right)' \mathbf{x}_t - \mathbf{y}_t \right) \circ \mathbf{X}_t \right)_{(3)} (\mathbf{A}_2 \otimes \mathbf{A}_1) \mathcal{G}_{(3)}' \right\|_F \\
\leq&\left\| \left(\frac{1}{T} \sum_{t=1}^{T} \left(\mathbf{A}_1 \mathcal{G}_{(1)} \left(\mathbf{A}_3^1 \otimes \mathbf{A}_2 \right)' \mathbf{x}_t - \mathbf{y}_t \right) \circ \mathbf{X}_t \right)_{(3)} \right. \\
&\left. - \left(\frac{1}{T} \sum_{t=1}^{T} \left(\mathbf{A}_1 \mathcal{G}_{(1)} \left(\mathbf{A}_3^2 \otimes \mathbf{A}_2 \right)' \mathbf{x}_t - \mathbf{y}_t \right) \circ \mathbf{X}_t \right)_{(3)} \right\|_F \nu \\
=&\left\| \frac{1}{T} \sum_{t=1}^{T} \mathbf{X}_t' \otimes \left(\mathbf{A}_1 \mathcal{G}_{(1)} \left(\mathbf{A}_3^1 \otimes \mathbf{A}_2 \right)' \mathbf{x}_t - \mathbf{y}_t \right)' \right. \\
&\left. - \frac{1}{T} \sum_{t=1}^{T} \mathbf{X}_t' \otimes \left(\mathbf{A}_1 \mathcal{G}_{(1)} \left(\mathbf{A}_3^2 \otimes \mathbf{A}_2 \right)' \mathbf{x}_t - \mathbf{y}_t \right)' \right\|_F \nu
\end{aligned}
\tag{S4.77}
$$

where $\nu := \sqrt{r_1 r_2 r_3} c$ and the last equality follows from Proposition 3. Similar to (S4.76), we have

$$
\begin{aligned}
&\left\| \frac{1}{T} \sum_{t=1}^{T} \mathbf{X}_t' \otimes \left(\mathbf{A}_1 \mathcal{G}_{(1)} \left(\mathbf{A}_3^1 \otimes \mathbf{A}_2 \right)' \mathbf{x}_t - \mathbf{y}_t \right)' \right. \\
&\left. - \frac{1}{T} \sum_{t=1}^{T} \mathbf{X}_t' \otimes \left(\mathbf{A}_1 \mathcal{G}_{(1)} \left(\mathbf{A}_3^2 \otimes \mathbf{A}_2 \right)' \mathbf{x}_t - \mathbf{y}_t \right)' \right\|_F \\
\leq& \frac{1}{T} \sum_{t=1}^{T} \|\mathbf{X}_t\|_F \left\| \left(\mathbf{A}_1 \mathcal{G}_{(1)} \left((\mathbf{A}_3^1 - \mathbf{A}_3^2) \otimes \mathbf{A}_2 \right)' \mathbf{x}_t \right)' \right\|_F \\
\leq& \nu c_1 \|\mathbf{A}_3^1 - \mathbf{A}_3^2\|_F,
\end{aligned}
\tag{S4.78}
$$

where $c_1 := \frac{1}{T} \sum_{t=1}^{T} \|\mathbf{x}_t\|_2^2$. Combining (S4.77) and (S4.78), we get that

$$
\|\nabla_{\mathbf{A}_3} Q(\llbracket \mathcal{G}; \mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3^1 \rrbracket) - \nabla_{\mathbf{A}_3} Q(\llbracket \mathcal{G}; \mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3^2 \rrbracket)\|_F \leq \nu^2 c_1 \|\mathbf{A}_3^1 - \mathbf{A}_3^2\|_F.
$$

Furthermore, similar to (S4.72), we have

$$\begin{aligned}&\|\nabla_{\mathbf{A}_3}\Psi(\mathcal{G},\mathbf{A}_1,\mathbf{A}_2,\mathbf{A}_3^1,\mathbf{U}_1,\mathbf{U}_2,\mathbf{U}_3)-\nabla_{\mathbf{A}_3}\Psi(\mathcal{G},\mathbf{A}_1,\mathbf{A}_2,\mathbf{A}_3^2,\mathbf{U}_1,\mathbf{U}_2,\mathbf{U}_3)\|_F\\ &\leq L_4\left\|\mathbf{A}_3^1-\mathbf{A}_3^2\right\|_F,\end{aligned}$$

where $L_4 := \nu^2 c_1 + \gamma_3$. This completes the proof. □

It can be easily verified from (S3.57) that

$$\begin{aligned}&\|\nabla_{\mathbf{U}_1}\Psi(\mathcal{G},\mathbf{A}_1,\mathbf{A}_2,\mathbf{A}_3,\mathbf{U}_1^1,\mathbf{U}_2,\mathbf{U}_3)-\nabla_{\mathbf{U}_1}\Psi(\mathcal{G},\mathbf{A}_1,\mathbf{A}_2,\mathbf{A}_3,\mathbf{U}_1^2,\mathbf{U}_2,\mathbf{U}_3)\|_F\\ &\leq \gamma_1\|\mathbf{U}_1^1-\mathbf{U}_1^2\|_F,\\ &\|\nabla_{\mathbf{U}_2}\Psi(\mathcal{G},\mathbf{A}_1,\mathbf{A}_2,\mathbf{A}_3,\mathbf{U}_1,\mathbf{U}_2^1,\mathbf{U}_3)-\nabla_{\mathbf{U}_2}\Psi(\mathcal{G},\mathbf{A}_1,\mathbf{A}_2,\mathbf{A}_3,\mathbf{U}_1,\mathbf{U}_2^2,\mathbf{U}_3)\|_F\\ &\leq \gamma_2\|\mathbf{U}_2^1-\mathbf{U}_2^2\|_F,\\ &\|\nabla_{\mathbf{U}_3}\Psi(\mathcal{G},\mathbf{A}_1,\mathbf{A}_2,\mathbf{A}_3,\mathbf{U}_1,\mathbf{U}_2,\mathbf{U}_3^1)-\nabla_{\mathbf{U}_3}\Psi(\mathcal{G},\mathbf{A}_1,\mathbf{A}_2,\mathbf{A}_3,\mathbf{U}_1,\mathbf{U}_2,\mathbf{U}_3^2)\|_F\\ &\leq \gamma_3\|\mathbf{U}_3^1-\mathbf{U}_3^2\|_F.\end{aligned} \tag{S4.79}$$

Lemma 12. *(Sufficient decrease) Let the sequence* $\{(\mathcal{G}^k,\mathbf{A}_1^k,\mathbf{A}_2^k,\mathbf{A}_3^k,\mathbf{U}_1^k,\mathbf{U}_2^k,\mathbf{U}_3^k)\}$ *be generated by Algorithm 1. Suppose that* $\rho_1 > L_1, \rho_2 > L_2, \rho_3 > L_3, \rho_4 > L_4, \rho_5 > \gamma_1, \rho_6 > \gamma_2, \rho_7 > \gamma_3$. *Then there exists a constant* $\bar{\rho} > 0$

such that the following inequality holds

$$F\left(\mathcal{G}^{k+1},\mathbf{A}_1^{k+1},\mathbf{A}_2^{k+1},\mathbf{A}_3^{k+1},\mathbf{U}_1^{k+1},\mathbf{U}_2^{k+1},\mathbf{U}_3^{k+1}\right)+\frac{\bar{\rho}}{2}\left(\|\mathcal{G}^{k+1}-\mathcal{G}^{k}\|_F^2\right.$$
$$+\|\mathbf{A}_1^{k+1}-\mathbf{A}_1^{k}\|_F^2+\|\mathbf{A}_2^{k+1}-\mathbf{A}_2^{k}\|_F^2+\|\mathbf{A}_3^{k+1}-\mathbf{A}_3^{k}\|_F^2+\|\mathbf{U}_1^{k+1}-\mathbf{U}_1^{k}\|_F^2$$
$$\left.+\|\mathbf{U}_2^{k+1}-\mathbf{U}_2^{k}\|_F^2+\|\mathbf{U}_3^{k+1}-\mathbf{U}_3^{k}\|_F^2\right)\leq\ F\left(\mathcal{G}^{k},\mathbf{A}_1^{k},\mathbf{A}_2^{k},\mathbf{A}_3^{k},\mathbf{U}_1^{k},\mathbf{U}_2^{k},\mathbf{U}_3^{k}\right), \tag{S4.80}$$

where F is defined in (S3.52) and $\bar{\rho}:=\min\{\rho_1-L_1,\rho_2-L_2,\rho_3-L_3,\rho_4-L_4,\rho_5-\gamma_1,\rho_6-\gamma_2,\rho_7-\gamma_3\}$.

Proof. It can be seen from (3.13) that

$$\beta\|\mathcal{G}^{k+1}\|_1+\delta_{\mathcal{D}}(\mathcal{G}^{k+1})+\left\langle\nabla_{\mathcal{G}}\Psi(\mathcal{G}^{k},\mathbf{A}_1^{k},\mathbf{A}_2^{k},\mathbf{A}_3^{k},\mathbf{U}_1^{k},\mathbf{U}_2^{k},\mathbf{U}_3^{k}),\mathcal{G}^{k+1}-\mathcal{G}^{k}\right\rangle$$
$$+\frac{\rho_1}{2}\|\mathcal{G}^{k+1}-\mathcal{G}^{k}\|_F^2\leq\beta\|\mathcal{G}^{k}\|_1+\delta_{\mathcal{D}}(\mathcal{G}^{k}). \tag{S4.81}$$

In addition, applying Lemma 8 and Lemma 1 of Bolte et al. (2014), we have

$$\Psi\left(\mathcal{G}^{k+1},\mathbf{A}_1^{k},\mathbf{A}_2^{k},\mathbf{A}_3^{k},\mathbf{U}_1^{k},\mathbf{U}_2^{k},\mathbf{U}_3^{k}\right)\leq\Psi\left(\mathcal{G}^{k},\mathbf{A}_1^{k},\mathbf{A}_2^{k},\mathbf{A}_3^{k},\mathbf{U}_1^{k},\mathbf{U}_2^{k},\mathbf{U}_3^{k}\right)$$
$$+\frac{L_1}{2}\|\mathcal{G}^{k+1}-\mathcal{G}^{k}\|_F^2+\left\langle\nabla_{\mathcal{G}}\Psi(\mathcal{G}^{k},\mathbf{A}_1^{k},\mathbf{A}_2^{k},\mathbf{A}_3^{k},\mathbf{U}_1^{k},\mathbf{U}_2^{k},\mathbf{U}_3^{k}),\mathcal{G}^{k+1}-\mathcal{G}^{k}\right\rangle. \tag{S4.82}$$

Combining (S4.81) and (S4.82), we obtain

$$\Psi\left(\mathcal{G}^{k+1},\mathbf{A}_1^{k},\mathbf{A}_2^{k},\mathbf{A}_3^{k},\mathbf{U}_1^{k},\mathbf{U}_2^{k},\mathbf{U}_3^{k}\right)+\beta\|\mathcal{G}^{k+1}\|_1+\delta_{\mathcal{D}}(\mathcal{G}^{k+1})$$
$$+\left\langle\nabla_{\mathcal{G}}\Psi(\mathcal{G}^{k},\mathbf{A}_1^{k},\mathbf{A}_2^{k},\mathbf{A}_3^{k},\mathbf{U}_1^{k},\mathbf{U}_2^{k},\mathbf{U}_3^{k}),\mathcal{G}^{k+1}-\mathcal{G}^{k}\right\rangle+\frac{\rho_1}{2}\|\mathcal{G}^{k+1}-\mathcal{G}^{k}\|_F^2$$
$$\leq\Psi\left(\mathcal{G}^{k},\mathbf{A}_1^{k},\mathbf{A}_2^{k},\mathbf{A}_3^{k},\mathbf{U}_1^{k},\mathbf{U}_2^{k},\mathbf{U}_3^{k}\right)++\frac{L_1}{2}\|\mathcal{G}^{k+1}-\mathcal{G}^{k}\|_F^2+\beta\|\mathcal{G}^{k}\|_1+\delta_{\mathcal{D}}(\mathcal{G}^{k})$$
$$+\left\langle\nabla_{\mathcal{G}}\Psi(\mathcal{G}^{k},\mathbf{A}_1^{k},\mathbf{A}_2^{k},\mathbf{A}_3^{k},\mathbf{U}_1^{k},\mathbf{U}_2^{k},\mathbf{U}_3^{k}),\mathcal{G}^{k+1}-\mathcal{G}^{k}\right\rangle,$$

which yields

$$\Psi\left(\mathcal{G}^{k+1},\mathbf{A}_1^k,\mathbf{A}_2^k,\mathbf{A}_3^k,\mathbf{U}_1^k,\mathbf{U}_2^k,\mathbf{U}_3^k\right)+\beta\|\mathcal{G}^{k+1}\|_1+\delta_{\mathcal{D}}(\mathcal{G}^{k+1})+\frac{\rho_1-L_1}{2}\|\mathcal{G}^{k+1}-\mathcal{G}^k\|_F^2$$
$$\leq\Psi\left(\mathcal{G}^{k},\mathbf{A}_1^k,\mathbf{A}_2^k,\mathbf{A}_3^k,\mathbf{U}_1^k,\mathbf{U}_2^k,\mathbf{U}_3^k\right)+\beta\|\mathcal{G}^{k}\|_1+\delta_{\mathcal{D}}(\mathcal{G}^{k}). \tag{S4.83}$$

By repeating the similar ideas and techniques, and combining Lemma 9-11 with (S4.79) and (3.13), we can obtain

$$\begin{aligned}
&\delta_{\mathfrak{B}_1}(\mathbf{A}_1^{k+1})+\Psi\left(\mathcal{G}^{k+1},\mathbf{A}_1^{k+1},\mathbf{A}_2^k,\mathbf{A}_3^k,\mathbf{U}_1^k,\mathbf{U}_2^k,\mathbf{U}_3^k\right)+\frac{\rho_2-L_2}{2}\|\mathbf{A}_1^{k+1}-\mathbf{A}_1^k\|_F^2\\
&\leq\delta_{\mathfrak{B}_1}(\mathbf{A}_1^{k})+\Psi\left(\mathcal{G}^{k+1},\mathbf{A}_1^{k},\mathbf{A}_2^k,\mathbf{A}_3^k,\mathbf{U}_1^k,\mathbf{U}_2^k,\mathbf{U}_3^k\right),\\
&\delta_{\mathfrak{B}_2}(\mathbf{A}_2^{k+1})+\Psi\left(\mathcal{G}^{k+1},\mathbf{A}_1^{k+1},\mathbf{A}_2^{k+1},\mathbf{A}_3^k,\mathbf{U}_1^k,\mathbf{U}_2^k,\mathbf{U}_3^k\right)+\frac{\rho_3-L_3}{2}\|\mathbf{A}_2^{k+1}-\mathbf{A}_2^k\|_F^2\\
&\leq\delta_{\mathfrak{B}_2}(\mathbf{A}_2^{k})+\Psi\left(\mathcal{G}^{k+1},\mathbf{A}_1^{k+1},\mathbf{A}_2^{k},\mathbf{A}_3^k,\mathbf{U}_1^k,\mathbf{U}_2^k,\mathbf{U}_3^k\right),\\
&\delta_{\mathfrak{B}_3}(\mathbf{A}_3^{k+1})+\Psi\left(\mathcal{G}^{k+1},\mathbf{A}_1^{k+1},\mathbf{A}_2^{k+1},\mathbf{A}_3^{k+1},\mathbf{U}_1^k,\mathbf{U}_2^k,\mathbf{U}_3^k\right)+\frac{\rho_4-L_4}{2}\|\mathbf{A}_3^{k+1}-\mathbf{A}_3^k\|_F^2\\
&\leq\delta_{\mathfrak{B}_3}(\mathbf{A}_3^{k})+\Psi\left(\mathcal{G}^{k+1},\mathbf{A}_1^{k+1},\mathbf{A}_2^{k+1},\mathbf{A}_3^{k},\mathbf{U}_1^k,\mathbf{U}_2^k,\mathbf{U}_3^k\right),\\
&\alpha_1\operatorname{tr}\left((\mathbf{U}_1^{k+1})'\mathbf{L}_1\mathbf{U}_1^{k+1}\right)+\Psi\left(\mathcal{G}^{k+1},\mathbf{A}_1^{k+1},\mathbf{A}_2^{k+1},\mathbf{A}_3^{k+1},\mathbf{U}_1^{k+1},\mathbf{U}_2^k,\mathbf{U}_3^k\right)\\
&+\frac{\rho_5-\gamma_1}{2}\|\mathbf{U}_1^{k+1}-\mathbf{U}_1^k\|_F^2\\
&\leq\alpha_1\operatorname{tr}\left((\mathbf{U}_1^{k})'\mathbf{L}_1\mathbf{U}_1^{k}\right)+\Psi\left(\mathcal{G}^{k+1},\mathbf{A}_1^{k+1},\mathbf{A}_2^{k+1},\mathbf{A}_3^{k+1},\mathbf{U}_1^{k},\mathbf{U}_2^k,\mathbf{U}_3^k\right),\\
&\alpha_2\operatorname{tr}\left((\mathbf{U}_2^{k+1})'\mathbf{L}_2\mathbf{U}_2^{k+1}\right)+\Psi\left(\mathcal{G}^{k+1},\mathbf{A}_1^{k+1},\mathbf{A}_2^{k+1},\mathbf{A}_3^{k+1},\mathbf{U}_1^{k+1},\mathbf{U}_2^{k+1},\mathbf{U}_3^k\right)\\
&+\frac{\rho_6-\gamma_2}{2}\|\mathbf{U}_2^{k+1}-\mathbf{U}_2^k\|_F^2\\
&\leq\alpha_2\operatorname{tr}\left((\mathbf{U}_2^{k})'\mathbf{L}_2\mathbf{U}_2^{k}\right)+\Psi\left(\mathcal{G}^{k+1},\mathbf{A}_1^{k+1},\mathbf{A}_2^{k+1},\mathbf{A}_3^{k+1},\mathbf{U}_1^{k+1},\mathbf{U}_2^k,\mathbf{U}_3^k\right),\\
&\alpha_3\operatorname{tr}\left((\mathbf{U}_3^{k+1})'\mathbf{L}_3\mathbf{U}_3^{k+1}\right)+\Psi\left(\mathcal{G}^{k+1},\mathbf{A}_1^{k+1},\mathbf{A}_2^{k+1},\mathbf{A}_3^{k+1},\mathbf{U}_1^{k+1},\mathbf{U}_2^{k+1},\mathbf{U}_3^{k+1}\right)\\
&+\frac{\rho_7-\gamma_3}{2}\|\mathbf{U}_3^{k+1}-\mathbf{U}_3^k\|_F^2\\
&\leq\alpha_3\operatorname{tr}\left((\mathbf{U}_3^{k})'\mathbf{L}_3\mathbf{U}_3^{k}\right)+\Psi\left(\mathcal{G}^{k+1},\mathbf{A}_1^{k+1},\mathbf{A}_2^{k+1},\mathbf{A}_3^{k+1},\mathbf{U}_1^{k+1},\mathbf{U}_2^{k+1},\mathbf{U}_3^{k}\right).
\end{aligned} \tag{S4.84}$$

This together with (S4.83) leads to

$$
\begin{aligned}
&\beta\|\mathcal{G}^{k+1}\|_1+\delta_{\mathcal{D}}(\mathcal{G}^{k+1})+\Psi\left(\mathcal{G}^{k+1},\mathbf{A}_1^{k+1},\mathbf{A}_2^{k+1},\mathbf{A}_3^{k+1},\mathbf{U}_1^{k+1},\mathbf{U}_2^{k+1},\mathbf{U}_3^{k+1}\right)\\
&+\frac{\rho_1-L_1}{2}\|\mathcal{G}^{k+1}-\mathcal{G}^k\|_F^2+\frac{\rho_2-L_2}{2}\|\mathbf{A}_1^{k+1}-\mathbf{A}_1^k\|_F^2+\frac{\rho_3-L_3}{2}\|\mathbf{A}_2^{k+1}-\mathbf{A}_2^k\|_F^2\\
&+\frac{\rho_4-L_4}{2}\|\mathbf{A}_3^{k+1}-\mathbf{A}_3^k\|_F^2+\frac{\rho_5-\gamma_1}{2}\|\mathbf{U}_1^{k+1}-\mathbf{U}_1^k\|_F^2+\frac{\rho_6-\gamma_2}{2}\|\mathbf{U}_2^{k+1}-\mathbf{U}_2^k\|_F^2\\
&+\frac{\rho_7-\gamma_3}{2}\|\mathbf{U}_3^{k+1}-\mathbf{U}_3^k\|_F^2+\sum_{i=1}^3\alpha_i\operatorname{tr}\left((\mathbf{U}_i^{k+1})'\mathbf{L}_i\mathbf{U}_i^{k+1}\right)+\sum_{i=1}^3\delta_{\mathfrak{B}_i}(\mathbf{A}_i^{k+1})\\
\le\ &\beta\|\mathcal{G}^k\|_1+\delta_{\mathcal{D}}(\mathcal{G}^k)+\Psi\left(\mathcal{G}^k,\mathbf{A}_1^k,\mathbf{A}_2^k,\mathbf{A}_3^k,\mathbf{U}_1^k,\mathbf{U}_2^k,\mathbf{U}_3^k\right)+\sum_{i=1}^3\delta_{\mathfrak{B}_i}(\mathbf{A}_i^k)\\
&+\sum_{i=1}^3\alpha_i\operatorname{tr}\left((\mathbf{U}_i^k)'\mathbf{L}_i\mathbf{U}_i^k\right).
\end{aligned}
$$

Denote $\bar{\rho}:=\min\{\rho_1-L_1,\rho_2-L_2,\rho_3-L_3,\rho_4-L_4,\rho_5-\gamma_1,\rho_6-\gamma_2,\rho_7-\gamma_3\}$, then one has

$$
\begin{aligned}
&\beta\|\mathcal{G}^{k+1}\|_1+\delta_{\mathcal{D}}(\mathcal{G}^{k+1})+\Psi\left(\mathcal{G}^{k+1},\mathbf{A}_1^{k+1},\mathbf{A}_2^{k+1},\mathbf{A}_3^{k+1},\mathbf{U}_1^{k+1},\mathbf{U}_2^{k+1},\mathbf{U}_3^{k+1}\right)\\
&+\frac{\bar{\rho}}{2}\left(\|\mathcal{G}^{k+1}-\mathcal{G}^k\|_F^2+\|\mathbf{A}_1^{k+1}-\mathbf{A}_1^k\|_F^2+\|\mathbf{A}_2^{k+1}-\mathbf{A}_2^k\|_F^2+\|\mathbf{A}_3^{k+1}-\mathbf{A}_3^k\|_F^2\right.\\
&\left.+\|\mathbf{U}_1^{k+1}-\mathbf{U}_1^k\|_F^2+\|\mathbf{U}_2^{k+1}-\mathbf{U}_2^k\|_F^2+\|\mathbf{U}_3^{k+1}-\mathbf{U}_3^k\|_F^2\right)\\
&+\sum_{i=1}^3\alpha_i\operatorname{tr}\left((\mathbf{U}_i^{k+1})'\mathbf{L}_i\mathbf{U}_i^{k+1}\right)+\sum_{i=1}^3\delta_{\mathfrak{B}_i}(\mathbf{A}_i^{k+1})\\
\le\ &\beta\|\mathcal{G}^k\|_1+\delta_{\mathcal{D}}(\mathcal{G}^k)+\Psi\left(\mathcal{G}^k,\mathbf{A}_1^k,\mathbf{A}_2^k,\mathbf{A}_3^k,\mathbf{U}_1^k,\mathbf{U}_2^k,\mathbf{U}_3^k\right)+\sum_{i=1}^3\delta_{\mathfrak{B}_i}(\mathbf{A}_i^k)\\
&+\sum_{i=1}^3\alpha_i\operatorname{tr}\left((\mathbf{U}_i^k)'\mathbf{L}_i\mathbf{U}_i^k\right).
\end{aligned}
$$

Recalling the definition of F in (S3.52) and the definition of Ψ in (S3.54),

the above inequality is equivalent to

$$
\begin{aligned}
&F\left(\mathcal{G}^{k+1}, \mathbf{A}_1^{k+1}, \mathbf{A}_2^{k+1}, \mathbf{A}_3^{k+1}, \mathbf{U}_1^{k+1}, \mathbf{U}_2^{k+1}, \mathbf{U}_3^{k+1}\right) + \frac{\bar{\rho}}{2}\left(\|\mathcal{G}^{k+1} - \mathcal{G}^k\|_F^2\right. \\
&+ \|\mathbf{A}_1^{k+1} - \mathbf{A}_1^k\|_F^2 + \|\mathbf{A}_2^{k+1} - \mathbf{A}_2^k\|_F^2 + \|\mathbf{A}_3^{k+1} - \mathbf{A}_3^k\|_F^2 + \|\mathbf{U}_1^{k+1} - \mathbf{U}_1^k\|_F^2 \\
&\left.+ \|\mathbf{U}_2^{k+1} - \mathbf{U}_2^k\|_F^2 + \|\mathbf{U}_3^{k+1} - \mathbf{U}_3^k\|_F^2\right) \\
&\leq F\left(\mathcal{G}^k, \mathbf{A}_1^k, \mathbf{A}_2^k, \mathbf{A}_3^k, \mathbf{U}_1^k, \mathbf{U}_2^k, \mathbf{U}_3^k\right).
\end{aligned} \tag{S4.85}
$$

This completes the proof. □

Lemma 13. *(Relative error condition) Let the sequence* $\{(\mathcal{G}^k, \mathbf{A}_1^k, \mathbf{A}_2^k, \mathbf{A}_3^k, \mathbf{U}_1^k, \mathbf{U}_2^k, \mathbf{U}_3^k)\}$ *be generated by Algorithm 1. Then there exist* $\mathcal{N}^{k+1} \in \partial F(\mathcal{G}^{k+1}, \mathbf{A}_1^{k+1}, \mathbf{A}_2^{k+1}, \mathbf{A}_3^{k+1}, \mathbf{U}_1^{k+1}, \mathbf{U}_2^{k+1}, \mathbf{U}_3^{k+1})$ *and a constant* $\bar{\delta} > 0$ *such that*

$$
\begin{aligned}
\|\mathcal{N}^{k+1}\|_F \leq \bar{\delta}\left(\right.&\|\mathcal{G}^{k+1} - \mathcal{G}^k\|_F + \|\mathbf{A}_1^{k+1} - \mathbf{A}_1^k\|_F + \|\mathbf{A}_2^{k+1} - \mathbf{A}_2^k\|_F \\
&+ \|\mathbf{A}_3^{k+1} - \mathbf{A}_3^k\|_F + \|\mathbf{U}_1^{k+1} - \mathbf{U}_1^k\|_F + \|\mathbf{U}_2^{k+1} - \mathbf{U}_2^k\|_F \\
&\left.+ \|\mathbf{U}_3^{k+1} - \mathbf{U}_3^k\|_F\right).
\end{aligned} \tag{S4.86}
$$

Proof. Firstly, we provide the first-order optimality conditions for the PALM

algorithm in (3.13):

$$\begin{cases} 0 \in \partial(\beta\|\mathcal{G}^{k+1}\|_1 + \delta_{\mathcal{D}}(\mathcal{G}^{k+1})) + \nabla_{\mathcal{G}}\Psi(\mathcal{G}^k, \mathbf{A}_1^k, \mathbf{A}_2^k, \mathbf{A}_3^k, \mathbf{U}_1^k, \mathbf{U}_2^k, \mathbf{U}_3^k) \\ \quad + \rho_1(\mathcal{G}^{k+1} - \mathcal{G}^k), \\ 0 \in \partial\delta_{\mathfrak{B}_1}(\mathbf{A}_1^{k+1}) + \nabla_{\mathbf{A}_1}\Psi(\mathcal{G}^{k+1}, \mathbf{A}_1^k, \mathbf{A}_2^k, \mathbf{A}_3^k, \mathbf{U}_1^k, \mathbf{U}_2^k, \mathbf{U}_3^k) \\ \quad + \rho_2(\mathbf{A}_1^{k+1} - \mathbf{A}_1^k), \\ 0 \in \partial\delta_{\mathfrak{B}_2}(\mathbf{A}_2^{k+1}) + \nabla_{\mathbf{A}_2}\Psi(\mathcal{G}^{k+1}, \mathbf{A}_1^{k+1}, \mathbf{A}_2^k, \mathbf{A}_3^k, \mathbf{U}_1^k, \mathbf{U}_2^k, \mathbf{U}_3^k) \\ \quad + \rho_3(\mathbf{A}_2^{k+1} - \mathbf{A}_2^k), \\ 0 \in \partial\delta_{\mathfrak{B}_3}(\mathbf{A}_3^{k+1}) + \nabla_{\mathbf{A}_3}\Psi(\mathcal{G}^{k+1}, \mathbf{A}_1^{k+1}, \mathbf{A}_2^{k+1}, \mathbf{A}_3^k, \mathbf{U}_1^k, \mathbf{U}_2^k, \mathbf{U}_3^k) \\ \quad + \rho_4(\mathbf{A}_3^{k+1} - \mathbf{A}_3^k), \\ 0 = 2\alpha_1\mathbf{L}_1\mathbf{U}_1^{k+1} + \nabla_{\mathbf{U}_1}\Psi(\mathcal{G}^{k+1}, \mathbf{A}_1^{k+1}, \mathbf{A}_2^{k+1}, \mathbf{A}_3^{k+1}, \mathbf{U}_1^k, \mathbf{U}_2^k, \mathbf{U}_3^k) \\ \quad + \rho_5(\mathbf{U}_1^{k+1} - \mathbf{U}_1^k), \\ 0 = 2\alpha_2\mathbf{L}_2\mathbf{U}_2^{k+1} + \nabla_{\mathbf{U}_2}\Psi(\mathcal{G}^{k+1}, \mathbf{A}_1^{k+1}, \mathbf{A}_2^{k+1}, \mathbf{A}_3^{k+1}, \mathbf{U}_1^{k+1}, \mathbf{U}_2^k, \mathbf{U}_3^k) \\ \quad + \rho_6(\mathbf{U}_2^{k+1} - \mathbf{U}_2^k), \\ 0 = 2\alpha_3\mathbf{L}_3\mathbf{U}_3^{k+1} + \nabla_{\mathbf{U}_3}\Psi(\mathcal{G}^{k+1}, \mathbf{A}_1^{k+1}, \mathbf{A}_2^{k+1}, \mathbf{A}_3^{k+1}, \mathbf{U}_1^{k+1}, \mathbf{U}_2^{k+1}, \mathbf{U}_3^k) \\ \quad + \rho_7(\mathbf{U}_3^{k+1} - \mathbf{U}_3^k). \end{cases} \tag{S4.87}$$

Denote

$$\begin{aligned} N_1 = {} & \nabla_{\mathcal{G}}\Psi(\mathcal{G}^{k+1}, \mathbf{A}_1^{k+1}, \mathbf{A}_2^{k+1}, \mathbf{A}_3^{k+1}, \mathbf{U}_1^{k+1}, \mathbf{U}_2^{k+1}, \mathbf{U}_3^{k+1}) \\ & - \nabla_{\mathcal{G}}\Psi(\mathcal{G}^k, \mathbf{A}_1^k, \mathbf{A}_2^k, \mathbf{A}_3^k, \mathbf{U}_1^k, \mathbf{U}_2^k, \mathbf{U}_3^k) - \rho_1(\mathcal{G}^{k+1} - \mathcal{G}^k), \end{aligned} \tag{S4.88}$$

and

$$
\begin{aligned}
N_2 &= \nabla_{\mathbf{A}_1}\Psi(\mathcal{G}^{k+1},\mathbf{A}_1^{k+1},\mathbf{A}_2^{k+1},\mathbf{A}_3^{k+1},\mathbf{U}_1^{k+1},\mathbf{U}_2^{k+1},\mathbf{U}_3^{k+1}) \\
&\quad - \nabla_{\mathbf{A}_1}\Psi(\mathcal{G}^{k+1},\mathbf{A}_1^{k},\mathbf{A}_2^{k},\mathbf{A}_3^{k},\mathbf{U}_1^{k},\mathbf{U}_2^{k},\mathbf{U}_3^{k}) - \rho_2(\mathbf{A}_1^{k+1}-\mathbf{A}_1^{k}), \\
N_3 &= \nabla_{\mathbf{A}_2}\Psi(\mathcal{G}^{k+1},\mathbf{A}_1^{k+1},\mathbf{A}_2^{k+1},\mathbf{A}_3^{k+1},\mathbf{U}_1^{k+1},\mathbf{U}_2^{k+1},\mathbf{U}_3^{k+1}) \\
&\quad - \nabla_{\mathbf{A}_2}\Psi(\mathcal{G}^{k+1},\mathbf{A}_1^{k+1},\mathbf{A}_2^{k},\mathbf{A}_3^{k},\mathbf{U}_1^{k},\mathbf{U}_2^{k},\mathbf{U}_3^{k}) - \rho_3(\mathbf{A}_2^{k+1}-\mathbf{A}_2^{k}), \\
N_4 &= \nabla_{\mathbf{A}_3}\Psi(\mathcal{G}^{k+1},\mathbf{A}_1^{k+1},\mathbf{A}_2^{k+1},\mathbf{A}_3^{k+1},\mathbf{U}_1^{k+1},\mathbf{U}_2^{k+1},\mathbf{U}_3^{k+1}) \\
&\quad - \nabla_{\mathbf{A}_3}\Psi(\mathcal{G}^{k+1},\mathbf{A}_1^{k+1},\mathbf{A}_2^{k+1},\mathbf{A}_3^{k},\mathbf{U}_1^{k},\mathbf{U}_2^{k},\mathbf{U}_3^{k}) - \rho_4(\mathbf{A}_3^{k+1}-\mathbf{A}_3^{k}), \\
N_5 &= \nabla_{\mathbf{U}_1}\Psi(\mathcal{G}^{k+1},\mathbf{A}_1^{k+1},\mathbf{A}_2^{k+1},\mathbf{A}_3^{k+1},\mathbf{U}_1^{k+1},\mathbf{U}_2^{k+1},\mathbf{U}_3^{k+1}) \\
&\quad - \nabla_{\mathbf{U}_1}\Psi(\mathcal{G}^{k+1},\mathbf{A}_1^{k+1},\mathbf{A}_2^{k+1},\mathbf{A}_3^{k+1},\mathbf{U}_1^{k},\mathbf{U}_2^{k},\mathbf{U}_3^{k}) - \rho_5(\mathbf{U}_1^{k+1}-\mathbf{U}_1^{k}), \\
N_6 &= \nabla_{\mathbf{U}_2}\Psi(\mathcal{G}^{k+1},\mathbf{A}_1^{k+1},\mathbf{A}_2^{k+1},\mathbf{A}_3^{k+1},\mathbf{U}_1^{k+1},\mathbf{U}_2^{k+1},\mathbf{U}_3^{k+1}) \\
&\quad - \nabla_{\mathbf{U}_2}\Psi(\mathcal{G}^{k+1},\mathbf{A}_1^{k+1},\mathbf{A}_2^{k+1},\mathbf{A}_3^{k+1},\mathbf{U}_1^{k+1},\mathbf{U}_2^{k},\mathbf{U}_3^{k}) - \rho_6(\mathbf{U}_2^{k+1}-\mathbf{U}_2^{k}), \\
N_7 &= \nabla_{\mathbf{U}_3}\Psi(\mathcal{G}^{k+1},\mathbf{A}_1^{k+1},\mathbf{A}_2^{k+1},\mathbf{A}_3^{k+1},\mathbf{U}_1^{k+1},\mathbf{U}_2^{k+1},\mathbf{U}_3^{k+1}) \\
&\quad - \nabla_{\mathbf{U}_3}\Psi(\mathcal{G}^{k+1},\mathbf{A}_1^{k+1},\mathbf{A}_2^{k+1},\mathbf{A}_3^{k+1},\mathbf{U}_1^{k+1},\mathbf{U}_2^{k+1},\mathbf{U}_3^{k}) - \rho_7(\mathbf{U}_3^{k+1}-\mathbf{U}_3^{k}).
\end{aligned}
\tag{S4.89}
$$

Then we get

$$
\begin{aligned}
N_1 &\in \partial_{\mathcal{G}} F(\mathcal{G}^{k+1},\mathbf{A}_1^{k+1},\mathbf{A}_2^{k+1},\mathbf{A}_3^{k+1},\mathbf{U}_1^{k+1},\mathbf{U}_2^{k+1},\mathbf{U}_3^{k+1}), \\
N_2 &\in \partial_{\mathbf{A}_1} F(\mathcal{G}^{k+1},\mathbf{A}_1^{k+1},\mathbf{A}_2^{k+1},\mathbf{A}_3^{k+1},\mathbf{U}_1^{k+1},\mathbf{U}_2^{k+1},\mathbf{U}_3^{k+1}), \\
N_3 &\in \partial_{\mathbf{A}_2} F(\mathcal{G}^{k+1},\mathbf{A}_1^{k+1},\mathbf{A}_2^{k+1},\mathbf{A}_3^{k+1},\mathbf{U}_1^{k+1},\mathbf{U}_2^{k+1},\mathbf{U}_3^{k+1}), \\
N_4 &\in \partial_{\mathbf{A}_3} F(\mathcal{G}^{k+1},\mathbf{A}_1^{k+1},\mathbf{A}_2^{k+1},\mathbf{A}_3^{k+1},\mathbf{U}_1^{k+1},\mathbf{U}_2^{k+1},\mathbf{U}_3^{k+1}),
\end{aligned}
$$

$$N_5 \in \partial_{\mathbf{U}_1} F(\mathcal{G}^{k+1}, \mathbf{A}_1^{k+1}, \mathbf{A}_2^{k+1}, \mathbf{A}_3^{k+1}, \mathbf{U}_1^{k+1}, \mathbf{U}_2^{k+1}, \mathbf{U}_3^{k+1}),$$

$$N_6 \in \partial_{\mathbf{U}_2} F(\mathcal{G}^{k+1}, \mathbf{A}_1^{k+1}, \mathbf{A}_2^{k+1}, \mathbf{A}_3^{k+1}, \mathbf{U}_1^{k+1}, \mathbf{U}_2^{k+1}, \mathbf{U}_3^{k+1}),$$

$$N_7 \in \partial_{\mathbf{U}_3} F(\mathcal{G}^{k+1}, \mathbf{A}_1^{k+1}, \mathbf{A}_2^{k+1}, \mathbf{A}_3^{k+1}, \mathbf{U}_1^{k+1}, \mathbf{U}_2^{k+1}, \mathbf{U}_3^{k+1}).$$

By Proposition 2.1 of Attouch et al. (2010), we obtain that

$$\begin{aligned} N^{k+1} :=\ & (N_1, N_2, N_3, N_4, N_5, N_6, N_7) \\ & \in \partial F(\mathcal{G}^{k+1}, \mathbf{A}_1^{k+1}, \mathbf{A}_2^{k+1}, \mathbf{A}_3^{k+1}, \mathbf{U}_1^{k+1}, \mathbf{U}_2^{k+1}, \mathbf{U}_3^{k+1}). \end{aligned}$$

Note that

$$\begin{aligned} \|N_1\|_F =& \Big\| \nabla_{\mathcal{G}} \Psi(\mathcal{G}^{k+1}, \mathbf{A}_1^{k+1}, \mathbf{A}_2^{k+1}, \mathbf{A}_3^{k+1}, \mathbf{U}_1^{k+1}, \mathbf{U}_2^{k+1}, \mathbf{U}_3^{k+1}) \\ & - \nabla_{\mathcal{G}} \Psi(\mathcal{G}^{k}, \mathbf{A}_1^{k}, \mathbf{A}_2^{k}, \mathbf{A}_3^{k}, \mathbf{U}_1^{k}, \mathbf{U}_2^{k}, \mathbf{U}_3^{k}) - \rho_1(\mathcal{G}^{k+1} - \mathcal{G}^k) \Big\|_F \\ \leq & \Big\| \nabla_{\mathcal{G}} \Psi(\mathcal{G}^{k+1}, \mathbf{A}_1^{k+1}, \mathbf{A}_2^{k+1}, \mathbf{A}_3^{k+1}, \mathbf{U}_1^{k+1}, \mathbf{U}_2^{k+1}, \mathbf{U}_3^{k+1}) \\ & - \nabla_{\mathcal{G}} \Psi(\mathcal{G}^{k}, \mathbf{A}_1^{k}, \mathbf{A}_2^{k}, \mathbf{A}_3^{k}, \mathbf{U}_1^{k}, \mathbf{U}_2^{k}, \mathbf{U}_3^{k}) \Big\|_F + \rho_1 \|\mathcal{G}^{k+1} - \mathcal{G}^k\|_F \\ = & \Big\| \nabla_{\mathcal{G}} Q([\![\mathcal{G}^{k+1}; \mathbf{A}_1^{k+1}, \mathbf{A}_2^{k+1}, \mathbf{A}_3^{k+1}]\!]) - \nabla_{\mathcal{G}} Q([\![\mathcal{G}^{k}; \mathbf{A}_1^{k}, \mathbf{A}_2^{k}, \mathbf{A}_3^{k}]\!]) \Big\|_F \\ & + \rho_1 \|\mathcal{G}^{k+1} - \mathcal{G}^k\|_F \\ = & \Big\| (\nabla_{\mathcal{G}} Q([\![\mathcal{G}^{k+1}; \mathbf{A}_1^{k+1}, \mathbf{A}_2^{k+1}, \mathbf{A}_3^{k+1}]\!])_{(1)} - (\nabla_{\mathcal{G}} Q([\![\mathcal{G}^{k}; \mathbf{A}_1^{k}, \mathbf{A}_2^{k}, \mathbf{A}_3^{k}]\!]))_{(1)} \Big\|_F \\ & + \rho_1 \|\mathcal{G}^{k+1} - \mathcal{G}^k\|_F, \end{aligned} \tag{S4.90}$$

where the second equality follows from (S3.57). Notice that (S3.53) can be rewritten as

$$Q([\![\mathcal{G}; \mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3]\!]) = \frac{1}{2T} \sum_{t=1}^{T} \left\| \mathbf{y}_t - \mathbf{A}_1 \mathcal{G}_{(1)} (\mathbf{A}_3 \otimes \mathbf{A}_2)' \mathbf{x}_t \right\|_2^2. \tag{S4.91}$$

Combining (S4.91) with Lemma 7, simple calculation gives

$$
\begin{aligned}
&\left\|\left(\nabla_{\mathcal{G}} Q([\![\mathcal{G}^{k+1};\mathbf{A}_1^{k+1},\mathbf{A}_2^{k+1},\mathbf{A}_3^{k+1}]\!])\right)_{(1)}-\left(\nabla_{\mathcal{G}} Q([\![\mathcal{G}^{k};\mathbf{A}_1^{k},\mathbf{A}_2^{k},\mathbf{A}_3^{k}]\!])\right)_{(1)}\right\|_F\\
=&\left\|\frac{1}{T}\sum_{t=1}^{T}(\mathbf{A}_1^{k+1})'\left(\mathbf{A}_1^{k+1}\mathcal{G}_{(1)}^{k+1}(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k+1})'\mathbf{x}_t-\mathbf{y}_t\right)\left((\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k+1})'\mathbf{x}_t\right)'\right.\\
&\left.-\frac{1}{T}\sum_{t=1}^{T}(\mathbf{A}_1^{k})'\left(\mathbf{A}_1^{k}\mathcal{G}_{(1)}^{k}(\mathbf{A}_3^{k}\otimes\mathbf{A}_2^{k})'\mathbf{x}_t-\mathbf{y}_t\right)\left((\mathbf{A}_3^{k}\otimes\mathbf{A}_2^{k})'\mathbf{x}_t\right)'\right\|_F\\
=&\left\|\frac{1}{T}\sum_{t=1}^{T}\left(\mathcal{G}_{(1)}^{k+1}(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k+1})'\mathbf{x}_t-(\mathbf{A}_1^{k+1})'\mathbf{y}_t\right)\left((\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k+1})'\mathbf{x}_t\right)'\right.\\
&\left.-\frac{1}{T}\sum_{t=1}^{T}\left(\mathcal{G}_{(1)}^{k}(\mathbf{A}_3^{k}\otimes\mathbf{A}_2^{k})'\mathbf{x}_t-(\mathbf{A}_1^{k})'\mathbf{y}_t\right)\left((\mathbf{A}_3^{k}\otimes\mathbf{A}_2^{k})'\mathbf{x}_t\right)'\right\|_F\\
=&\left\|\frac{1}{T}\sum_{t=1}^{T}\mathcal{G}_{(1)}^{k+1}(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k+1})'\mathbf{x}_t\mathbf{x}_t'(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k+1})\right.\\
&-\frac{1}{T}\sum_{t=1}^{T}(\mathbf{A}_1^{k+1})'\mathbf{y}_t\mathbf{x}_t'(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k+1})+\frac{1}{T}\sum_{t=1}^{T}(\mathbf{A}_1^{k})'\mathbf{y}_t\mathbf{x}_t'(\mathbf{A}_3^{k}\otimes\mathbf{A}_2^{k})\\
&\left.-\frac{1}{T}\sum_{t=1}^{T}\mathcal{G}_{(1)}^{k}(\mathbf{A}_3^{k}\otimes\mathbf{A}_2^{k})'\mathbf{x}_t\mathbf{x}_t'(\mathbf{A}_3^{k}\otimes\mathbf{A}_2^{k})\right\|_F\\
\le&\left\|\frac{1}{T}\sum_{t=1}^{T}\mathcal{G}_{(1)}^{k+1}(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k+1})'\mathbf{x}_t\mathbf{x}_t'(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k+1})\right.\\
&\left.-\frac{1}{T}\sum_{t=1}^{T}\mathcal{G}_{(1)}^{k}(\mathbf{A}_3^{k}\otimes\mathbf{A}_2^{k})'\mathbf{x}_t\mathbf{x}_t'(\mathbf{A}_3^{k}\otimes\mathbf{A}_2^{k})\right\|_F\\
&+\left\|\frac{1}{T}\sum_{t=1}^{T}(\mathbf{A}_1^{k})'\mathbf{y}_t\mathbf{x}_t'(\mathbf{A}_3^{k}\otimes\mathbf{A}_2^{k})-\frac{1}{T}\sum_{t=1}^{T}(\mathbf{A}_1^{k+1})'\mathbf{y}_t\mathbf{x}_t'(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k+1})\right\|_F\\
\le&\frac{1}{T}\sum_{t=1}^{T}\left\|\mathcal{G}_{(1)}^{k+1}(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k+1})'\mathbf{x}_t\mathbf{x}_t'(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k+1})\right.\\
&\left.-\mathcal{G}_{(1)}^{k}(\mathbf{A}_3^{k}\otimes\mathbf{A}_2^{k})'\mathbf{x}_t\mathbf{x}_t'(\mathbf{A}_3^{k}\otimes\mathbf{A}_2^{k})\right\|_F\\
&+\frac{1}{T}\sum_{t=1}^{T}\left\|(\mathbf{A}_1^{k})'\mathbf{y}_t\mathbf{x}_t'(\mathbf{A}_3^{k}\otimes\mathbf{A}_2^{k})-(\mathbf{A}_1^{k+1})'\mathbf{y}_t\mathbf{x}_t'(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k+1})\right\|_F,
\end{aligned}
\tag{S4.92}
$$

where the second equality results from $(\mathbf{A}_1^{k+1})'\mathbf{A}_1^{k+1} = \mathbf{I}_{r_i}$. Next, we start by analyzing the second term on the right side of (S4.92). Note that

$$\begin{aligned}
&\left\|(\mathbf{A}_1^k)'\mathbf{y}_t\mathbf{x}_t'(\mathbf{A}_3^k \otimes \mathbf{A}_2^k) - (\mathbf{A}_1^{k+1})'\mathbf{y}_t\mathbf{x}_t'(\mathbf{A}_3^{k+1} \otimes \mathbf{A}_2^{k+1})\right\|_F \\
\leq\ &\left\|(\mathbf{A}_1^k)'\mathbf{y}_t\mathbf{x}_t'(\mathbf{A}_3^k \otimes \mathbf{A}_2^k) - (\mathbf{A}_1^{k+1})'\mathbf{y}_t\mathbf{x}_t'(\mathbf{A}_3^k \otimes \mathbf{A}_2^k)\right. \\
&\quad + (\mathbf{A}_1^{k+1})'\mathbf{y}_t\mathbf{x}_t'(\mathbf{A}_3^k \otimes \mathbf{A}_2^k) - (\mathbf{A}_1^{k+1})'\mathbf{y}_t\mathbf{x}_t'(\mathbf{A}_3^k \otimes \mathbf{A}_2^{k+1}) \\
&\quad \left. + (\mathbf{A}_1^{k+1})'\mathbf{y}_t\mathbf{x}_t'(\mathbf{A}_3^k \otimes \mathbf{A}_2^{k+1}) - (\mathbf{A}_1^{k+1})'\mathbf{y}_t\mathbf{x}_t'(\mathbf{A}_3^{k+1} \otimes \mathbf{A}_2^{k+1})\right\|_F \\
\leq\ &\left\|(\mathbf{A}_1^k)'\mathbf{y}_t\mathbf{x}_t'(\mathbf{A}_3^k \otimes \mathbf{A}_2^k) - (\mathbf{A}_1^{k+1})'\mathbf{y}_t\mathbf{x}_t'(\mathbf{A}_3^k \otimes \mathbf{A}_2^k)\right\|_F \\
&\quad + \left\|(\mathbf{A}_1^{k+1})'\mathbf{y}_t\mathbf{x}_t'(\mathbf{A}_3^k \otimes \mathbf{A}_2^k) - (\mathbf{A}_1^{k+1})'\mathbf{y}_t\mathbf{x}_t'(\mathbf{A}_3^k \otimes \mathbf{A}_2^{k+1})\right\|_F \\
&\quad + \left\|(\mathbf{A}_1^{k+1})'\mathbf{y}_t\mathbf{x}_t'(\mathbf{A}_3^k \otimes \mathbf{A}_2^{k+1}) - (\mathbf{A}_1^{k+1})'\mathbf{y}_t\mathbf{x}_t'(\mathbf{A}_3^{k+1} \otimes \mathbf{A}_2^{k+1})\right\|_F .
\end{aligned} \tag{S4.93}$$

By the Cauchy-Schwarz inequality, one can get that

$$\begin{aligned}
&\left\|(\mathbf{A}_1^k)'\mathbf{y}_t\mathbf{x}_t'(\mathbf{A}_3^k \otimes \mathbf{A}_2^k) - (\mathbf{A}_1^{k+1})'\mathbf{y}_t\mathbf{x}_t'(\mathbf{A}_3^k \otimes \mathbf{A}_2^k)\right\|_F \\
=\ &\left\|(\mathbf{A}_1^k - \mathbf{A}_1^{k+1})'\mathbf{y}_t\mathbf{x}_t'(\mathbf{A}_3^k \otimes \mathbf{A}_2^k)\right\|_F \\
\leq\ &\left\|(\mathbf{A}_1^k - \mathbf{A}_1^{k+1})'\right\|_F \left\|\mathbf{y}_t\mathbf{x}_t'(\mathbf{A}_3^k \otimes \mathbf{A}_2^k)\right\|_F \\
\leq\ &\|\mathbf{A}_1^k - \mathbf{A}_1^{k+1}\|_F \|\mathbf{y}_t\mathbf{x}_t'\|_F \|\mathbf{A}_3^k \otimes \mathbf{A}_2^k\| \\
\leq\ &\|\mathbf{A}_1^k - \mathbf{A}_1^{k+1}\|_F \|\mathbf{y}_t\mathbf{x}_t'\|_F,
\end{aligned} \tag{S4.94}$$

where the second inequality holds by Lemma 2.1 of Chen et al. (2021) and

the last inequality holds by Proposition 2. One can similarly obtain that

$$\begin{aligned}
&\left\|(\mathbf{A}_1^{k+1})'\mathbf{y}_t\mathbf{x}_t'(\mathbf{A}_3^k\otimes\mathbf{A}_2^k)-(\mathbf{A}_1^{k+1})'\mathbf{y}_t\mathbf{x}_t'(\mathbf{A}_3^k\otimes\mathbf{A}_2^{k+1})\right\|_F\\
&\le\|\mathbf{y}_t\mathbf{x}_t'\|_F\|\mathbf{A}_2^k-\mathbf{A}_2^{k+1}\|_F,\\
&\left\|(\mathbf{A}_1^{k+1})'\mathbf{y}_t\mathbf{x}_t'(\mathbf{A}_3^k\otimes\mathbf{A}_2^{k+1})-(\mathbf{A}_1^{k+1})'\mathbf{y}_t\mathbf{x}_t'(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k+1})\right\|_F\\
&\le\|\mathbf{y}_t\mathbf{x}_t'\|_F\|\mathbf{A}_3^k-\mathbf{A}_3^{k+1}\|_F.
\end{aligned}$$

Taking this together with (S4.93) and (S4.94), we can deduce

$$\begin{aligned}
&\left\|(\mathbf{A}_1^{k})'\mathbf{y}_t\mathbf{x}_t'(\mathbf{A}_3^k\otimes\mathbf{A}_2^k)-(\mathbf{A}_1^{k+1})'\mathbf{y}_t\mathbf{x}_t'(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k+1})\right\|_F\\
&\le\|\mathbf{y}_t\mathbf{x}_t'\|_F\left(\|\mathbf{A}_1^k-\mathbf{A}_1^{k+1}\|_F+\|\mathbf{A}_2^k-\mathbf{A}_2^{k+1}\|_F+\|\mathbf{A}_3^k-\mathbf{A}_3^{k+1}\|_F\right).
\end{aligned}\tag{S4.95}$$

Summing up the above inequality over $t=1,\ldots,T$, we get that

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^T\left\|(\mathbf{A}_1^{k})'\mathbf{y}_t\mathbf{x}_t'(\mathbf{A}_3^k\otimes\mathbf{A}_2^k)-(\mathbf{A}_1^{k+1})'\mathbf{y}_t\mathbf{x}_t'(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k+1})\right\|_F\\
&\le\frac{1}{T}\sum_{t=1}^T\|\mathbf{y}_t\mathbf{x}_t'\|_F\left(\|\mathbf{A}_1^k-\mathbf{A}_1^{k+1}\|_F+\|\mathbf{A}_2^k-\mathbf{A}_2^{k+1}\|_F+\|\mathbf{A}_3^k-\mathbf{A}_3^{k+1}\|_F\right)\\
&=c_2\left(\|\mathbf{A}_1^k-\mathbf{A}_1^{k+1}\|_F+\|\mathbf{A}_2^k-\mathbf{A}_2^{k+1}\|_F+\|\mathbf{A}_3^k-\mathbf{A}_3^{k+1}\|_F\right),
\end{aligned}\tag{S4.96}$$

where

$$c_2:=\frac{1}{T}\sum_{t=1}^T\|\mathbf{y}_t\mathbf{x}_t'\|_F.\tag{S4.97}$$

Now we analyze the first term on the right side of (S4.92). It is easily

seen that

$$
\begin{aligned}
&\left\|\mathcal{G}_{(1)}^{k+1}(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k+1})'\mathbf{x}_t\mathbf{x}_t'(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k+1})\right.\\
&\qquad\left.-\,\mathcal{G}_{(1)}^{k}(\mathbf{A}_3^{k}\otimes\mathbf{A}_2^{k})'\mathbf{x}_t\mathbf{x}_t'(\mathbf{A}_3^{k}\otimes\mathbf{A}_2^{k})\right\|_F\\
=\;&\left\|\mathcal{G}_{(1)}^{k+1}(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k+1})'\mathbf{x}_t\mathbf{x}_t'(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k+1})\right.\\
&\qquad-\,\mathcal{G}_{(1)}^{k}(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k+1})'\mathbf{x}_t\mathbf{x}_t'(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k+1})\\
&\qquad+\,\mathcal{G}_{(1)}^{k}(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k+1})'\mathbf{x}_t\mathbf{x}_t'(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k+1})\\
&\qquad\left.-\,\mathcal{G}_{(1)}^{k}(\mathbf{A}_3^{k}\otimes\mathbf{A}_2^{k})'\mathbf{x}_t\mathbf{x}_t'(\mathbf{A}_3^{k}\otimes\mathbf{A}_2^{k})\right\|_F\\
\le\;&\left\|\mathcal{G}_{(1)}^{k+1}(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k+1})'\mathbf{x}_t\mathbf{x}_t'(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k+1})\right.\\
&\qquad\left.-\,\mathcal{G}_{(1)}^{k}(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k+1})'\mathbf{x}_t\mathbf{x}_t'(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k+1})\right\|_F\\
&\quad+\left\|\mathcal{G}_{(1)}^{k}(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k+1})'\mathbf{x}_t\mathbf{x}_t'(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k+1})\right.\\
&\qquad\left.-\,\mathcal{G}_{(1)}^{k}(\mathbf{A}_3^{k}\otimes\mathbf{A}_2^{k})'\mathbf{x}_t\mathbf{x}_t'(\mathbf{A}_3^{k}\otimes\mathbf{A}_2^{k})\right\|_F.
\end{aligned}
\tag{S4.98}
$$

Applying the Cauchy-Schwarz inequality to (S4.98) yields

$$
\begin{aligned}
&\left\|\mathcal{G}_{(1)}^{k+1}(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k+1})'\mathbf{x}_t\mathbf{x}_t'(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k+1})\right.\\
&-\left.\mathcal{G}_{(1)}^{k}(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k+1})'\mathbf{x}_t\mathbf{x}_t'(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k+1})\right\|_F\\
=\;&\left\|(\mathcal{G}_{(1)}^{k+1}-\mathcal{G}_{(1)}^{k})(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k+1})'\mathbf{x}_t\mathbf{x}_t'(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k+1})\right\|_F\\
\le\;&\|\mathcal{G}_{(1)}^{k+1}-\mathcal{G}_{(1)}^{k}\|_F\left\|(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k+1})'\mathbf{x}_t\mathbf{x}_t'(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k+1})\right\|_F\\
=\;&\|\mathcal{G}^{k+1}-\mathcal{G}^{k}\|_F\left\|(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k+1})'\mathbf{x}_t\right\|_2^2\\
\le\;&\|\mathcal{G}^{k+1}-\mathcal{G}^{k}\|_F\left(\|\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k+1}\|\,\|\mathbf{x}_t\|_2\right)^2\\
\le\;&\|\mathcal{G}^{k+1}-\mathcal{G}^{k}\|_F\,\|\mathbf{x}_t\|_2^2,
\end{aligned}
\tag{S4.99}
$$

where the second inequality follows from Example 5.6.6 of Horn and Johnson (2013).

On the other hand, for the second term of the right hand inequality in

(S4.98), we obtain that

$$
\begin{aligned}
&\left\|\mathcal{G}_{(1)}^k(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k+1})'\mathbf{x}_t\mathbf{x}_t'(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k+1})-\mathcal{G}_{(1)}^k(\mathbf{A}_3^{k}\otimes\mathbf{A}_2^{k})'\mathbf{x}_t\mathbf{x}_t'(\mathbf{A}_3^{k}\otimes\mathbf{A}_2^{k})\right\|_F\\
\le&\ \|\mathcal{G}_{(1)}^k\|_F\left\|(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k+1})'\mathbf{x}_t\mathbf{x}_t'(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k+1})-(\mathbf{A}_3^{k}\otimes\mathbf{A}_2^{k})'\mathbf{x}_t\mathbf{x}_t'(\mathbf{A}_3^{k}\otimes\mathbf{A}_2^{k})\right\|_F\\
=&\ \|\mathcal{G}_{(1)}^k\|_F\left\|(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k+1})'\mathbf{x}_t\mathbf{x}_t'(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k+1})-(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k+1})'\mathbf{x}_t\mathbf{x}_t'(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k})\right.\\
&+(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k+1})'\mathbf{x}_t\mathbf{x}_t'(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k})-(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k+1})'\mathbf{x}_t\mathbf{x}_t'(\mathbf{A}_3^{k}\otimes\mathbf{A}_2^{k})\\
&+(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k+1})'\mathbf{x}_t\mathbf{x}_t'(\mathbf{A}_3^{k}\otimes\mathbf{A}_2^{k})-(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k})'\mathbf{x}_t\mathbf{x}_t'(\mathbf{A}_3^{k}\otimes\mathbf{A}_2^{k})\\
&\left.+(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k})'\mathbf{x}_t\mathbf{x}_t'(\mathbf{A}_3^{k}\otimes\mathbf{A}_2^{k})-(\mathbf{A}_3^{k}\otimes\mathbf{A}_2^{k})'\mathbf{x}_t\mathbf{x}_t'(\mathbf{A}_3^{k}\otimes\mathbf{A}_2^{k})\right\|_F\\
\le&\ \nu\left\|(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k+1})'\mathbf{x}_t\mathbf{x}_t'(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k+1})-(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k+1})'\mathbf{x}_t\mathbf{x}_t'(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k})\right\|_F\\
&+\nu\left\|(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k+1})'\mathbf{x}_t\mathbf{x}_t'(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k})-(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k+1})'\mathbf{x}_t\mathbf{x}_t'(\mathbf{A}_3^{k}\otimes\mathbf{A}_2^{k})\right\|_F\\
&+\nu\left\|(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k+1})'\mathbf{x}_t\mathbf{x}_t'(\mathbf{A}_3^{k}\otimes\mathbf{A}_2^{k})-(\mathbf{A}_3^{k}\otimes\mathbf{A}_2^{k})'\mathbf{x}_t\mathbf{x}_t'(\mathbf{A}_3^{k}\otimes\mathbf{A}_2^{k})\right\|_F,
\end{aligned}
\tag{S4.100}
$$

where the second inequality follows from the triangle inequality and $\|\mathcal{G}_{(1)}^k\|_F \le \nu := \sqrt{r_1r_2r_3}c$. Similar to (S4.75), we have

$$
\begin{aligned}
&\left\|(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k+1})'\mathbf{x}_t\mathbf{x}_t'(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k+1})-(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k+1})'\mathbf{x}_t\mathbf{x}_t'(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k})\right\|_F\\
=&\left\|(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k+1})'\mathbf{x}_t\mathbf{x}_t'\left(\mathbf{A}_3^{k+1}\otimes(\mathbf{A}_2^{k+1}-\mathbf{A}_2^{k})\right)\right\|_F\\
\le&\left\|\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k+1}\right\|\left\|\mathbf{x}_t\mathbf{x}_t'\left(\mathbf{A}_3^{k+1}\otimes(\mathbf{A}_2^{k+1}-\mathbf{A}_2^{k})\right)\right\|_F\\
\le&\left\|\mathbf{x}_t\mathbf{x}_t'\left(\mathbf{A}_3^{k+1}\otimes(\mathbf{A}_2^{k+1}-\mathbf{A}_2^{k})\right)\right\|_F\\
\le&\|\mathbf{x}_t\mathbf{x}_t'\|_F\|\mathbf{A}_3^{k+1}\otimes(\mathbf{A}_2^{k+1}-\mathbf{A}_2^{k})\|\\
=&\|\mathbf{x}_t\mathbf{x}_t'\|_F\|\mathbf{A}_3^{k+1}\|\|\mathbf{A}_2^{k+1}-\mathbf{A}_2^{k}\|\le\|\mathbf{x}_t\mathbf{x}_t'\|_F\|\mathbf{A}_2^{k+1}-\mathbf{A}_2^{k}\|_F.
\end{aligned}
\tag{S4.101}
$$

Similarly, we can get that

$$\begin{aligned}
&\left\|(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k+1})'\mathbf{x}_t\mathbf{x}_t'(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k})-(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k+1})'\mathbf{x}_t\mathbf{x}_t'(\mathbf{A}_3^{k}\otimes\mathbf{A}_2^{k})\right\|_F\\
\leq&\left\|\mathbf{x}_t\mathbf{x}_t'\right\|_F\left\|\mathbf{A}_3^{k+1}-\mathbf{A}_3^{k}\right\|_F,\\
&\left\|(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k+1})'\mathbf{x}_t\mathbf{x}_t'(\mathbf{A}_3^{k}\otimes\mathbf{A}_2^{k})-(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k})'\mathbf{x}_t\mathbf{x}_t'(\mathbf{A}_3^{k}\otimes\mathbf{A}_2^{k})\right\|_F\\
\leq&\left\|\mathbf{x}_t\mathbf{x}_t'\right\|_F\left\|\mathbf{A}_2^{k+1}-\mathbf{A}_2^{k}\right\|_F,\\
&\left\|(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k})'\mathbf{x}_t\mathbf{x}_t'(\mathbf{A}_3^{k}\otimes\mathbf{A}_2^{k})-(\mathbf{A}_3^{k}\otimes\mathbf{A}_2^{k})'\mathbf{x}_t\mathbf{x}_t'(\mathbf{A}_3^{k}\otimes\mathbf{A}_2^{k})\right\|_F\\
\leq&\left\|\mathbf{x}_t\mathbf{x}_t'\right\|_F\left\|\mathbf{A}_3^{k+1}-\mathbf{A}_3^{k}\right\|_F.
\end{aligned}\tag{S4.102}$$

Substituting (S4.101) and (S4.102) into (S4.100), we immediately establish

$$\begin{aligned}
&\left\|\mathcal{G}_{(1)}^{k}(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k+1})'\mathbf{x}_t\mathbf{x}_t'(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k+1})\right.\\
&\left.-\mathcal{G}_{(1)}^{k}(\mathbf{A}_3^{k}\otimes\mathbf{A}_2^{k})'\mathbf{x}_t\mathbf{x}_t'(\mathbf{A}_3^{k}\otimes\mathbf{A}_2^{k})\right\|_F\\
\leq&\ 2\nu\left\|\mathbf{x}_t\mathbf{x}_t'\right\|_F\left(\left\|\mathbf{A}_2^{k+1}-\mathbf{A}_2^{k}\right\|_F+\left\|\mathbf{A}_3^{k+1}-\mathbf{A}_3^{k}\right\|_F\right).
\end{aligned}\tag{S4.103}$$

Plugging (S4.103) and (S4.99) into (S4.98), we have

$$\begin{aligned}
&\left\|\mathcal{G}_{(1)}^{k+1}(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k+1})'\mathbf{x}_t\mathbf{x}_t'(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k+1})\right.\\
&\left.-\mathcal{G}_{(1)}^{k}(\mathbf{A}_3^{k}\otimes\mathbf{A}_2^{k})'\mathbf{x}_t\mathbf{x}_t'(\mathbf{A}_3^{k}\otimes\mathbf{A}_2^{k})\right\|_F\\
\leq&\left\|\mathcal{G}^{k+1}-\mathcal{G}^{k}\right\|_F\left\|\mathbf{x}_t\right\|_2^2+2\nu\left\|\mathbf{x}_t\mathbf{x}_t'\right\|_F\left(\left\|\mathbf{A}_2^{k+1}-\mathbf{A}_2^{k}\right\|_F+\left\|\mathbf{A}_3^{k+1}-\mathbf{A}_3^{k}\right\|_F\right).
\end{aligned}\tag{S4.104}$$

Summing up the above inequality over $t = 1, \ldots, T$, one can obtain that

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\Big\|\mathcal{G}_{(1)}^{k+1}(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k+1})'\mathbf{x}_t\mathbf{x}_t'(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k+1}) \\
&\qquad\qquad - \mathcal{G}_{(1)}^{k}(\mathbf{A}_3^{k}\otimes\mathbf{A}_2^{k})'\mathbf{x}_t\mathbf{x}_t'(\mathbf{A}_3^{k}\otimes\mathbf{A}_2^{k})\Big\|_F \\
&\le \frac{1}{T}\sum_{t=1}^{T}\|\mathcal{G}^{k+1}-\mathcal{G}^{k}\|_F\|\mathbf{x}_t\|_2^2 + 2\nu\frac{1}{T}\sum_{t=1}^{T}\|\mathbf{x}_t\|_2^2\|\mathbf{A}_2^{k+1}-\mathbf{A}_2^{k}\|_F \qquad \text{(S4.105)} \\
&\quad + 2\nu\frac{1}{T}\sum_{t=1}^{T}\|\mathbf{x}_t\|_2^2\|\mathbf{A}_3^{k+1}-\mathbf{A}_3^{k}\|_F \\
&= c_1\|\mathcal{G}^{k+1}-\mathcal{G}^{k}\|_F + 2\nu c_1\|\mathbf{A}_2^{k+1}-\mathbf{A}_2^{k}\|_F + 2\nu c_1\|\mathbf{A}_3^{k+1}-\mathbf{A}_3^{k}\|_F,
\end{aligned}$$

where

$$c_1 := \frac{1}{T}\sum_{t=1}^{T}\|\mathbf{x}_t\|_2^2. \qquad \text{(S4.106)}$$

Substituting (S4.105) and (S4.96) into (S4.92) establishes

$$\begin{aligned}
&\Big\|\big(\nabla_{\mathcal{G}}Q([\![\mathcal{G}^{k+1};\mathbf{A}_1^{k+1},\mathbf{A}_2^{k+1},\mathbf{A}_3^{k+1}]\!])\big)_{(1)} - \big(\nabla_{\mathcal{G}}Q([\![\mathcal{G}^{k};\mathbf{A}_1^{k},\mathbf{A}_2^{k},\mathbf{A}_3^{k}]\!])\big)_{(1)}\Big\|_F \\
&\le c_2\left(\|\mathbf{A}_1^{k}-\mathbf{A}_1^{k+1}\|_F + \|\mathbf{A}_2^{k}-\mathbf{A}_2^{k+1}\|_F + \|\mathbf{A}_3^{k}-\mathbf{A}_3^{k+1}\|_F\right) + c_1\|\mathcal{G}^{k+1}-\mathcal{G}^{k}\|_F \\
&\quad + 2\nu c_1\|\mathbf{A}_2^{k+1}-\mathbf{A}_2^{k}\|_F + 2\nu c_1\|\mathbf{A}_3^{k+1}-\mathbf{A}_3^{k}\|_F.
\end{aligned}$$

This taken together with (S4.90) leads to

$$\begin{aligned}
\|N_1\|_F &\le c_2(\|\mathbf{A}_1^{k}-\mathbf{A}_1^{k+1}\|_F + \|\mathbf{A}_2^{k}-\mathbf{A}_2^{k+1}\|_F + \|\mathbf{A}_3^{k}-\mathbf{A}_3^{k+1}\|_F) + c_1\|\mathcal{G}^{k+1}-\mathcal{G}^{k}\|_F \\
&\quad + 2\nu c_1\|\mathbf{A}_2^{k+1}-\mathbf{A}_2^{k}\|_F + 2\nu c_1\|\mathbf{A}_3^{k+1}-\mathbf{A}_3^{k}\|_F + \rho_1\|\mathcal{G}^{k+1}-\mathcal{G}^{k}\|_F \\
&= (\rho_1+c_1)\|\mathcal{G}^{k+1}-\mathcal{G}^{k}\|_F + c_2\|\mathbf{A}_1^{k+1}-\mathbf{A}_1^{k}\|_F \\
&\quad + (c_2+2\nu c_1)\|\mathbf{A}_2^{k+1}-\mathbf{A}_2^{k}\|_F + (c_2+2\nu c_1)\|\mathbf{A}_3^{k+1}-\mathbf{A}_3^{k}\|_F.
\end{aligned} \qquad \text{(S4.107)}$$

Similar to the above analysis about the upper bound of N_1, we can easily obtain that

$$
\begin{aligned}
&\|N_2\|_F \\
=&\big\|\nabla_{\mathbf{A}_1}\Psi\left(\mathcal{G}^{k+1},\mathbf{A}_1^{k+1},\mathbf{A}_2^{k+1},\mathbf{A}_3^{k+1},\mathbf{U}_1^{k+1},\mathbf{U}_2^{k+1},\mathbf{U}_3^{k+1}\right)\\
&-\nabla_{\mathbf{A}_1}\Psi\left(\mathcal{G}^{k+1},\mathbf{A}_1^{k},\mathbf{A}_2^{k},\mathbf{A}_3^{k},\mathbf{U}_1^{k},\mathbf{U}_2^{k},\mathbf{U}_3^{k}\right)-\rho_2(\mathbf{A}_1^{k+1}-\mathbf{A}_1^{k})\big\|_F\\
\le&\big\|\nabla_{\mathbf{A}_1}\Psi\left(\mathcal{G}^{k+1},\mathbf{A}_1^{k+1},\mathbf{A}_2^{k+1},\mathbf{A}_3^{k+1},\mathbf{U}_1^{k+1},\mathbf{U}_2^{k+1},\mathbf{U}_3^{k+1}\right)\\
&-\nabla_{\mathbf{A}_1}\Psi\left(\mathcal{G}^{k+1},\mathbf{A}_1^{k},\mathbf{A}_2^{k},\mathbf{A}_3^{k},\mathbf{U}_1^{k},\mathbf{U}_2^{k},\mathbf{U}_3^{k}\right)\big\|_F+\rho_2\|\mathbf{A}_1^{k+1}-\mathbf{A}_1^{k}\|_F\\
=&\big\|\nabla_{\mathbf{A}_1}Q([\![\mathcal{G}^{k+1};\mathbf{A}_1^{k+1},\mathbf{A}_2^{k+1},\mathbf{A}_3^{k+1}]\!])-\gamma_1(\mathbf{U}_1^{k+1}-\mathbf{A}_1^{k+1})\\
&-\nabla_{\mathbf{A}_1}Q([\![\mathcal{G}^{k+1};\mathbf{A}_1^{k},\mathbf{A}_2^{k},\mathbf{A}_3^{k}]\!])+\gamma_1(\mathbf{U}_1^{k}-\mathbf{A}_1^{k})\big\|_F+\rho_2\|\mathbf{A}_1^{k+1}-\mathbf{A}_1^{k}\|_F\\
\le&\big\|\nabla_{\mathbf{A}_1}Q([\![\mathcal{G}^{k+1};\mathbf{A}_1^{k+1},\mathbf{A}_2^{k+1},\mathbf{A}_3^{k+1}]\!])-\nabla_{\mathbf{A}_1}Q([\![\mathcal{G}^{k+1};\mathbf{A}_1^{k},\mathbf{A}_2^{k},\mathbf{A}_3^{k}]\!])\big\|_F\\
&+\gamma_1\|\mathbf{U}_1^{k+1}-\mathbf{U}_1^{k}\|_F+(\gamma_1+\rho_2)\|\mathbf{A}_1^{k+1}-\mathbf{A}_1^{k}\|_F.
\end{aligned}
\tag{S4.108}
$$

where the second equality follows from (S3.57) and the last inequality follows from the triangle inequality. Taking (S3.56) and (S4.65) together yields

$$
\begin{aligned}
&\big\|\nabla_{\mathbf{A}_1}Q([\![\mathcal{G}^{k+1};\mathbf{A}_1^{k+1},\mathbf{A}_2^{k+1},\mathbf{A}_3^{k+1}]\!])-\nabla_{\mathbf{A}_1}Q([\![\mathcal{G}^{k+1};\mathbf{A}_1^{k},\mathbf{A}_2^{k},\mathbf{A}_3^{k}]\!])\big\|_F\\
=&\Big\|(\nabla Q([\![\mathcal{G}^{k+1};\mathbf{A}_1^{k+1},\mathbf{A}_2^{k+1},\mathbf{A}_3^{k+1}]\!]))_{(1)}(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k+1})(\mathcal{G}_{(1)}^{k+1})'\\
&-(\nabla Q([\![\mathcal{G}^{k+1};\mathbf{A}_1^{k},\mathbf{A}_2^{k},\mathbf{A}_3^{k}]\!]))_{(1)}(\mathbf{A}_3^{k}\otimes\mathbf{A}_2^{k})(\mathcal{G}_{(1)}^{k+1})'\Big\|_F.
\end{aligned}
$$

By Lemma 7, we obtain

$$
\begin{aligned}
&\Big\|(\nabla Q([\![\mathcal{G}^{k+1};\mathbf{A}_1^{k+1},\mathbf{A}_2^{k+1},\mathbf{A}_3^{k+1}]\!]))_{(1)}(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k+1})(\mathcal{G}_{(1)}^{k+1})'\\
&\quad-(\nabla Q([\![\mathcal{G}^{k+1};\mathbf{A}_1^{k},\mathbf{A}_2^{k},\mathbf{A}_3^{k}]\!]))_{(1)}(\mathbf{A}_3^{k}\otimes\mathbf{A}_2^{k})(\mathcal{G}_{(1)}^{k+1})'\Big\|_F\\
=&\Big\|\frac{1}{T}\sum_{t=1}^{T}\Big(\mathbf{A}_1^{k+1}\mathcal{G}_{(1)}^{k+1}(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k+1})'\mathbf{x}_t-\mathbf{y}_t\Big)\mathbf{x}_t'(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k+1})(\mathcal{G}_{(1)}^{k+1})'\\
&\quad-\frac{1}{T}\sum_{t=1}^{T}\Big(\mathbf{A}_1^{k}\mathcal{G}_{(1)}^{k+1}(\mathbf{A}_3^{k}\otimes\mathbf{A}_2^{k})'\mathbf{x}_t-\mathbf{y}_t\Big)\mathbf{x}_t'(\mathbf{A}_3^{k}\otimes\mathbf{A}_2^{k})(\mathcal{G}_{(1)}^{k+1})'\Big\|_F\\
\le&\frac{1}{T}\sum_{t=1}^{T}\Big\|\mathbf{A}_1^{k+1}\mathcal{G}_{(1)}^{k+1}(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k+1})'\mathbf{x}_t\mathbf{x}_t'(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k+1})(\mathcal{G}_{(1)}^{k+1})'\\
&\quad-\mathbf{y}_t\mathbf{x}_t'(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k+1})(\mathcal{G}_{(1)}^{k+1})'+\mathbf{y}_t\mathbf{x}_t'(\mathbf{A}_3^{k}\otimes\mathbf{A}_2^{k})(\mathcal{G}_{(1)}^{k+1})'\\
&\quad-\mathbf{A}_1^{k}\mathcal{G}_{(1)}^{k+1}(\mathbf{A}_3^{k}\otimes\mathbf{A}_2^{k})'\mathbf{x}_t\mathbf{x}_t'(\mathbf{A}_3^{k}\otimes\mathbf{A}_2^{k})(\mathcal{G}_{(1)}^{k+1})'\Big\|_F\\
\le&\frac{1}{T}\sum_{t=1}^{T}\Big\|\mathbf{A}_1^{k+1}\mathcal{G}_{(1)}^{k+1}(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k+1})'\mathbf{x}_t\mathbf{x}_t'(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k+1})(\mathcal{G}_{(1)}^{k+1})'\\
&\quad-\mathbf{A}_1^{k}\mathcal{G}_{(1)}^{k+1}(\mathbf{A}_3^{k}\otimes\mathbf{A}_2^{k})'\mathbf{x}_t\mathbf{x}_t'(\mathbf{A}_3^{k}\otimes\mathbf{A}_2^{k})(\mathcal{G}_{(1)}^{k+1})'\Big\|_F\\
&\quad+\frac{1}{T}\sum_{t=1}^{T}\Big\|\mathbf{y}_t\mathbf{x}_t'\Big((\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k+1})(\mathcal{G}_{(1)}^{k+1})'-(\mathbf{A}_3^{k}\otimes\mathbf{A}_2^{k})(\mathcal{G}_{(1)}^{k+1})'\Big)\Big\|_F\\
=&\frac{1}{T}\sum_{t=1}^{T}\Big\|\mathbf{A}_1^{k+1}\mathcal{G}_{(1)}^{k+1}(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k+1})'\mathbf{x}_t\mathbf{x}_t'(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k+1})(\mathcal{G}_{(1)}^{k+1})'\\
&\quad-\mathbf{A}_1^{k}\mathcal{G}_{(1)}^{k+1}(\mathbf{A}_3^{k}\otimes\mathbf{A}_2^{k})'\mathbf{x}_t\mathbf{x}_t'(\mathbf{A}_3^{k}\otimes\mathbf{A}_2^{k})(\mathcal{G}_{(1)}^{k+1})'\Big\|_F\\
&\quad+\frac{1}{T}\sum_{t=1}^{T}\Big\|\mathbf{y}_t\mathbf{x}_t'\Big((\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k+1})-(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k})\\
&\quad+(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k})-(\mathbf{A}_3^{k}\otimes\mathbf{A}_2^{k})\Big)\mathcal{G}_{(1)}^{k+1})'\Big\|_F.
\end{aligned}
\tag{S4.109}
$$

First, let us consider the second term on the right-hand side of the above

inequality. Note that

$$\begin{aligned}
&\Big\|\mathbf{y}_t\mathbf{x}_t'\Big((\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k+1})-(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k})+(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k})\\
&\quad-(\mathbf{A}_3^{k}\otimes\mathbf{A}_2^{k})\Big)\mathcal{G}_{(1)}^{k+1)'}\Big\|_F\\
\le&\Big\|\mathbf{y}_t\mathbf{x}_t'\left((\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k+1})-(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k}))\right)\mathcal{G}_{(1)}^{k+1)'}\Big\|_F\\
&+\Big\|\mathbf{y}_t\mathbf{x}_t'\left((\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k})-(\mathbf{A}_3^{k}\otimes\mathbf{A}_2^{k}))\right)\mathcal{G}_{(1)}^{k+1)'}\Big\|_F\\
=&\Big\|\mathbf{y}_t\mathbf{x}_t'\left(\mathbf{A}_3^{k+1}\otimes(\mathbf{A}_2^{k+1}-\mathbf{A}_2^{k})\right)(\mathcal{G}_{(1)}^{k+1)'}\Big\|_F\\
&+\Big\|\mathbf{y}_t\mathbf{x}_t'\left((\mathbf{A}_3^{k+1}-\mathbf{A}_3^{k})\otimes\mathbf{A}_2^{k}\right)(\mathcal{G}_{(1)}^{k+1)'}\Big\|_F.
\end{aligned}\tag{S4.110}$$

By using the same discussion in (S4.75), we can obtain that

$$\begin{aligned}
&\Big\|\mathbf{y}_t\mathbf{x}_t'\left(\mathbf{A}_3^{k+1}\otimes(\mathbf{A}_2^{k+1}-\mathbf{A}_2^{k})\right)(\mathcal{G}_{(1)}^{k+1)'}\Big\|_F\le\nu\left\|\mathbf{y}_t\mathbf{x}_t'\right\|_F\|\mathbf{A}_2^{k+1}-\mathbf{A}_2^{k}\|_F,\\
&\Big\|\mathbf{y}_t\mathbf{x}_t'\left((\mathbf{A}_3^{k+1}-\mathbf{A}_3^{k})\otimes\mathbf{A}_2^{k}\right)(\mathcal{G}_{(1)}^{k+1)'}\Big\|_F\le\nu\left\|\mathbf{y}_t\mathbf{x}_t'\right\|_F\|\mathbf{A}_3^{k+1}-\mathbf{A}_3^{k}\|_F.
\end{aligned}\tag{S4.111}$$

Substituting (S4.111) into (S4.110) establishes

$$\begin{aligned}
&\Big\|\mathbf{y}_t\mathbf{x}_t'\left((\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k+1})(\mathcal{G}_{(1)}^{k+1)'}-(\mathbf{A}_3^{k}\otimes\mathbf{A}_2^{k})(\mathcal{G}_{(1)}^{k+1)'}\right)\Big\|_F\\
\le&\,\nu\left\|\mathbf{y}_t\mathbf{x}_t'\right\|_F\|\mathbf{A}_2^{k+1}-\mathbf{A}_2^{k}\|_F+\nu\left\|\mathbf{y}_t\mathbf{x}_t'\right\|_F\|\mathbf{A}_3^{k+1}-\mathbf{A}_3^{k}\|_F,
\end{aligned}$$

which indicates that

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\Big\|\mathbf{y}_t\mathbf{x}_t'\left((\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k+1})(\mathcal{G}_{(1)}^{k+1)'}-(\mathbf{A}_3^{k}\otimes\mathbf{A}_2^{k})(\mathcal{G}_{(1)}^{k+1)'}\right)\Big\|_F\\
\le&\frac{1}{T}\sum_{t=1}^{T}\nu\left\|\mathbf{y}_t\mathbf{x}_t'\right\|_F\|\mathbf{A}_2^{k+1}-\mathbf{A}_2^{k}\|_F+\frac{1}{T}\sum_{t=1}^{T}\nu\left\|\mathbf{y}_t\mathbf{x}_t'\right\|_F\|\mathbf{A}_3^{k+1}-\mathbf{A}_3^{k}\|_F\\
=&\,c_2\nu\|\mathbf{A}_2^{k+1}-\mathbf{A}_2^{k}\|_F+c_2\nu\|\mathbf{A}_3^{k+1}-\mathbf{A}_3^{k}\|_F,
\end{aligned}\tag{S4.112}$$

where c_2 is defined in (S4.97).

For the first term on the right-hand side of (S4.109), a similar technique as that of (S4.98) yields

$$\begin{aligned}
&\Big\|\mathbf{A}_1^{k+1}\mathcal{G}_{(1)}^{k+1}(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k+1})'\mathbf{x}_t\mathbf{x}_t'(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k+1})(\mathcal{G}_{(1)}^{k+1})' \\
&\quad - \mathbf{A}_1^{k}\mathcal{G}_{(1)}^{k+1}(\mathbf{A}_3^{k}\otimes\mathbf{A}_2^{k})'\mathbf{x}_t\mathbf{x}_t'(\mathbf{A}_3^{k}\otimes\mathbf{A}_2^{k})(\mathcal{G}_{(1)}^{k+1})'\Big\|_F \\
&\le \|\mathbf{A}_1^{k+1}-\mathbf{A}_1^{k}\|_F\nu^2\left\|\mathbf{x}_t\mathbf{x}_t'\right\|_F + \nu^2\left\|\mathbf{x}_t\mathbf{x}_t'\right\|_F(2\|\mathbf{A}_2^{k+1}-\mathbf{A}_2^{k}\|_F + 2\|\mathbf{A}_3^{k+1}-\mathbf{A}_3^{k}\|_F).
\end{aligned}$$

Summing up the above inequality over $t=1,\ldots,T$, one can easily obtain that

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\Big\|\mathbf{A}_1^{k+1}\mathcal{G}_{(1)}^{k+1}(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k+1})'\mathbf{x}_t\mathbf{x}_t'(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k+1})(\mathcal{G}_{(1)}^{k+1})' \\
&\quad - \mathbf{A}_1^{k}\mathcal{G}_{(1)}^{k+1}(\mathbf{A}_3^{k}\otimes\mathbf{A}_2^{k})'\mathbf{x}_t\mathbf{x}_t'(\mathbf{A}_3^{k}\otimes\mathbf{A}_2^{k})(\mathcal{G}_{(1)}^{k+1})'\Big\|_F \\
&\le \frac{1}{T}\sum_{t=1}^{T}\|\mathbf{A}_1^{k+1}-\mathbf{A}_1^{k}\|_F\nu^2\left\|\mathbf{x}_t\mathbf{x}_t'\right\|_F + \frac{1}{T}\sum_{t=1}^{T}2\nu^2\left\|\mathbf{x}_t\mathbf{x}_t'\right\|_F\|\mathbf{A}_2^{k+1}-\mathbf{A}_2^{k}\|_F \\
&\quad + \frac{1}{T}\sum_{t=1}^{T}2\nu^2\left\|\mathbf{x}_t\mathbf{x}_t'\right\|_F\|\mathbf{A}_3^{k+1}-\mathbf{A}_3^{k}\|_F \\
&= \nu^2c_1\|\mathbf{A}_1^{k+1}-\mathbf{A}_1^{k}\|_F + 2\nu^2c_1\|\mathbf{A}_2^{k+1}-\mathbf{A}_2^{k}\|_F + 2\nu^2c_1\|\mathbf{A}_3^{k+1}-\mathbf{A}_3^{k}\|_F,
\end{aligned} \tag{S4.113}$$

where c_1 is defined in (S4.106). Plugging (S4.113) and (S4.112) into (S4.109), we conclude that

$$\begin{aligned}
&\left\|\nabla_{\mathbf{A}_1}Q([\![\mathcal{G}^{k+1};\mathbf{A}_1^{k+1},\mathbf{A}_2^{k+1},\mathbf{A}_3^{k+1}]\!]) - \nabla_{\mathbf{A}_1}Q([\![\mathcal{G}^{k+1};\mathbf{A}_1^{k},\mathbf{A}_2^{k},\mathbf{A}_3^{k}]\!])\right\|_F \\
&\le \nu c_2\|\mathbf{A}_2^{k+1}-\mathbf{A}_2^{k}\|_F + \nu c_2\|\mathbf{A}_3^{k+1}-\mathbf{A}_3^{k}\|_F + \nu^2c_1\|\mathbf{A}_1^{k+1}-\mathbf{A}_1^{k}\|_F \\
&\quad + 2\nu^2c_1\|\mathbf{A}_2^{k+1}-\mathbf{A}_2^{k}\|_F + 2\nu^2c_1\|\mathbf{A}_3^{k+1}-\mathbf{A}_3^{k}\|_F.
\end{aligned}$$

Combining the above inequality and (S4.108), we arrive at

$$
\begin{aligned}
&\|N_2\|_F \\
=\ &\big\|\nabla_{\mathbf{A}_1}\Psi(\mathcal{G}^{k+1},\mathbf{A}_1^{k+1},\mathbf{A}_2^{k+1},\mathbf{A}_3^{k+1},\mathbf{U}_1^{k+1},\mathbf{U}_2^{k+1},\mathbf{U}_3^{k+1}) \\
&-\nabla_{\mathbf{A}_1}\Psi(\mathcal{G}^{k+1},\mathbf{A}_1^{k},\mathbf{A}_2^{k},\mathbf{A}_3^{k},\mathbf{U}_1^{k},\mathbf{U}_2^{k},\mathbf{U}_3^{k})-\rho_2(\mathbf{A}_1^{k+1}-\mathbf{A}_1^{k})\big\|_F \\
\le\ &\nu c_2\|\mathbf{A}_2^{k+1}-\mathbf{A}_2^{k}\|_F+\nu c_2\|\mathbf{A}_3^{k+1}-\mathbf{A}_3^{k}\|_F+\nu^2 c_1\|\mathbf{A}_1^{k+1}-\mathbf{A}_1^{k}\|_F \\
&+2\nu^2 c_1\|\mathbf{A}_2^{k+1}-\mathbf{A}_2^{k}\|_F+2\nu^2 c_1\|\mathbf{A}_3^{k+1}-\mathbf{A}_3^{k}\|_F+\gamma_1\|\mathbf{U}_1^{k+1}-\mathbf{U}_1^{k}\|_F \\
&+(\rho_2+\gamma_1)\|\mathbf{A}_1^{k+1}-\mathbf{A}_1^{k}\|_F \\
=\ &(\nu^2 c_1+\rho_2+\gamma_1)\|\mathbf{A}_1^{k+1}-\mathbf{A}_1^{k}\|_F+(\nu c_2+2\nu^2 c_1)\|\mathbf{A}_2^{k+1}-\mathbf{A}_2^{k}\|_F \\
&+(\nu c_2+2\nu^2 c_1)\|\mathbf{A}_3^{k+1}-\mathbf{A}_3^{k}\|_F+\gamma_1\|\mathbf{U}_1^{k+1}-\mathbf{U}_1^{k}\|_F.
\end{aligned}
\tag{S4.114}
$$

Similar to the analysis of N_2, we have

$$
\begin{aligned}
&\|N_3\|_F \\
=\ &\big\|\nabla_{\mathbf{A}_2}\Psi(\mathcal{G}^{k+1},\mathbf{A}_1^{k+1},\mathbf{A}_2^{k+1},\mathbf{A}_3^{k+1},\mathbf{U}_1^{k+1},\mathbf{U}_2^{k+1},\mathbf{U}_3^{k+1}) \\
&-\nabla_{\mathbf{A}_2}\Psi(\mathcal{G}^{k+1},\mathbf{A}_1^{k+1},\mathbf{A}_2^{k},\mathbf{A}_3^{k},\mathbf{U}_1^{k},\mathbf{U}_2^{k},\mathbf{U}_3^{k})-\rho_3(\mathbf{A}_2^{k+1}-\mathbf{A}_2^{k})\big\|_F \\
\le\ &\big\|\nabla_{\mathbf{A}_2}Q(\llbracket\mathcal{G}^{k+1};\mathbf{A}_1^{k+1},\mathbf{A}_2^{k+1},\mathbf{A}_3^{k+1}\rrbracket)-\nabla_{\mathbf{A}_2}Q(\llbracket\mathcal{G}^{k+1};\mathbf{A}_1^{k+1},\mathbf{A}_2^{k},\mathbf{A}_3^{k}\rrbracket)\big\|_F \\
&+\gamma_2\|\mathbf{U}_2^{k+1}-\mathbf{U}_2^{k}\|_F+(\gamma_2+\rho_3)\|\mathbf{A}_2^{k+1}-\mathbf{A}_2^{k}\|_F.
\end{aligned}
\tag{S4.115}
$$

Note that (S3.56) together with (S3.55) yields

$$
\begin{aligned}
&\left\|\nabla_{\mathbf{A}_2} Q([\![\mathcal{G}^{k+1};\mathbf{A}_1^{k+1},\mathbf{A}_2^{k+1},\mathbf{A}_3^{k+1}]\!]) - \nabla_{\mathbf{A}_2} Q([\![\mathcal{G}^{k+1};\mathbf{A}_1^{k+1},\mathbf{A}_2^{k},\mathbf{A}_3^{k}]\!])\right\|_F \\
&= \Bigg\| \left(\frac{1}{T}\sum_{t=1}^{T}(\mathbf{A}_1^{k+1}\mathcal{G}_{(1)}^{k+1}(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k+1})'\mathbf{x}_t - \mathbf{y}_t)\circ\mathbf{X}_t\right)_{(2)}(\mathbf{A}_1^{k+1}\otimes\mathbf{A}_3^{k+1})(\mathcal{G}_{(2)}^{k+1})' \\
&\quad - \left(\frac{1}{T}\sum_{t=1}^{T}(\mathbf{A}_1^{k+1}\mathcal{G}_{(1)}^{k+1}(\mathbf{A}_3^{k}\otimes\mathbf{A}_2^{k})'\mathbf{x}_t - \mathbf{y}_t)\circ\mathbf{X}_t\right)_{(2)}(\mathbf{A}_1^{k+1}\otimes\mathbf{A}_3^{k})(\mathcal{G}_{(2)}^{k+1})'\Bigg\|_F \\
&= \Bigg\| \frac{1}{T}\sum_{t=1}^{T}\mathbf{X}_t\otimes\left(\mathbf{A}_1^{k+1}\mathcal{G}_{(1)}^{k+1}(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k+1})'\mathbf{x}_t - \mathbf{y}_t\right)'(\mathbf{A}_1^{k+1}\otimes\mathbf{A}_3^{k+1})(\mathcal{G}_{(2)}^{k+1})' \\
&\quad - \frac{1}{T}\sum_{t=1}^{T}\mathbf{X}_t\otimes\left(\mathbf{A}_1^{k+1}\mathcal{G}_{(1)}^{k+1}(\mathbf{A}_3^{k}\otimes\mathbf{A}_2^{k})'\mathbf{x}_t - \mathbf{y}_t\right)'(\mathbf{A}_1^{k+1}\otimes\mathbf{A}_3^{k})(\mathcal{G}_{(2)}^{k+1})'\Bigg\|_F \\
&\leq \frac{1}{T}\sum_{t=1}^{T}\Bigg\|\mathbf{X}_t\otimes\left(\mathbf{A}_1^{k+1}\mathcal{G}_{(1)}^{k+1}(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k+1})'\mathbf{x}_t - \mathbf{y}_t\right)'(\mathbf{A}_1^{k+1}\otimes\mathbf{A}_3^{k+1})(\mathcal{G}_{(2)}^{k+1})' \\
&\quad - \mathbf{X}_t\otimes\left(\mathbf{A}_1^{k+1}\mathcal{G}_{(1)}^{k+1}(\mathbf{A}_3^{k}\otimes\mathbf{A}_2^{k})'\mathbf{x}_t - \mathbf{y}_t\right)'(\mathbf{A}_1^{k+1}\otimes\mathbf{A}_3^{k})(\mathcal{G}_{(2)}^{k+1})'\Bigg\|_F \\
&= \frac{1}{T}\sum_{t=1}^{T}\Bigg\|\mathbf{X}_t\otimes\left(\mathbf{A}_1^{k+1}\mathcal{G}_{(1)}^{k+1}(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k+1})'\mathbf{x}_t - \mathbf{y}_t\right)'\mathbf{A}_1^{k+1}\otimes\mathbf{A}_3^{k+1})(\mathcal{G}_{(2)}^{k+1})' \\
&\quad - \mathbf{X}_t\otimes\left(\mathbf{A}_1^{k+1}\mathcal{G}_{(1)}^{k+1}(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k})'\mathbf{x}_t - \mathbf{y}_t\right)'(\mathbf{A}_1^{k+1}\otimes\mathbf{A}_3^{k+1})(\mathcal{G}_{(2)}^{k+1})' \\
&\quad + \mathbf{X}_t\otimes\left(\mathbf{A}_1^{k+1}\mathcal{G}_{(1)}^{k+1}(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k})'\mathbf{x}_t - \mathbf{y}_t\right)'(\mathbf{A}_1^{k+1}\otimes\mathbf{A}_3^{k+1})(\mathcal{G}_{(2)}^{k+1})' \\
&\quad - \mathbf{X}_t\otimes\left(\mathbf{A}_1^{k+1}\mathcal{G}_{(1)}^{k+1}(\mathbf{A}_3^{k}\otimes\mathbf{A}_2^{k})'\mathbf{x}_t - \mathbf{y}_t\right)'(\mathbf{A}_1^{k+1}\otimes\mathbf{A}_3^{k+1})(\mathcal{G}_{(2)}^{k+1})' \\
&\quad + \mathbf{X}_t\otimes\left(\mathbf{A}_1^{k+1}\mathcal{G}_{(1)}^{k+1}(\mathbf{A}_3^{k}\otimes\mathbf{A}_2^{k})'\mathbf{x}_t - \mathbf{y}_t\right)'(\mathbf{A}_1^{k+1}\otimes\mathbf{A}_3^{k+1})(\mathcal{G}_{(2)}^{k+1})' \\
&\quad - \mathbf{X}_t\otimes\left(\mathbf{A}_1^{k+1}\mathcal{G}_{(1)}^{k+1}(\mathbf{A}_3^{k}\otimes\mathbf{A}_2^{k})'\mathbf{x}_t - \mathbf{y}_t\right)'(\mathbf{A}_1^{k+1}\otimes\mathbf{A}_3^{k})(\mathcal{G}_{(2)}^{k+1})'\Bigg\|_F,
\end{aligned}
\tag{S4.116}
$$

where the second equality holds by Proposition 3. Then we have

$$
\begin{aligned}
&\left\|\nabla_{\mathbf{A}_2} Q([\![\mathcal{G}^{k+1};\mathbf{A}_1^{k+1},\mathbf{A}_2^{k+1},\mathbf{A}_3^{k+1}]\!]) - \nabla_{\mathbf{A}_2} Q([\![\mathcal{G}^{k+1};\mathbf{A}_1^{k+1},\mathbf{A}_2^{k},\mathbf{A}_3^{k}]\!])\right\|_F \\
\leq\ & \frac{1}{T}\sum_{t=1}^{T}\Big\|\mathbf{X}_t\otimes(\mathbf{A}_1^{k+1}\mathcal{G}_{(1)}^{k+1}(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k+1})'\mathbf{x}_t-\mathbf{y}_t)'\mathbf{A}_1^{k+1}\otimes\mathbf{A}_3^{k+1})(\mathcal{G}_{(2)}^{k+1})' \\
&\quad-\mathbf{X}_t\otimes(\mathbf{A}_1^{k+1}\mathcal{G}_{(1)}^{k+1}(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k})'\mathbf{x}_t-\mathbf{y}_t)'(\mathbf{A}_1^{k+1}\otimes\mathbf{A}_3^{k+1})(\mathcal{G}_{(2)}^{k+1})'\Big\|_F \\
&\quad+\frac{1}{T}\sum_{t=1}^{T}\Big\|\mathbf{X}_t\otimes(\mathbf{A}_1^{k+1}\mathcal{G}_{(1)}^{k+1}(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k})'\mathbf{x}_t-\mathbf{y}_t)'(\mathbf{A}_1^{k+1}\otimes\mathbf{A}_3^{k+1})(\mathcal{G}_{(2)}^{k+1})' \\
&\quad-\mathbf{X}_t\otimes(\mathbf{A}_1^{k+1}\mathcal{G}_{(1)}^{k+1}(\mathbf{A}_3^{k}\otimes\mathbf{A}_2^{k})'\mathbf{x}_t-\mathbf{y}_t)'(\mathbf{A}_1^{k+1}\otimes\mathbf{A}_3^{k+1})(\mathcal{G}_{(2)}^{k+1})'\Big\|_F \\
&\quad+\frac{1}{T}\sum_{t=1}^{T}\Big\|\mathbf{X}_t\otimes(\mathbf{A}_1^{k+1}\mathcal{G}_{(1)}^{k+1}(\mathbf{A}_3^{k}\otimes\mathbf{A}_2^{k})'\mathbf{x}_t-\mathbf{y}_t)'(\mathbf{A}_1^{k+1}\otimes\mathbf{A}_3^{k+1})(\mathcal{G}_{(2)}^{k+1})' \\
&\quad-\mathbf{X}_t\otimes(\mathbf{A}_1^{k+1}\mathcal{G}_{(1)}^{k+1}(\mathbf{A}_3^{k}\otimes\mathbf{A}_2^{k})'\mathbf{x}_t-\mathbf{y}_t)'(\mathbf{A}_1^{k+1}\otimes\mathbf{A}_3^{k})(\mathcal{G}_{(2)}^{k+1})'\Big\|_F \\
\leq\ & \frac{1}{T}\sum_{t=1}^{T}\Big\|\mathbf{X}_t\otimes\Big(\mathbf{A}_1^{k+1}\mathcal{G}_{(1)}^{k+1}(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k+1})'\mathbf{x}_t-\mathbf{y}_t\Big)'\mathbf{A}_1^{k+1}\otimes\mathbf{A}_3^{k+1})(\mathcal{G}_{(2)}^{k+1})' \\
&\quad-\mathbf{X}_t\otimes\Big(\mathbf{A}_1^{k+1}\mathcal{G}_{(1)}^{k+1}(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k})'\mathbf{x}_t-\mathbf{y}_t\Big)'(\mathbf{A}_1^{k+1}\otimes\mathbf{A}_3^{k+1})(\mathcal{G}_{(2)}^{k+1})'\Big\|_F \\
&\quad+\frac{1}{T}\sum_{t=1}^{T}\Big\|\mathbf{X}_t\otimes\Big(\mathbf{A}_1^{k+1}\mathcal{G}_{(1)}^{k+1}(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k})'\mathbf{x}_t-\mathbf{y}_t\Big)'(\mathbf{A}_1^{k+1}\otimes\mathbf{A}_3^{k+1})(\mathcal{G}_{(2)}^{k+1})' \\
&\quad-\mathbf{X}_t\otimes\Big(\mathbf{A}_1^{k+1}\mathcal{G}_{(1)}^{k+1}(\mathbf{A}_3^{k}\otimes\mathbf{A}_2^{k})'\mathbf{x}_t-\mathbf{y}_t\Big)'(\mathbf{A}_1^{k+1}\otimes\mathbf{A}_3^{k+1})(\mathcal{G}_{(2)}^{k+1})'\Big\|_F \\
&\quad+\frac{1}{T}\sum_{t=1}^{T}\Big\|\mathbf{X}_t\otimes\Big(\mathbf{A}_1^{k+1}\mathcal{G}_{(1)}^{k+1}(\mathbf{A}_3^{k}\otimes\mathbf{A}_2^{k})'\mathbf{x}_t-\mathbf{y}_t\Big)'(\mathbf{A}_1^{k+1}\otimes\mathbf{A}_3^{k+1})(\mathcal{G}_{(2)}^{k+1})' \\
&\quad-\mathbf{X}_t\otimes\Big(\mathbf{A}_1^{k+1}\mathcal{G}_{(1)}^{k+1}(\mathbf{A}_3^{k}\otimes\mathbf{A}_2^{k})'\mathbf{x}_t-\mathbf{y}_t\Big)'(\mathbf{A}_1^{k+1}\otimes\mathbf{A}_3^{k})(\mathcal{G}_{(2)}^{k+1})'\Big\|_F .
\end{aligned}
\tag{S4.117}
$$

Now we focus on the first term on the right-hand side of the above inequal-

ity.

$$
\begin{aligned}
&\Big\|\mathbf{X}_t \otimes \Big(\mathbf{A}_1^{k+1}\mathcal{G}_{(1)}^{k+1}(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k+1})'\mathbf{x}_t - \mathbf{y}_t\Big)'(\mathbf{A}_1^{k+1}\otimes\mathbf{A}_3^{k+1})(\mathcal{G}_{(2)}^{k+1})' \\
&\quad - \mathbf{X}_t \otimes \Big(\mathbf{A}_1^{k+1}\mathcal{G}_{(1)}^{k+1}(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k}\,)'\mathbf{x}_t - \mathbf{y}_t\Big)'(\mathbf{A}_1^{k+1}\otimes\mathbf{A}_3^{k+1})(\mathcal{G}_{(2)}^{k+1})'\Big\|_F \\
\le\ &\Big\|\mathbf{X}_t \otimes \Big(\mathbf{A}_1^{k+1}\mathcal{G}_{(1)}^{k+1}(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k+1})'\mathbf{x}_t - \mathbf{y}_t\Big)' \\
&\quad - \mathbf{X}_t \otimes \Big(\mathbf{A}_1^{k+1}\mathcal{G}_{(1)}^{k+1}(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k}\,)'\mathbf{x}_t - \mathbf{y}_t\Big)'\Big\|_F \Big\|(\mathbf{A}_1^{k+1}\otimes\mathbf{A}_3^{k+1})(\mathcal{G}_{(2)}^{k+1})'\Big\|_F \\
\le\ &\Big\|\mathbf{X}_t \otimes \Big(\mathbf{A}_1^{k+1}\mathcal{G}_{(1)}^{k+1}(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k+1})'\mathbf{x}_t - \mathbf{y}_t\Big)' \\
&\quad - \mathbf{X}_t \otimes \Big(\mathbf{A}_1^{k+1}\mathcal{G}_{(1)}^{k+1}(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k}\,)'\mathbf{x}_t - \mathbf{y}_t\Big)'\Big\|_F \nu \\
=\ &\nu\Big\|\mathbf{X}_t \otimes \Big(\mathbf{x}_t'(\mathbf{A}_3^{k+1}\otimes(\mathbf{A}_2^{k+1}-\mathbf{A}_2^{k}))(\mathcal{G}_{(1)}^{k+1})'(\mathbf{A}_1^{k+1})'\Big)\Big\|_F,
\end{aligned}
$$

where the second inequality follows from (S4.69). Using (S4.75) further leads to

$$
\begin{aligned}
&\nu\Big\|\mathbf{X}_t \otimes \Big(\mathbf{x}_t'\left(\mathbf{A}_3^{k+1}\otimes(\mathbf{A}_2^{k+1}-\mathbf{A}_2^{k})\right)(\mathcal{G}_{(1)}^{k+1})'(\mathbf{A}_1^{k+1})'\Big)\Big\|_F \\
=\ &\nu\left\|\mathbf{X}_t\right\|_F\Big\|\Big(\mathbf{x}_t'(\mathbf{A}_3^{k+1}\otimes(\mathbf{A}_2^{k+1}-\mathbf{A}_2^{k}))(\mathcal{G}_{(1)}^{k+1})'(\mathbf{A}_1^{k+1})'\Big)\Big\|_F \\
\le\ &\nu\left\|\mathbf{X}_t\right\|_F\left\|\mathbf{x}_t\right\|_2\nu\|\mathbf{A}_2^{k+1}-\mathbf{A}_2^{k}\|_F = \ \nu^2\left\|\mathbf{x}_t\right\|_2^2\|\mathbf{A}_2^{k+1}-\mathbf{A}_2^{k}\|_F,
\end{aligned}
$$

where the first equality holds by Proposition 1 and the second equality follows from the fact that $\|\mathbf{x}_t\|_2 = \|\mathbf{X}_t\|_F$. Summing the aforementioned inequality from $t+1$ to T yields

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\nu\Big\|\mathbf{X}_t \otimes \Big(\mathbf{x}_t'\left(\mathbf{A}_3^{k+1}\otimes(\mathbf{A}_2^{k+1}-\mathbf{A}_2^{k})\right)(\mathcal{G}_{(1)}^{k+1})'(\mathbf{A}_1^{k+1})'\Big)\Big\|_F \\
\le\ &\frac{1}{T}\sum_{t=1}^{T}\nu^2\left\|\mathbf{x}_t\right\|_2^2\|\mathbf{A}_2^{k+1}-\mathbf{A}_2^{k}\|_F = \ \nu^2 c_1\|\mathbf{A}_2^{k+1}-\mathbf{A}_2^{k}\|_F,
\end{aligned}
\tag{S4.118}
$$

where c_1 is defined in (S4.106). Similarly, we can get that

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\Big\|\mathbf{X}_t\otimes\Big(\mathbf{A}_1^{k+1}\mathcal{G}_{(1)}^{k+1}(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k})'\mathbf{x}_t-\mathbf{y}_t\Big)'(\mathbf{A}_1^{k+1}\otimes\mathbf{A}_3^{k+1})(\mathcal{G}_{(2)}^{k+1})'\\
&-\mathbf{X}_t\otimes\Big(\mathbf{A}_1^{k+1}\mathcal{G}_{(1)}^{k+1}(\mathbf{A}_3^{k}\otimes\mathbf{A}_2^{k})'\mathbf{x}_t-\mathbf{y}_t\Big)'(\mathbf{A}_1^{k+1}\otimes\mathbf{A}_3^{k+1})(\mathcal{G}_{(2)}^{k+1})'\Big\|_F\\
&\le\nu^2c_1\|\mathbf{A}_3^{k+1}-\mathbf{A}_3^{k}\|_F,\\
&\frac{1}{T}\sum_{t=1}^{T}\Big\|\mathbf{X}_t\otimes\Big(\mathbf{A}_1^{k+1}\mathcal{G}_{(1)}^{k+1}(\mathbf{A}_3^{k}\otimes\mathbf{A}_2^{k})'\mathbf{x}_t-\mathbf{y}_t\Big)'(\mathbf{A}_1^{k+1}\otimes\mathbf{A}_3^{k+1})(\mathcal{G}_{(2)}^{k+1})'\\
&-\mathbf{X}_t\otimes\Big(\mathbf{A}_1^{k+1}\mathcal{G}_{(1)}^{k+1}(\mathbf{A}_3^{k}\otimes\mathbf{A}_2^{k})'\mathbf{x}_t-\mathbf{y}_t\Big)'(\mathbf{A}_1^{k+1}\otimes\mathbf{A}_3^{k})(\mathcal{G}_{(2)}^{k+1})'\Big\|_F\\
&\le\nu^2c_1\|\mathbf{A}_3^{k+1}-\mathbf{A}_3^{k}\|_F.
\end{aligned}\tag{S4.119}$$

Plugging (S4.119) and (S4.118) into (S4.117) yields

$$\begin{aligned}
&\big\|\nabla_{\mathbf{A}_2}Q([\![\mathcal{G}^{k+1};\mathbf{A}_1^{k+1},\mathbf{A}_2^{k+1},\mathbf{A}_3^{k+1}]\!])-\nabla_{\mathbf{A}_2}Q([\![\mathcal{G}^{k+1};\mathbf{A}_1^{k+1},\mathbf{A}_2^{k},\mathbf{A}_3^{k}]\!])\big\|_F\\
&\le\nu^2c_1\|\mathbf{A}_2^{k+1}-\mathbf{A}_2^{k}\|_F+2\nu^2c_1\|\mathbf{A}_3^{k+1}-\mathbf{A}_3^{k}\|_F.
\end{aligned}\tag{S4.120}$$

Combining (S4.120) and (S4.115), we obtain that

$$\begin{aligned}
\|N_3\|_F&\le\nu^2c_1\|\mathbf{A}_2^{k+1}-\mathbf{A}_2^{k}\|_F+2\nu^2c_1\|\mathbf{A}_3^{k+1}-\mathbf{A}_3^{k}\|_F+\gamma_2\|\mathbf{U}_2^{k+1}-\mathbf{U}_2^{k}\|_F\\
&\quad+(\gamma_2+\rho_3)\|\mathbf{A}_2^{k+1}-\mathbf{A}_2^{k}\|_F\\
&=(\nu^2c_1+\gamma_2+\rho_3)\|\mathbf{A}_2^{k+1}-\mathbf{A}_2^{k}\|_F+2\nu^2c_1\|\mathbf{A}_3^{k+1}-\mathbf{A}_3^{k}\|_F\\
&\quad+\gamma_2\|\mathbf{U}_2^{k+1}-\mathbf{U}_2^{k}\|_F.
\end{aligned}\tag{S4.121}$$

Regarding N_4 defined in (S4.88), we have

$$\begin{aligned}
&\|N_4\|_F \\
=\ &\Big\|\nabla_{\mathbf{A}_3}\Psi(\mathcal{G}^{k+1},\mathbf{A}_1^{k+1},\mathbf{A}_2^{k+1},\mathbf{A}_3^{k+1},\mathbf{U}_1^{k+1},\mathbf{U}_2^{k+1},\mathbf{U}_3^{k+1}) \\
&\quad -\nabla_{\mathbf{A}_3}\Psi(\mathcal{G}^{k+1},\mathbf{A}_1^{k+1},\mathbf{A}_2^{k+1},\mathbf{A}_3^{k},\mathbf{U}_1^{k},\mathbf{U}_2^{k},\mathbf{U}_3^{k})-\rho_4(\mathbf{A}_3^{k+1}-\mathbf{A}_3^{k})\Big\|_F \\
\le\ &\Big\|\nabla_{\mathbf{A}_3}Q([\![\mathcal{G}^{k+1};\mathbf{A}_1^{k+1},\mathbf{A}_2^{k+1},\mathbf{A}_3^{k+1}]\!])-\nabla_{\mathbf{A}_3}Q([\![\mathcal{G}^{k+1};\mathbf{A}_1^{k+1},\mathbf{A}_2^{k+1},\mathbf{A}_3^{k}]\!])\Big\|_F \\
&\quad +\gamma_3\|\mathbf{U}_3^{k+1}-\mathbf{U}_3^{k}\|_F+(\gamma_3+\rho_4)\|\mathbf{A}_3^{k+1}-\mathbf{A}_3^{k}\|_F.
\end{aligned} \tag{S4.122}$$

By (S3.56), (S3.55) and Proposition 3, we can get

$$\begin{aligned}
&\Big\|\nabla_{\mathbf{A}_3}Q([\![\mathcal{G}^{k+1};\mathbf{A}_1^{k+1},\mathbf{A}_2^{k+1},\mathbf{A}_3^{k+1}]\!])-\nabla_{\mathbf{A}_3}Q([\![\mathcal{G}^{k+1};\mathbf{A}_1^{k+1},\mathbf{A}_2^{k+1},\mathbf{A}_3^{k}]\!])\Big\|_F \\
=\ &\Bigg\|\left(\frac{1}{T}\sum_{t=1}^{T}(\mathbf{A}_1^{k+1}\mathcal{G}_{(1)}^{k+1}(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k+1})'\mathbf{x}_t-\mathbf{y}_t)\circ\mathbf{X}_t\right)_{(3)}(\mathbf{A}_2^{k+1}\otimes\mathbf{A}_1^{k+1})(\mathcal{G}_{(3)}^{k+1})' \\
&\quad -\left(\frac{1}{T}\sum_{t=1}^{T}(\mathbf{A}_1^{k+1}\mathcal{G}_{(1)}^{k+1}(\mathbf{A}_3^{k}\otimes\mathbf{A}_2^{k+1})'\mathbf{x}_t-\mathbf{y}_t)\circ\mathbf{X}_t\right)_{(3)}(\mathbf{A}_2^{k+1}\otimes\mathbf{A}_1^{k+1})(\mathcal{G}_{(3)}^{k+1})'\Bigg\|_F \\
=\ &\Bigg\|\frac{1}{T}\sum_{t=1}^{T}\mathbf{X}_t'\otimes\left(\mathbf{A}_1^{k+1}\mathcal{G}_{(1)}^{k+1}(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k+1})'\mathbf{x}_t-\mathbf{y}_t\right)'(\mathbf{A}_2^{k+1}\otimes\mathbf{A}_1^{k+1})(\mathcal{G}_{(3)}^{k+1})' \\
&\quad -\frac{1}{T}\sum_{t=1}^{T}\mathbf{X}_t'\otimes\left(\mathbf{A}_1^{k+1}\mathcal{G}_{(1)}^{k+1}(\mathbf{A}_3^{k}\otimes\mathbf{A}_2^{k+1})'\mathbf{x}_t-\mathbf{y}_t\right)'(\mathbf{A}_2^{k+1}\otimes\mathbf{A}_1^{k+1})(\mathcal{G}_{(3)}^{k+1})'\Bigg\|_F \\
\le\ &\frac{1}{T}\sum_{t=1}^{T}\Bigg\|\mathbf{X}_t'\otimes\left(\mathbf{A}_1^{k+1}\mathcal{G}_{(1)}^{k+1}(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k+1})'\mathbf{x}_t-\mathbf{y}_t\right)'(\mathbf{A}_2^{k+1}\otimes\mathbf{A}_1^{k+1})(\mathcal{G}_{(3)}^{k+1})' \\
&\quad -\mathbf{X}_t'\otimes\left(\mathbf{A}_1^{k+1}\mathcal{G}_{(1)}^{k+1}(\mathbf{A}_3^{k}\otimes\mathbf{A}_2^{k+1})'\mathbf{x}_t-\mathbf{y}_t\right)'(\mathbf{A}_2^{k+1}\otimes\mathbf{A}_1^{k+1})(\mathcal{G}_{(3)}^{k+1})'\Bigg\|_F \\
\le\ &\frac{1}{T}\sum_{t=1}^{T}\Bigg\|\mathbf{X}_t'\otimes\left(\mathbf{A}_1^{k+1}\mathcal{G}_{(1)}^{k+1}(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k+1})'\mathbf{x}_t-\mathbf{y}_t\right)' \\
&\quad -\mathbf{X}_t'\otimes\left(\mathbf{A}_1^{k}\mathcal{G}_{(1)}^{k+1}(\mathbf{A}_3^{k}\otimes\mathbf{A}_2^{k+1})'\mathbf{x}_t-\mathbf{y}_t\right)'\Bigg\|_F\Big\|(\mathbf{A}_2^{k+1}\otimes\mathbf{A}_1^{k+1})(\mathcal{G}_{(3)}^{k+1})'\Big\|_F \\
\le\ &\frac{1}{T}\sum_{t=1}^{T}\Bigg\|\mathbf{X}_t'\otimes\left(\mathbf{A}_1^{k+1}\mathcal{G}_{(1)}^{k+1}(\mathbf{A}_3^{k+1}\otimes\mathbf{A}_2^{k+1})'\mathbf{x}_t-\mathbf{y}_t\right)' \\
&\quad -\mathbf{X}_t'\otimes\left(\mathbf{A}_1^{k+1}\mathcal{G}_{(1)}^{k+1}(\mathbf{A}_3^{k}\otimes\mathbf{A}_2^{k+1})'\mathbf{x}_t-\mathbf{y}_t\right)'\Bigg\|_F\nu,
\end{aligned} \tag{S4.123}$$

where the last equality holds by (S4.69). Using similar techniques as those of (S4.75) to obtain

$$\begin{aligned}
&\Big\|\mathbf{X}_t' \otimes (\mathbf{A}_1^{k+1}\mathcal{G}_{(1)}^{k+1}(\mathbf{A}_3^{k+1} \otimes \mathbf{A}_2^{k+1})'\mathbf{x}_t - \mathbf{y}_t)' \\
&\quad -\mathbf{X}_t' \otimes (\mathbf{A}_1^{k+1}\mathcal{G}_{(1)}^{k+1}(\mathbf{A}_3^{k} \otimes \mathbf{A}_2^{k+1})'\mathbf{x}_t - \mathbf{y}_t)'\Big\|_F \\
=&\Big\|\mathbf{X}_t' \otimes (\mathbf{A}_1^{k+1}\mathcal{G}_{(1)}^{k+1}((\mathbf{A}_3^{k+1} - \mathbf{A}_3^{k}) \otimes \mathbf{A}_2^{k+1})'\mathbf{x}_t)'\Big\|_F \\
\leq& \|\mathbf{x}_t\|_2^2\, \nu \|\mathbf{A}_3^{k+1} - \mathbf{A}_3^{k}\|_F.
\end{aligned}$$

Combining the above inequality with (S4.123) to get

$$\begin{aligned}
&\big\|\nabla_{\mathbf{A}_3} Q([\![\mathcal{G}^{k+1}; \mathbf{A}_1^{k+1}, \mathbf{A}_2^{k+1}, \mathbf{A}_3^{k+1}]\!]) - \nabla_{\mathbf{A}_3} Q([\![\mathcal{G}^{k+1}; \mathbf{A}_1^{k+1}, \mathbf{A}_2^{k+1}, \mathbf{A}_3^{k}]\!])\big\|_F \\
\leq& \frac{1}{T}\sum_{t=1}^{T} \|\mathbf{x}_t\|_2^2\, \nu^2 \|\mathbf{A}_3^{k+1} - \mathbf{A}_3^{k}\|_F \\
=& \nu^2 c_1 \|\mathbf{A}_3^{k+1} - \mathbf{A}_3^{k}\|_F.
\end{aligned} \tag{S4.124}$$

Substituting (S4.124) into (S4.122) yields

$$\begin{aligned}
\|N_4\|_F &\leq \nu^2 c_1 \|\mathbf{A}_3^{k+1} - \mathbf{A}_3^{k}\|_F + \gamma_3 \|\mathbf{U}_3^{k+1} - \mathbf{U}_3^{k}\|_F + (\gamma_3 + \rho_4)\|\mathbf{A}_3^{k+1} - \mathbf{A}_3^{k}\|_F \\
&= (\nu^2 c_1 + \gamma_3 + \rho_4)\|\mathbf{A}_3^{k+1} - \mathbf{A}_3^{k}\|_F + \gamma_3 \|\mathbf{U}_3^{k+1} - \mathbf{U}_3^{k}\|_F.
\end{aligned} \tag{S4.125}$$

For N_5 in (S4.88), it follows from (S3.57) that

$$\begin{aligned}\|N_5\|_F = & \big\|\nabla_{\mathbf{U}_1}\Psi(\mathcal{G}^{k+1},\mathbf{A}_1^{k+1},\mathbf{A}_2^{k+1},\mathbf{A}_3^{k+1},\mathbf{U}_1^{k+1},\mathbf{U}_2^{k+1},\mathbf{U}_3^{k+1}) \\ & -\nabla_{\mathbf{U}_1}\Psi(\mathcal{G}^{k+1},\mathbf{A}_1^{k+1},\mathbf{A}_2^{k+1},\mathbf{A}_3^{k+1},\mathbf{U}_1^{k},\mathbf{U}_2^{k},\mathbf{U}_3^{k}) \\ & -\rho_5(\mathbf{U}_1^{k+1}-\mathbf{U}_1^{k})\big\|_F \\ = & \big\|\gamma_1(\mathbf{U}_1^{k+1}-\mathbf{A}_1^{k+1})-\gamma_1(\mathbf{U}_1^{k}-\mathbf{A}_1^{k+1})-\rho_5(\mathbf{U}_1^{k+1}-\mathbf{U}_1^{k})\big\|_F \\ = & \big\|\gamma_1(\mathbf{U}_1^{k+1}-\mathbf{U}_1^{k})-\rho_5(\mathbf{U}_1^{k+1}-\mathbf{U}_1^{k})\big\|_F \\ \le & \|\gamma_1(\mathbf{U}_1^{k+1}-\mathbf{U}_1^{k})\|_F+\|\rho_5(\mathbf{U}_1^{k+1}-\mathbf{U}_1^{k})\|_F \\ = & (\gamma_1+\rho_5)\|\mathbf{U}_1^{k+1}-\mathbf{U}_1^{k}\|_F.\end{aligned} \tag{S4.126}$$

Similarly, we can easily obtain that

$$\begin{aligned}\|N_6\|_F = & \big\|\nabla_{\mathbf{U}_2}\Psi(\mathcal{G}^{k+1},\mathbf{A}_1^{k+1},\mathbf{A}_2^{k+1},\mathbf{A}_3^{k+1},\mathbf{U}_1^{k+1},\mathbf{U}_2^{k+1},\mathbf{U}_3^{k+1}) \\ & -\nabla_{\mathbf{U}_2}\Psi(\mathcal{G}^{k+1},\mathbf{A}_1^{k+1},\mathbf{A}_2^{k+1},\mathbf{A}_3^{k+1},\mathbf{U}_1^{k+1},\mathbf{U}_2^{k},\mathbf{U}_3^{k})-\rho_6(\mathbf{U}_2^{k+1}-\mathbf{U}_2^{k})\big\|_F \\ \le & (\gamma_2+\rho_6)\|\mathbf{U}_2^{k+1}-\mathbf{U}_2^{k}\|_F, \\ \|N_7\|_F = & \big\|\nabla_{\mathbf{U}_3}\Psi(\mathcal{G}^{k+1},\mathbf{A}_1^{k+1},\mathbf{A}_2^{k+1},\mathbf{A}_3^{k+1},\mathbf{U}_1^{k+1},\mathbf{U}_2^{k+1},\mathbf{U}_3^{k+1}) \\ & -\nabla_{\mathbf{U}_3}\Psi(\mathcal{G}^{k+1},\mathbf{A}_1^{k+1},\mathbf{A}_2^{k+1},\mathbf{A}_3^{k+1},\mathbf{U}_1^{k+1},\mathbf{U}_2^{k+1},\mathbf{U}_3^{k}) \\ & -\rho_7(\mathbf{U}_3^{k+1}-\mathbf{U}_3^{k})\big\|_F \\ \le & (\gamma_3+\rho_7)\|\mathbf{U}_3^{k+1}-\mathbf{U}_3^{k}\|_F.\end{aligned} \tag{S4.127}$$

Observe that

$$\|N^{k+1}\|_F = \|\,(N_1, N_2, N_3, N_4, N_5, N_6, N_7)\,\|_F$$

$$\leq \|N_1\|_F + \|N_2\|_F + \|N_3\|_F + \|N_4\|_F + \|N_5\|_F + \|N_6\|_F + \|N_7\|_F.$$

By taking (S4.107) collectively with (S4.114), (S4.121), (S4.125), (S4.126) and (S4.127), we get

$$\begin{aligned}
\|N^{k+1}\|_F \leq\ & (\rho_1 + c_1)\|\mathcal{G}^{k+1} - \mathcal{G}^k\|_F + (c_2 + \nu^2 c_1 + \rho_2 + \gamma_1)\|\mathbf{A}_1^k - \mathbf{A}_1^{k+1}\|_F \\
& + (c_1 + 2\nu c_1 + \nu c_2 + 3\nu^2 c_1 + \gamma_2 + \rho_3)\|\mathbf{A}_2^k - \mathbf{A}_2^{k+1}\|_F \\
& + (c_2 + 2\nu c_1 + \nu c_2 + 5\nu^2 c_1 + \gamma_3 + \rho_4)\|\mathbf{A}_3^k - \mathbf{A}_3^{k+1}\|_F \\
& + (2\gamma_1 + \rho_5)\|\mathbf{U}_1^{k+1} - \mathbf{U}_1^k\|_F + (2\gamma_2 + \rho_6)\|\mathbf{U}_2^{k+1} - \mathbf{U}_2^k\|_F \\
& + (2\gamma_3 + \rho_7)\|\mathbf{U}_3^{k+1} - \mathbf{U}_3^k\|_F \\
\leq\ & \bar{\delta}\left(\|\mathcal{G}^{k+1} - \mathcal{G}^k\|_F + \|\mathbf{A}_1^{k+1} - \mathbf{A}_1^k\|_F + \|\mathbf{A}_2^{k+1} - \mathbf{A}_2^k\|_F \right. \\
& \left. + \|\mathbf{A}_3^{k+1} - \mathbf{A}_3^k\|_F + \|\mathbf{U}_1^{k+1} - \mathbf{U}_1^k\|_F + \|\mathbf{U}_2^{k+1} - \mathbf{U}_2^k\|_F + \|\mathbf{U}_3^{k+1} - \mathbf{U}_3^k\|_F\right),
\end{aligned}$$

where $\bar{\delta} := \max\{\rho_1 + c_1, c_2 + \nu^2 c_1 + \rho_2 + \gamma_1, c_1 + 2\nu c_1 + \nu c_2 + 3\nu^2 c_1 + \gamma_2 + \rho_3, c_2 + \nu c_2 + 2\nu c_1 + 5\nu^2 c_1 + \gamma_3 + \rho_4, 2\gamma_1 + \rho_5, 2\gamma_2 + \rho_6, 2\gamma_3 + \rho_7\}$. This completes the proof. □

S4.2 Proof of Theorem 2

Now, by combining Lemmas 12 and 13, we establish the global convergence of Algorithm 1 in the following theorem.

Proof. It can be easily seen from (2.5) that $\mathrm{tr}(\mathbf{U}_i'\mathbf{L}_i\mathbf{U}_i) \geq 0$ for $i = 1, 2, 3$, which in conjunction with the definition of $F(\mathcal{G}, \mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3, \mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3)$ in (S3.52) immediately establishes $F \geq 0$. In addition, note that $F(\mathcal{G}, \mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3, \mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3)$ tends to infinity as $\|\mathcal{G}\|_F$, $\|\mathbf{A}_i\|_F$ or $\|\mathbf{U}_i\|_F$ tends to infinity, where $i = 1, 2, 3$. Therefore, we obtain that F is coercive. Note that F is the monotonically decreasing by Lemma 12, we can get that the sequence $\{(\mathcal{G}^k, \mathbf{A}_1^k, \mathbf{A}_2^k, \mathbf{A}_3^k, \mathbf{U}_1^k, \mathbf{U}_2^k, \mathbf{U}_3^k)\}$ is bounded. Then there exists a subsequence $\{(\mathcal{G}^{k_j}, \mathbf{A}_1^{k_j}, \mathbf{A}_2^{k_j}, \mathbf{A}_3^{k_j}, \mathbf{U}_1^{k_j}, \mathbf{U}_2^{k_j}, \mathbf{U}_3^{k_j})\}$ such that the sequence $\{(\mathcal{G}^{k_j}, \mathbf{A}_1^{k_j}, \mathbf{A}_2^{k_j}, \mathbf{A}_3^{k_j}, \mathbf{U}_1^{k_j}, \mathbf{U}_2^{k_j}, \mathbf{U}_3^{k_j})\}$ converges to $\{(\mathcal{G}^*, \mathbf{A}_1^*, \mathbf{A}_2^*, \mathbf{A}_3^*, \mathbf{U}_1^*, \mathbf{U}_2^*, \mathbf{U}_3^*)\}$ as j tends to infinity. It is worth noting that $\delta_{\mathfrak{B}_i}(\mathbf{A}_i), i = 1, 2, 3$, and $\delta_{\mathcal{D}}(\mathcal{G})$ in F are discontinuous, whereas all other functions in F are continuous. Moreover, $\mathbf{A}_i^k$ and $\mathcal{G}^k$ satisfy $(\mathbf{A}_i^k)'\mathbf{A}_i^k = \mathbf{I}_{r_i}, i = 1, 2, 3$, and $\|\mathcal{G}^k\|_\infty \leq c$ in each iteration, which result in $(\mathbf{A}_i^*)'\mathbf{A}_i^* = \mathbf{I}_{r_i}, i = 1, 2, 3, \|\mathcal{G}^*\|_\infty \leq c$. Hence, $\delta_{\mathfrak{B}_i}(\mathbf{A}_i^{k_j})$ tends to $\delta_{\mathfrak{B}_i}(\mathbf{A}_i^*)$ and $\delta_{\mathcal{D}}(\mathcal{G}^{k_j})$ tends to $\delta_{\mathcal{D}}(\mathcal{G}^*)$ as j tends to infinity. As a result, $F(\mathcal{G}^{k_j}, \mathbf{A}_1^{k_j}, \mathbf{A}_2^{k_j}, \mathbf{A}_3^{k_j}, \mathbf{U}_1^{k_j}, \mathbf{U}_2^{k_j}, \mathbf{U}_3^{k_j})$ tends to $F(\mathcal{G}^*, \mathbf{A}_1^*, \mathbf{A}_2^*, \mathbf{A}_3^*, \mathbf{U}_1^*, \mathbf{U}_2^*, \mathbf{U}_3^*)$ as j tends to infinity.

Notice that $\Psi(\mathcal{G}, \mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3, \mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3)$ defined in (S3.54) and $\delta_{\mathcal{D}}(\mathcal{G})$ are Kurdyka-Łojasiewicz (KL) function (Bolte et al., 2014). For any $i = 1, 2, 3$, $\delta_{\mathfrak{B}_i}(\mathbf{A}_i)$ is a KL function since $\mathfrak{B}_i$ is a semi-algebra set (Qiu et al., 2024). In addition, it follows from Example 2 of Bolte et al. (2014) that the

function $\mathrm{tr}(\mathbf{U}_i'\mathbf{L}_i\mathbf{U}_i)$ is also a KL function since it is a real analytic function. Then one can deduce that $F(\mathcal{G}, \mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3, \mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3)$ is a KL function. Combining with Lemma 12, Lemma 13 and Theorem 3.9 of Attouch et al. (2013), we can establish that $\{\mathcal{G}^k, \mathbf{A}_1^k, \mathbf{A}_2^k, \mathbf{A}_3^k, \mathbf{U}_1^k, \mathbf{U}_2^k, \mathbf{U}_3^k\}$ converges to a critical point of $F(\mathcal{G}, \mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3, \mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3)$. This completes the proof. □

S5 Additional Numerical Experiments

S5.1 Synthetic Data

Next we compare STDGR with Lasso, NNM, LRTLT, TVART, and SHORR for synthetic data. For the noise term, we set $\boldsymbol{\varepsilon}_t \overset{\text{i.i.d.}}{\sim} N(\mathbf{0}, 0.25\mathbf{I}_m)$. Additionally, the factor matrices $\mathbf{A}_i$ are constructed as follows (Belkin and Niyogi, 2001): we randomly generate weight matrices $\mathbf{Z}_i$ and get the corresponding Laplacian matrices based on (2.6). Then the first r_i eigenvectors associated with the smallest eigenvalues are used to obtain $\mathbf{A}_i$. In spectral clustering, the entries of $\mathbf{A}_i$ measure the similarities between data samples and clusters, as they indicate how likely data samples belong to specific clusters (Wang et al., 2014). Therefore, this construction of $\mathbf{A}_i$ can reflect the similarity of the rows of $\mathbf{A}_i$. Two cases are considered, where $\mathcal{G}$ is a

diagonal cube with superdiagonal elements being nonzero. Case a: $(m, p) = (70, 12)$, Tucker rank $(r_1, r_2, r_3) = (5, 5, 5)$, and $(\mathcal{G}_{111}, \mathcal{G}_{222}, \mathcal{G}_{333}, \mathcal{G}_{444}, \mathcal{G}_{555}) = (0.2, 0.2, 0.2, 0.1, 0.1)$. Case b: $(m, p) = (100, 20)$, Tucker rank $(r_1, r_2, r_3) = (7, 7, 7)$, and $(\mathcal{G}_{111}, \mathcal{G}_{222}, \mathcal{G}_{333}, \mathcal{G}_{444}, \mathcal{G}_{555}, \mathcal{G}_{666}, \mathcal{G}_{777}) = (0.1, 0.1, 0.1, 0.05, 0.05, 0.05, 0.05)$. For each simulation experiment, the reported estimation error is averaged over $K = 10$ independent replications. We also report the standard error (SE) of the mean estimation error for each method. Let $E^{(i)} = \|\widehat{\mathcal{W}}_{\mathrm{e}}^{(i)} - \mathcal{W}\|_F$ denote the estimation error in the i-th independent replication, where $\widehat{\mathcal{W}}_{\mathrm{e}}^{(i)}$ is the estimator of the coefficient tensor in the i-th independent replication and $\mathcal{W}$ is the ground truth coefficient tensor. The mean estimation error and its SE are defined as follows: $\bar{E} = \frac{1}{K}\sum_{i=1}^{K} E^{(i)}, \mathrm{SE} = \sqrt{\frac{1}{K(K-1)}\sum_{i=1}^{K}\left(E^{(i)} - \bar{E}\right)^2}$.

In Figure 1, we show the estimation error $\|\widehat{\mathcal{W}} - \mathcal{W}\|_F$ and SE versus different sample sizes T of different methods for synthetic data under the above two settings of (m, p). It can be observed from the first row of Figure 1 that the estimation errors of different methods decrease as the size T increases. Moreover, the errors obtained by STDGR are lower than those obtained by other methods, where the similarity of $\mathbf{A}_i$ is exploited in STDGR. Additionally, the SHORR outperforms Lasso, NNM, LRTLT, TVART in terms of estimation errors, which shows the Tucker decomposi-

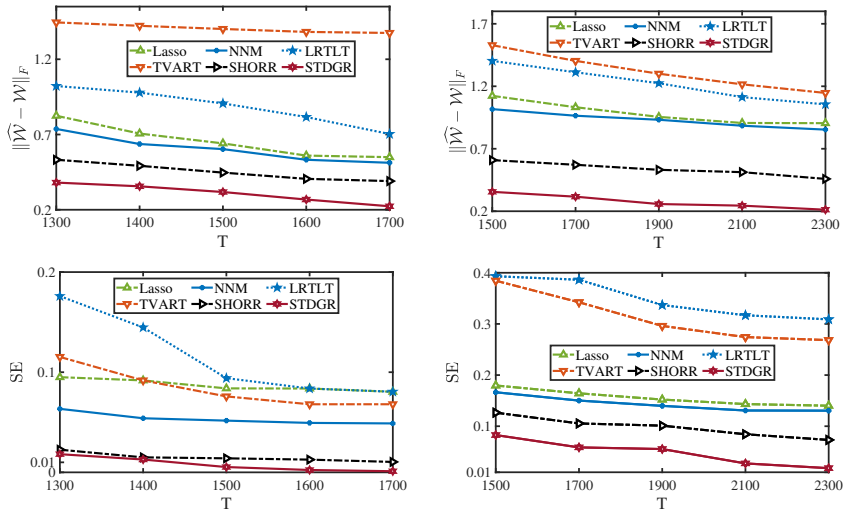


Figure 1: The estimation error $\|\widehat{\mathcal{W}} - \mathcal{W}\|_F$ and SE versus different sample size T of different methods for synthetic data under two settings of (m, p). First row: Estimation error versus T. Second row: SE versus T. (a) Case a. (b) Case b.

tion based method performs better than the matrix or CP based methods. Moreover, it can be observed from the second row of Figure 1 that the SE of different methods generally decrease as T increases. Moreover, the SE obtained by STDGR are lower than those obtained by other methods in both settings. Additionally, SHORR has lower SE than Lasso, NNM, LRTLT, and TVART in the testing experiments.

S5.2 Traffic Flow Dataset

In Figure 2, we show the MSE ($\times 10^{-2}$) versus number of training samples of different methods for the Hangzhou and PeMS04 datasets, where the num-

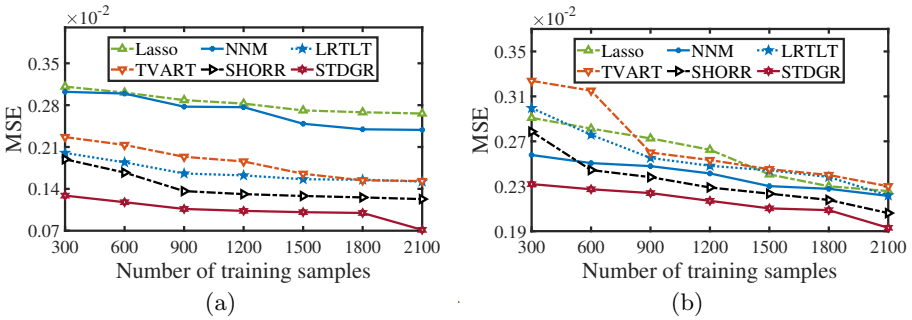
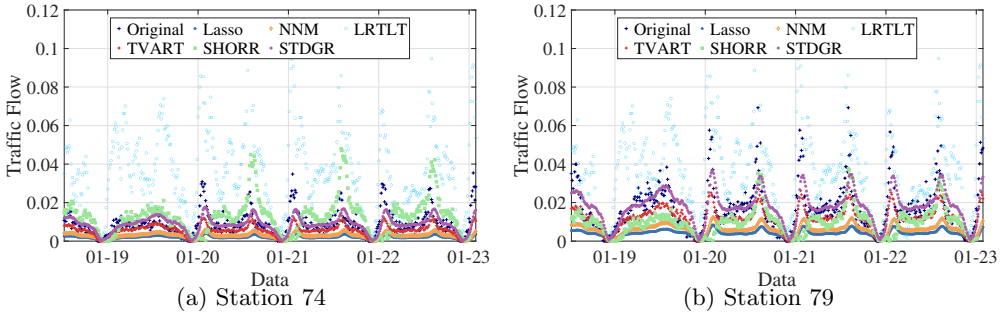


Figure 2: MSE versus number of training samples of different methods for the two traffic flow datasets.

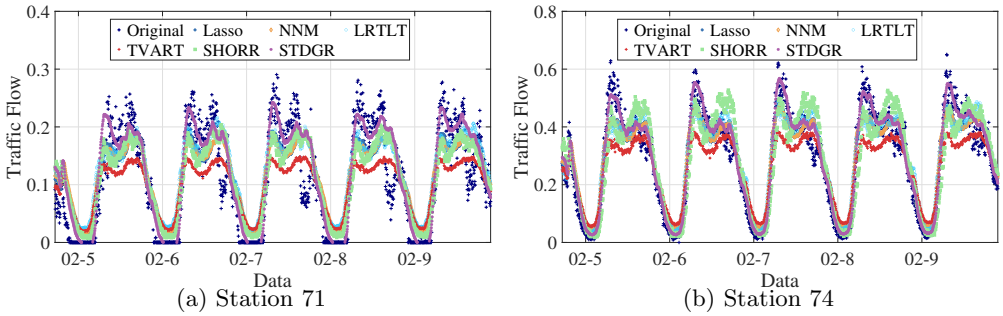
ber of training samples varies from 300 to 2100 with step size 300 and the number of testing samples is 300. As can be seen that the MSE of all methods generally decreases as the number of the training samples increases. And the STDGR consistently achieves the lowest MSE for all testing cases and exhibits great improvement when the number of training samples increases. Moreover, the SHORR performs better than Lasso, NNM, LRTLT, and TVART for most cases, especially for large number of training samples. These results demonstrate that the STDGR obtains higher forecasting accuracy compared with the competition approaches, even when the number of training samples is relatively small.

In Figures 3 and 4, we show the traffic flow prediction of different methods at different time points for the Hangzhou and PeMS04 datasets, respectively, where the 74th and 79th metro stations of the Hangzhou dataset are shown in Figure 3 and the 71st and 74th sensors of the PeMS04 dataset



(a) Station 74 (b) Station 79

Figure 3: Traffic flow prediction of different methods at different time points for the 74th and 79th metro stations of the Hangzhou dataset, respectively.



(a) Station 71 (b) Station 74

Figure 4: Traffic flow prediction of different methods at different time points for the 71st and 74th sensors of the PeMS04 dataset, respectively.

are shown in Figure 4. Here 80% samples are used for training and the remaining 20% samples are used for testing for the two datasets. It can be observed that the STDGR can fit the original traffic flow better compared with other methods for different sensors or stations. Furthermore, the STDGR can predict the trends of traffic flows better for the two datasets, which exhibits the superior performance of STDGR for prediction.

S5.3 Comprehensive Climate Dataset

In this subsection, we test the performance of STDGR for the climate dataset. The comprehensive climate dataset (CCD) contains 17 climate variables of North America from 1990 to 2002. The data are collected and preprocessed by five federal agencies and interpolated on a 2.5×2.5 degree grid, which results in 125 observation locations for each variable. For each location, 156 time points (12 months per year $\times$ 13 years) are recorded. The dataset is publicly available at `https://viterbi-web.usc.edu/~liu32/data/NA-1990-2002-Monthly.csv`. All series are performed de-seasonalization by removing seasonal averages and then standardized to zero mean and unit variance. For the CCD, we choose eight variables in the experiments, including Carbon dioxide (CO2), methane (CH4), Carbon monoxide (CO), hydrogen (H2), wet day, temperature minimum (TMN), temperature mean (TMP), and global horizontal radiation (GLO).

For the VAR(p) model, the order is simply set to $p = 2$ for all methods, except for TVART, for which $p = 1$ is employed. We set $\gamma_1 = \gamma_2 = \gamma_3$ in each case and choose them from the set $\{50, 80, 90, 100, 140, 150, 180, 300, 350, 1000, 2000, 3000\}$ to get the best forecasting performance. By the condition in Theorem 2, we set $\rho_i = \bar{a}_1 L_i$ for $i = 1, 2, 3, 4$ and $\rho_{j+4} = \bar{a}_2 \gamma_j$ for $j = 1, 2, 3$ to guarantee the convergence of PALM Algorithm, where $\bar{a}_1$ is

Table 1: MSE of different methods for CCD with eight variables.

Variable	Lasso	NNM	LRTLT	TVART	SHORR	STDGR
CO2	0.9787	0.9775	1.0161	1.1477	0.9774	**0.9614**
CH4	1.1942	1.1977	1.1619	1.2653	1.1655	**1.1107**
CO	0.8832	1.0420	0.8972	0.9267	0.8374	**0.7986**
H2	0.5687	0.5652	0.5575	0.6191	0.5299	**0.4730**
Wet	0.8850	0.8788	1.1835	1.0781	0.8747	**0.8645**
TMN	0.7620	0.7741	0.9386	0.8626	0.7416	**0.7332**
TMP	0.7720	0.7711	0.9829	0.8553	0.7584	**0.7388**
GLO	0.9603	0.9424	1.1006	0.9956	0.9236	**0.9005**

chosen from the set $\{1.001, 1.1\}$ and $\bar{\alpha}_2$ is chosen from the set $\{10, 50\}$. The parameter c defined in (2.4) is set to 1 and β is set to 0.001. For simplicity, we set $\alpha_1 = \alpha_2 = \alpha_3 = \bar{\alpha}$ in all experiments, where $\bar{\alpha}$ is selected from the set $\{0.001, 0.4, 1, 2, 4, 5, 6, 10, 12, 14, 100\}$ to obtain the best forecasting performance in our experiments.

In Table 1, we report the MSE of different methods for the variables CO2, CH4, CO, H2, wet day, TMN, TMP, and GLO, where the first 60% time series of each variable are used for training and the resting 40% time series are used for testing. The best results are highlighted in bold. It can be seen from this table that the STDGR achieves lower MSE than other

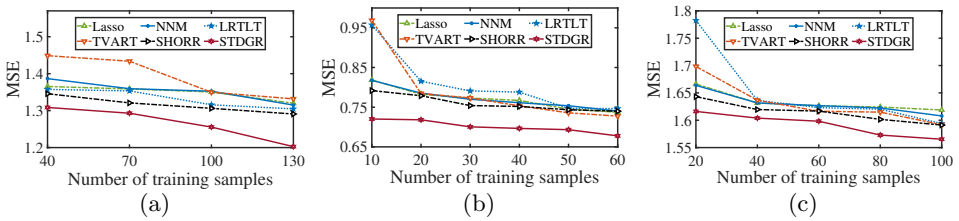


Figure 5: MSE versus number of training samples of different methods for the CCD.

methods. Furthermore, the SHORR outperforms Lasso, NNM, LRTLT, and TVART in terms of MSE. The improvement of STDGR is very impressive compared with SHORR for CH4, CO, H2.

In Figure 5, we show the MSE of different methods versus the number of training samples for CO2, CH4, and H2, where the number of testing samples are 20 for CO2 and 10 for CH4, H2, respectively. We can observe that the MSE of different methods decreases as the number of training samples increase for the three variables, which demonstrates that larger number of training samples can improve the forecasting accuracy for VAR. Moreover, the MSE obtained by STDGR is lower than those obtained by other methods, which indicates the superior performance of STDGR for forecasting.

S5.4 Macroeconomic Dataset

In this subsection, we test the performance of STDGR for the macroeconomic dataset, which contains 40 quarterly macroeconomic variables of

the United States from 1959 to 2007 with 194 time points for each variable (Koop, 2013). The dataset is publicly available at `https://www.stlouisfed.org/research/economists/mccracken/fred-databases`. The variables are seasonally adjusted except for the financial series, and transformed to stationary with zero mean and unit variance (Wang et al., 2022). For the macroeconomic dataset, the graph regularization term incorporates the structural relationships among economic variables. Since macroeconomic indicators are often interconnected through economic mechanisms, variables with similar economic characteristics are encouraged to have similar latent representations. Therefore, the graph regularization reduces estimation variability and improves the model's ability to capture meaningful dynamic dependencies among macroeconomic variables.

We set $p = 4$ in the VAR(p) model and $\gamma_1 = \gamma_2 = \gamma_3 = 0.001$ for simplicity. As suggested in Theorem 2, the penalty parameters are simply set to $\rho_i = \bar{a}_1 L_i$ for $i = 1, \ldots, 4$ and $\rho_{j+4} = \bar{a}_2 \gamma_j$ for $j = 1, 2, 3$, where $\bar{a}_1 = 1.001$ and $\bar{a}_2 = 10$. Besides, c is set to 1 for simplicity. For the parameters β, α_i in the STDGR model, we set $\beta = 0.001$, $\alpha_1 = \alpha_2 = \alpha_3 = 0.01$. Moreover, the Tucker ranks of SHORR and STDGR are chosen as the same for this dataset.

Table 2 reports the MSE, running time (in seconds), and iteration

counts of Lasso, NNM, LRTLT, TVART, SHORR, and STDGR, where we use 80% and 90% samples for training and the remaining samples for testing, respectively. It can be seen from this table that the MSE achieved by STDGR is lower than those achieved by other methods for the testing two cases, which shows that the forecasting accuracy of STDGR is better than other comparison methods. Furthermore, SHORR achieves the second-lowest MSE in both settings, but requires more running time than Lasso, NNM, LRTLT, and STDGR. Although the STDGR takes more running time than Lasso, NNM, LRTLT, TVART, the MSE obtained by STDGR is lower than these methods.

Table 2: Comparisons of MSE, running time (in seconds) and iteration counts of different methods for the macroeconomic dataset.

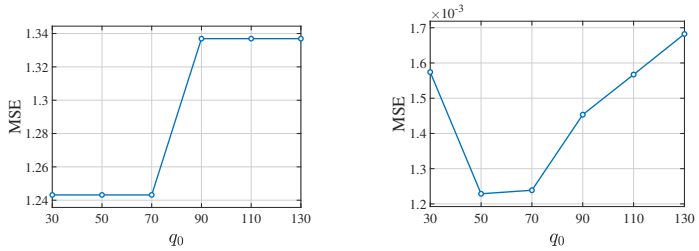
Index	Training size	Lasso	NNM	LRTLT	TVART	SHORR	STDGR
MSE	80%	1.5911	1.3897	1.5675	1.5123	0.8259	**0.6805**
	90%	0.9973	0.8642	0.8005	1.0742	0.6707	**0.5951**
Time	80%	0.05	0.08	0.68	4.98	2.78	0.99
	90%	0.03	0.06	0.64	6.85	5.67	1.17
Iteration	80%	104/200	78/100	46/100	185/500	46/100	30/200
	90%	75/200	52/100	45/100	252/500	82/100	35/200

S6 Discussions

S6.1 Parameters Analysis

In order to guarantee the consistency of Tucker rank selection in (4.14) (see details in Section 4.1), we need $\bar{c} = o_p(r_1 mp/T)$ (Wang et al., 2022, Theorem 3). Therefore, we recommend $\bar{c} = \sqrt{mp\log(T)/(q_0 T)}$, where $q_0 > 0$ is a given constant. Now we give the sensitive analysis of different choices of q_0 for the CH_4 variable and Hangzhou dataset, where 65% samples are used for training and the remaining 35% samples are used for validation in the two datasets. Here q_0 varies from 30 to 130 with step size 20. In Figure 6, we show the MSE of STDGR versus different q_0 for the CH_4 variable and Hangzhou dataset. It can be seen from this figure that the MSE of STDGR remains stable when q_0 varies from 30 to 70, while the MSE increases when q_0 is larger than 70 for the CH_4 variable. For the Hangzhou dataset, the MSE obtained by STDGR decrease when $q_0 \leq 50$, and then increases when $q_0 > 50$. For simplicity, we set $q_0 = 50$ in all experiments, i.e., $\bar{c} = \sqrt{mp\log(T)/(50T)}$.

Next we analyze the influence of parameters γ_1, γ_2, and γ_3 for STDGR. We test the performance of STDGR with different parameters for the CH_4 variable of the CCD dataset in Figure 7, where two parameters are varied

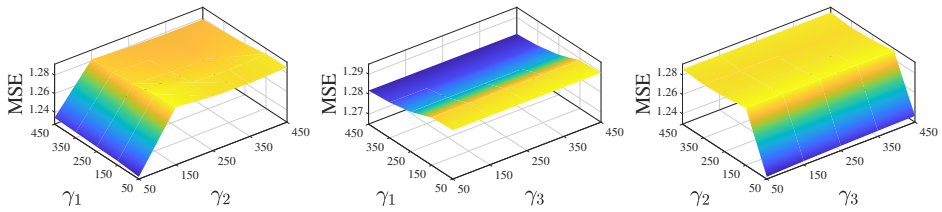


(a) (b)

Figure 6: MSE of STDGR versus different q_0 in $\bar{c} = \sqrt{mp\log(T)/(q_0 T)}$ for the CH_4 variable in the CCD dataset and Hangzhou dataset. (a) CH_4. (b) Hangzhou.

from 50 to 450 with step size 100, while the remaining parameter is set to 150. Here 65% samples are used for training and the remaining 35% samples are used for validation. It can be seen from Figures 7 (a) and (c) that the MSE varies greatly when γ_2 increases from 50 to 150 but remains stable when $\gamma_2 \geq 150$. Figure 7(b) shows that the MSE varies slightly for different values of γ_1 and γ_3. Since the MSE of STDGR remains stable for $\gamma_i \geq 150, i = 1, 2, 3$, we simply set $\gamma_1 = \gamma_2 = \gamma_3$ in all experiments. The detailed ranges of γ_i for each dataset will be given in the revised manuscript

According to Theorem 2 in the main file, we need $\rho_1 > L_1, \rho_2 > L_2, \rho_3 > L_3, \rho_4 > L_4, \rho_5 > \gamma_1, \rho_6 > \gamma_2, \rho_7 > \gamma_3$ to guarantee the convergence of PALM algorithm, where $L_1 = \frac{1}{T}\sum_{t=1}^{T} \|\mathbf{x}_t\|_2^2, L_j = r_1 r_2 r_3 c^2 L_1 + \gamma_{j-1}, j = 2, 3, 4$. For the sake of simplicity, we set $\rho_i = \bar{a}_1 L_i$ for $i = 1, 2, 3, 4$ and $\rho_{j+4} = \bar{a}_2 \gamma_j$ for $j = 1, 2, 3$, where $\bar{a}_1, \bar{a}_2 > 1$. Therefore, instead of directly tuning ρ_i, we



(a) (b) (c)

Figure 7: MSE versus different parameter pairs of STDGR for the CH_4 variable in the CCD dataset. (a) MSE versus (γ_1, γ_2) with $\gamma_3 = 150$. (b) MSE versus (γ_1, γ_3) with $\gamma_2 = 150$. (c) MSE versus (γ_2, γ_3) with $\gamma_1 = 150$.

tune $\bar{a}_1$ and $\bar{a}_2$ by grid search. In Figure 8, we show the MSE of STDGR for different parameter pairs $(\bar{a}_1, \bar{a}_2)$ for the CH_4 variable in the CCD dataset, where 65% samples are used for training and the remaining samples are used for validation. Here $\bar{a}_1 \in \{1.0001, 1.001, 1.01, 1.1, 10\}$, and $\bar{a}_2$ varies from 10 to 90 with step size 20. As shown in Figure 8, the MSE varies slightly with $\bar{a}_1$, whereas it varies greatly for different $\bar{a}_2$. Therefore, we will set $\bar{a}_1$ to be fixed except for the traffic dataset, and $\bar{a}_2$ will be chosen from a candidate set. The detailed ranges of $\bar{a}_1, \bar{a}_2$ for each dataset will be given in the revised manuscript

Now we give the sensitive analysis of parameters $\beta, \bar{\alpha}, \epsilon$ for the CH_4 variable from the CCD dataset and the Hangzhou traffic dataset. For each dataset, 65% samples are used for training and the remaining 35% samples are used for validation. In Figures 9 and 10, we show the MSE versus

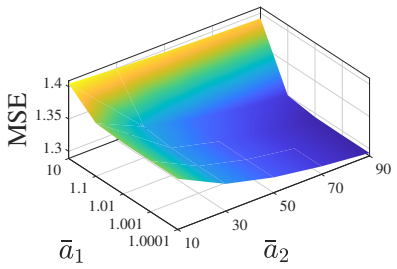
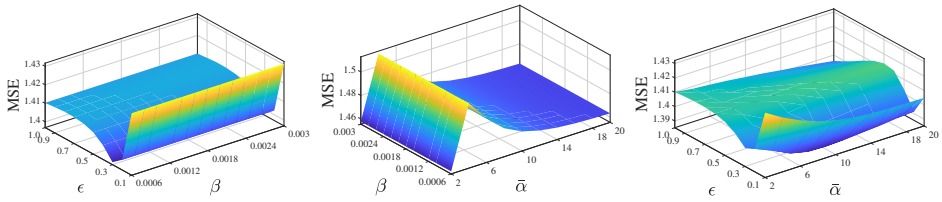


Figure 8: MSE of STDGR versus different parameter pairs $(\bar{a}_1, \bar{a}_2)$ for the CH_4 variable in the CCD dataset.

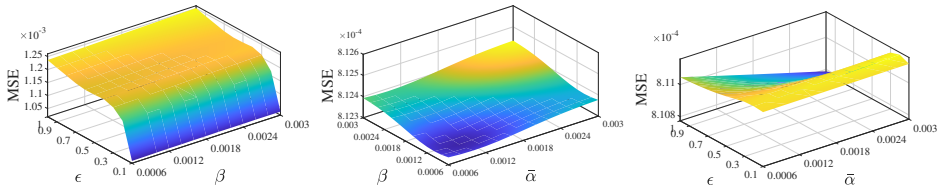
different parameters pairs for the CH_4 variable from the CCD and Hangzhou datasets, where β varies from 0.0006 to 0.003 with step size 0.0002, and ϵ varies from 0.1 to 1 with step size 0.1. For $\bar{\alpha}$, it varies from 2 to 20 with step size 2 for CCD and from 0.0006 to 0.003 with step size 0.0002 for Hangzhou dataset. It can be seen from Figures 9(a), (b) and 10(b) that, the MSE of STDGR are stable for different β in both datasets. For simplicity, we set $\beta = 0.001$ in all experiments. Moreover, we can see from Figures 9(c) and 10(a) that the MSE obtained by STDGR varies slightly for different ϵ, which shows that ϵ is not too sensitive for STDGR. In the experiments, we simply set $\epsilon = 0.2$. However, for parameter $\bar{\alpha}$, the MSE of STDGR varies greatly for different values, which can be seen from Figures 9(b), (c) and 10(c). Therefore, we will choose $\bar{\alpha}$ from some ranges to get the best results of STDGR, where the detailed ranges will be given for each dataset.

For parameters α_1, α_2, and α_3, we test different choices of two param-



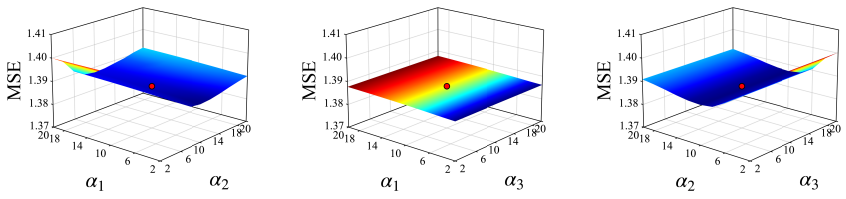
(a) (b) (c)

Figure 9: MSE versus different parameters pairs of STDGR for the CH_4 variable in the CCD dataset. (a) MSE versus (β, ϵ). (b) MSE versus $(\bar{\alpha}, \beta)$. (c) MSE versus $(\bar{\alpha}, \epsilon)$.



(a) (b) (c)

Figure 10: MSE versus different parameters pairs of STDGR for the Hangzhou dataset. (a) MSE versus (β, ϵ). (b) MSE versus $(\bar{\alpha}, \beta)$. (c) MSE versus $(\bar{\alpha}, \epsilon)$.



(a) (b) (c)

Figure 11: MSE versus different parameters pairs of STDGR for the CH_4 variable in the CCD dataset. (a) MSE versus (α_1, α_2) with $\alpha_3 = 10$. (b) MSE versus (α_1, α_3) with $\alpha_2 = 10$. (c) MSE versus (α_2, α_3) with $\alpha_1 = 10$.

eters from 2 to 20 with step size 2, where the other parameter is set to 10. In Figure 11, we show the MSE of STDGR versus different parameters pairs for the CH_4 variable from the CCD dataset. As shown in Figure 11, the MSE is low when $\alpha_1 = \alpha_2 = \alpha_3 = 10$ for the three cases. Furthermore, when we choose the three parameters to be the same, the MSE will be small. These results support the use of the common parameter choice $\alpha_1 = \alpha_2 = \alpha_3 = \bar{\alpha}$. This restriction reduces computational cost of parameter tuning while retaining competitive forecasting performance. Therefore, we set $\alpha_1 = \alpha_2 = \alpha_3 = \bar{\alpha}$ in the experiments for the sake of simplicity.

S6.2 Choice of Lag Order p

In the experiments, the lag order is determined through a preliminary sensitivity analysis before conducting the subsequent prediction experiments. Specifically, we test the prediction MSE and computational time (in seconds) over a range of candidate lag orders and select the suitable lag order based on both prediction error and computational time.

We compare the MSE and computational time of STDGR by varying p from 1 to 14 with step size 1 for the macroeconomic, PeMS04, and CCD datasets. For each dataset, three cases are tested, i.e., the first 60%, (70%, or 80%) samples are used for training, and the remaining samples are used

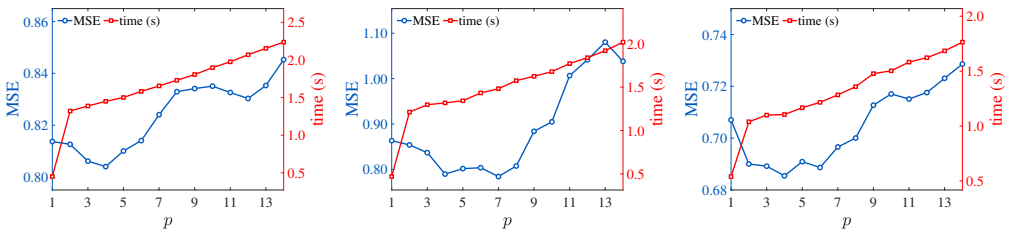


Figure 12: MSE and computational time of STDGR with different lag orders p for the macroeconomic dataset. (a) 60% training samples. (b) 70% training samples. (c) 80% training samples.

to evaluate the prediction performance. The MSE and computational time versus lag orders for different datasets are presented in Figures 12–15.

In Figure 12, we show the MSE and computational time of STDGR with different lag orders for the macroeconomic dataset, where 60%, 70%, and 80% samples are used for training, respectively. It can be seen from this figure that STDGR achieves the lowest prediction MSE when $p = 4$ for the two cases with 60% and 80% training samples. Moreover, STDGR achieves the lowest MSE when $p = 7$ for the case with 70% training samples. In fact, the performance of STDGR is close to each other for 70% training samples when $p = 4$ and $p = 7$, while more computation time is required for $p = 7$. For the three cases, the computational time increases as p increases. In order to balance the computational time and prediction accuracy, we

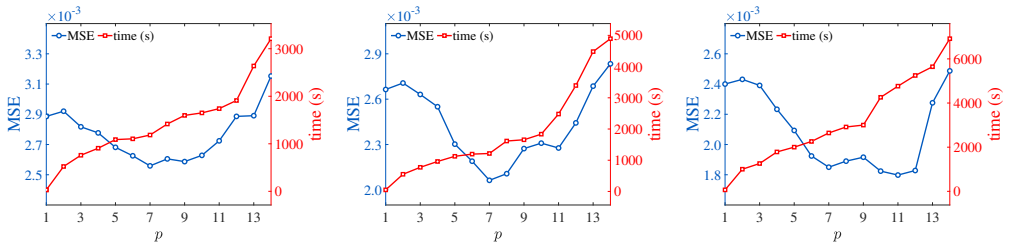
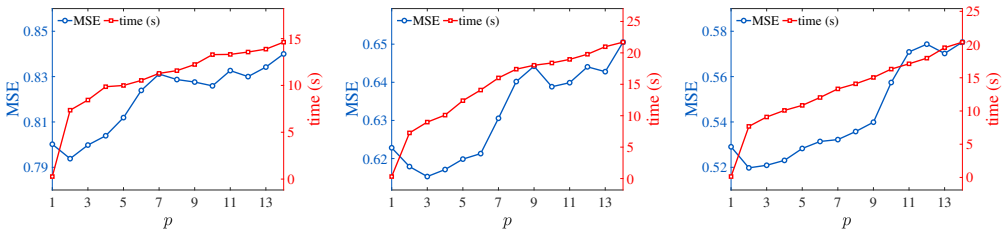


Figure 13: MSE and computational time of STDGR with different lag orders p for the PeMS04 dataset. (a) 60% training samples. (b) 70% training samples. (c) 80% training samples.

simply choose $p = 4$ for the macroeconomic dataset.

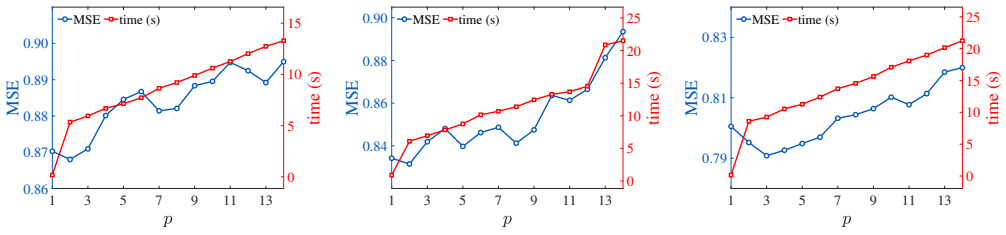
In Figure 13, we show the MSE and computational time of STDGR with different lag orders for the PeMS04 dataset, where 60%, 70%, and 80% samples are used for training, respectively. We can see that STDGR achieves the lowest prediction MSE when $p = 7$ for the two cases with 60% and 70% training samples. For 80% training samples, STDGR achieves the lowest prediction MSE when $p = 11$, which is slightly smaller than that when $p = 7$. However, the computational time required by STDGR with $p = 11$ is approximately twice that of STDGR with $p = 7$. For simplicity, we choose $p = 7$ for the PeMS04 dataset in order to balance the computational time and prediction accuracy.

In Figures 14 and 15, we show the MSE and computational time of



(a) (b) (c)

Figure 14: MSE and computational time of STDGR with different lag orders p for the CO variable in the CCD dataset. (a) 60% training samples. (b) 70% training samples. (c) 80% training samples.



(a) (b) (c)

Figure 15: MSE and computational time of STDGR with different lag orders p for the Wet variable in the CCD dataset. (a) 60% training samples. (b) 70% training samples. (c) 80% training samples.

STDGR with different lag orders for the CCD dataset (CO and Wet variables), where 60%, 70%, and 80% samples are used for training, respectively. For the CO variable, STDGR obtains the lowest prediction MSE when $p = 2$ for 60% and 80% training samples, and obtains the lowest MSE when $p = 3$ for 70% training samples. For the Wet variable, STDGR obtains the lowest

MSE when $p = 2$ for 60% and 70% training samples, and obtains the lowest MSE when $p = 3$ for 80% training samples. However, the computational time increases as the lag order p increases for different number of training samples. Therefore, we simply select $p = 2$ for the CCD dataset in order to balance the prediction accuracy and computational time.

S6.3 Performance of Different Initialization

To assess the sensitivity of STDGR to the initial estimator, we compare the nuclear norm minimization (NNM) initialization with two alternatives: Lasso initialization and random initialization. For random initialization, we construct a random tensor with Tucker form as follows:

$$\widehat{\mathcal{W}}_{\text{Rand}} = \mathcal{G}^{(0)} \times_1 \mathbf{A}_1^{(0)} \times_2 \mathbf{A}_2^{(0)} \times_3 \mathbf{A}_3^{(0)},$$

where the entries of the core tensor $\mathcal{G}^{(0)}$ are independently generated from $N(0, 1)$, and each factor matrix $\mathbf{A}_i^{(0)}$ is a random matrix with orthogonal columns. For the Lasso initialization, we use the model in Basu and Michailidis (2015) to get an initial estimator $\widehat{\mathcal{W}}_{\text{Lasso}}$.

Table 3 reports the forecasting MSEs obtained by STDGR with three different initialization strategies for four variables in the CCD dataset, where the Tucker rank is prespecified as $\mathbf{r}^{(0)} = (10, 10, 2)$ for the random strategy. For each variable, 80% samples are used for training and the re-

Table 3: Forecasting performance of STDGR under different initialization strategies for four variables in the CCD dataset.

Variable	Random	Lasso	NNM
CH_4	1.2879	1.1674	**0.8459**
TMN	1.0860	0.9689	**0.6943**
TMP	1.0005	1.0251	**0.7089**
GLO	1.2381	1.0945	**0.8607**

maining 20% samples are used for testing. It can be seen from Table 3 that the MSEs obtained by STDGR with NNM based initialization are lower than those obtained with the Lasso and random tensor initializations for the four variables in the CCD dataset. Meanwhile, the Lasso initialization achieves better forecasting performance than the random initialization in most cases.

S6.4 Performance of Different Choices of Tucker Rank

We further test the performance of STDGR for different choices of Tucker rank. We compare the ridge-type ratio estimator (the method used in the manuscript) with the following alternative: a truncation-based rank-selection rule based on singular value thresholding Wang et al. (2024) (de-

noted by Truncation). Specifically, we first obtain the initial estimate $\widehat{\mathcal{W}}_{\mathrm{NN}}$ based on nuclear norm minimization method and the estimate the rank according to the unfolding matrices of $\widehat{\mathcal{W}}_{\mathrm{NN}}$, i.e.,

$$\widehat{r}_i = \sum_j \mathbb{I}\left\{\sigma_j\left((\widehat{\mathcal{W}}_{\mathrm{NN}})_{(i)}\right) > \gamma\right\}, \qquad \gamma = \lambda_{\mathrm{NN}}/2,$$

where $\mathbb{I}\{\cdot\}$ denotes the 0-1 indicator function, and λ_{NN} is the regularization parameter in the nuclear norm minimization model.

Now we conduct numerical experiments in Table 4 for the above choices of Tucker rank by using the PeMS04 dataset and the CO variable from the CCD dataset. For each dataset, 80% samples are used for training, and the remaining 20% samples are used for testing. As shown in Table 4, the ridge-type ratio estimator achieves slightly smaller forecasting MSE for both datasets than the truncation method. Note that the Tucker rank selection of ridge-type ratio estimator is rank consistent. This experiment indicates that STDGR is effective by using a rank-consistent selection procedure.

S6.5 Different Construction of Graphs

Now we compare the performance of STDGR under different constructions of graphs. Except for the construction based on an initial estimator of NNM, we also construct a graph Laplacian matrix based on traffic-network distances for the PeMS04 dataset or geographical distances for the CCD

Table 4: Performance of STDGR with different choices of Tucker rank.

	CO		PeMS04	
Method	Rank	MSE	Rank	MSE
Truncation	$(11, 16, 2)$	0.4789	$(7, 8, 2)$	0.002111
Ridge-type ratio	$(3, 6, 1)$	0.4433	$(32, 13, 4)$	0.001830

dataset.

For PeMS04, the graph Laplacian matrices $\mathbf{L}_1$ and $\mathbf{L}_2$, corresponding to the factor matrices $\mathbf{A}_1$ and $\mathbf{A}_2$, respectively, are constructed using the traffic network dataset. Specifically, the distance r_{ij} between sensors i and j has already been calculated, which is available at `https://ieee-dataport.org/documents/pems03-and-pems04`. The weighted adjacency matrix $\mathbf{Z} = (z_{ij})$ is constructed as $z_{ij} = \exp(-\frac{r_{ij}^2}{2\epsilon^2})$, where $\epsilon = 0.2$ in the experiments. The degree of sensor i is computed by $d_i = \sum_j z_{ij}$, and the corresponding diagonal matrix is defined as $\mathbf{D} = \mathrm{diag}(d_1, \ldots, d_m)$, where m denotes the number of sensors in PeMS04. The factor matrices $\mathbf{A}_1$ and $\mathbf{A}_2$ can be interpreted as the loadings of the response and predictor factors, respectively. Both the response and predictor factors are related to the m sensors in PeMS04. Therefore, the same weighted adjacency matrix $\mathbf{Z}$ is used to construct the graph Laplacian matrices for $\mathbf{A}_1$ and $\mathbf{A}_2$, i.e., $\mathbf{L}_1 = \mathbf{L}_2 = \mathbf{D} - \mathbf{Z}$.

For the CCD dataset, the distance r_{ij} between locations i and j is computed from their latitude and longitude coordinates using the Haversine formula, which follows from Bahadori et al. (2014). The weighted adjacency matrix $\mathbf{Z} = (z_{ij})$ is constructed as $z_{ij} = \exp(-\frac{r_{ij}^2}{2\epsilon^2})$, where $\epsilon = 0.2$ in the experiments. The degree of location i is computed by $d_i = \sum_j z_{ij}$, and the corresponding degree matrix is defined as $\mathbf{D}_{\text{geo}} = \text{diag}(d_1, \ldots, d_m)$, where m denotes the number of locations in the CCD dataset. Then the corresponding graph Laplacian matrix is constructed as $\mathbf{L}_{\text{geo}} = \mathbf{D}_{\text{geo}} - \mathbf{Z}$. Similarly to PeMS04, the response and predictor factors are related to the m locations in the CCD dataset. Therefore, the same weighted adjacency matrix $\mathbf{Z}$ is used to construct the graph Laplacian matrices for $\mathbf{A}_1$ and $\mathbf{A}_2$, i.e., $\mathbf{L}_1 = \mathbf{L}_2 = \mathbf{L}_{\text{geo}}$.

For both CCD and PeMS04, the factor matrix $\mathbf{A}_3$ corresponds to the lag dimension. Following Belkin and Niyogi (2001), we construct an adjacency matrix, where the weight between two lag orders is set to one if and only if they are adjacent, and zero otherwise. Specifically,

$$(\mathbf{Z}_{\text{lag}})_{ij} = \begin{cases} 1, & \text{if } |i-j| = 1, \\ 0, & \text{otherwise.} \end{cases}$$

The degree of lag order i is computed by $d_i = \sum_j (\mathbf{Z}_{\text{lag}})_{ij}$, and the corresponding diagonal degree matrix is defined as $\mathbf{D}_{\text{lag}} = \text{diag}(d_1, \ldots, d_p)$. Then

Table 5: MSE of STDGR with different constructions of graphs for the CCD (Wet and TMN variables) and PeMS04 datasets.

Dataset	Distance-based	Initialization-based
Wet	0.8550	**0.8238**
TMN	0.7401	**0.7209**
PeMS04	**0.002105**	0.002221

the lag-order graph Laplacian matrix is constructed as $\mathbf{L}_{\text{lag}} = \mathbf{D}_{\text{lag}} - \mathbf{Z}_{\text{lag}}$.

In Table 5, we report the prediction MSE of STDGR with different constructions of graphs, where the constructions are based on an initial estimator of NNM for both CCD (Wet and TMN variables) and PeMS04 datasets, and traffic-network distances for the PeMS04 dataset or geographical distances for the CCD dataset (Wet and TMN variables). Here 65% samples are used for training and the remaining 35% samples are used for testing. It can be seen from this table that the MSEs obtained by STDGR with the initialization-based graph Laplacian matrices are lower than those obtained by STDGR with the distance-based graph Laplacian matrices for Wet and TMN variables, which shows that the construction of graph based on an initial estimator of NNM is effective for STDGR. More importantly, the distance information of many real-world datasets is not available, such

as the Hangzhou dataset. In order to unify the construction of graph Laplacian matrices for different datasets, we use the initial estimator of NNM in the experiments.

S6.6 Ablation Analysis

In this subsection, we show the effectiveness of STDGR for both the sparse Tucker decomposition and graph regularization. Specifically, we compare with the Tucker decomposition model with sparse core tensor (called TDSC for short), which uses the sparsity constraint on the core tensor rather than on the factor matrices compared with Wang et al. (2022) and does not impose the graph regularization. Moreover, we also compare with the Tucker decomposition and graph regularization without sparse core tensor constraint (called TDGR for short). And the SHORR is also compared to show the effectiveness of the sparsity constraint on the core tensor rather than on the factor matrices in Tucker decomposition.

Table 6: Comparisons of MSE ($\times 10^{-3}$) of SHORR, TDSC, TDGR, STDGR for the Hangzhou dataset.

Training size	SHORR	TDSC	TDGR	STDGR
75%	0.4475	0.3941	0.4782	0.3532
85%	0.3714	0.3370	0.3921	0.2743

In Table 6, we report the MSE ($\times 10^{-3}$) of SHORR, TDSC, TDGR, and STDGR for the Hangzhou dataset, where 75% and 85% samples are used for training and the remaining samples are used for testing, respectively. It can be seen from this table that the MSE of TDSC is lower than that of SHORR, which shows the superiority of imposing sparsity on the core tensor rather than on the factor matrices. Moreover, the SHORR and TDSC perform better than TDGR for the testing cases, which implies that the sparsity constraints on the core tensor or factor matrices are more effective than the graph regularization. Besides, the STDGR substantially outperforms other three methods. This demonstrates that the sparsity constraint on the core tensor and graph regularization on the factor matrices are both effective for VAR modeling.

S6.7 Performance of STDGR With Uniform Noise

To further assess the practical robustness of the proposed STDGR method, we add additional numerical experiments under non-Gaussian noise settings. Specifically, for the noise term, we set $\boldsymbol{\varepsilon}_t = (\varepsilon_{t1}, \ldots, \varepsilon_{tm})'$, where each element of $\boldsymbol{\varepsilon}_t$ is independent and follows a uniform distribution on $[-1, 1]$, i.e., $\varepsilon_{tj} \overset{\text{i.i.d.}}{\sim} \text{Uniform}(-1, 1)$, $j = 1, \ldots, m$. Additionally, the factor matrices $\mathbf{A}_i$ are constructed as follows (Belkin and Niyogi, 2001): we

first generate the weight matrices $\mathbf{Z}_i$ randomly and construct the corresponding Laplacian matrices as $\mathbf{L}_i = \mathbf{D}_i - \mathbf{Z}_i$, where $\mathbf{D}_i$ is the corresponding degree matrix. Then the columns of $\mathbf{A}_i$ are chosen as the r_i eigenvectors of $\mathbf{L}_i$ associated with its r_i smallest eigenvalues. In spectral clustering, the entries of $\mathbf{A}_i$ measure the similarities between data samples and clusters, as they indicate how likely data samples belong to specific clusters (Wang et al., 2014). Therefore, this construction of $\mathbf{A}_i$ can reflect the similarity of the rows of $\mathbf{A}_i$. The sample size T is set to 1000. Four cases are considered: Case c. $(m, p) = (50, 10)$. Case d. $(m, p) = (150, 10)$. Case e. $(m, p) = (50, 15)$. Case f. $(m, p) = (150, 15)$. In each case, we set the Tucker rank as $(r_1, r_2, r_3) = (9, 9, 9)$ and specify the core tensor as a diagonal tensor, where the nonzero superdiagonal entries are set as $(\mathcal{G}_{111}, \mathcal{G}_{222}, \mathcal{G}_{333}, \mathcal{G}_{444}, \mathcal{G}_{555}, \mathcal{G}_{666}, \mathcal{G}_{777}, \mathcal{G}_{888}, \mathcal{G}_{999}) = (0.5, 1, 0.1, 0.5, 1, 0.1, 0.5, 1, 0.1)$. Due to the randomness, all experiments are performed 10 times and we take the average estimation error as the last result. The estimation errors (i.e., $\|\widehat{\mathcal{W}} - \mathcal{W}\|_F$) of Lasso, NNM, LRTLT, TVART, SHORR, and STDGR for the four cases are reported in Table 7.

As shown in Table 7, STDGR achieves the lowest estimation error for each case, while SHORR achieves the second lowest estimation error. In most cases, the Lasso and NNM yield lower estimation errors than LRTLT

Table 7: Estimation errors of different methods under the noise with uniform distribution.

(m, p)	Lasso	NNM	LRTLT	TVART	SHORR	STDGR
$(50, 10)$	0.8200	0.8097	1.1254	0.8160	0.7265	**0.6491**
$(150, 10)$	1.6049	1.5431	1.8799	1.8770	1.1388	**0.8522**
$(50, 15)$	0.8037	0.9902	2.2396	1.0030	0.7333	**0.6598**
$(150, 15)$	1.8401	1.7900	3.4170	1.9364	1.2043	**0.9713**

and TVART. This experiment provides a preliminary empirical evidence that STDGR performs well under the uniform noise setting.

Bibliography

Attouch, H., J. Bolte, P. Redont, and A. Soubeyran (2010). Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the Kurdyka-Łojasiewicz inequality. *Math. Oper. Res. 35*(2), 438–457.

Attouch, H., J. Bolte, and B. F. Svaiter (2013). Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized Gauss-Seidel methods. *Math. Program. 137*(1-2), 91–129.

Bahadori, M. T., Q. R. Yu, and Y. Liu (2014). Fast multivariate spatio-temporal analysis via low rank tensor learning. *Advances in Neural Information Processing Systems 27*, 3491–3499.

Basu, S. and G. Michailidis (2015). Regularized estimation in sparse high-dimensional time series models. *Ann. Statist. 43*(4), 1535–1567.

Beck, A. (2017). *First-order methods in optimization*. Philadelphia: Society for Industrial and Applied Mathematics.

Belkin, M. and P. Niyogi (2001). Laplacian eigenmaps and spectral techniques for embedding and clustering. In *Proceedings of the 15th International Conference on Neural Information Processing Systems*, pp. 585–591.

Bolte, J., S. Sabach, and M. Teboulle (2014). Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Math. Program. 146*(1-2), 459–494.

Chen, Y., Y. Chi, J. Fan, and C. Ma (2021). Spectral methods for data science: A statistical perspective. *Foundations and Trends® in Machine Learning 14*(5), 566–806.

Golub, G. H. and C. F. Van Loan (2013). *Matrix Computations, 4th Edition*. Philadelphia, PA: Johns Hopkins University Press.

Gueridi, D. and F. Kittaneh (2022). Inequalities for the kronecker product of matrices. *Ann. Funct. Anal. 13*(3), 50.

Han, R., R. Willett, and A. R. Zhang (2022). An optimal statistical and computational framework for generalized tensor estimation. *Ann. Statist. 50*(1), 1–29.

Hirn, M. (2021). Lecture 01: Introduction to spectral graph theory. Technical report.

Horn, R. A. and C. R. Johnson (2013). *Matrix Analysis* (2nd ed.). New York: Cambridge University Press.

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

Koop, G. M. (2013). Forecasting with medium and large bayesian vars. *J. Appl. Econometrics 28*(2), 177–203.

Masini, R. P., M. C. Medeiros, and E. F. Mendes (2022). Regularized estimation of high-dimensional vector autoregressions with weakly dependent innovations. *J. Time Series Anal. 43*(4), 532–557.

Qiu, D., B. Yang, and X. Zhang (2024). Robust tensor completion via dictionary learning and generalized nonconvex regularization for visual

data recovery. *IEEE Trans. Circuits Syst. Video Technol. 34*(11), 11026–11039.

Rockafellar, R. T. and R. J. B. Wets (1998). *Variational Analysis*. New York: Springer.

Stewart, G. W. and J.-g. Sun (1990). *Matrix Perturbation Theory*. New York: Academic Press.

Wang, D., X. Zhang, G. Li, and R. Tsay (2022). High-dimensional vector autoregression with common response and predictor factors. *arXiv:2203.15170*.

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

Wang, D., Y. Zheng, H. Lian, and G. Li (2022). High-dimensional vector autoregressive time series modeling via tensor decomposition. *J. Amer. Statist. Assoc. 117*(539), 1338–1356.

Wang, H., T. Liu, R. Li, M. X. Cheng, T. Zhao, and J. Gao (2024). Roselora: Row and column-wise sparse low-rank adaptation of pre-trained language model for knowledge editing and fine-tuning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 996–1008. Association for Computational Linguistics.

Wang, H., C. Weng, and J. Yuan (2014). Multi-feature spectral clustering with minimax optimization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4106–4113.

Yang, K., X. Dong, and L. Toni (2019). Error analysis on graph laplacian regularized estimator. *arXiv:1902.03720*.

Yu, Y., T. Wang, and R. J. Samworth (2015). A useful variant of the davis–kahan theorem for statisticians. *Biometrika 102*(2), 315–323.

Zou, H., T. Hastie, and R. Tibshirani (2006). Sparse principal component analysis. *J. Comput. Graph. Statist. 15*(2), 265–286.