Supplementary material:
Semiparametric Nonlinear Bipartite Graph Representation Learning with Provable Guarantees

A. Formulas of gradients and Hessian

For future references, we provide explicit formulas of the gradient and the Hessian for loss (3). We introduce some definitions beforehand. Let us denote each column of weight matrices as $\mathbf{U} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_r)$ and $\mathbf{V} = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_r)$ (similar for $\mathbf{U}^\star$, $\mathbf{V}^\star$). To simplify notations, for a sequence of vectors $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n$, we let $(\boldsymbol{a}_i)_{i=1}^n = (\boldsymbol{a}_1; \ldots; \boldsymbol{a}_n)$ be the long vector by stacking them up; for a sequence of matrices $\mathbf{A}_1, \ldots, \mathbf{A}_n$, we let $\text{diag}\big((\mathbf{A}_i)_{i=1}^n\big)$ be the block diagonal matrix with each block being specified by $\mathbf{A}_i$ sequentially. Moreover, we define the following quantities: $\forall k, l \in [m]$ and $\forall i \in [r]$,

$$
\begin{aligned}
\boldsymbol{d}_{ki} &= \phi_1'(\boldsymbol{u}_i^T \mathbf{x}_{k_1}) \phi_2(\boldsymbol{v}_i^T \mathbf{z}_{k_2}) \mathbf{x}_{k_1}, & \boldsymbol{d}_{li}' &= \phi_1'(\boldsymbol{u}_i^T \mathbf{x}_{l_1}') \phi_2(\boldsymbol{v}_i^T \mathbf{z}_{l_2}') \mathbf{x}_{l_1}', \\
\boldsymbol{p}_{ki} &= \phi_1(\boldsymbol{u}_i^T \mathbf{x}_{k_1}) \phi_2'(\boldsymbol{v}_i^T \mathbf{z}_{k_2}) \mathbf{z}_{k_2}, & \boldsymbol{p}_{li}' &= \phi_1(\boldsymbol{u}_i^T \mathbf{x}_{l_1}') \phi_2'(\boldsymbol{v}_i^T \mathbf{z}_{l_2}') \mathbf{z}_{l_2}', \\
\boldsymbol{Q}_{ki} &= \phi_1''(\boldsymbol{u}_i^T \mathbf{x}_{k_1}) \phi_2(\boldsymbol{v}_i^T \mathbf{z}_{k_2}) \mathbf{x}_{k_1} \mathbf{x}_{k_1}^T, & \boldsymbol{Q}_{li}' &= \phi_1''(\boldsymbol{u}_i^T \mathbf{x}_{l_1}') \phi_2(\boldsymbol{v}_i^T \mathbf{z}_{l_2}') \mathbf{x}_{l_1}' \mathbf{x}_{l_1}'^T, \\
\boldsymbol{R}_{ki} &= \phi_1(\boldsymbol{u}_i^T \mathbf{x}_{k_1}) \phi_2''(\boldsymbol{v}_i^T \mathbf{z}_{k_2}) \mathbf{z}_{k_2} \mathbf{z}_{k_2}^T, & \boldsymbol{R}_{li}' &= \phi_1(\boldsymbol{u}_i^T \mathbf{x}_{l_1}') \phi_2''(\boldsymbol{v}_i^T \mathbf{z}_{l_2}') \mathbf{z}_{l_2}' \mathbf{z}_{l_2}'^T, \\
\boldsymbol{S}_{ki} &= \phi_1'(\boldsymbol{u}_i^T \mathbf{x}_{k_1}) \phi_2'(\boldsymbol{v}_i^T \mathbf{z}_{k_2}) \mathbf{x}_{k_1} \mathbf{z}_{k_2}^T, & \boldsymbol{S}_{li}' &= \phi_1'(\boldsymbol{u}_i^T \mathbf{x}_{l_1}') \phi_2'(\boldsymbol{v}_i^T \mathbf{z}_{l_2}') \mathbf{x}_{l_1}' \mathbf{z}_{l_2}'^T.
\end{aligned}
$$

The quantities on the left part are vectors or matrices calculated by using samples in Ω, which is indexed by k, while the quantities on the right part are calculated by using samples in Ω', which is indexed by l. We should mention that ϕ_i', ϕ_i'' are the first derivative and the second derivative of the activation function ϕ_i (if ϕ_i is ReLU then $\phi_i'' = 0$), while superscript of $\mathbf{x}_{l_1}'$ (and $\mathbf{z}_{l_2}'$) means the sample is from Ω' (i.e. the sample index l is always used with superscript $(\cdot)'$). In addition, we define two scalars as

$$
A_{kl} = \frac{(y_k - y_l')^2 \cdot \exp\big((y_k - y_l')(\boldsymbol{\Theta}_{k_1 k_2} - \boldsymbol{\Theta}_{l_1 l_2}')\big)}{\big(1 + \exp\big((y_k - y_l')(\boldsymbol{\Theta}_{k_1 k_2} - \boldsymbol{\Theta}_{l_1 l_2}')\big)\big)^2}, \quad B_{kl} = \frac{y_k - y_l'}{1 + \exp\big((y_k - y_l')(\boldsymbol{\Theta}_{k_1 k_2} - \boldsymbol{\Theta}_{l_1 l_2}')\big)}.
$$

We define $A_{kl}^\star$, $B_{kl}^\star$ as above by replacing $\boldsymbol{\Theta}_{k_1 k_2}$ with $\boldsymbol{\Theta}_{k_1 k_2}^\star$ and $\boldsymbol{\Theta}_{l_1 l_2}'$ with $\boldsymbol{\Theta}_{l_1 l_2}^{\star\prime}$.

With above definitions and by simple calculations, one can show the gradient is given by

$$
\begin{aligned}
\nabla_{\mathbf{U}} \mathcal{L}(\mathbf{U}, \mathbf{V}) &= \left(\frac{\partial \mathcal{L}(\mathbf{U}, \mathbf{V})}{\partial \boldsymbol{u}_1}, \ldots, \frac{\partial \mathcal{L}(\mathbf{U}, \mathbf{V})}{\partial \boldsymbol{u}_r} \right) \text{ with } \frac{\partial \mathcal{L}(\mathbf{U}, \mathbf{V})}{\partial \boldsymbol{u}_i} = -\frac{1}{m^2} \sum_{k,l=1}^m B_{kl} \left(\boldsymbol{d}_{ki} - \boldsymbol{d}_{li}' \right), \\
\nabla_{\mathbf{V}} \mathcal{L}(\mathbf{U}, \mathbf{V}) &= \left(\frac{\partial \mathcal{L}(\mathbf{U}, \mathbf{V})}{\partial \boldsymbol{v}_1}, \ldots, \frac{\partial \mathcal{L}(\mathbf{U}, \mathbf{V})}{\partial \boldsymbol{v}_r} \right) \text{ with } \frac{\partial \mathcal{L}(\mathbf{U}, \mathbf{V})}{\partial \boldsymbol{v}_i} = -\frac{1}{m^2} \sum_{k,l=1}^m B_{kl} \left(\boldsymbol{p}_{ki} - \boldsymbol{p}_{li}' \right).
\end{aligned}
\tag{5}
$$

Furthermore, $\forall i, j \in [r]$, one can show

$$
\begin{aligned}
\frac{\partial^2 \mathcal{L}(\mathbf{U}, \mathbf{V})}{\partial \boldsymbol{u}_i \partial \boldsymbol{u}_j} &= \frac{1}{m^2} \sum_{k,l=1}^m A_{kl} \left(\boldsymbol{d}_{ki} - \boldsymbol{d}_{li}' \right) \left(\boldsymbol{d}_{kj} - \boldsymbol{d}_{lj}' \right)^T - \frac{\delta_{ij}}{m^2} \sum_{k,l=1}^m B_{kl} \left(\boldsymbol{Q}_{ki} - \boldsymbol{Q}_{li}' \right), \\
\frac{\partial^2 \mathcal{L}(\mathbf{U}, \mathbf{V})}{\partial \boldsymbol{u}_i \partial \boldsymbol{v}_j} &= \frac{1}{m^2} \sum_{k,l=1}^m A_{kl} \left(\boldsymbol{d}_{ki} - \boldsymbol{d}_{li}' \right) \left(\boldsymbol{p}_{kj} - \boldsymbol{p}_{lj}' \right)^T - \frac{\delta_{ij}}{m^2} \sum_{k,l=1}^m B_{kl} \left(\boldsymbol{S}_{ki} - \boldsymbol{S}_{li}' \right), \\
\frac{\partial^2 \mathcal{L}(\mathbf{U}, \mathbf{V})}{\partial \boldsymbol{v}_i \partial \boldsymbol{v}_j} &= \frac{1}{m^2} \sum_{k,l=1}^m A_{kl} \left(\boldsymbol{p}_{ki} - \boldsymbol{p}_{li}' \right) \left(\boldsymbol{p}_{kj} - \boldsymbol{p}_{lj}' \right)^T - \frac{\delta_{ij}}{m^2} \sum_{k,l=1}^m B_{kl} \left(\boldsymbol{R}_{ki} - \boldsymbol{R}_{li}' \right).
\end{aligned}
$$

To combine all blocks and form the Hessian matrix, we will vectorize weight matrices and further define long vectors $\boldsymbol{d}_k = (\boldsymbol{d}_{ki})_{i=1}^r$, $\boldsymbol{p}_k = (\boldsymbol{p}_{ki})_{i=1}^r$, $\boldsymbol{d}_l' = (\boldsymbol{d}_{li}')_{i=1}^r$, $\boldsymbol{p}_l' = (\boldsymbol{p}_{li}')_{i=1}^r$, and block diagonal matrices $\boldsymbol{Q}_k = \text{diag}\left((\boldsymbol{Q}_{ki})_{i=1}^r\right)$, $\boldsymbol{R}_k = \text{diag}\left((\boldsymbol{R}_{ki})_{i=1}^r\right)$, $\boldsymbol{S}_k = \text{diag}\left((\boldsymbol{S}_{ki})_{i=1}^r\right)$ (similar for $\boldsymbol{Q}_l'$, $\boldsymbol{R}_l'$, $\boldsymbol{S}_l'$). Then, the Hessian matrix $\nabla^2 \mathcal{L}(\mathbf{U}, \mathbf{V}) \in \mathbb{R}^{r(d_1+d_2) \times r(d_1+d_2)}$ is

$$
\nabla^2 \mathcal{L}(\mathbf{U}, \mathbf{V}) = \begin{pmatrix} \left(\frac{\partial^2 \mathcal{L}}{\partial \boldsymbol{u}_i \partial \boldsymbol{u}_j} \right)_{i,j} & \left(\frac{\partial^2 \mathcal{L}}{\partial \boldsymbol{u}_i \partial \boldsymbol{v}_j} \right)_{i,j} \\ \left(\frac{\partial^2 \mathcal{L}}{\partial \boldsymbol{v}_i \partial \boldsymbol{u}_j} \right)_{i,j} & \left(\frac{\partial^2 \mathcal{L}}{\partial \boldsymbol{v}_i \partial \boldsymbol{v}_j} \right)_{i,j} \end{pmatrix}
$$

$$= \frac{1}{m^2}\sum_{k,l=1}^{m} A_{kl}\cdot\begin{pmatrix}\boldsymbol{d}_k-\boldsymbol{d}'_l\\ \boldsymbol{p}_k-\boldsymbol{p}'_l\end{pmatrix}\begin{pmatrix}\boldsymbol{d}_k-\boldsymbol{d}'_l\\ \boldsymbol{p}_k-\boldsymbol{p}'_l\end{pmatrix}^T - \frac{1}{m^2}\sum_{k,l=1}^{m} B_{kl}\cdot\begin{pmatrix}\boldsymbol{Q}_k-\boldsymbol{Q}'_l & \boldsymbol{S}_k-\boldsymbol{S}'_l\\ \boldsymbol{S}_k^T-\boldsymbol{S}_l'^T & \boldsymbol{R}_k-\boldsymbol{R}'_l\end{pmatrix}. \tag{6}$$

For all quantities defined above, we add superscript $(\cdot)^\star$ to denote the underlying true quantities, which are obtained by replacing $\mathbf{U},\mathbf{V}$ with true weight matrices $\mathbf{U}^\star,\mathbf{V}^\star$. For example, we have $A^\star_{kl}, B^\star_{kl}, \boldsymbol{d}^\star_{ki}, \boldsymbol{p}^\star_{ki}, \boldsymbol{Q}^\star_{ki}, \boldsymbol{R}^\star_{ki}, \boldsymbol{S}^\star_{ki}, \boldsymbol{d}^\star_k, \boldsymbol{p}^\star_k$. We simplify the notation further by dropping the subscripts of sample index. We let A, B, $\boldsymbol{d},\boldsymbol{q}$, $\boldsymbol{d}'$, $\boldsymbol{q}'$, …, and their corresponding $(\cdot)^\star$ version, denote general references of corresponding quantities, which may be computed by using any samples in $\mathcal{D}$ and $\mathcal{D}'$ (see Assumption 2). We stress that all samples in $\mathcal{D}$ and $\mathcal{D}'$ have the same distribution, so that $\boldsymbol{d}_1,\ldots,\boldsymbol{d}_m\sim\boldsymbol{d}$, $\boldsymbol{p}_1,\ldots,\boldsymbol{p}_m\sim\boldsymbol{p}$, with $\boldsymbol{d}$ and $\boldsymbol{d}'$, and $\boldsymbol{p}$ and $\boldsymbol{p}'$ independent from each other.

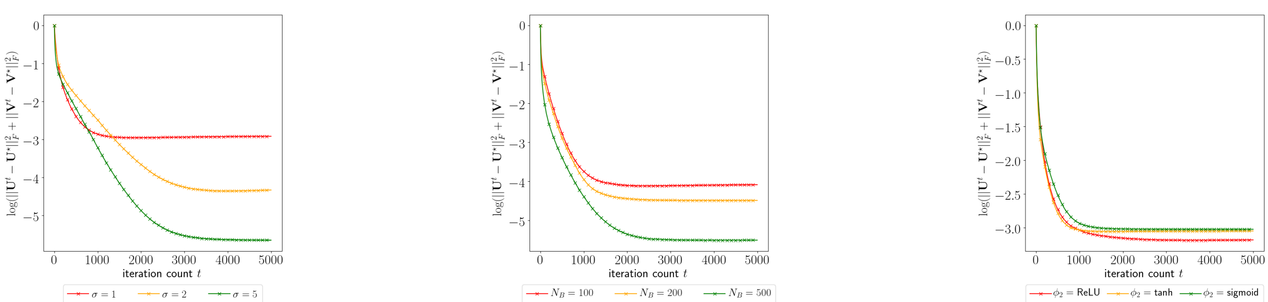
For $i=1,2$, we let $q_i=1$ if ϕ_i is ReLU and $q_i=0$ if $\phi_i\in\{\text{sigmoid},\tanh\}$. Thus,

$$|\phi_i(x)|\le|x|^{q_i},\quad \forall i=1,2. \tag{7}$$

We also let $q=q_1\vee q_2$ and $q'=q_1q_2$.

B. Local Linear Convergence

We verify the local linear convergence of GD on synthetic data sets sampled with ReLU activation functions. We fix $d=d_1=d_2=10$ and $r=3$. The features $\{\mathbf{x}_i,\mathbf{x}'_i\}_{i\in[n_1]}$, $\{\mathbf{z}_j,\mathbf{z}'_j\}_{j\in[n_2]}$, are independently sampled from a Gaussian distribution. We fix $n_1=n_2=400$ and the number of observations $m=2000$. We randomly initialize $(\mathbf{U}^0,\mathbf{V}^0)$ near the ground truth $(\mathbf{U}^\star,\mathbf{V}^\star)$ with fixed error in Frobenius norm. In particular, we fix $||\mathbf{U}^0-\mathbf{U}^\star||_F^2+||\mathbf{V}^0-\mathbf{V}^\star||_F^2=1$. For the Gaussian model, $y\sim\mathcal{N}(\boldsymbol{\Theta}\cdot\sigma^2,\sigma^2)$. For the binomial model, $y\sim B\left((N_B,\frac{\exp(\boldsymbol{\Theta})}{1+\exp(\boldsymbol{\Theta})}\right)$. For Poisson model, $y\sim\text{Pois}(\exp(\boldsymbol{\Theta}))$. To introduce some variations, as well as to verify that our model allows for two separate neural networks, we let $\phi_1=\text{ReLU}$ and $\phi_2\in\{\text{ReLU, sigmoid, tanh}\}$. The estimation error during training process is shown in Figure 3, which verifies the linear convergence rate of GD before reaching the local minima.



(a) Gaussian Model (b) Binomial Model (c) Poisson Model

Figure 3: Local linear convergence of gradient descent on synthetic data sets.

C. Main Lemmas

We summarize lemmas that are required to prove main theorems.

Lemma 7. *For any $k,l\in[m]$, we have that the conditional expectation given all covariates* $\mathbb{E}\left[B^\star_{kl}\mid \mathbf{x}_{k_1},\mathbf{z}_{k_2},\mathbf{x}'_{l_1},\mathbf{z}'_{l_2}\right]=0$.

Lemma 8. *Under Assumptions 1 and 2, there exists a constant $C>0$, independent of $\mathbf{U}^\star$, $\mathbf{V}^\star$, such that:*

(1) if ϕ_1, $\phi_2\in\{\text{sigmoid},\tanh\}$, then

$$\lambda_{\min}\left(\mathbb{E}\left[\begin{pmatrix}\boldsymbol{d}^\star-\boldsymbol{d}^{\star\prime}\\ \boldsymbol{p}^\star-\boldsymbol{p}^{\star\prime}\end{pmatrix}\begin{pmatrix}\boldsymbol{d}^\star-\boldsymbol{d}^{\star\prime}\\ \boldsymbol{p}^\star-\boldsymbol{p}^{\star\prime}\end{pmatrix}^T\right]\right)\ge\frac{C}{\bar\kappa(\mathbf{U}^\star)\bar\kappa(\mathbf{V}^\star)\max(\|\mathbf{U}^\star\|_2^2,\|\mathbf{V}^\star\|_2^2)};$$

(2) if either ϕ_1 or ϕ_2 is ReLU, then by fixing the first row of $\mathbf{U}^\star$ (i.e. treating it as known),

$$\lambda_{\min}\left(\mathbb{E}\left[\begin{pmatrix}\boldsymbol{d}^\star-\boldsymbol{d}^{\star\prime}\\ \boldsymbol{p}^\star-\boldsymbol{p}^{\star\prime}\end{pmatrix}\begin{pmatrix}\boldsymbol{d}^\star-\boldsymbol{d}^{\star\prime}\\ \boldsymbol{p}^\star-\boldsymbol{p}^{\star\prime}\end{pmatrix}^T\right]\right)\ge\frac{C\|\mathbf{e}_1^T\mathbf{U}^\star\|_{\min}^2}{\bar\kappa(\mathbf{U}^\star)\bar\kappa(\mathbf{V}^\star)\max(\|\mathbf{U}^\star\|_2^2,\|\mathbf{V}^\star\|_2^2)(1+\|\mathbf{e}_1^T\mathbf{U}^\star\|_2)^2},$$

where $\mathbf{e}_1 = (1, 0, \ldots, 0) \in \mathbb{R}^{d_1}$.

Lemma 9. *Let* $\nabla^2 \mathcal{L}_1(\mathbf{U}, \mathbf{V}) \coloneqq \frac{1}{m^2}\sum_{k,l=1}^m \mathbf{H}_{1,k,l}$ *where*

$$\mathbf{H}_{1,k,l} = A_{kl} \cdot \begin{pmatrix} \boldsymbol{d}_k - \boldsymbol{d}'_l \\ \boldsymbol{p}_k - \boldsymbol{p}'_l \end{pmatrix} \begin{pmatrix} \boldsymbol{d}_k - \boldsymbol{d}'_l \\ \boldsymbol{p}_k - \boldsymbol{p}'_l \end{pmatrix}^T .$$

Suppose Assumptions 1 and 2 hold. For any $s \geq 1$*, if*

$$m \wedge n_1 \wedge n_2 \gtrsim s(d_1 + d_2) \left\{\log\left(r(d_1 + d_2)\right)\right\}^{1+2q},$$

then

$$\|\nabla^2 \mathcal{L}_1(\mathbf{U}, \mathbf{V}) - \mathbb{E}\left[\nabla^2 \mathcal{L}_1(\mathbf{U}^\star, \mathbf{V}^\star)\right]\|_2 \lesssim \beta^3 r^{\frac{3(1-q)}{2}} \left(\|\mathbf{V}^\star\|_F^{3q} + \|\mathbf{U}^\star\|_F^{3q}\right) \cdot \left(\sqrt{\frac{s(d_1 + d_2)\log\left(r(d_1 + d_2)\right)}{m \wedge n_1 \wedge n_2}} + \left(\|\mathbf{U} - \mathbf{U}^\star\|_F^2 + \|\mathbf{V} - \mathbf{V}^\star\|_F^2\right)^{\frac{2-q}{4}}\right),$$

with probability at least $1 - 1/(d_1 + d_2)^s$.

Lemma 10. *Let* $\nabla^2 \mathcal{L}_2(\mathbf{U}, \mathbf{V}) \coloneqq \frac{1}{m^2}\sum_{k,l=1}^m \mathbf{H}_{2,k,l}$ *where*

$$\mathbf{H}_{2,k,l} = B_{kl} \begin{pmatrix} \boldsymbol{Q}_k - \boldsymbol{Q}'_l & \boldsymbol{S}_k - \boldsymbol{S}'_l \\ \boldsymbol{S}_k^T - \boldsymbol{S}_l'^T & \boldsymbol{R}_k - \boldsymbol{R}'_l \end{pmatrix}.$$

Suppose Assumptions 1 and 2 hold. For any $s \geq 1$*, if*

$$m \wedge n_1 \wedge n_2 \gtrsim s(d_1 + d_2) \left\{\log\left(r(d_1 + d_2)\right)\right\}^{1+q-q'},$$

then

$$\|\nabla^2 \mathcal{L}_2(\mathbf{U}, \mathbf{V}) - \mathbb{E}\left[\nabla^2 \mathcal{L}_2(\mathbf{U}^\star, \mathbf{V}^\star)\right]\|_2 \lesssim \beta^2 r^{\frac{1-q}{2}} \left(\|\mathbf{V}^\star\|_F^{2q} + \|\mathbf{U}^\star\|_F^{2q}\right) \cdot \left(\sqrt{\frac{s(d_1 + d_2)\log\left(r(d_1 + d_2)\right)}{m \wedge n_1 \wedge n_2}} + \left(\|\mathbf{U} - \mathbf{U}^\star\|_F^2 + \|\mathbf{V} - \mathbf{V}^\star\|_F^2\right)^{\frac{2-q}{4}}\right),$$

with probability at least $1 - 1/(d_1 + d_2)^s$.

Lemma 11. *Under Assumption 2,*

$$\|\mathbb{E}[\nabla^2 \mathcal{L}(\mathbf{U}^\star, \mathbf{V}^\star)]\|_2 \lesssim \beta^2 r^{1-q} \left(\|\mathbf{V}^\star\|_F^2 + \|\mathbf{U}^\star\|_F^2\right)^q .$$

D. Proofs of Main Lemmas

D.1. Proof of Lemma 7

For any pair (y_k, y'_l), let R_{kl} denote the rank statistics, and $y_{(\cdot)}^{kl}$ denote the order statistics. We have

$$\mathbb{E}\left[B_{kl}^\star \mid \mathbf{x}_{k_1}, \mathbf{z}_{k_2}, \mathbf{x}'_{l_1}, \mathbf{z}'_{l_2}\right] = \mathbb{E}\left[\mathbb{E}\left[B_{kl}^\star \mid \mathbf{x}_{k_1}, \mathbf{z}_{k_2}, \mathbf{x}'_{l_1}, \mathbf{z}'_{l_2}, y_{(\cdot)}^{kl}\right] \mid \mathbf{x}_{k_1}, \mathbf{z}_{k_2}, \mathbf{x}'_{l_1}, \mathbf{z}'_{l_2}\right].$$

Moreover, as shown in (2),

$$\begin{aligned} P(R_{kl} | y_{(\cdot)}^{kl}, \mathbf{x}_{k_1}, \mathbf{z}_{k_2}, \mathbf{x}'_{l_1}, \mathbf{z}'_{l_2}) &= \frac{\exp\left(y_k \boldsymbol{\Theta}_{k_1k_2}^\star + y'_l \boldsymbol{\Theta}_{l_1l_2}^{\star\prime}\right)}{\exp\left(y_k \boldsymbol{\Theta}_{k_1k_2}^\star + y'_l \boldsymbol{\Theta}_{l_1l_2}^{\star\prime}\right) + \exp\left(y_k \boldsymbol{\Theta}_{l_1l_2}^{\star\prime} + y'_l \boldsymbol{\Theta}_{k_1k_2}^\star\right)} \\ &= \frac{1}{1 + \exp\left(-(y_k - y'_l)(\boldsymbol{\Theta}_{k_1k_2}^\star - \boldsymbol{\Theta}_{l_1l_2}^{\star\prime})\right)}. \end{aligned}$$

Thus,

$$\begin{aligned}
\mathbb{E}\big[B^\star_{kl} \mid \mathbf{x}_{k_1}, \mathbf{z}_{k_2}, \mathbf{x}'_{l_1}, \mathbf{z}'_{l_2}, y^{kl}_{(\cdot)}\big] &= \frac{y_k - y'_l}{1+\exp\left((y_k - y'_l)(\boldsymbol{\Theta}^\star_{k_1k_2} - \boldsymbol{\Theta}^{\star\prime}_{l_1l_2})\right)} \cdot P(R_{kl}|y^{kl}_{(\cdot)}, \mathbf{x}_{k_1}, \mathbf{z}_{k_2}, \mathbf{x}'_{l_1}, \mathbf{z}'_{l_2}) \\
&\quad + \frac{y'_l - y_k}{1+\exp\left((y'_l - y_k)(\boldsymbol{\Theta}^\star_{k_1k_2} - \boldsymbol{\Theta}^{\star\prime}_{l_1l_2})\right)} \left(1 - P(R_{kl}|y^{kl}_{(\cdot)}, \mathbf{x}_{k_1}, \mathbf{z}_{k_2}, \mathbf{x}'_{l_1}, \mathbf{z}'_{l_2})\right) \\
&= \frac{y_k - y'_l}{\left(1+\exp\left((y_k - y'_l)(\boldsymbol{\Theta}^\star_{k_1k_2} - \boldsymbol{\Theta}^{\star\prime}_{l_1l_2})\right)\right)\left(1+\exp\left(-(y_k - y'_l)(\boldsymbol{\Theta}^\star_{k_1k_2} - \boldsymbol{\Theta}^{\star\prime}_{l_1l_2})\right)\right)} \\
&\quad + \frac{y'_l - y_k}{\left(1+\exp\left(-(y_k - y'_l)(\boldsymbol{\Theta}^\star_{k_1k_2} - \boldsymbol{\Theta}^{\star\prime}_{l_1l_2})\right)\right)\left(1+\exp\left((y_k - y'_l)(\boldsymbol{\Theta}^\star_{k_1k_2} - \boldsymbol{\Theta}^{\star\prime}_{l_1l_2})\right)\right)} \\
&= 0,
\end{aligned}$$

which completes the proof.

D.2. Proof of Lemma 8

When $\boldsymbol{d}^{\star\prime} = \mathbf{0}$, $\boldsymbol{p}^{\star\prime} = \mathbf{0}$, and $\phi_1(x) = \phi_2(x)$, Lemma D.1 in Zhong et al. (2018) established a similar result. We prove a generalization of their result here. We first introduce additional notations.

Suppose QR decompositions of $\mathbf{U}^\star, \mathbf{V}^\star$ are $\mathbf{U}^\star = \mathbf{Q}_1\mathbf{R}_1$ and $\mathbf{V}^\star = \mathbf{Q}_2\mathbf{R}_2$, respectively, with $\mathbf{Q}_i \in \mathbb{R}^{d_i \times r}$ and $\mathbf{R}_i \in \mathbb{R}^{r\times r}$ for $i = 1, 2$. Let $\mathbf{Q}_i^\perp \in \mathbb{R}^{d_i\times(d_i-r)}$ be the orthogonal complement of $\mathbf{Q}_i$. For any vectors $\boldsymbol{a} = (\boldsymbol{a}_1; \ldots; \boldsymbol{a}_r)$ and $\boldsymbol{b} = (\boldsymbol{b}_1; \ldots; \boldsymbol{b}_r)$ such that $\boldsymbol{a}_p \in \mathbb{R}^{d_1}$, $\boldsymbol{b}_p \in \mathbb{R}^{d_2}$ for $p \in [r]$ and $\|\boldsymbol{a}\|_2^2 + \|\boldsymbol{b}\|_2^2 = 1$, we express each component by $\boldsymbol{a}_p = \mathbf{Q}_1\boldsymbol{r}_{1p} + \mathbf{Q}_1^\perp\boldsymbol{s}_{1p}$ and $\boldsymbol{b}_p = \mathbf{Q}_2\boldsymbol{r}_{2p} + \mathbf{Q}_2^\perp\boldsymbol{s}_{2p}$, and let $\boldsymbol{r}_i = (\boldsymbol{r}_{i1}, \ldots, \boldsymbol{r}_{ir}) \in \mathbb{R}^{r\times r}$ and $\boldsymbol{s}_i = (\boldsymbol{s}_{i1}, \ldots, \boldsymbol{s}_{ir}) \in \mathbb{R}^{(d_i-r)\times r}$. Further, we let $\boldsymbol{t}_i = (\boldsymbol{t}_{i1}, \ldots, \boldsymbol{t}_{ir}) \in \mathbb{R}^{r\times r}$ with $\boldsymbol{t}_{ip} = \mathbf{R}_i^{-1}\boldsymbol{r}_{ip}$, and also let $\bar{\boldsymbol{t}}_i \in \mathbb{R}^{r\times r}$ denote the matrix that replaces the diagonal entries of $\boldsymbol{t}_i$ by 0. Lastly, for $i = 1, 2$ and variable $x \sim \mathcal{N}(0, 1)$, we define following quantities

$$\tau_{i,j,k} = \mathbb{E}[(\phi_i(x))^j x^k], \qquad \tau'_{i,j,k} = \mathbb{E}[(\phi'_i(x))^j x^k], \qquad \tau''_i = \mathbb{E}[\phi_i(x)\phi'_i(x)x].$$

Using the above notations,

$$\begin{aligned}
&\begin{pmatrix}\boldsymbol{a}^T & \boldsymbol{b}^T\end{pmatrix} \mathbb{E}\left[\begin{pmatrix}\boldsymbol{d}^\star - \boldsymbol{d}^{\star\prime} \\ \boldsymbol{p}^\star - \boldsymbol{p}^{\star\prime}\end{pmatrix}\begin{pmatrix}\boldsymbol{d}^\star - \boldsymbol{d}^{\star\prime} \\ \boldsymbol{p}^\star - \boldsymbol{p}^{\star\prime}\end{pmatrix}^T\right]\begin{pmatrix}\boldsymbol{a} \\ \boldsymbol{b}\end{pmatrix} \\
&= \mathbb{E}\Bigg[\bigg(\sum_{p=1}^r \left(\phi'_1(\boldsymbol{u}_p^{\star T}\mathbf{x})\phi_2(\boldsymbol{v}_p^{\star T}\mathbf{z})\boldsymbol{a}_p^T\mathbf{x} + \phi_1(\boldsymbol{u}_p^{\star T}\mathbf{x})\phi'_2(\boldsymbol{v}_p^{\star T}\mathbf{z})\boldsymbol{b}_p^T\mathbf{z}\right) \\
&\qquad - \sum_{p=1}^r \left(\phi'_1(\boldsymbol{u}_p^{\star T}\mathbf{x}')\phi_2(\boldsymbol{v}_p^{\star T}\mathbf{z}')\boldsymbol{a}_p^T\mathbf{x}' + \phi_1(\boldsymbol{u}_p^{\star T}\mathbf{x}')\phi'_2(\boldsymbol{v}_p^{\star T}\mathbf{z}')\boldsymbol{b}_p^T\mathbf{z}'\right)\bigg)^2\Bigg] \\
&= 2\mathrm{Var}\bigg(\sum_{p=1}^r \left(\phi'_1(\boldsymbol{u}_p^{\star T}\mathbf{x})\phi_2(\boldsymbol{v}_p^{\star T}\mathbf{z})\boldsymbol{a}_p^T\mathbf{x} + \phi_1(\boldsymbol{u}_p^{\star T}\mathbf{x})\phi'_2(\boldsymbol{v}_p^{\star T}\mathbf{z})\boldsymbol{b}_p^T\mathbf{z}\right)\bigg). \qquad (8)
\end{aligned}$$

Plugging the expression of each component of $\boldsymbol{a}$, $\boldsymbol{b}$ in (8),

$$\begin{aligned}
&\frac{1}{2}\begin{pmatrix}\boldsymbol{a}^T & \boldsymbol{b}^T\end{pmatrix} \mathbb{E}\left[\begin{pmatrix}\boldsymbol{d}^\star - \boldsymbol{d}^{\star\prime} \\ \boldsymbol{p}^\star - \boldsymbol{p}^{\star\prime}\end{pmatrix}\begin{pmatrix}\boldsymbol{d}^\star - \boldsymbol{d}^{\star\prime} \\ \boldsymbol{p}^\star - \boldsymbol{p}^{\star\prime}\end{pmatrix}^T\right]\begin{pmatrix}\boldsymbol{a} \\ \boldsymbol{b}\end{pmatrix} \\
&= \mathrm{Var}\bigg(\sum_{p=1}^r \left(\phi'_1(\boldsymbol{u}_p^{\star T}\mathbf{x})\phi_2(\boldsymbol{v}_p^{\star T}\mathbf{z})\mathbf{x}^T\mathbf{Q}_1\boldsymbol{r}_{1p} + \phi_1(\boldsymbol{u}_p^{\star T}\mathbf{x})\phi'_2(\boldsymbol{v}_p^{\star T}\mathbf{z})\mathbf{z}^T\mathbf{Q}_2\boldsymbol{r}_{2p}\right) \\
&\qquad + \sum_{p=1}^r \phi'_1(\boldsymbol{u}_p^{\star T}\mathbf{x})\phi_2(\boldsymbol{v}_p^{\star T}\mathbf{z})\mathbf{x}^T\mathbf{Q}_1^\perp\boldsymbol{s}_{1p} + \sum_{p=1}^r \phi_1(\boldsymbol{u}_p^{\star T}\mathbf{x})\phi'_2(\boldsymbol{v}_p^{\star T}\mathbf{z})\mathbf{z}^T\mathbf{Q}_2^\perp\boldsymbol{s}_{2p}\bigg) \\
&= \mathrm{Var}\bigg(\sum_{p=1}^r \left(\phi'_1(\boldsymbol{u}_p^{\star T}\mathbf{x})\phi_2(\boldsymbol{v}_p^{\star T}\mathbf{z})\mathbf{x}^T\mathbf{Q}_1\boldsymbol{r}_{1p} + \phi_1(\boldsymbol{u}_p^{\star T}\mathbf{x})\phi'_2(\boldsymbol{v}_p^{\star T}\mathbf{z})\mathbf{z}^T\mathbf{Q}_2\boldsymbol{r}_{2p}\right)\bigg)
\end{aligned}$$

$$
\begin{aligned}
&+\operatorname{Var}\left(\sum_{p=1}^{r}\phi_1'(\boldsymbol{u}_p^{\star T}\mathbf{x})\phi_2(\boldsymbol{v}_p^{\star T}\mathbf{z})\mathbf{x}^T\mathbf{Q}_1^{\perp}\boldsymbol{s}_{1p}\right)+\operatorname{Var}\left(\sum_{p=1}^{r}\phi_1(\boldsymbol{u}_p^{\star T}\mathbf{x})\phi_2'(\boldsymbol{v}_p^{\star T}\mathbf{z})\mathbf{z}^T\mathbf{Q}_2^{\perp}\boldsymbol{s}_{2p}\right)\\
=&\operatorname{Var}\left(\sum_{p=1}^{r}\left(\phi_1'(\boldsymbol{u}_p^{\star T}\mathbf{x})\phi_2(\boldsymbol{v}_p^{\star T}\mathbf{z})\mathbf{x}^T\mathbf{Q}_1\boldsymbol{r}_{1p}+\phi_1(\boldsymbol{u}_p^{\star T}\mathbf{x})\phi_2'(\boldsymbol{v}_p^{\star T}\mathbf{z})\mathbf{z}^T\mathbf{Q}_2\boldsymbol{r}_{2p}\right)\right)\\
&+\mathbb{E}\left[\left(\sum_{p=1}^{r}\phi_1'(\boldsymbol{u}_p^{\star T}\mathbf{x})\phi_2(\boldsymbol{v}_p^{\star T}\mathbf{z})\mathbf{x}^T\mathbf{Q}_1^{\perp}\boldsymbol{s}_{1p}\right)^2\right]+\mathbb{E}\left[\left(\sum_{p=1}^{r}\phi_1(\boldsymbol{u}_p^{\star T}\mathbf{x})\phi_2'(\boldsymbol{v}_p^{\star T}\mathbf{z})\mathbf{z}^T\mathbf{Q}_2^{\perp}\boldsymbol{s}_{2p}\right)^2\right]\\
=:&\mathcal{I}_1+\mathcal{I}_2+\mathcal{I}_3,
\end{aligned}
\tag{9}
$$

where the second equality is due to the independence among $\mathbf{x}^T\mathbf{Q}_1\boldsymbol{r}_{1p}$, $\mathbf{x}^T\mathbf{Q}_1^{\perp}\boldsymbol{s}_{1p}$, $\mathbf{z}^T\mathbf{Q}_2\boldsymbol{r}_{2p}$ and $\mathbf{z}^T\mathbf{Q}_2^{\perp}\boldsymbol{s}_{2p}$; the third equality is due to the fact that the last two terms have mean zero. By Lemma 12, there exists a constant C_1 not depending on $(\mathbf{U}^\star,\mathbf{V}^\star)$ such that

$$
\mathcal{I}_2+\mathcal{I}_3\geq\frac{C_1}{\bar{\kappa}(\mathbf{U}^\star)\bar{\kappa}(\mathbf{V}^\star)}\left(\|\boldsymbol{s}_1\|_F^2+\|\boldsymbol{s}_2\|_F^2\right). \tag{10}
$$

For term $\mathcal{I}_1$, let us denote the inside variable as

$$
g(\mathbf{U}^{\star T}\mathbf{x},\mathbf{V}^{\star T}\mathbf{z})=\sum_{p=1}^{r}\left(\phi_1'(\boldsymbol{u}_p^{\star T}\mathbf{x})\phi_2(\boldsymbol{v}_p^{\star T}\mathbf{z})\mathbf{x}^T\mathbf{U}^\star\boldsymbol{t}_{1p}+\phi_1(\boldsymbol{u}_p^{\star T}\mathbf{x})\phi_2'(\boldsymbol{v}_p^{\star T}\mathbf{z})\mathbf{z}^T\mathbf{V}^\star\boldsymbol{t}_{2p}\right).
$$

Using Lemma 19, Assumption 1, and independence among $\mathbf{x},\mathbf{x}',\mathbf{z},\mathbf{z}'$,

$$
\begin{aligned}
\mathcal{I}_1&=\operatorname{Var}(g(\mathbf{U}^{\star T}\mathbf{x},\mathbf{V}^{\star T}\mathbf{z}))=\frac{1}{2}\mathbb{E}\left[\left(g(\mathbf{U}^{\star T}\mathbf{x},\mathbf{V}^{\star T}\mathbf{z})-g(\mathbf{U}^{\star T}\mathbf{x}',\mathbf{V}^{\star T}\mathbf{z}')\right)^2\right]\\
&\geq\frac{1}{2\bar{\kappa}(\mathbf{U}^\star)\bar{\kappa}(\mathbf{V}^\star)}\mathbb{E}\left[\left(g(\bar{\mathbf{x}},\bar{\mathbf{z}})-g(\bar{\mathbf{x}}',\bar{\mathbf{z}}')\right)^2\right]=\frac{1}{\bar{\kappa}(\mathbf{U}^\star)\bar{\kappa}(\mathbf{V}^\star)}\operatorname{Var}(g(\bar{\mathbf{x}},\bar{\mathbf{z}})).
\end{aligned}
\tag{11}
$$

Here $\bar{\mathbf{x}},\bar{\mathbf{x}}',\bar{\mathbf{z}},\bar{\mathbf{z}}'$ are standard Gaussian random vectors with dimension r. With some abuse of notations we let $\mathbf{x},\mathbf{z}$ denote two independent Gaussian vectors, whose dimensions may be d_1,d_2, or r, which are clear from the context. By the definition of $g(\cdot,\cdot)$,

$$
g(\mathbf{x},\mathbf{z})=\sum_{p=1}^{r}\left(\phi_1'(\mathbf{x}_p)\phi_2(\mathbf{z}_p)\mathbf{x}^T\boldsymbol{t}_{1p}+\phi_1(\mathbf{x}_p)\phi_2'(\mathbf{z}_p)\mathbf{z}^T\boldsymbol{t}_{2p}\right).
$$

Therefore,

$$
\mathbb{E}\left[g(\mathbf{x},\mathbf{z})\right]=\tau_{1,1,1}'\tau_{2,1,0}\operatorname{Trace}(\boldsymbol{t}_1)+\tau_{1,1,0}\tau_{2,1,1}'\operatorname{Trace}(\boldsymbol{t}_2) \tag{12}
$$

and

$$
\begin{aligned}
\mathbb{E}\left[g^2(\mathbf{x},\mathbf{z})\right]=&\,\mathbb{E}\left[\left(\sum_{p=1}^{r}\phi_1'(\mathbf{x}_p)\phi_2(\mathbf{z}_p)\mathbf{x}^T\boldsymbol{t}_{1p}\right)^2\right]+\mathbb{E}\left[\left(\sum_{p=1}^{r}\phi_1(\mathbf{x}_p)\phi_2'(\mathbf{z}_p)\mathbf{z}^T\boldsymbol{t}_{2p}\right)^2\right]\\
&+2\sum_{1\leq p,q\leq r}\mathbb{E}\left[\phi_1'(\mathbf{x}_p)\phi_2(\mathbf{z}_p)\phi_1(\mathbf{x}_q)\phi_2'(\mathbf{z}_q)\mathbf{x}^T\boldsymbol{t}_{1p}\mathbf{z}^T\boldsymbol{t}_{2q}\right]=:\mathcal{I}_4+\mathcal{I}_5+2\mathcal{I}_6.
\end{aligned}
\tag{13}
$$

From Lemma 13, we have

$$
\begin{aligned}
\mathcal{I}_4=&\left(\tau_{2,2,0}\tau_{1,2,0}'-\tau_{2,1,0}^2(\tau_{1,1,0}')^2\right)\|\bar{\boldsymbol{t}}_1\|_F^2+\tau_{2,1,0}^2(\tau_{1,1,1}')^2\operatorname{Trace}(\bar{\boldsymbol{t}}_1^2)+\tau_{2,1,0}^2(\tau_{1,1,0}')^2\|\bar{\boldsymbol{t}}_1\mathbf{1}\|_2^2\\
&+2\tau_{2,1,0}^2\tau_{1,1,2}'\tau_{1,1,0}'\mathbf{1}^T\bar{\boldsymbol{t}}_1^T\operatorname{diag}(\boldsymbol{t}_1)+\tau_{2,1,0}^2(\tau_{1,1,1}')^2(\mathbf{1}^T\operatorname{diag}(\boldsymbol{t}_1))^2+\left(\tau_{2,2,0}\tau_{1,2,2}'-\tau_{2,1,0}^2(\tau_{1,1,1}')^2\right)\|\operatorname{diag}(\boldsymbol{t}_1)\|_2^2,\\
\mathcal{I}_5=&\left(\tau_{1,2,0}\tau_{2,2,0}'-\tau_{1,1,0}^2(\tau_{2,1,0}')^2\right)\|\bar{\boldsymbol{t}}_2\|_F^2+\tau_{1,1,0}^2(\tau_{2,1,1}')^2\operatorname{Trace}(\bar{\boldsymbol{t}}_2^2)+\tau_{1,1,0}^2(\tau_{2,1,0}')^2\|\bar{\boldsymbol{t}}_2\mathbf{1}\|_2^2\\
&+2\tau_{1,1,0}^2\tau_{2,1,2}'\tau_{2,1,0}'\mathbf{1}^T\bar{\boldsymbol{t}}_2^T\operatorname{diag}(\boldsymbol{t}_2)+\tau_{1,1,0}^2(\tau_{2,1,1}')^2(\mathbf{1}^T\operatorname{diag}(\boldsymbol{t}_2))^2+\left(\tau_{1,2,0}\tau_{2,2,2}'-\tau_{1,1,0}^2(\tau_{2,1,1}')^2\right)\|\operatorname{diag}(\boldsymbol{t}_2)\|_2^2,
\end{aligned}
$$

and

$$\begin{aligned}\mathcal{I}_6 &= \left(\tau_1''\tau_2'' - \tau_{1,1,0}\tau_{2,1,0}\tau'_{1,1,1}\tau'_{2,1,1}\right)\mathrm{diag}(\boldsymbol{t}_1)^T\mathrm{diag}(\boldsymbol{t}_2) + \tau_{1,1,1}\tau_{2,1,1}\tau'_{1,1,0}\tau'_{2,1,0}\mathrm{Trace}(\bar{\boldsymbol{t}}_1\bar{\boldsymbol{t}}_2)\\ &+ \tau_{1,1,0}\tau_{2,1,0}\tau'_{1,1,1}\tau'_{2,1,1}\mathbf{1}^T\mathrm{diag}(\boldsymbol{t}_1)\mathrm{diag}(\boldsymbol{t}_2)^T\mathbf{1} + \tau_{1,1,0}\tau_{2,1,1}\tau'_{1,1,1}\tau'_{2,1,0}\mathbf{1}^T\bar{\boldsymbol{t}}_2^T\mathrm{diag}(\boldsymbol{t}_1)\\ &+ \tau_{1,1,1}\tau_{2,1,0}\tau'_{1,1,0}\tau'_{2,1,1}\mathbf{1}^T\bar{\boldsymbol{t}}_1^T\mathrm{diag}(\boldsymbol{t}_2).\end{aligned}$$

Using the fact that

$$\mathrm{Trace}(\bar{\boldsymbol{t}}_1^2) = \frac{1}{2}\|\bar{\boldsymbol{t}}_1 + \bar{\boldsymbol{t}}_1^T\|_F^2 - \|\bar{\boldsymbol{t}}_1\|_F^2, \qquad 2\mathrm{Trace}(\bar{\boldsymbol{t}}_1\bar{\boldsymbol{t}}_2) = \|\bar{\boldsymbol{t}}_1 + \bar{\boldsymbol{t}}_2^T\|_F^2 - \|\bar{\boldsymbol{t}}_1\|_F^2 - \|\bar{\boldsymbol{t}}_2\|_F^2,$$

it follows from (12) and (13) that

$$\begin{aligned}&\mathrm{Var}(g(\mathbf{x},\mathbf{z})) = \mathbb{E}\left[g^2(\mathbf{x},\mathbf{z})\right] - \left(\mathbb{E}\left[g(\mathbf{x},\mathbf{z})\right]\right)^2\\ &= \tau_{1,1,1}\tau_{2,1,1}\tau'_{1,1,0}\tau'_{2,1,0}\|\bar{\boldsymbol{t}}_1 + \bar{\boldsymbol{t}}_2^T\|_F^2 + \frac{1}{2}\tau_{2,1,0}^2(\tau'_{1,1,1})^2\|\bar{\boldsymbol{t}}_1 + \bar{\boldsymbol{t}}_1^T\|_F^2 + \frac{1}{2}\tau_{1,1,0}^2(\tau'_{2,1,1})^2\|\bar{\boldsymbol{t}}_2 + \bar{\boldsymbol{t}}_2^T\|_F^2\\ &\quad + \left(\tau_{2,2,0}\tau'_{1,2,0} - \tau_{2,1,0}^2(\tau'_{1,1,0})^2 - \tau_{1,1,1}\tau_{2,1,1}\tau'_{1,1,0}\tau'_{2,1,0} - \tau_{2,1,0}^2(\tau'_{1,1,1})^2\right)\|\bar{\boldsymbol{t}}_1\|_F^2\\ &\quad + \left(\tau_{1,2,0}\tau'_{2,2,0} - \tau_{1,1,0}^2(\tau'_{2,1,0})^2 - \tau_{1,1,1}\tau_{2,1,1}\tau'_{1,1,0}\tau'_{2,1,0} - \tau_{1,1,0}^2(\tau'_{2,1,1})^2\right)\|\bar{\boldsymbol{t}}_2\|_F^2\\ &\quad + \|\tau_{2,1,0}\tau'_{1,1,0}\bar{\boldsymbol{t}}_1\mathbf{1} + \tau_{2,1,0}\tau'_{1,1,2}\mathrm{diag}(\boldsymbol{t}_1) + \tau_{1,1,1}\tau'_{2,1,1}\mathrm{diag}(\boldsymbol{t}_2)\|_2^2\\ &\quad + \|\tau_{1,1,0}\tau'_{2,1,0}\bar{\boldsymbol{t}}_2\mathbf{1} + \tau_{1,1,0}\tau'_{2,1,2}\mathrm{diag}(\boldsymbol{t}_2) + \tau_{2,1,1}\tau'_{1,1,1}\mathrm{diag}(\boldsymbol{t}_1)\|_2^2\\ &\quad + \left(\tau_{2,2,0}\tau'_{1,2,2} - \tau_{2,1,0}^2(\tau'_{1,1,1})^2 - \tau_{2,1,0}^2(\tau'_{1,1,2})^2 - \tau_{2,1,1}^2(\tau'_{1,1,1})^2\right)\|\mathrm{diag}(\boldsymbol{t}_1)\|_2^2\\ &\quad + \left(\tau_{1,2,0}\tau'_{2,2,2} - \tau_{1,1,0}^2(\tau'_{2,1,1})^2 - \tau_{1,1,0}^2(\tau'_{2,1,2})^2 - \tau_{1,1,1}^2(\tau'_{2,1,1})^2\right)\|\mathrm{diag}(\boldsymbol{t}_2)\|_2^2\\ &\quad + 2\left(\tau_1''\tau_2'' - \tau_{1,1,0}\tau_{2,1,0}\tau'_{1,1,1}\tau'_{2,1,1} - \tau_{2,1,0}\tau_{1,1,1}\tau'_{1,1,2}\tau'_{2,1,1} - \tau_{1,1,0}\tau_{2,1,1}\tau'_{2,1,2}\tau'_{1,1,1}\right)\mathrm{diag}(\boldsymbol{t}_1)^T\mathrm{diag}(\boldsymbol{t}_2). \qquad (14)\end{aligned}$$

By Lemma 14 we obtain a lower bound $\mathrm{Var}(g(\mathbf{x},\mathbf{z}))$, which in turn gives the lower bound on $\mathcal{I}_1$ by combining with (11). We have two cases.

Case 1. By Lemma 14 (1), we plug the lower bound of (14) into (11) and have that, for some constant $C_2 > 0$ not depending on $(\mathbf{U}^\star, \mathbf{V}^\star)$,

$$\begin{aligned}\mathcal{I}_1 &\geq \frac{C_2}{\bar{\kappa}(\mathbf{U}^\star)\bar{\kappa}(\mathbf{V}^\star)}\left(\|\boldsymbol{t}_1\|_F^2 + \|\boldsymbol{t}_2\|_F^2\right) = \frac{C_2}{\bar{\kappa}(\mathbf{U}^\star)\bar{\kappa}(\mathbf{V}^\star)}\left(\|\mathbf{R}_1^{-1}\boldsymbol{r}_1\|_F^2 + \|\mathbf{R}_2^{-1}\boldsymbol{r}_2\|_F^2\right)\\ &\geq \frac{C_2}{\bar{\kappa}(\mathbf{U}^\star)\bar{\kappa}(\mathbf{V}^\star)\max(\|\mathbf{U}^\star\|_2^2, \|\mathbf{V}^\star\|_2^2)}\left(\|\boldsymbol{r}_1\|_F^2 + \|\boldsymbol{r}_2\|_F^2\right).\end{aligned}$$

Combining the above display with (9) and (10),

$$\begin{aligned}&\begin{pmatrix}\boldsymbol{a}^T & \boldsymbol{b}^T\end{pmatrix}\mathbb{E}\left[\begin{pmatrix}\boldsymbol{d}^\star - \boldsymbol{d}^{\star\prime}\\ \boldsymbol{p}^\star - \boldsymbol{p}^{\star\prime}\end{pmatrix}\begin{pmatrix}\boldsymbol{d}^\star - \boldsymbol{d}^{\star\prime}\\ \boldsymbol{p}^\star - \boldsymbol{p}^{\star\prime}\end{pmatrix}^T\right]\begin{pmatrix}\boldsymbol{a}\\ \boldsymbol{b}\end{pmatrix}\\ &\qquad\qquad \geq \frac{\min(C_1, C_2)}{\bar{\kappa}(\mathbf{U}^\star)\bar{\kappa}(\mathbf{V}^\star)\max(\|\mathbf{U}^\star\|_2^2, \|\mathbf{V}^\star\|_2^2)}\left(\|\boldsymbol{r}_1\|_F^2 + \|\boldsymbol{r}_2\|_F^2 + \|\boldsymbol{s}_1\|_F^2 + \|\boldsymbol{s}_2\|_F^2\right).\end{aligned}$$

Minimizing over the set $\{(\boldsymbol{a}, \boldsymbol{b}) : \|\boldsymbol{a}\|_F^2 + \|\boldsymbol{b}\|_F^2 = 1\}$ on both sides, we have

$$\lambda_{\min}\left(\mathbb{E}\left[\begin{pmatrix}\boldsymbol{d}^\star - \boldsymbol{d}^{\star\prime}\\ \boldsymbol{p}^\star - \boldsymbol{p}^{\star\prime}\end{pmatrix}\begin{pmatrix}\boldsymbol{d}^\star - \boldsymbol{d}^{\star\prime}\\ \boldsymbol{p}^\star - \boldsymbol{p}^{\star\prime}\end{pmatrix}^T\right]\right) \geq \frac{\min(C_1, C_2)}{\bar{\kappa}(\mathbf{U}^\star)\bar{\kappa}(\mathbf{V}^\star)\max(\|\mathbf{U}^\star\|_2^2, \|\mathbf{V}^\star\|_2^2)}. \qquad (15)$$

Case 2. By Lemma 14 (2), we plug the lower bound of (14) into (11) and have that, for some constant $C_3 > 0$,

$$\mathcal{I}_1 \geq \frac{C_3}{\bar{\kappa}(\mathbf{U}^\star)\bar{\kappa}(\mathbf{V}^\star)}\left(\|\bar{\boldsymbol{t}}_1\|_F^2 + \|\bar{\boldsymbol{t}}_2\|_F^2 + \|\mathrm{diag}(\boldsymbol{t}_1) + \mathrm{diag}(\boldsymbol{t}_2)\|_2^2\right).$$

Combining with (9) and (10),

$$\begin{pmatrix}\boldsymbol{a}^T & \boldsymbol{b}^T\end{pmatrix}\mathbb{E}\left[\begin{pmatrix}\boldsymbol{d}^\star-\boldsymbol{d}^{\star\prime}\\ \boldsymbol{p}^\star-\boldsymbol{p}^{\star\prime}\end{pmatrix}\begin{pmatrix}\boldsymbol{d}^\star-\boldsymbol{d}^{\star\prime}\\ \boldsymbol{p}^\star-\boldsymbol{p}^{\star\prime}\end{pmatrix}^T\right]\begin{pmatrix}\boldsymbol{a}\\ \boldsymbol{b}\end{pmatrix}$$
$$\geq \frac{\min(C_1,C_3)}{\bar{\kappa}(\mathbf{U}^\star)\bar{\kappa}(\mathbf{V}^\star)}\left(\|\bar{\boldsymbol{t}}_1\|_F^2+\|\bar{\boldsymbol{t}}_2\|_F^2+\|\mathrm{diag}(\boldsymbol{t}_1)+\mathrm{diag}(\boldsymbol{t}_2)\|_2^2+\|\boldsymbol{s}_1\|_F^2+\|\boldsymbol{s}_2\|_F^2\right).$$

Since the first row of $\mathbf{U}^\star$ is fixed, we minimize over the set $\{(\boldsymbol{a},\boldsymbol{b}):\|\boldsymbol{a}\|_F^2+\|\boldsymbol{b}\|_F^2=1,\mathbf{e}_1^T\boldsymbol{a}_p=0,\forall p\in[r]\}$. Equivalently, the right hand side is minimizing the following optimization problem

$$\begin{aligned}\gamma_{\mathbf{U}^\star}\coloneqq\min_{\boldsymbol{t}_1,\boldsymbol{t}_2,\boldsymbol{s}_1,\boldsymbol{s}_2}\quad & \|\bar{\boldsymbol{t}}_1\|_F^2+\|\bar{\boldsymbol{t}}_2\|_F^2+\|\mathrm{diag}(\boldsymbol{t}_1)+\mathrm{diag}(\boldsymbol{t}_2)\|_2^2+\|\boldsymbol{s}_1\|_F^2+\|\boldsymbol{s}_2\|_F^2\\ \text{s.t.}\quad & \mathbf{R}_1\boldsymbol{t}_1=\boldsymbol{r}_1,\quad \mathbf{R}_2\boldsymbol{t}_2=\boldsymbol{r}_2,\\ & \|\boldsymbol{r}_1\|_F^2+\|\boldsymbol{r}_2\|_F^2+\|\boldsymbol{s}_1\|_F^2+\|\boldsymbol{s}_2\|_F^2=1,\\ & \mathbf{e}_1^T\mathbf{Q}_1\boldsymbol{r}_1+\mathbf{e}_1^T\mathbf{Q}_1^\perp\boldsymbol{s}_1=\mathbf{0}.\end{aligned}$$

By Theorem D.6. in Zhong et al. (2018),

$$\gamma_{\mathbf{U}^\star}\geq\frac{\|\mathbf{e}_1^T\mathbf{U}^\star\|_{\min}^2}{36\max(\|\mathbf{U}^\star\|_2^2,\|\mathbf{V}^\star\|_2^2)(1+\|\mathbf{e}_1^T\mathbf{U}^\star\|_2)^2}.$$

Thus,

$$\lambda_{\min}\left(\mathbb{E}\left[\begin{pmatrix}\boldsymbol{d}^\star-\boldsymbol{d}^{\star\prime}\\ \boldsymbol{p}^\star-\boldsymbol{p}^{\star\prime}\end{pmatrix}\begin{pmatrix}\boldsymbol{d}^\star-\boldsymbol{d}^{\star\prime}\\ \boldsymbol{p}^\star-\boldsymbol{p}^{\star\prime}\end{pmatrix}^T\right]\right)\geq\frac{\min(C_1,C_3)\|\mathbf{e}_1^T\mathbf{U}^\star\|_{\min}^2}{36\bar{\kappa}(\mathbf{U}^\star)\bar{\kappa}(\mathbf{V}^\star)\max(\|\mathbf{U}^\star\|_2^2,\|\mathbf{V}^\star\|_2^2)(1+\|\mathbf{e}_1^T\mathbf{U}^\star\|_2)^2}. \tag{16}$$

Combing (15) and (16) together completes the proof.

D.3. Proof of Lemma 9

The concentration is shown by taking expectation hierarchically. In particular, we let $\nabla^2\bar{\mathcal{L}}_1(\mathbf{U},\mathbf{V})\coloneqq\mathbb{E}\big[\nabla^2\mathcal{L}_1(\mathbf{U},\mathbf{V})\mid\mathcal{D},\mathcal{D}'\big]$, where the expectation is over the random sampling of the entries from $\mathcal{D}$ and $\mathcal{D}'$. Then, we know $\mathbb{E}[\nabla^2\bar{\mathcal{L}}_1(\mathbf{U},\mathbf{V})]=\mathbb{E}[\nabla^2\mathcal{L}_1(\mathbf{U},\mathbf{V})]$. Moreover,

$$\begin{aligned}\|\nabla^2\mathcal{L}_1(\mathbf{U},\mathbf{V})-\mathbb{E}\left[\nabla^2\mathcal{L}_1(\mathbf{U}^\star,\mathbf{V}^\star)\right]\|&\leq\|\nabla^2\mathcal{L}_1(\mathbf{U},\mathbf{V})-\nabla^2\bar{\mathcal{L}}_1(\mathbf{U},\mathbf{V})\|+\|\nabla^2\bar{\mathcal{L}}_1(\mathbf{U},\mathbf{V})-\mathbb{E}\left[\nabla^2\bar{\mathcal{L}}_1(\mathbf{U},\mathbf{V})\right]\|\\&\quad+\|\mathbb{E}\left[\nabla^2\mathcal{L}_1(\mathbf{U},\mathbf{V})\right]-\mathbb{E}\left[\nabla^2\mathcal{L}_1(\mathbf{U}^\star,\mathbf{V}^\star)\right]\|\\&=:\mathcal{J}_1+\mathcal{J}_2+\mathcal{J}_3.\end{aligned}$$

Using Lemma 15, for all $s\geq1$

$$P\left(\mathcal{J}_1+\mathcal{J}_2\gtrsim\beta^2r^{1-q}\sqrt{\frac{s(d_1+d_2)\log\left(r(d_1+d_2)\right)}{m\wedge n_1\wedge n_2}}\left(\|\mathbf{V}\|_F^{2q}+\|\mathbf{U}\|_F^{2q}\right)\right)\lesssim\frac{1}{(d_1+d_2)^s}.$$

By Lemma 17,

$$\mathcal{J}_3\lesssim\beta^3r^{\frac{3(1-q)}{2}}\left(\|\mathbf{V}^\star\|_F^{3q}+\|\mathbf{U}^\star\|_F^{3q}\right)\left(\|\mathbf{U}-\mathbf{U}^\star\|_F^{1-q/2}+\|\mathbf{V}-\mathbf{V}^\star\|_F^{1-q/2}\right).$$

Combining the above two displays, using the fact that

$$\|\mathbf{V}\|_F^{2q}+\|\mathbf{U}\|_F^{2q}\lesssim\|\mathbf{V}-\mathbf{V}^\star\|_F^{2q}+\|\mathbf{U}-\mathbf{U}^\star\|_F^{2q}+\|\mathbf{V}^\star\|_F^{2q}+\|\mathbf{U}^\star\|_F^{2q},$$

and dropping higher order terms, we know that, with probability at least $1-1/(d_1+d_2)^s$,

$$\begin{aligned}&\|\nabla^2\mathcal{L}_1(\mathbf{U},\mathbf{V})-\mathbb{E}\left[\nabla^2\mathcal{L}_1(\mathbf{U}^\star,\mathbf{V}^\star)\right]\|_2\\&\quad\lesssim\beta^3r^{\frac{3(1-q)}{2}}\left(\|\mathbf{V}^\star\|_F^{3q}+\|\mathbf{U}^\star\|_F^{3q}\right)\left(\sqrt{\frac{s(d_1+d_2)\log\left(r(d_1+d_2)\right)}{m\wedge n_1\wedge n_2}}+\|\mathbf{U}-\mathbf{U}^\star\|_F^{1-q/2}+\|\mathbf{V}-\mathbf{V}^\star\|_F^{1-q/2}\right).\end{aligned}$$

Noting that $\|\mathbf{U}-\mathbf{U}^\star\|_F^{1-q/2}+\|\mathbf{V}-\mathbf{V}^\star\|_F^{1-q/2}\lesssim\left(\|\mathbf{U}-\mathbf{U}^\star\|_F^2+\|\mathbf{V}-\mathbf{V}^\star\|_F^2\right)^{\frac{2-q}{4}}$ completes the proof.

D.4. Proof of Lemma 10

The proof is similar to that of Lemma 9. We define $\nabla^2 \bar{\mathcal{L}}_2(\mathbf{U}, \mathbf{V}) = \mathbb{E}[\nabla^2 \mathcal{L}_2(\mathbf{U}, \mathbf{V}) \mid \mathcal{D}, \mathcal{D}']$. Then,

$$\begin{aligned}
\|\nabla^2 \mathcal{L}_2(\mathbf{U}, \mathbf{V}) - &\mathbb{E}\left[\nabla^2 \mathcal{L}_2(\mathbf{U}^\star, \mathbf{V}^\star)\right]\| \\
&\leq \|\nabla^2 \mathcal{L}_2(\mathbf{U}, \mathbf{V}) - \nabla^2 \bar{\mathcal{L}}_2(\mathbf{U}, \mathbf{V})\| + \|\nabla^2 \bar{\mathcal{L}}_2(\mathbf{U}, \mathbf{V}) - \mathbb{E}\left[\nabla^2 \bar{\mathcal{L}}_2(\mathbf{U}, \mathbf{V})\right]\| \\
&\quad + \|\mathbb{E}\left[\nabla^2 \mathcal{L}_2(\mathbf{U}, \mathbf{V})\right] - \mathbb{E}\left[\nabla^2 \mathcal{L}_2(\mathbf{U}^\star, \mathbf{V}^\star)\right]\| \\
&:= \mathcal{T}_1 + \mathcal{T}_2 + \mathcal{T}_3.
\end{aligned}$$

Using Lemma 16 and noting that $\|\mathbf{V}\|_2^{q_2(1-q_1)} + \|\mathbf{U}\|_2^{q_1(1-q_2)} \leq \|\mathbf{V}\|_2^q + \|\mathbf{U}\|_2^q$, for all $s \geq 1$,

$$P\left(\mathcal{T}_1 + \mathcal{T}_2 \gtrsim \beta \sqrt{\frac{s(d_1+d_2)\log\left(r(d_1+d_2)\right)}{m \wedge n_1 \wedge n_2}} \left(\|\mathbf{V}\|_2^q + \|\mathbf{U}\|_2^q\right)\right) \lesssim \frac{1}{(d_1+d_2)^s}.$$

Using Lemma 18,

$$\mathcal{T}_3 \lesssim \beta^2 r^{\frac{1-q}{2}} \left(\|\mathbf{V}^\star\|_F^{2q} + \|\mathbf{U}^\star\|_F^{2q}\right) \left(\|\mathbf{U} - \mathbf{U}^\star\|_F^{1-q/2} + \|\mathbf{V} - \mathbf{V}^\star\|_F^{1-q/2}\right).$$

Combining the last two displays, we complete the proof.

D.5. Proof of Lemma 11

The Hessian, given in Appendix A, can be decomposed as

$$\mathbb{E}[\nabla^2 \mathcal{L}(\mathbf{U}^\star, \mathbf{V}^\star)] = \mathbb{E}[\nabla^2 \mathcal{L}_1(\mathbf{U}^\star, \mathbf{V}^\star)] + \mathbb{E}[\nabla^2 \mathcal{L}_2(\mathbf{U}^\star, \mathbf{V}^\star)] = \mathbb{E}[\nabla^2 \mathcal{L}_1(\mathbf{U}^\star, \mathbf{V}^\star)].$$

The last equality is due to Lemma 7. By (33),

$$\|\mathbb{E}[\nabla^2 \mathcal{L}_1(\mathbf{U}^\star, \mathbf{V}^\star)]\|_2 \lesssim \beta^2 \left(\|\mathbf{V}^\star\|_F^{2q_2} r^{1-q_2} + \|\mathbf{U}^\star\|_F^{2q_1} r^{1-q_1}\right) \lesssim \beta^2 r^{1-q} \left(\|\mathbf{V}^\star\|_F^2 + \|\mathbf{U}^\star\|_F^2\right)^q.$$

This completes the proof.

E. Proofs of Main Theorems

E.1. Proof of Theorem 3

We take $\mathbb{E}\left[\frac{\partial \mathcal{L}(\mathbf{U}^\star, \mathbf{V}^\star)}{\partial \mathbf{U}}\right]$ as an example and $\mathbb{E}\left[\frac{\partial \mathcal{L}(\mathbf{U}^\star, \mathbf{V}^\star)}{\partial \mathbf{V}}\right]$ can be proved similarly. For any $i \in [r]$, by the formula in (5) in Appendix A,

$$\begin{aligned}
\frac{\partial \mathcal{L}(\mathbf{U}^\star, \mathbf{V}^\star)}{\partial \boldsymbol{u}_i} &= -\frac{1}{m^2} \sum_{k,l=1}^m B_{kl}^\star \left(\boldsymbol{d}_{ki}^\star - \boldsymbol{d}_{li}^{\star\prime}\right) \\
&= -\mathbb{E}\left[\frac{1}{m^2} \sum_{k,l=1}^m \mathbb{E}[B_{kl}^\star \mid \mathbf{x}_{k_1}, \mathbf{z}_{k_2}, \mathbf{x}'_{l_1}, \mathbf{z}'_{l_2}] \cdot (\boldsymbol{d}_{ki}^\star - \boldsymbol{d}_{li}^{\star\prime})\right] = \mathbf{0},
\end{aligned}$$

where, for the second term from the end, the outer expectation is taken over randomness in sampling of all covariate, and the last equality is due to Lemma 7. Doing same derivation for each column and we obtain $\mathbb{E}\left[\nabla_{\mathbf{U}} \mathcal{L}(\mathbf{U}^\star, \mathbf{V}^\star)\right] = \mathbf{0}$. Similarly $\mathbb{E}\left[\nabla_{\mathbf{V}} \mathcal{L}(\mathbf{U}^\star, \mathbf{V}^\star)\right] = \mathbf{0}$.

E.2. Proof of Theorem 4

Recall the formula for the Hessian matrix in (6). The second term has zero expectation at $(\mathbf{U}^\star, \mathbf{V}^\star)$ by Lemma 7. Therefore,

$$\mathbb{E}\left[\nabla^2 \mathcal{L}(\mathbf{U}^\star, \mathbf{V}^\star)\right] = \mathbb{E}\left[A^\star \cdot \begin{pmatrix} \boldsymbol{d}^\star - \boldsymbol{d}^{\star\prime} \\ \boldsymbol{p}^\star - \boldsymbol{p}^{\star\prime} \end{pmatrix} \begin{pmatrix} \boldsymbol{d}^\star - \boldsymbol{d}^{\star\prime} \\ \boldsymbol{p}^\star - \boldsymbol{p}^{\star\prime} \end{pmatrix}^T\right] \tag{17}$$

where, as introduced in Appendix A, $A^\star = \frac{(y-y')^2\exp\left((y-y')(\mathbf{\Theta}^\star-\mathbf{\Theta}^{\star\prime})\right)}{(1+\exp((y-y')(\mathbf{\Theta}^\star-\mathbf{\Theta}^{\star\prime})))^2}$, $\boldsymbol{d}^\star = \left(\phi_1'(\boldsymbol{u}_i^{\star T}\mathbf{x})\phi_2(\boldsymbol{v}_i^{\star T}\mathbf{z})\mathbf{x}\right)_{i=1}^r$, $\boldsymbol{p}^\star = \left(\phi_1(\boldsymbol{u}_i^{\star T}\mathbf{x})\phi_2'(\boldsymbol{v}_i^{\star T}\mathbf{z})\mathbf{z}\right)_{i=1}^r$, and $(y,\mathbf{x},\mathbf{z})$ and $(y',\mathbf{x}',\mathbf{z}')$ are two independent samples from $\mathcal{D}$ and $\mathcal{D}'$, respectively. By Assumption 2, $|\mathbf{\Theta}^\star|\vee|\mathbf{\Theta}^{\star\prime}|\leq\alpha$. Thus, $|(y-y')(\mathbf{\Theta}^\star-\mathbf{\Theta}^{\star\prime})|\leq 2\alpha|y-y'|$. Using the symmetry and monotonicity of $\psi(x)$ defined in Assumption 2,

$$\frac{\exp\left((y-y')(\mathbf{\Theta}^\star-\mathbf{\Theta}^{\star\prime})\right)}{\left(1+\exp\left((y-y')(\mathbf{\Theta}^\star-\mathbf{\Theta}^{\star\prime})\right)\right)^2} = \psi\left(|(y-y')(\mathbf{\Theta}^\star-\mathbf{\Theta}^{\star\prime})|\right)\geq\psi(2\alpha|y-y'|).$$

Therefore, $A^\star\geq(y-y')^2\psi(2\alpha|y-y'|)$. Taking conditional expectation in (17),

$$\begin{aligned}\mathbb{E}\left[\nabla^2\mathcal{L}(\mathbf{U}^\star,\mathbf{V}^\star)\right]\succeq&\mathbb{E}\left[(y-y')^2\psi(2\alpha|y-y'|)\cdot\begin{pmatrix}\boldsymbol{d}^\star-\boldsymbol{d}^{\star\prime}\\\boldsymbol{p}^\star-\boldsymbol{p}^{\star\prime}\end{pmatrix}\begin{pmatrix}\boldsymbol{d}^\star-\boldsymbol{d}^{\star\prime}\\\boldsymbol{p}^\star-\boldsymbol{p}^{\star\prime}\end{pmatrix}^T\right]\\ =&\mathbb{E}\left[\mathbb{E}\left[(y-y')^2\psi(2\alpha|y-y'|)\mid\mathbf{x},\mathbf{z},\mathbf{x}',\mathbf{z}'\right]\begin{pmatrix}\boldsymbol{d}^\star-\boldsymbol{d}^{\star\prime}\\\boldsymbol{p}^\star-\boldsymbol{p}^{\star\prime}\end{pmatrix}\begin{pmatrix}\boldsymbol{d}^\star-\boldsymbol{d}^{\star\prime}\\\boldsymbol{p}^\star-\boldsymbol{p}^{\star\prime}\end{pmatrix}^T\right]\\ =&\mathbb{E}\left[M_\alpha(\mathbf{\Theta}^\star,\mathbf{\Theta}^{\star\prime})\cdot\begin{pmatrix}\boldsymbol{d}^\star-\boldsymbol{d}^{\star\prime}\\\boldsymbol{p}^\star-\boldsymbol{p}^{\star\prime}\end{pmatrix}\begin{pmatrix}\boldsymbol{d}^\star-\boldsymbol{d}^{\star\prime}\\\boldsymbol{p}^\star-\boldsymbol{p}^{\star\prime}\end{pmatrix}^T\right].\end{aligned}$$

Here $M_\alpha(\mathbf{\Theta}^\star,\mathbf{\Theta}^{\star\prime})$ is defined in Assumption 2. Note that $|\mathbf{\Theta}^\star|\vee|\mathbf{\Theta}^{\star\prime}|\leq\alpha$ and $M_\alpha(\cdot,\cdot)$ is strictly positive in the area $[-\alpha,\alpha]\times[-\alpha,\alpha]$. Since $M_\alpha(\cdot,\cdot)$ is a continuous function, it attains its minimum value in the compact support. Define

$$\gamma_\alpha=\inf_{[-\alpha,\alpha]\times[-\alpha,\alpha]}M_\alpha(\mathbf{\Theta}_1,\mathbf{\Theta}_2)>0,$$

we further have

$$\mathbb{E}\left[\nabla^2\mathcal{L}(\mathbf{U}^\star,\mathbf{V}^\star)\right]\succeq\gamma_\alpha\mathbb{E}\left[\begin{pmatrix}\boldsymbol{d}^\star-\boldsymbol{d}^{\star\prime}\\\boldsymbol{p}^\star-\boldsymbol{p}^{\star\prime}\end{pmatrix}\begin{pmatrix}\boldsymbol{d}^\star-\boldsymbol{d}^{\star\prime}\\\boldsymbol{p}^\star-\boldsymbol{p}^{\star\prime}\end{pmatrix}^T\right]. \tag{18}$$

Here, γ_α depends on α reciprocally. Combining (18) with Lemma 8, we finish the proof.

E.3. Proof of Theorem 5

Define

$$\begin{aligned}\nabla^2\mathcal{L}_1(\mathbf{U},\mathbf{V})=&\frac{1}{m^2}\sum_{k,l=1}^m A_{kl}\cdot\begin{pmatrix}\boldsymbol{d}_k-\boldsymbol{d}_l'\\\boldsymbol{p}_k-\boldsymbol{p}_l'\end{pmatrix}\begin{pmatrix}\boldsymbol{d}_k-\boldsymbol{d}_l'\\\boldsymbol{p}_k-\boldsymbol{p}_l'\end{pmatrix}^T,\\ \nabla^2\mathcal{L}_2(\mathbf{U},\mathbf{V})=&\frac{1}{m^2}\sum_{k,l=1}^m B_{kl}\begin{pmatrix}\boldsymbol{Q}_k-\boldsymbol{Q}_l' & \boldsymbol{S}_k-\boldsymbol{S}_l'\\\boldsymbol{S}_k^T-\boldsymbol{S}_l'^T & \boldsymbol{R}_k-\boldsymbol{R}_l'\end{pmatrix}.\end{aligned}$$

Then, we know from (6) that $\nabla^2\mathcal{L}(\mathbf{U},\mathbf{V})=\nabla^2\mathcal{L}_1(\mathbf{U},\mathbf{V})+\nabla^2\mathcal{L}_2(\mathbf{U},\mathbf{V})$. Thus,

$$\begin{aligned}&\|\nabla^2\mathcal{L}(\mathbf{U},\mathbf{V})-\mathbb{E}\left[\nabla^2\mathcal{L}(\mathbf{U}^\star,\mathbf{V}^\star)\right]\|_2\\ &\quad\leq\|\nabla^2\mathcal{L}_1(\mathbf{U},\mathbf{V})-\mathbb{E}\left[\nabla^2\mathcal{L}_1(\mathbf{U}^\star,\mathbf{V}^\star)\right]\|_2+\|\nabla^2\mathcal{L}_2(\mathbf{U},\mathbf{V})-\mathbb{E}\left[\nabla^2\mathcal{L}_2(\mathbf{U}^\star,\mathbf{V}^\star)\right]\|_2.\end{aligned}$$

Combining Lemmas 9 and 10, we know the second term only contributes the higher order error. Thus, for all $s\geq1$, with probability at least $1-1/(d_1+d_2)^s$,

$$\begin{aligned}&\|\nabla^2\mathcal{L}(\mathbf{U},\mathbf{V})-\mathbb{E}\left[\nabla^2\mathcal{L}(\mathbf{U}^\star,\mathbf{V}^\star)\right]\|_2\\ &\quad\lesssim\beta^3r^{\frac{3(1-q)}{2}}\left(\|\mathbf{V}^\star\|_F^{3q}+\|\mathbf{U}^\star\|_F^{3q}\right)\left(\sqrt{\frac{s(d_1+d_2)\log\left(r(d_1+d_2)\right)}{m\wedge n_1\wedge n_2}}+\|\mathbf{U}-\mathbf{U}^\star\|_F^{1-q/2}+\|\mathbf{V}-\mathbf{V}^\star\|_F^{1-q/2}\right).\end{aligned}$$

E.4. Proof of Theorem 6

We first bound $\nabla^2\mathcal{L}(\mathbf{U}_1,\mathbf{V}_1)-\nabla^2\mathcal{L}(\mathbf{U}_2,\mathbf{V}_2)$ for any $(\mathbf{U}_1,\mathbf{V}_1)$, $(\mathbf{U}_2,\mathbf{V}_2)\in\mathcal{B}(\mathbf{U}^\star,\mathbf{V}^\star)$. Note that

$$\begin{aligned}\|\nabla^2\mathcal{L}(\mathbf{U}_1,\mathbf{V}_1)-&\nabla^2\mathcal{L}(\mathbf{U}_2,\mathbf{V}_2)\|_2\\ &\leq\|\nabla^2\mathcal{L}(\mathbf{U}_1,\mathbf{V}_1)-\mathbb{E}[\nabla^2\mathcal{L}(\mathbf{U}_1,\mathbf{V}_1)]\|_2+\|\nabla^2\mathcal{L}(\mathbf{U}_2,\mathbf{V}_2)-\mathbb{E}[\nabla^2\mathcal{L}(\mathbf{U}_2,\mathbf{V}_2)]\|_2\\ &\quad+\|\mathbb{E}[\nabla^2\mathcal{L}(\mathbf{U}_1,\mathbf{V}_1)]-\mathbb{E}[\nabla^2\mathcal{L}(\mathbf{U}_2,\mathbf{V}_2)]\|_2.\end{aligned}$$

Using the same derivation as in Lemmas 15, 16, 17, and 18, we can show that with probability $1-1/(d_1+d_2)^s$,

$$\begin{aligned}&\|\nabla^2\mathcal{L}(\mathbf{U}_1,\mathbf{V}_1)-\mathbb{E}[\nabla^2\mathcal{L}(\mathbf{U}_1,\mathbf{V}_1)]\|_2\lesssim\beta^2r^{1-q}\sqrt{\frac{s(d_1+d_2)\log\left(r(d_1+d_2)\right)}{m\wedge n_1\wedge n_2}}\left(\|\mathbf{U}_1\|_F^{2q}+\|\mathbf{V}_1\|_F^{2q}\right),\\ &\|\nabla^2\mathcal{L}(\mathbf{U}_2,\mathbf{V}_2)-\mathbb{E}[\nabla^2\mathcal{L}(\mathbf{U}_2,\mathbf{V}_2)]\|_2\lesssim\beta^2r^{1-q}\sqrt{\frac{s(d_1+d_2)\log\left(r(d_1+d_2)\right)}{m\wedge n_1\wedge n_2}}\left(\|\mathbf{U}_2\|_F^{2q}+\|\mathbf{V}_2\|_F^{2q}\right),\\ &\|\mathbb{E}[\nabla^2\mathcal{L}(\mathbf{U}_1,\mathbf{V}_1)]-\mathbb{E}[\nabla^2\mathcal{L}(\mathbf{U}_2,\mathbf{V}_2)]\|_2\lesssim\beta^3r^{\frac{3(1-q)}{2}}\left(\|\mathbf{U}_2\|_F^{3q}+\|\mathbf{V}_2\|_F^{3q}\right)\left(\|\mathbf{U}_1-\mathbf{U}_2\|_F^2+\|\mathbf{V}_1-\mathbf{V}_2\|_F^2\right)^{\frac{2-q}{4}}.\end{aligned}$$

Noting that $\|\mathbf{U}\|_F^2+\|\mathbf{V}\|_F^2\lesssim\|\mathbf{U}^\star\|_F^2+\|\mathbf{V}^\star\|_F^2$ for $(\mathbf{U},\mathbf{V})\in\mathcal{B}(\mathbf{U}^\star,\mathbf{V}^\star)$, we then have

$$\begin{aligned}&\|\nabla^2\mathcal{L}(\mathbf{U}_1,\mathbf{V}_1)-\nabla^2\mathcal{L}(\mathbf{U}_2,\mathbf{V}_2)\|_2\\ &\quad\lesssim\beta^3r^{\frac{3(1-q)}{2}}\left(\|\mathbf{U}^\star\|_F^{3q}+\|\mathbf{V}^\star\|_F^{3q}\right)\left(\sqrt{\frac{s(d_1+d_2)\log\left(r(d_1+d_2)\right)}{m\wedge n_1\wedge n_2}}+\left(\|\mathbf{U}_1-\mathbf{U}_2\|_F^2+\|\mathbf{V}_1-\mathbf{V}_2\|_F^2\right)^{\frac{2-q}{4}}\right).\end{aligned}$$

Define $\Upsilon^\star=C_{\mathcal{B}}\beta^3r^{\frac{3(1-q)}{2}}\left(\|\mathbf{U}^\star\|_F^{3q}+\|\mathbf{V}^\star\|_F^{3q}\right)$ for sufficiently large constant $C_{\mathcal{B}}$. For any two points $(\mathbf{U}_1,\mathbf{V}_1)$, $(\mathbf{U}_2,\mathbf{V}_2)\in\mathcal{B}_R(\mathbf{U}^\star,\mathbf{V}^\star)$, if their distance satisfies

$$\|\mathbf{U}_1-\mathbf{U}_2\|_F^2+\|\mathbf{V}_1-\mathbf{V}_2\|_F^2\leq\left(\frac{\lambda^\star_{\min}}{20\Upsilon^\star}\right)^{\frac{4}{2-q}},$$

and the sample sizes m,n_1,n_2 satisfy (which is implied by the condition in Theorem 5)

$$m\wedge n_1\wedge n_2\geq\left(\frac{20\Upsilon^\star}{\lambda^\star_{\min}}\right)^2s(d_1+d_2)\log(r(d_1+d_2)),$$

then we know

$$\|\nabla^2\mathcal{L}(\mathbf{U}_1,\mathbf{V}_1)-\nabla^2\mathcal{L}(\mathbf{U}_2,\mathbf{V}_2)\|_2\leq\frac{\lambda^\star_{\min}}{10}.\tag{19}$$

Next, we consider a neighborhood of $(\mathbf{U}^\star,\mathbf{V}^\star)$ with radius $(\frac{\lambda^\star_{\min}}{4\Upsilon^\star})^{\frac{2}{2-q}}$, that is

$$\|\mathbf{U}-\mathbf{U}^\star\|_F^2+\|\mathbf{V}-\mathbf{V}^\star\|_F^2\leq(\frac{\lambda^\star_{\min}}{4\Upsilon^\star})^{\frac{4}{2-q}}.$$

For any $(\mathbf{U},\mathbf{V})$ in this neighborhood, by Weyl's theorem (Weyl, 1912), we can show

$$\begin{aligned}\lambda_{\min}(\nabla^2\mathcal{L}(\mathbf{U},\mathbf{V}))\geq&\lambda_{\min}(\mathbb{E}[\nabla^2\mathcal{L}(\mathbf{U}^\star,\mathbf{V}^\star)])-\|\nabla^2\mathcal{L}(\mathbf{U},\mathbf{V})-\mathbb{E}[\nabla^2\mathcal{L}(\mathbf{U}^\star,\mathbf{V}^\star)]\|_2\\ \geq&\lambda^\star_{\min}-\lambda^\star_{\min}/2\geq\lambda^\star_{\min}/2.\end{aligned}$$

Similarly, $\lambda_{\max}(\nabla^2\mathcal{L}(\mathbf{U},\mathbf{V}))\leq3\lambda^\star_{\max}/2$ where, by Lemma 11, $\lambda^\star_{\max}=\beta^2r^{1-q}\left(\|\mathbf{V}^\star\|_F^2+\|\mathbf{U}^\star\|_F^2\right)^q$. We consider doing one-step GD at $(\mathbf{U},\mathbf{V})$. Let

$$\mathbf{U}'=\mathbf{U}-\eta\nabla_{\mathbf{U}}\mathcal{L}(\mathbf{U},\mathbf{V})\quad\text{and}\quad\mathbf{V}'=\mathbf{V}-\eta\nabla_{\mathbf{V}}\mathcal{L}(\mathbf{U},\mathbf{V}).$$

Suppose the continuous line from $(\mathbf{U}, \mathbf{V})$ to $(\mathbf{U}^\star, \mathbf{V}^\star)$ is parameterized by $\xi \in [0,1]$ with $\mathbf{U}_\xi = \mathbf{U}^\star + \xi(\mathbf{U} - \mathbf{U}^\star)$ and $\mathbf{V}_\xi = \mathbf{V}^\star + \xi(\mathbf{V} - \mathbf{V}^\star)$. Let $\Xi = \{\xi_1, \ldots, \xi_{|\Xi|}\}$ be a $(\frac{1}{5})^{\frac{4}{2-q}}$-net of interval $[0,1]$ with $|\Xi| = 5^{\frac{4}{2-q}} \le 5^4$, and accordingly, we define $(\mathbf{U}_i, \mathbf{V}_i) = (\mathbf{U}_{\xi_i}, \mathbf{V}_{\xi_i})$ for $i \in [|\Xi|]$ and have set $\mathcal{S} = \{(\mathbf{U}_1, \mathbf{V}_1), \ldots, (\mathbf{U}_{|\Xi|}, \mathbf{V}_{|\Xi|})\}$. Taking the union bound over $\mathcal{S}$,

$$P\left(\exists(\mathbf{U},\mathbf{V}) \in \mathcal{S}, \lambda_{\min}(\nabla^2\mathcal{L}(\mathbf{U},\mathbf{V})) \le \frac{\lambda^\star_{\min}}{2} \text{ or } \lambda_{\max}(\nabla^2\mathcal{L}(\mathbf{U},\mathbf{V})) \ge \frac{3\lambda^\star_{\max}}{2}\right) \lesssim \frac{1}{(d_1+d_2)^s}. \tag{20}$$

Furthermore, since Ξ is a net of $[0,1]$, for any $\xi \in [0,1]$ there exists $\xi' \in [|\Xi|]$ such that

$$\|\mathbf{U}_\xi - \mathbf{U}_{\xi'}\|_F^2 + \|\mathbf{V}_\xi - \mathbf{V}_{\xi'}\|_F^2 \le \left(\frac{\lambda^\star_{\min}}{20\Upsilon^\star}\right)^{\frac{4}{2-q}}.$$

Thus, by (19), (20), and Weyl's theorem, we obtain

$$\begin{aligned}
\lambda_{\min}(\nabla^2\mathcal{L}(\mathbf{U}_\xi,\mathbf{V}_\xi)) \ge& \frac{\lambda^\star_{\min}}{2} - \frac{\lambda^\star_{\min}}{10} = \frac{2\lambda^\star_{\min}}{5},\\
\lambda_{\max}(\nabla^2\mathcal{L}(\mathbf{U}_\xi,\mathbf{V}_\xi)) \le& \frac{3\lambda^\star_{\max}}{2} + \frac{\lambda^\star_{\min}}{10} \le \frac{8\lambda^\star_{\max}}{5}.
\end{aligned}$$

With this,

$$\begin{aligned}
&\|\mathbf{U}' - \mathbf{U}^\star\|_F^2 + \|\mathbf{V}' - \mathbf{V}^\star\|_F^2\\
&\quad = \|\mathbf{U} - \mathbf{U}^\star\|_F^2 + \|\mathbf{V} - \mathbf{V}^\star\|_F^2 + \eta^2\|\nabla\mathcal{L}(\mathbf{U},\mathbf{V})\|_F^2\\
&\qquad - 2\eta \underbrace{\text{vec}\begin{pmatrix}\mathbf{U}-\mathbf{U}^\star\\ \mathbf{V}-\mathbf{V}^\star\end{pmatrix}^T\left(\int_0^1 \nabla^2\mathcal{L}(\mathbf{U}_\xi,\mathbf{V}_\xi)d\xi\right)\text{vec}\begin{pmatrix}\mathbf{U}-\mathbf{U}^\star\\ \mathbf{V}-\mathbf{V}^\star\end{pmatrix}}_{\mathbf{H}(\mathbf{U},\mathbf{V})}\\
&\quad \le \|\mathbf{U} - \mathbf{U}^\star\|_F^2 + \|\mathbf{V} - \mathbf{V}^\star\|_F^2 + \left(\frac{8\eta^2\lambda^\star_{\max}}{5} - 2\eta\right)\mathbf{H}(\mathbf{U},\mathbf{V}).
\end{aligned}$$

The last inequality is from Theorem 3 and the fact that $\|\nabla\mathcal{L}(\mathbf{U}^\star,\mathbf{V}^\star) - \mathbb{E}[\nabla\mathcal{L}(\mathbf{U}^\star,\mathbf{V}^\star)]\|_F$ only contributes higher-order terms by concentration. Let $\eta = 1/\lambda^\star_{\max}$, then

$$\begin{aligned}
\|\mathbf{U}' - \mathbf{U}^\star\|_F^2 + \|\mathbf{V}' - \mathbf{V}^\star\|_F^2 &\le \|\mathbf{U} - \mathbf{U}^\star\|_F^2 + \|\mathbf{V} - \mathbf{V}^\star\|_F^2 - \frac{2}{5\lambda^\star_{\max}}\mathbf{H}(\mathbf{U},\mathbf{V})\\
&\le (1 - \frac{\lambda^\star_{\min}}{7\lambda^\star_{\max}})\big(\|\mathbf{U} - \mathbf{U}^\star\|_F^2 + \|\mathbf{V} - \mathbf{V}^\star\|_F^2\big),
\end{aligned}$$

which completes the proof.

F. Complementary Lemmas

In this section, we list intermediate results required for proving lemmas in Appendix C. Notations in each lemma are introduced in the proofs of the corresponding lemmas.

Lemma 12. *Under conditions Lemma 8, there exists a constant $C > 0$ not depending on $\mathbf{U}^\star$, $\mathbf{V}^\star$ such that*

$$\mathcal{I}_2 \ge \frac{C}{\bar{\kappa}(\mathbf{U}^\star)\bar{\kappa}(\mathbf{V}^\star)}\|\boldsymbol{s}_1\|_F^2, \qquad \mathcal{I}_3 \ge \frac{C}{\bar{\kappa}(\mathbf{U}^\star)\bar{\kappa}(\mathbf{V}^\star)}\|\boldsymbol{s}_2\|_F^2.$$

Lemma 13. *Under conditions of Lemma 8, we have*

$$\begin{aligned}
\mathcal{I}_4 =& \left(\tau_{2,2,0}\tau'_{1,2,0} - \tau^2_{2,1,0}(\tau'_{1,1,0})^2\right)\|\bar{\boldsymbol{t}}_1\|_F^2 + \tau^2_{2,1,0}(\tau'_{1,1,1})^2\text{Trace}(\bar{\boldsymbol{t}}_1^2) + \tau^2_{2,1,0}(\tau'_{1,1,0})^2\|\bar{\boldsymbol{t}}_1\mathbf{1}\|_2^2\\
&+ 2\tau^2_{2,1,0}\tau'_{1,1,2}\tau'_{1,1,0}\mathbf{1}^T\bar{\boldsymbol{t}}_1^T\text{diag}(\boldsymbol{t}_1) + \tau^2_{2,1,0}(\tau'_{1,1,1})^2(\mathbf{1}^T\text{diag}(\boldsymbol{t}_1))^2 + \left(\tau_{2,2,0}\tau'_{1,2,2} - \tau^2_{2,1,0}(\tau'_{1,1,1})^2\right)\|\text{diag}(\boldsymbol{t}_1)\|_2^2,\\
\mathcal{I}_5 =& \left(\tau_{1,2,0}\tau'_{2,2,0} - \tau^2_{1,1,0}(\tau'_{2,1,0})^2\right)\|\bar{\boldsymbol{t}}_2\|_F^2 + \tau^2_{1,1,0}(\tau'_{2,1,1})^2\text{Trace}(\bar{\boldsymbol{t}}_2^2) + \tau^2_{1,1,0}(\tau'_{2,1,0})^2\|\bar{\boldsymbol{t}}_2\mathbf{1}\|_2^2
\end{aligned}$$

$$+ 2\tau_{1,1,0}^2\tau_{2,1,2}'\tau_{2,1,0}'\mathbf{1}^T\bar{\boldsymbol{t}}_2^T\mathrm{diag}(\boldsymbol{t}_2) + \tau_{1,1,0}^2(\tau_{2,1,1}')^2(\mathbf{1}^T\mathrm{diag}(\boldsymbol{t}_2))^2 + \left(\tau_{1,2,0}\tau_{2,2,2}' - \tau_{1,1,0}^2(\tau_{2,1,1}')^2\right)\|\mathrm{diag}(\boldsymbol{t}_2)\|_2^2,$$

and

$$\begin{aligned}\mathcal{I}_6 = &\left(\tau_1''\tau_2'' - \tau_{1,1,0}\tau_{2,1,0}\tau_{1,1,1}'\tau_{2,1,1}'\right)\mathrm{diag}(\boldsymbol{t}_1)^T\mathrm{diag}(\boldsymbol{t}_2) + \tau_{1,1,1}\tau_{2,1,1}\tau_{1,1,0}'\tau_{2,1,0}'\mathrm{Trace}(\bar{\boldsymbol{t}}_1\bar{\boldsymbol{t}}_2)\\ &+ \tau_{1,1,0}\tau_{2,1,0}\tau_{1,1,1}'\tau_{2,1,1}'\mathbf{1}^T\mathrm{diag}(\boldsymbol{t}_1)\mathrm{diag}(\boldsymbol{t}_2)^T\mathbf{1} + \tau_{1,1,0}\tau_{2,1,1}\tau_{1,1,1}'\tau_{2,1,0}'\mathbf{1}^T\bar{\boldsymbol{t}}_2^T\mathrm{diag}(\boldsymbol{t}_1)\\ &+ \tau_{1,1,1}\tau_{2,1,0}\tau_{1,1,0}'\tau_{2,1,1}'\mathbf{1}^T\bar{\boldsymbol{t}}_1^T\mathrm{diag}(\boldsymbol{t}_2).\end{aligned}$$

Lemma 14. *Under conditions of Lemma 8, there exists constant $C > 0$ not depending on $\mathbf{U}^\star$, $\mathbf{V}^\star$ such that:*

(1) if ϕ_1, $\phi_2 \in \{sigmoid, tanh\}$, then

$$Var(g(\mathbf{x},\mathbf{z})) \geq C\left(\|\boldsymbol{t}_1\|_F^2 + \|\boldsymbol{t}_2\|_F^2\right);$$

(2) if either ϕ_1 or ϕ_2 is ReLU, then

$$Var(g(\mathbf{x},\mathbf{z})) \geq C\left(\|\bar{\boldsymbol{t}}_1\|_F^2 + \|\bar{\boldsymbol{t}}_2\|_F^2 + \|\mathrm{diag}(\boldsymbol{t}_1) + \mathrm{diag}(\boldsymbol{t}_2)\|_2^2\right).$$

Lemma 15. *Under conditions of Lemma 9, we have*

$$P\left(\mathcal{J}_1 \gtrsim \beta^2\sqrt{\frac{s(d_1+d_2)\log(r(d_1+d_2))}{m}}\left(\|\mathbf{V}\|_F^{2q_2}r^{1-q_2} + \|\mathbf{U}\|_F^{2q_1}r^{1-q_1}\right)\right) \lesssim \frac{1}{(d_1+d_2)^s},$$

$$P\left(\mathcal{J}_2 \gtrsim \beta^2\sqrt{\frac{s(d_1+d_2)\log(r(d_1+d_2))}{n_1\wedge n_2}}\left(\|\mathbf{V}\|_F^{2q_2}r^{1-q_2} + \|\mathbf{U}\|_F^{2q_1}r^{1-q_1}\right)\right) \lesssim \frac{1}{(d_1+d_2)^s},$$

where $\{q_i\}_{i=1,2}$ are defined in Appendix A (see (7)*).*

Lemma 16. *Under conditions of Lemma 10, we have*

$$P\left(\mathcal{T}_1 \gtrsim \beta\sqrt{\frac{s(d_1+d_2)\log(r(d_1+d_2))}{m}}\left(\|\mathbf{V}\|_2^{q_2(1-q_1)} + \|\mathbf{U}\|_2^{q_1(1-q_2)}\right)\right) \lesssim \frac{1}{(d_1+d_2)^s},$$

$$P\left(\mathcal{T}_2 \gtrsim \beta\sqrt{\frac{s(d_1+d_2)\log(r(d_1+d_2))}{n_1\wedge n_2}}\left(\|\mathbf{V}\|_2^{q_2(1-q_1)} + \|\mathbf{U}\|_2^{q_1(1-q_2)}\right)\right) \lesssim \frac{1}{(d_1+d_2)^s}.$$

Lemma 17. *Under conditions of Lemma 9, we have*

$$\mathcal{J}_3 \lesssim \beta^3 r^{\frac{3(1-q)}{2}}\left(\|\mathbf{V}^\star\|_F^{3q} + \|\mathbf{U}^\star\|_F^{3q}\right)\left(\|\mathbf{U}-\mathbf{U}^\star\|_F^{1-q/2} + \|\mathbf{V}-\mathbf{V}^\star\|_F^{1-q/2}\right).$$

Lemma 18. *Under conditions of Lemma 10, we have*

$$\mathcal{T}_3 \lesssim \beta^2 r^{\frac{1-q}{2}}\left(\|\mathbf{V}^\star\|_F^{2q} + \|\mathbf{U}^\star\|_F^{2q}\right)\left(\|\mathbf{U}-\mathbf{U}^\star\|_F^{1-q/2} + \|\mathbf{V}-\mathbf{V}^\star\|_F^{1-q/2}\right).$$

G. Proofs of Other Lemmas

We present proofs of lemmas in Appendix F.

G.1. Proof of Lemma 12

By symmetry, we only show the proof for $\mathcal{I}_2$. By the definition of $\mathcal{I}_2$ in (9),

$$\mathcal{I}_2 = \mathbb{E}\left[\left(\sum_{p=1}^r \phi_1'(\boldsymbol{u}_p^{\star T}\mathbf{x})\phi_2(\boldsymbol{v}_p^{\star T}\mathbf{z})\mathbf{x}^T\mathbf{Q}_1^\perp \boldsymbol{s}_{1p}\right)^2\right]$$

$$
\begin{aligned}
=&\sum_{p=1}^{r}\mathbb{E}\big[(\phi_1'(\boldsymbol{u}_p^{\star T}\mathbf{x}))^2(\phi_2(\boldsymbol{v}_p^{\star T}\mathbf{z}))^2\boldsymbol{s}_{1p}^T(\mathbf{Q}_1^{\perp})^T\mathbf{x}\mathbf{x}^T\mathbf{Q}_1^{\perp}\boldsymbol{s}_{1p}\big]\\
&+\sum_{1\le p\neq q\le r}\mathbb{E}\big[\phi_1'(\boldsymbol{u}_p^{\star T}\mathbf{x})\phi_1'(\boldsymbol{u}_q^{\star T}\mathbf{x})\phi_2(\boldsymbol{v}_p^{\star T}\mathbf{z})\phi_2(\boldsymbol{v}_q^{\star T}\mathbf{z})\boldsymbol{s}_{1q}^T(\mathbf{Q}_1^{\perp})^T\mathbf{x}\mathbf{x}^T\mathbf{Q}_1^{\perp}\boldsymbol{s}_{1p}\big]\\
=&\sum_{p=1}^{r}\mathbb{E}\big[(\phi_1'(\boldsymbol{u}_p^{\star T}\mathbf{x}))^2(\phi_2(\boldsymbol{v}_p^{\star T}\mathbf{z}))^2\boldsymbol{s}_{1p}^T\boldsymbol{s}_{1p}\big]+\sum_{1\le p\neq q\le r}\mathbb{E}\big[\phi_1'(\boldsymbol{u}_p^{\star T}\mathbf{x})\phi_1'(\boldsymbol{u}_q^{\star T}\mathbf{x})\phi_2(\boldsymbol{v}_p^{\star T}\mathbf{z})\phi_2(\boldsymbol{v}_q^{\star T}\mathbf{z})\boldsymbol{s}_{1q}^T\boldsymbol{s}_{1p}\big]\\
=&\mathbb{E}\big[\big\|\sum_{p=1}^{r}\phi_1'(\boldsymbol{u}_p^{\star T}\mathbf{x})\phi_2(\boldsymbol{v}_p^{\star T}\mathbf{z})\boldsymbol{s}_{1p}\big\|^2\big]\\
\geq&\frac{1}{\bar{\kappa}(\mathbf{U}^\star)\bar{\kappa}(\mathbf{V}^\star)}\mathbb{E}\big[\big\|\sum_{p=1}^{r}\phi_1'(\mathbf{x}_p)\phi_2(\mathbf{z}_p)\boldsymbol{s}_{1p}\big\|_2^2\big].
\end{aligned}
\tag{21}
$$

Here the third equality is due to the independence among $\boldsymbol{u}_p^{\star T}\mathbf{x}$, $\mathbf{x}^T\mathbf{Q}_1^{\perp}$ and $\mathbf{z}$; the last inequality is from Lemma 19 and Assumption 1. Further,

$$
\begin{aligned}
\mathbb{E}\big[\big\|\sum_{p=1}^{r}\phi_1'(\mathbf{x}_p)\phi_2(\mathbf{z}_p)\boldsymbol{s}_{1p}\big\|_2^2\big]=&\sum_{p,q=1}^{r}\mathbb{E}\left[\phi_1'(\mathbf{x}_p)\phi_2(\mathbf{z}_p)\phi_1'(\mathbf{x}_q)\phi_2(\mathbf{z}_q)\boldsymbol{s}_{1p}^T\boldsymbol{s}_{1q}\right]\\
=&\tau_{1,2,0}'\tau_{2,2,0}\sum_{p=1}^{r}\|\boldsymbol{s}_{1p}\|^2+(\tau_{1,1,0}')^2(\tau_{2,1,0})^2\sum_{1\le p\neq q\le r}\boldsymbol{s}_{1p}^T\boldsymbol{s}_{1q}\\
=&\tau_{1,2,0}'\tau_{2,2,0}\|\boldsymbol{s}_1\|_F^2+(\tau_{1,1,0}')^2(\tau_{2,1,0})^2\left(\|\boldsymbol{s}_1\mathbf{1}\|_2^2-\|\boldsymbol{s}_1\|_F^2\right)\\
\geq&\left(\tau_{1,2,0}'\tau_{2,2,0}-(\tau_{1,1,0}')^2(\tau_{2,1,0})^2\right)\|\boldsymbol{s}_1\|_F^2.
\end{aligned}
$$

Combining with (21),

$$
\mathcal{I}_2\geq\frac{\tau_{1,2,0}'\tau_{2,2,0}-(\tau_{1,1,0}')^2(\tau_{2,1,0})^2}{\bar{\kappa}(\mathbf{U}^\star)\bar{\kappa}(\mathbf{V}^\star)}\|\boldsymbol{s}_1\|_F^2.
$$

Note that $\tau_{1,2,0}'>(\tau_{1,1,0}')^2$ and $\tau_{2,2,0}>(\tau_{2,1,0})^2$ for all activation functions in $\{\text{sigmoid},\tanh,\text{ReLU}\}$. Thus, for some constant $C>0$ we have $\mathcal{I}_2\geq\frac{C}{\bar{\kappa}(\mathbf{U}^\star)\bar{\kappa}(\mathbf{V}^\star)}\|\boldsymbol{s}_1\|_F^2$. Similarly, we can show

$$
\mathcal{I}_3\geq\frac{\tau_{1,2,0}\tau_{2,2,0}'-(\tau_{1,1,0})^2(\tau_{2,1,0}')^2}{\bar{\kappa}(\mathbf{U}^\star)\bar{\kappa}(\mathbf{V}^\star)}\|\boldsymbol{s}_2\|_F^2\geq\frac{C}{\bar{\kappa}(\mathbf{U}^\star)\bar{\kappa}(\mathbf{V}^\star)}\|\boldsymbol{s}_2\|_F^2.
$$

This completes the proof.

G.2. Proof of Lemma 13

By symmetry, we only show the proof for $\mathcal{I}_4$ and $\mathcal{I}_5$ can be proved analogously. By the definition of $\mathcal{I}_4$ in (13),

$$
\begin{aligned}
\mathcal{I}_4=&\mathbb{E}\left[\left(\sum_{p=1}^{r}\phi_1'(\mathbf{x}_p)\phi_2(\mathbf{z}_p)\mathbf{x}^T\boldsymbol{t}_{1p}\right)^2\right]\\
=&\sum_{p=1}^{r}\mathbb{E}\left[(\phi_1'(\mathbf{x}_p))^2(\phi_2(\mathbf{z}_p))^2\boldsymbol{t}_{1p}^T\mathbf{x}\mathbf{x}^T\boldsymbol{t}_{1p}\right]+\sum_{1\le p\neq q\le r}\mathbb{E}\left[\phi_1'(\mathbf{x}_p)\phi_2(\mathbf{z}_p)\phi_1'(\mathbf{x}_q)\phi_2(\mathbf{z}_q)\boldsymbol{t}_{1p}^T\mathbf{x}\mathbf{x}^T\boldsymbol{t}_{1q}\right]\\
=&\tau_{2,2,0}\sum_{p=1}^{r}\mathbb{E}\left[(\phi_1'(\mathbf{x}_p))^2\boldsymbol{t}_{1p}^T\mathbf{x}\mathbf{x}^T\boldsymbol{t}_{1p}\right]+\tau_{2,1,0}^2\sum_{1\le p\neq q\le r}\mathbb{E}\left[\phi_1'(\mathbf{x}_p)\phi_1'(\mathbf{x}_q)\boldsymbol{t}_{1p}^T\mathbf{x}\mathbf{x}^T\boldsymbol{t}_{1q}\right]\\
:=&\tau_{2,2,0}\mathcal{I}_{41}+\tau_{2,1,0}^2\mathcal{I}_{42}.
\end{aligned}
\tag{22}
$$

By simple derivations, we let $\boldsymbol{t}_{1pp}=[\boldsymbol{t}_{1p}]_p$ be the p-th entry of $\boldsymbol{t}_{1p}$, and have

$$
\mathcal{I}_{41}=(\tau_{1,2,2}'-\tau_{1,2,0}')\sum_{p=1}^{r}\boldsymbol{t}_{1pp}^2+\tau_{1,2,0}'\sum_{p=1}^{r}\|\boldsymbol{t}_{1p}\|_2^2=(\tau_{1,2,2}'-\tau_{1,2,0}')\|\text{diag}(\boldsymbol{t}_1)\|_2^2+\tau_{1,2,0}'\|\boldsymbol{t}_1\|_F^2;
\tag{23}
$$

$$\mathcal{I}_{42} = \sum_{1\le p\neq q\le r}\left((\tau'_{1,1,1})^2(\boldsymbol{t}_{1pp}\boldsymbol{t}_{1qq}+\boldsymbol{t}_{1pq}\boldsymbol{t}_{1qp}) + \tau'_{1,1,2}\tau'_{1,1,0}(\boldsymbol{t}_{1pp}\boldsymbol{t}_{1qp}+\boldsymbol{t}_{1pq}\boldsymbol{t}_{1qq}) + (\tau'_{1,1,0})^2\sum_{\substack{k=1\\k\neq p,q}}^{r}\boldsymbol{t}_{1pk}\boldsymbol{t}_{1qk}\right)$$
$$= \sum_{1\le p\neq q\le r}\big((\tau'_{1,1,1})^2(\boldsymbol{t}_{1pp}\boldsymbol{t}_{1qq}+\boldsymbol{t}_{1pq}\boldsymbol{t}_{1qp}) + (\tau'_{1,1,2}\tau'_{1,1,0}-(\tau'_{1,1,0})^2)(\boldsymbol{t}_{1pp}\boldsymbol{t}_{1qp}+\boldsymbol{t}_{1pq}\boldsymbol{t}_{1qq}) + (\tau'_{1,1,0})^2\boldsymbol{t}_{1p}^T\boldsymbol{t}_{1q}\big).$$

Moreover, for each component of $\mathcal{I}_{42}$ we have

$$\sum_{1\le p\neq q\le r}\boldsymbol{t}_{1pp}\boldsymbol{t}_{1qq}+\boldsymbol{t}_{1pq}\boldsymbol{t}_{1qp} = (\mathbf{1}^T\mathrm{diag}(\boldsymbol{t}_1))^2 + \mathrm{Trace}(\boldsymbol{t}_1^2) - 2\|\mathrm{diag}(\boldsymbol{t}_1)\|_2^2,$$
$$\sum_{1\le p\neq q\le r}\boldsymbol{t}_{1pp}\boldsymbol{t}_{1qp}+\boldsymbol{t}_{1pq}\boldsymbol{t}_{1qq} = 2\sum_{1\le p\neq q\le r}\boldsymbol{t}_{1pp}\boldsymbol{t}_{1qp} = 2\big(\mathbf{1}^T\boldsymbol{t}_1^T\mathrm{diag}(\boldsymbol{t}_1) - \|\mathrm{diag}(\boldsymbol{t}_1)\|_2^2\big),$$
$$\sum_{1\le p\neq q\le r}\boldsymbol{t}_{1p}^T\boldsymbol{t}_{1q} = \|\boldsymbol{t}_1\mathbf{1}\|_2^2 - \|\boldsymbol{t}_1\|_F^2.$$

Plugging into the formula of $\mathcal{I}_{42}$,

$$\mathcal{I}_{42} =(\tau'_{1,1,1})^2\left((\mathbf{1}^T\mathrm{diag}(\boldsymbol{t}_1))^2 + \mathrm{Trace}(\boldsymbol{t}_1^2) - 2\|\mathrm{diag}(\boldsymbol{t}_1)\|_2^2\right) + (\tau'_{1,1,0})^2\left(\|\boldsymbol{t}_1\mathbf{1}\|_2^2 - \|\boldsymbol{t}_1\|_F^2\right)$$
$$+ 2(\tau'_{1,1,2}\tau'_{1,1,0}-(\tau'_{1,1,0})^2)\left(\mathbf{1}^T\boldsymbol{t}_1^T\mathrm{diag}(\boldsymbol{t}_1) - \|\mathrm{diag}(\boldsymbol{t}_1)\|_2^2\right). \tag{24}$$

Combining (22), (23), (24) together,

$$\mathcal{I}_4 =\tau_{2,2,0}\left((\tau'_{1,2,2}-\tau'_{1,2,0})\|\mathrm{diag}(\boldsymbol{t}_1)\|_2^2 + \tau'_{1,2,0}\|\boldsymbol{t}_1\|_F^2\right) + \tau_{2,1,0}^2\Bigg((\tau'_{1,1,0})^2(\|\boldsymbol{t}_1\mathbf{1}\|_2^2 - \|\boldsymbol{t}_1\|_F^2)$$
$$+ 2\left(\tau'_{1,1,2}\tau'_{1,1,0}-(\tau'_{1,1,0})^2\right)\left(\mathbf{1}^T\boldsymbol{t}_1^T\mathrm{diag}(\boldsymbol{t}_1) - \|\mathrm{diag}(\boldsymbol{t}_1)\|_2^2\right) + (\tau'_{1,1,1})^2\big((\mathbf{1}^T\mathrm{diag}(\boldsymbol{t}_1))^2$$
$$+ \mathrm{Trace}(\boldsymbol{t}_1^2) - 2\|\mathrm{diag}(\boldsymbol{t}_1)\|_2^2\big)\Bigg)$$
$$= \left(\tau_{2,2,0}\tau'_{1,2,0} - \tau_{2,1,0}^2(\tau'_{1,1,0})^2\right)\|\boldsymbol{t}_1\|_F^2 + \tau_{2,1,0}^2(\tau'_{1,1,0})^2\|\boldsymbol{t}_1\mathbf{1}\|_2^2 + \tau_{2,1,0}^2(\tau'_{1,1,1})^2\left(\mathbf{1}^T\mathrm{diag}(\boldsymbol{t}_1)\right)^2$$
$$+ \tau_{2,1,0}^2(\tau'_{1,1,1})^2\mathrm{Trace}(\boldsymbol{t}_1^2) + 2\left(\tau_{2,1,0}^2\tau'_{1,1,2}\tau'_{1,1,0} - \tau_{2,1,0}^2(\tau'_{1,1,0})^2\right)\mathbf{1}^T\boldsymbol{t}_1^T\mathrm{diag}(\boldsymbol{t}_1) + \big(\tau_{2,2,0}\tau'_{1,2,2}$$
$$- \tau_{2,2,0}\tau'_{1,2,0} - 2\tau_{2,1,0}^2\tau'_{1,1,2}\tau'_{1,1,0} + 2\tau_{2,1,0}^2(\tau'_{1,1,0})^2 - 2\tau_{2,1,0}^2(\tau'_{1,1,1})^2\big)\|\mathrm{diag}(\boldsymbol{t}_1)\|_2^2.$$

Recall that $\bar{\boldsymbol{t}}_i \in \mathbb{R}^{r\times r}$, $i=1,2$, denotes the matrix that replaces the diagonal entries of $\boldsymbol{t}_i$ by 0. Therefore, the above display can be further simplified as

$$\mathcal{I}_4 = \left(\tau_{2,2,0}\tau'_{1,2,0} - \tau_{2,1,0}^2(\tau'_{1,1,0})^2\right)\|\bar{\boldsymbol{t}}_1\|_F^2 + \tau_{2,1,0}^2(\tau'_{1,1,1})^2\mathrm{Trace}(\bar{\boldsymbol{t}}_1^2) + \tau_{2,1,0}^2(\tau'_{1,1,0})^2\|\bar{\boldsymbol{t}}_1\mathbf{1}\|_2^2$$
$$+ 2\tau_{2,1,0}^2\tau'_{1,1,2}\tau'_{1,1,0}\mathbf{1}^T\bar{\boldsymbol{t}}_1^T\mathrm{diag}(\boldsymbol{t}_1) + \tau_{2,1,0}^2(\tau'_{1,1,1})^2(\mathbf{1}^T\mathrm{diag}(\boldsymbol{t}_1))^2 + \left(\tau_{2,2,0}\tau'_{1,2,2} - \tau_{2,1,0}^2(\tau'_{1,1,1})^2\right)\|\mathrm{diag}(\boldsymbol{t}_1)\|_2^2.$$

This completes the proof for $\mathcal{I}_4$. $\mathcal{I}_5$ can be obtained analogously by changing the role of ϕ_1 and ϕ_2. By the definition of $\mathcal{I}_6$ in (13),

$$\mathcal{I}_6 =\sum_{p=1}^{r}\mathbb{E}\left[\phi_1'(\mathbf{x}_p)\phi_1(\mathbf{x}_p)\mathbf{x}^T\boldsymbol{t}_{1p}\right]\mathbb{E}\left[\phi_2'(\mathbf{z}_p)\phi_2(\mathbf{z}_p)\mathbf{z}^T\boldsymbol{t}_{2p}\right]$$
$$+ \sum_{1\le p\neq q\le r}\mathbb{E}\left[\phi_1'(\mathbf{x}_p)\phi_1(\mathbf{x}_q)\mathbf{x}^T\boldsymbol{t}_{1p}\right]\mathbb{E}\left[\phi_2'(\mathbf{z}_q)\phi_2(\mathbf{z}_p)\mathbf{z}^T\boldsymbol{t}_{2q}\right]$$
$$=\tau_1''\tau_2''\sum_{p=1}^{r}\boldsymbol{t}_{1pp}\boldsymbol{t}_{2pp} + \sum_{1\le p\neq q\le r}\left(\tau_{1,1,0}\tau'_{1,1,1}\boldsymbol{t}_{1pp} + \tau'_{1,1,0}\tau_{1,1,1}\boldsymbol{t}_{1pq}\right)\left(\tau_{2,1,1}\tau'_{2,1,0}\boldsymbol{t}_{2qp} + \tau'_{2,1,1}\tau_{2,1,0}\boldsymbol{t}_{2qq}\right)$$
$$=\tau_1''\tau_2''\mathrm{diag}(\boldsymbol{t}_1)^T\mathrm{diag}(\boldsymbol{t}_2) + \tau_{1,1,0}\tau'_{1,1,1}\tau_{2,1,1}\tau'_{2,1,0}\big(\mathbf{1}^T\boldsymbol{t}_2^T\mathrm{diag}(\boldsymbol{t}_1) - \mathrm{diag}(\boldsymbol{t}_1)^T\mathrm{diag}(\boldsymbol{t}_2)\big)$$
$$+ \tau'_{1,1,0}\tau_{1,1,1}\tau'_{2,1,1}\tau_{2,1,0}\big(\mathbf{1}^T\boldsymbol{t}_1^T\mathrm{diag}(\boldsymbol{t}_2) - \mathrm{diag}(\boldsymbol{t}_1)^T\mathrm{diag}(\boldsymbol{t}_2)\big)$$
$$+ \tau_{1,1,0}\tau'_{1,1,1}\tau'_{2,1,1}\tau_{2,1,0}\big(\mathbf{1}^T\mathrm{diag}(\boldsymbol{t}_1)\mathrm{diag}(\boldsymbol{t}_2)^T\mathbf{1} - \mathrm{diag}(\boldsymbol{t}_1)^T\mathrm{diag}(\boldsymbol{t}_2)\big)$$

$$
\begin{aligned}
&\quad + \tau'_{1,1,0}\tau_{1,1,1}\tau_{2,1,1}\tau'_{2,1,0}\big(\mathrm{Trace}(\boldsymbol{t}_1\boldsymbol{t}_2) - \mathrm{diag}(\boldsymbol{t}_1)^T\mathrm{diag}(\boldsymbol{t}_2)\big) \\
&= \big(\tau''_1\tau''_2 - \tau_{1,1,0}\tau_{2,1,0}\tau'_{1,1,1}\tau'_{2,1,1}\big)\,\mathrm{diag}(\boldsymbol{t}_1)^T\mathrm{diag}(\boldsymbol{t}_2) + \tau_{1,1,1}\tau_{2,1,1}\tau'_{1,1,0}\tau'_{2,1,0}\mathrm{Trace}(\bar{\boldsymbol{t}}_1\bar{\boldsymbol{t}}_2) \\
&\quad + \tau_{1,1,0}\tau_{2,1,0}\tau'_{1,1,1}\tau'_{2,1,1}\mathbf{1}^T\mathrm{diag}(\boldsymbol{t}_1)\mathrm{diag}(\boldsymbol{t}_2)^T\mathbf{1} + \tau_{1,1,0}\tau_{2,1,1}\tau'_{1,1,1}\tau'_{2,1,0}\mathbf{1}^T\bar{\boldsymbol{t}}_2^T\mathrm{diag}(\boldsymbol{t}_1) \\
&\quad + \tau_{1,1,1}\tau_{2,1,0}\tau'_{1,1,0}\tau'_{2,1,1}\mathbf{1}^T\bar{\boldsymbol{t}}_1^T\mathrm{diag}(\boldsymbol{t}_2).
\end{aligned}
$$

This completes the proof

G.3. Proof of Lemma 14

Proof of (1). By symmetry of activation functions, $\tau'_{i,1,1} = 0$. Thus, plugging into (14) and we have

$$
\begin{aligned}
&\mathrm{Var}(g(\mathbf{x},\mathbf{z})) \\
&= \tau_{1,1,1}\tau_{2,1,1}\tau'_{1,1,0}\tau'_{2,1,0}\|\bar{\boldsymbol{t}}_1 + \bar{\boldsymbol{t}}_2^T\|_F^2 + \big(\tau_{2,2,0}\tau'_{1,2,0} - \tau_{2,1,0}^2(\tau'_{1,1,0})^2 - \tau_{1,1,1}\tau_{2,1,1}\tau'_{1,1,0}\tau'_{2,1,0}\big)\,\|\bar{\boldsymbol{t}}_1\|_F^2 \\
&\quad + \big(\tau_{1,2,0}\tau'_{2,2,0} - \tau_{1,1,0}^2(\tau'_{2,1,0})^2 - \tau_{1,1,1}\tau_{2,1,1}\tau'_{1,1,0}\tau'_{2,1,0}\big)\,\|\bar{\boldsymbol{t}}_2\|_F^2 + 2\tau''_1\tau''_2\mathrm{diag}(\boldsymbol{t}_1)^T\mathrm{diag}(\boldsymbol{t}_2) \\
&\quad + \|\tau_{2,1,0}\tau'_{1,1,0}\bar{\boldsymbol{t}}_1\mathbf{1} + \tau_{2,1,0}\tau'_{1,1,2}\mathrm{diag}(\boldsymbol{t}_1)\|_2^2 + \|\tau_{1,1,0}\tau'_{2,1,0}\bar{\boldsymbol{t}}_2\mathbf{1} + \tau_{1,1,0}\tau'_{2,1,2}\mathrm{diag}(\boldsymbol{t}_2)\|_2^2 \\
&\quad + \big(\tau_{2,2,0}\tau'_{1,2,2} - \tau_{2,1,0}^2(\tau'_{1,1,2})^2\big)\,\|\mathrm{diag}(\boldsymbol{t}_1)\|_2^2 + \big(\tau_{1,2,0}\tau'_{2,2,2} - \tau_{1,1,0}^2(\tau'_{2,1,2})^2\big)\,\|\mathrm{diag}(\boldsymbol{t}_2)\|_2^2 \\
&\geq \tau_{1,1,1}\tau_{2,1,1}\tau'_{1,1,0}\tau'_{2,1,0}\|\bar{\boldsymbol{t}}_1 + \bar{\boldsymbol{t}}_2^T\|_F^2 + \rho_1\left(\|\bar{\boldsymbol{t}}_1\|_F^2 + \|\bar{\boldsymbol{t}}_2\|_F^2\right) + \tau''_1\tau''_2\|\mathrm{diag}(\boldsymbol{t}_1) + \mathrm{diag}(\boldsymbol{t}_2)\|_2^2 \\
&\quad + \big(\tau_{2,2,0}\tau'_{1,2,2} - \tau_{2,1,0}^2(\tau'_{1,1,2})^2 - \tau''_1\tau''_2\big)\,\|\mathrm{diag}(\boldsymbol{t}_1)\|_2^2 + \big(\tau_{1,2,0}\tau'_{2,2,2} - \tau_{1,1,0}^2(\tau'_{2,1,2})^2 - \tau''_1\tau''_2\big)\,\|\mathrm{diag}(\boldsymbol{t}_2)\|_2^2 \\
&\geq \tau_{1,1,1}\tau_{2,1,1}\tau'_{1,1,0}\tau'_{2,1,0}\|\bar{\boldsymbol{t}}_1 + \bar{\boldsymbol{t}}_2^T\|_F^2 + \rho_1\left(\|\bar{\boldsymbol{t}}_1\|_F^2 + \|\bar{\boldsymbol{t}}_2\|_F^2\right) + \tau''_1\tau''_2\|\mathrm{diag}(\boldsymbol{t}_1) + \mathrm{diag}(\boldsymbol{t}_2)\|_2^2 \\
&\quad + \rho_2\left(\|\mathrm{diag}(\boldsymbol{t}_1)\|_2^2 + \|\mathrm{diag}(\boldsymbol{t}_2)\|_2^2\right),
\end{aligned}
$$

where, for $j = 1, 2$, $i = 1, 2$ and $\bar{i} = 3 - i$, $\rho_j = \rho_{j1} \wedge \rho_{j2}$ with

$$
\begin{aligned}
\rho_{1i} =& \tau_{\bar{i},2,0}\tau'_{i,2,0} - \tau_{\bar{i},1,0}^2(\tau'_{i,1,0})^2 - \tau_{1,1,1}\tau_{2,1,1}\tau'_{1,1,0}\tau'_{2,1,0}, \\
\rho_{2i} =& \tau_{\bar{i},2,0}\tau'_{i,2,2} - \tau_{\bar{i},1,0}^2(\tau'_{i,1,2})^2 - \tau''_1\tau''_2.
\end{aligned}
$$

Further, by Stein's identity (Stein, 1972), $\tau_{i,1,1} = \tau'_{i,1,0}$. We can also numerically check that $\tau''_1, \tau''_2, \rho_1, \rho_2 > 0$. Therefore, the above display leads to

$$
\mathrm{Var}(g(\mathbf{x},\mathbf{z})) \geq \min(\rho_1, \rho_2)\left(\|\boldsymbol{t}_1\|_F^2 + \|\boldsymbol{t}_2\|_F^2\right).
$$

Proof of (2). Without loss of generality, we assume ϕ_1 is ReLU. Then, $\tau_{1,1,1} = \tau_{1,2,0} = \tau'_{1,1,0} = \tau'_{1,2,0} = \tau'_{1,1,2} = \tau'_{1,2,2} = \tau''_1 = 1/2$ and $\tau_{1,1,0} = \tau'_{1,1,1} = 1/\sqrt{2\pi}$. Thus, plugging into (14) and we have

$$
\begin{aligned}
&\mathrm{Var}(g(\mathbf{x},\mathbf{z})) \\
&= \frac{(\tau'_{2,1,0})^2}{4}\|\bar{\boldsymbol{t}}_1 + \bar{\boldsymbol{t}}_2^T\|_F^2 + \frac{\tau_{2,1,0}^2}{4\pi}\|\bar{\boldsymbol{t}}_1 + \bar{\boldsymbol{t}}_1^T\|_F^2 + \frac{(\tau'_{2,1,1})^2}{4\pi}\|\bar{\boldsymbol{t}}_2 + \bar{\boldsymbol{t}}_2^T\|_F^2 \\
&\quad + \frac{1}{2}\big(\tau_{2,2,0} - \frac{\pi+2}{2\pi}\tau_{2,1,0}^2 - \frac{1}{2}(\tau'_{2,1,0})^2\big)\|\bar{\boldsymbol{t}}_1\|_F^2 + \frac{1}{2}\big(\tau'_{2,2,0} - \frac{\pi+2}{2\pi}(\tau'_{2,1,0})^2 - \frac{1}{\pi}(\tau'_{2,1,1})^2\big)\|\bar{\boldsymbol{t}}_2\|_F^2 \\
&\quad + \frac{1}{4}\|\tau_{2,1,0}\bar{\boldsymbol{t}}_1\mathbf{1} + \tau_{2,1,0}\mathrm{diag}(\boldsymbol{t}_1) + \tau'_{2,1,1}\mathrm{diag}(\boldsymbol{t}_2)\|_2^2 + \frac{1}{2\pi}\|\tau'_{2,1,0}\bar{\boldsymbol{t}}_2\mathbf{1} + \tau'_{2,1,2}\mathrm{diag}(\boldsymbol{t}_2) + \tau_{2,1,1}\mathrm{diag}(\boldsymbol{t}_1)\|_2^2 \\
&\quad + \frac{1}{2}\left\{\big(\tau_{2,2,0} - \frac{\pi+2}{2\pi}\tau_{2,1,0}^2 - \frac{1}{\pi}(\tau'_{2,1,0})^2\big)\|\mathrm{diag}(\boldsymbol{t}_1)\|_2^2 + \big(\tau'_{2,2,2} - \frac{\pi+2}{2\pi}(\tau'_{2,1,1})^2 - \frac{1}{\pi}(\tau'_{2,1,2})^2\big)\|\mathrm{diag}(\boldsymbol{t}_2)\|_2^2\right\} \\
&\quad + \big(\tau''_2 - \frac{\pi+2}{2\pi}\tau_{2,1,0}\tau'_{2,1,1} - \frac{1}{\pi}\tau'_{2,1,0}\tau'_{2,1,2}\big)\mathrm{diag}(\boldsymbol{t}_1)^T\mathrm{diag}(\boldsymbol{t}_2) \\
&\geq \frac{1}{2}\left\{\big(\tau_{2,2,0} - \frac{\pi+2}{2\pi}\tau_{2,1,0}^2 - \frac{1}{2}(\tau'_{2,1,0})^2\big)\|\bar{\boldsymbol{t}}_1\|_F^2 + \big(\tau'_{2,2,0} - \frac{\pi+2}{2\pi}(\tau'_{2,1,0})^2 - \frac{1}{\pi}(\tau'_{2,1,1})^2\big)\|\bar{\boldsymbol{t}}_2\|_F^2\right. \\
&\quad \left. + \big(\tau_{2,2,0} - \frac{\pi+2}{2\pi}\tau_{2,1,0}^2 - \frac{1}{\pi}(\tau'_{2,1,0})^2\big)\|\mathrm{diag}(\boldsymbol{t}_1)\|_2^2 + \big(\tau'_{2,2,2} - \frac{\pi+2}{2\pi}(\tau'_{2,1,1})^2 - \frac{1}{\pi}(\tau'_{2,1,2})^2\big)\|\mathrm{diag}(\boldsymbol{t}_2)\|_2^2\right\} \\
&\quad + \big(\tau''_2 - \frac{\pi+2}{2\pi}\tau_{2,1,0}\tau'_{2,1,1} - \frac{1}{\pi}\tau'_{2,1,0}\tau'_{2,1,2}\big)\mathrm{diag}(\boldsymbol{t}_1)^T\mathrm{diag}(\boldsymbol{t}_2).
\end{aligned}
$$

Define

$$\begin{aligned}\rho_3 =&\big(\tau_{2,2,0} - \frac{\pi+2}{2\pi}\tau_{2,1,0}^2 - \frac{1}{2}(\tau_{2,1,0}')^2\big) \wedge \big(\tau_{2,2,0}' - \frac{\pi+2}{2\pi}(\tau_{2,1,0}')^2 - \frac{1}{\pi}(\tau_{2,1,1}')^2\big),\\ \rho_4 =&\big(\tau_{2,2,0} - \frac{\pi+2}{2\pi}\tau_{2,1,0}^2 - \frac{1}{\pi}(\tau_{2,1,0}')^2\big) \wedge \big(\tau_{2,2,2}' - \frac{\pi+2}{2\pi}(\tau_{2,1,1}')^2 - \frac{1}{\pi}(\tau_{2,1,2}')^2\big) \wedge \big(\tau_2'' \\ &- \frac{\pi+2}{2\pi}\tau_{2,1,0}\tau_{2,1,1}' - \frac{1}{\pi}\tau_{2,1,0}'\tau_{2,1,2}'\big).\end{aligned}$$

Then, we can numerically check $\rho_3, \rho_4 > 0$ when $\phi_2 \in \{\text{sigmoid}, \tanh, \text{ReLU}\}$ and hence

$$\text{Var}(g(\mathbf{x},\mathbf{z})) \geq \frac{\min(\rho_3,\rho_4)}{2}\left(\|\bar{\boldsymbol{t}}_1\|_F^2 + \|\bar{\boldsymbol{t}}_2\|_F^2 + \|\text{diag}(\boldsymbol{t}_1) + \text{diag}(\boldsymbol{t}_2)\|_2^2\right).$$

This completes the proof.

G.4. Proof of Lemma 15

Proof of $\mathcal{J}_1$. For any two samples $(y,\mathbf{x},\mathbf{z}) \in \mathcal{D}$ and $(y',\mathbf{x}',\mathbf{z}') \in \mathcal{D}'$, let us define

$$\mathbf{H}_1\left((\mathbf{x},\mathbf{z}),(\mathbf{x}',\mathbf{z}')\right) = \frac{(y-y')^2\exp\left((y-y')(\boldsymbol{\Theta}-\boldsymbol{\Theta}')\right)}{\left(1+\exp\left((y-y')(\boldsymbol{\Theta}-\boldsymbol{\Theta}')\right)\right)^2}\cdot\begin{pmatrix}\boldsymbol{d}-\boldsymbol{d}'\\ \boldsymbol{p}-\boldsymbol{p}'\end{pmatrix}\begin{pmatrix}\boldsymbol{d}-\boldsymbol{d}'\\ \boldsymbol{p}-\boldsymbol{p}'\end{pmatrix}^T,$$

where $\boldsymbol{\Theta} = \langle\phi_1(\mathbf{U}^T\mathbf{x}), \phi_2(\mathbf{V}^T\mathbf{z})\rangle$. To ease notations, we suppress the evaluation sample of $\mathbf{H}_1$. We apply Lemma 22 to bound $\mathcal{J}_1$. We first check all conditions of Lemma 22. By Assumption 2 and symmetry of $(\boldsymbol{d},\boldsymbol{p})$ and $(\boldsymbol{d}',\boldsymbol{p}')$,

$$\begin{aligned}\|\mathbf{H}_1\|_2 \leq& 4\beta^2\left\|\begin{pmatrix}\boldsymbol{d}-\boldsymbol{d}'\\ \boldsymbol{p}-\boldsymbol{p}'\end{pmatrix}\begin{pmatrix}\boldsymbol{d}-\boldsymbol{d}'\\ \boldsymbol{p}-\boldsymbol{p}'\end{pmatrix}^T\right\|_2 \leq 16\beta^2\left(\boldsymbol{d}^T\boldsymbol{d} + \boldsymbol{p}^T\boldsymbol{p}\right)\\ =&16\beta^2\left(\sum_{p=1}^r\left(\phi_1'(\boldsymbol{u}_p^T\mathbf{x})\right)^2\left(\phi_2(\boldsymbol{v}_p^T\mathbf{z})\right)^2\mathbf{x}^T\mathbf{x} + \left(\phi_1(\boldsymbol{u}_p^T\mathbf{x})\right)^2\left(\phi_2'(\boldsymbol{v}_p^T\mathbf{z})\right)^2\mathbf{z}^T\mathbf{z}\right)\\ \leq&16\beta^2\left(\sum_{p=1}^r\left(\phi_2(\boldsymbol{v}_p^T\mathbf{z})\right)^2\mathbf{x}^T\mathbf{x} + \left(\phi_1(\boldsymbol{u}_p^T\mathbf{x})\right)^2\mathbf{z}^T\mathbf{z}\right).\end{aligned}$$

The last inequality is due to the fact that $|\phi_i'| \leq 1$ for activation functions in $\{\text{sigmoid}, \tanh, \text{ReLU}\}$. Using (7), we further obtain

$$\begin{aligned}\|\mathbf{H}_1\|_2 \leq& 16\beta^2\left(\sum_{p=1}^r(\mathbf{z}^T\boldsymbol{v}_p\boldsymbol{v}_p^T\mathbf{z})^{q_2}\cdot\mathbf{x}^T\mathbf{x} + (\mathbf{x}^T\boldsymbol{u}_p\boldsymbol{u}_p^T\mathbf{x})^{q_1}\cdot\mathbf{z}^T\mathbf{z}\right)\\ =&16\beta^2\left(\left(\mathbf{z}^T\mathbf{V}\mathbf{V}^T\mathbf{z}\right)^{q_2}r^{1-q_2}\cdot\mathbf{x}^T\mathbf{x} + \left(\mathbf{x}^T\mathbf{U}\mathbf{U}^T\mathbf{x}\right)^{q_1}r^{1-q_1}\cdot\mathbf{z}^T\mathbf{z}\right).\end{aligned} \tag{25}$$

By Lemma 20, $\forall s > 0$

$$P\bigg(\max_{(\mathbf{x},\mathbf{z})\in\mathcal{D}\cup\mathcal{D}'}(\mathbf{z}^T\mathbf{V}\mathbf{V}^T\mathbf{z})^{q_2}r^{1-q_2}\cdot\mathbf{x}^T\mathbf{x} \gtrsim (\|\mathbf{V}\|_F + \sqrt{s\log n_2}\|\mathbf{V}\|_2)^{2q_2}r^{1-q_2}\cdot(\sqrt{d_1}+\sqrt{s\log n_1})^2\bigg) \lesssim \frac{1}{(n_1\wedge n_2)^s}.$$

Thus, we can bound the second term in (25) similarly and have

$$\begin{aligned}P\bigg(\max_{\mathcal{D}\cup\mathcal{D}'}\|\mathbf{H}_1\|_2 \gtrsim& \beta^2\underbrace{\big((\|\mathbf{V}\|_F + \sqrt{s\log n_2}\|\mathbf{V}\|_2)^{2q_2}r^{1-q_2}\cdot(\sqrt{d_1}+\sqrt{s\log n_1})^2 + (\|\mathbf{U}\|_F + \sqrt{s\log n_1}\|\mathbf{U}\|_2)^{2q_1}r^{1-q_1}\cdot(\sqrt{d_2}+\sqrt{s\log n_2})^2\big)}_{\nu_1(\mathcal{J}_1)}\bigg) \lesssim \frac{1}{(n_1\wedge n_2)^s}.\end{aligned} \tag{26}$$

We next verify the second condition in Lemma 22. By the symmetry of $\mathbf{H}_1$, we only need bound the following quantity

$$\begin{aligned}
&\frac{1}{n_1^2 n_2^2}\sum_{(\mathbf{x},\mathbf{z})\in\mathcal{D}}\sum_{(\mathbf{x}',\mathbf{z}')\in\mathcal{D}'}\mathbf{H}_1\big((\mathbf{x},\mathbf{z}),(\mathbf{x}',\mathbf{z}')\big)\mathbf{H}_1\big((\mathbf{x},\mathbf{z}),(\mathbf{x}',\mathbf{z}')\big)^T\\
=&\frac{1}{n_1^2 n_2^2}\sum_{(\mathbf{x},\mathbf{z})\in\mathcal{D}}\sum_{(\mathbf{x}',\mathbf{z}')\in\mathcal{D}'}\frac{(y-y')^4\exp(2(y-y')(\mathbf{\Theta}-\mathbf{\Theta}'))}{(1+\exp((y-y')(\mathbf{\Theta}-\mathbf{\Theta}')))^4}\left\|\begin{pmatrix}\boldsymbol{d}-\boldsymbol{d}'\\ \boldsymbol{p}-\boldsymbol{p}'\end{pmatrix}\right\|_2^2\begin{pmatrix}\boldsymbol{d}-\boldsymbol{d}'\\ \boldsymbol{p}-\boldsymbol{p}'\end{pmatrix}\begin{pmatrix}\boldsymbol{d}-\boldsymbol{d}'\\ \boldsymbol{p}-\boldsymbol{p}'\end{pmatrix}^T\\
\preceq&\frac{64\beta^4}{n_1^2 n_2^2}\sum_{(\mathbf{x},\mathbf{z})\in\mathcal{D}}\sum_{(\mathbf{x}',\mathbf{z}')\in\mathcal{D}'}\big((\boldsymbol{d}^T\boldsymbol{d}+\boldsymbol{p}^T\boldsymbol{p})+(\boldsymbol{d}'^T\boldsymbol{d}'+\boldsymbol{p}'^T\boldsymbol{p}')\big)\cdot\left(\begin{pmatrix}\boldsymbol{d}\\ \boldsymbol{p}\end{pmatrix}\begin{pmatrix}\boldsymbol{d}\\ \boldsymbol{p}\end{pmatrix}^T+\begin{pmatrix}\boldsymbol{d}'\\ \boldsymbol{p}'\end{pmatrix}\begin{pmatrix}\boldsymbol{d}'\\ \boldsymbol{p}'\end{pmatrix}^T\right)\\
=&\frac{128\beta^4}{n_1n_2}\sum_{(\mathbf{x},\mathbf{z})\in\mathcal{D}}\big(\boldsymbol{d}^T\boldsymbol{d}+\boldsymbol{p}^T\boldsymbol{p}\big)\cdot\begin{pmatrix}\boldsymbol{d}\\ \boldsymbol{p}\end{pmatrix}\begin{pmatrix}\boldsymbol{d}\\ \boldsymbol{p}\end{pmatrix}^T+\frac{128\beta^4}{n_1n_2}\sum_{(\mathbf{x},\mathbf{z})\in\mathcal{D}}\big(\boldsymbol{d}^T\boldsymbol{d}+\boldsymbol{p}^T\boldsymbol{p}\big)\cdot\frac{1}{n_1n_2}\sum_{(\mathbf{x}',\mathbf{z}')\in\mathcal{D}'}\begin{pmatrix}\boldsymbol{d}'\\ \boldsymbol{p}'\end{pmatrix}\begin{pmatrix}\boldsymbol{d}'\\ \boldsymbol{p}'\end{pmatrix}^T\\
=:&128\beta^4\mathcal{J}_{11}+128\beta^4\mathcal{J}_{12}.
\end{aligned}\tag{27}$$

We only bound $\mathcal{J}_{11}$ as an example. $\mathcal{J}_{12}$ can be bounded in the same sketch.

Step 1. Bound $\|\mathbb{E}[\mathcal{J}_{11}]\|_2$. For any vectors $\boldsymbol{a}=(\boldsymbol{a}_1;\ldots;\boldsymbol{a}_r)$ and $\boldsymbol{b}=(\boldsymbol{b}_1;\ldots;\boldsymbol{b}_r)$ such that $\boldsymbol{a}_p\in\mathbb{R}^{d_1}$, $\boldsymbol{b}_p\in\mathbb{R}^{d_2}$ for $p\in[r]$ and $\|\boldsymbol{a}\|_2^2+\|\boldsymbol{b}\|_2^2=1$,

$$\begin{aligned}
\left|\begin{pmatrix}\boldsymbol{a} & \boldsymbol{b}\end{pmatrix}\mathbb{E}[\mathcal{J}_{11}]\begin{pmatrix}\boldsymbol{a}\\ \boldsymbol{b}\end{pmatrix}\right| =&\mathbb{E}\Bigg[\bigg(\sum_{p=1}^r(\phi_1'(\boldsymbol{u}_p^T\mathbf{x}))^2(\phi_2(\boldsymbol{v}_p^T\mathbf{z}))^2\mathbf{x}^T\mathbf{x}+(\phi_1(\boldsymbol{u}_p^T\mathbf{x}))^2(\phi_2'(\boldsymbol{v}_p^T\mathbf{z}))^2\mathbf{z}^T\mathbf{z}\bigg)\\
&\cdot\bigg(\sum_{i=1}^r\phi_1'(\boldsymbol{u}_i^T\mathbf{x})\phi_2(\boldsymbol{v}_i^T\mathbf{z})\boldsymbol{a}_i^T\mathbf{x}+\sum_{j=1}^r\phi_1(\boldsymbol{u}_j^T\mathbf{x})\phi_2'(\boldsymbol{v}_j^T\mathbf{z})\boldsymbol{b}_j^T\mathbf{z}\bigg)^2\Bigg]\\
\leq&\mathbb{E}\Bigg[\big((\mathbf{z}^T\mathbf{V}\mathbf{V}^T\mathbf{z})^{q_2}r^{1-q_2}\cdot\mathbf{x}^T\mathbf{x}+(\mathbf{x}^T\mathbf{U}\mathbf{U}^T\mathbf{x})^{q_1}r^{1-q_1}\cdot\mathbf{z}^T\mathbf{z}\big)\\
&\cdot\bigg(\sum_{i,j=1}^r|\mathbf{z}^T\boldsymbol{v}_i\boldsymbol{v}_j^T\mathbf{z}|^{q_2}|\mathbf{x}^T\boldsymbol{a}_i\boldsymbol{a}_j^T\mathbf{x}|+2\sum_{i,j=1}^r|\mathbf{x}^T\boldsymbol{a}_i|\cdot|\boldsymbol{u}_j^T\mathbf{x}|^{q_1}\cdot|\mathbf{z}^T\boldsymbol{b}_j|\cdot|\boldsymbol{v}_i^T\mathbf{z}|^{q_2}\\
&+\sum_{i,j=1}^r|\mathbf{x}^T\boldsymbol{u}_i\boldsymbol{u}_j^T\mathbf{x}|^{q_1}|\mathbf{z}^T\boldsymbol{b}_i\boldsymbol{b}_j^T\mathbf{z}|\bigg)\Bigg].
\end{aligned}$$

By Lemma 21 and we have

$$\left|\begin{pmatrix}\boldsymbol{a} & \boldsymbol{b}\end{pmatrix}\mathbb{E}[\mathcal{J}_{11}]\begin{pmatrix}\boldsymbol{a}\\ \boldsymbol{b}\end{pmatrix}\right|\lesssim\Big(d_1r^{1-q_2}\|\mathbf{V}\|_F^{2q_2}+d_2r^{1-q_1}\|\mathbf{U}\|_F^{2q_1}\Big)\left(\sum_{i=1}^r\|\boldsymbol{a}_i\|_2\|\boldsymbol{v}_i\|_2^{q_2}+\|\boldsymbol{b}_i\|_2\|\boldsymbol{u}_i\|_2^{q_1}\right)^2.$$

Maximizing over set $\{(\boldsymbol{a},\boldsymbol{b}):\|\boldsymbol{a}\|_2^2+\|\boldsymbol{b}\|_2^2=1\}$ on both sides and we get

$$\|\mathbb{E}[\mathcal{J}_{11}]\|_2\lesssim\Big(d_1\|\mathbf{V}\|_F^{2q_2}r^{1-q_2}+d_2\|\mathbf{U}\|_F^{2q_1}r^{1-q_1}\Big)\Big(\|\mathbf{V}\|_F^{2q_2}r^{1-q_2}+\|\mathbf{U}\|_F^{2q_1}r^{1-q_1}\Big).\tag{28}$$

Step 2. Bound $\|\mathcal{J}_{11}-\mathbb{E}[\mathcal{J}_{11}]\|_2$. We apply Lemma 24. Let us first define the random matrix

$$\mathbf{J}_{11}(\mathbf{x},\mathbf{z}):=\big(\boldsymbol{d}^T\boldsymbol{d}+\boldsymbol{p}^T\boldsymbol{p}\big)\cdot\begin{pmatrix}\boldsymbol{d}\\ \boldsymbol{p}\end{pmatrix}\begin{pmatrix}\boldsymbol{d}\\ \boldsymbol{p}\end{pmatrix}^T.$$

For the condition (a) in Lemma 24, we note that

$$\begin{aligned}
\|\mathbf{J}_{11}(\mathbf{x},\mathbf{z})\|_2=&(\boldsymbol{d}^T\boldsymbol{d}+\boldsymbol{p}^T\boldsymbol{p})^2\leq\Big(\sum_{p=1}^r(\mathbf{z}^T\boldsymbol{v}_p\boldsymbol{v}_p^T\mathbf{z})^{q_2}\mathbf{x}^T\mathbf{x}+\sum_{p=1}^r(\mathbf{x}^T\boldsymbol{u}_p\boldsymbol{u}_p^T\mathbf{x})^{q_1}\mathbf{z}^T\mathbf{z}\Big)^2\\
=&\big(r^{1-q_2}(\mathbf{z}^T\mathbf{V}\mathbf{V}^T\mathbf{z})^{q_2}\mathbf{x}^T\mathbf{x}+r^{1-q_1}(\mathbf{x}^T\mathbf{U}\mathbf{U}^T\mathbf{x})^{q_1}\mathbf{z}^T\mathbf{z}\big)^2.
\end{aligned}$$

By Lemma 20, for any constants $K_1^{(1,1)} \wedge K_2^{(1,1)} \wedge K_3^{(1,1)} \geq 1$ (in what follows we may keep using such notation, where the first superscript indexes the function $\{\mathcal{L}_i\}_{i=1,2}$ we are dealing with; the second superscript indexes the times we have used for this notation),

$$P\left(\|\mathbf{J}_{11}(\mathbf{x},\mathbf{z})\|_2 \gtrsim (K_3^{(1,1)})^2\left(d_1(K_2^{(1,1)})^{q_2}\|\mathbf{V}\|_F^{2q_2}r^{1-q_2} + d_2(K_1^{(1,1)})^{q_1}\|\mathbf{U}\|_F^{2q_1}r^{1-q_1}\right)^2\right)$$
$$\leq 2\exp\left(-(d_1\wedge d_2)K_3^{(1,1)}\right) + q_2\exp\left(-\frac{\|\mathbf{V}\|_F^2 K_2^{(1,1)}}{\|\mathbf{V}\|_2^2}\right) + q_1\exp\left(-\frac{\|\mathbf{U}\|_F^2 K_1^{(1,1)}}{\|\mathbf{U}\|_2^2}\right). \quad (29)$$

For the condition (b) in Lemma 24, we apply the inequalities in Lemma 21 and have

$$\begin{aligned}
\|\mathbb{E}[\mathbf{J}_{11}(\mathbf{x},\mathbf{z})\mathbf{J}_{11}(\mathbf{x},\mathbf{z})^T]\|_2 &= \max_{\|\boldsymbol{a}\|_F^2+\|\boldsymbol{b}\|_F^2=1}\mathbb{E}\big[\big(\boldsymbol{d}^T\boldsymbol{d}+\boldsymbol{p}^T\boldsymbol{p}\big)^3\big(\boldsymbol{a}^T\boldsymbol{d}+\boldsymbol{b}^T\boldsymbol{p}\big)^2\big]\\
&\leq \max_{\|\boldsymbol{a}\|_F^2+\|\boldsymbol{b}\|_F^2=1}\mathbb{E}\big[\big(r^{1-q_2}(\mathbf{z}^T\mathbf{V}\mathbf{V}^T\mathbf{z})^{q_2}\mathbf{x}^T\mathbf{x} + r^{1-q_1}(\mathbf{x}^T\mathbf{U}\mathbf{U}^T\mathbf{x})^{q_1}\mathbf{z}^T\mathbf{z}\big)^3\\
&\qquad\cdot\big(\sum_{i=1}^r|\boldsymbol{v}_i^T\mathbf{z}|^{q_2}|\boldsymbol{a}_i^T\mathbf{x}| + \sum_{j=1}^r|\boldsymbol{u}_j^T\mathbf{x}|^{q_1}|\boldsymbol{b}_j^T\mathbf{z}|\big)^2\big]\\
&\lesssim \left(d_1\|\mathbf{V}\|_F^{2q_2}r^{1-q_2} + d_2\|\mathbf{U}\|_F^{2q_1}r^{1-q_1}\right)^3\max_{\|\boldsymbol{a}\|_F^2+\|\boldsymbol{b}\|_F^2=1}\big(\sum_{i=1}^r\|\boldsymbol{a}_i\|_2\|\boldsymbol{v}_i\|_2^{q_2} + \|\boldsymbol{b}_i\|_2\|\boldsymbol{u}_i\|_2^{q_1}\big)^2\\
&\lesssim \left(d_1\|\mathbf{V}\|_F^{2q_2}r^{1-q_2} + d_2\|\mathbf{U}\|_F^{2q_1}r^{1-q_1}\right)^3\left(\|\mathbf{V}\|_F^{2q_2}r^{1-q_2} + \|\mathbf{U}\|_F^{2q_1}r^{1-q_1}\right). \quad (30)
\end{aligned}$$

For the condition (c) in Lemma 24, we consider the following quantity for any unit vector $(\boldsymbol{a};\boldsymbol{b})$:

$$\begin{aligned}
&\mathbb{E}\big[\big(\boldsymbol{d}^T\boldsymbol{d}+\boldsymbol{p}^T\boldsymbol{p}\big)^2\big(\boldsymbol{a}^T\boldsymbol{d}+\boldsymbol{b}^T\boldsymbol{p}\big)^4\big]\\
&\leq\mathbb{E}\big[\big(r^{1-q_2}(\mathbf{z}^T\mathbf{V}\mathbf{V}^T\mathbf{z})^{q_2}\mathbf{x}^T\mathbf{x} + r^{1-q_1}(\mathbf{x}^T\mathbf{U}\mathbf{U}^T\mathbf{x})^{q_1}\mathbf{z}^T\mathbf{z}\big)^2\big(\sum_{i=1}^r|\boldsymbol{v}_i^T\mathbf{z}|^{q_2}|\boldsymbol{a}_i^T\mathbf{x}| + \sum_{j=1}^r|\boldsymbol{u}_j^T\mathbf{x}|^{q_1}|\boldsymbol{b}_j^T\mathbf{z}|\big)^4\big]\\
&\lesssim\left(d_1\|\mathbf{V}\|_F^{2q_2}r^{1-q_2} + d_2\|\mathbf{U}\|_F^{2q_1}r^{1-q_1}\right)^2\left(\|\mathbf{V}\|_F^{2q_2}r^{1-q_2} + \|\mathbf{U}\|_F^{2q_1}r^{1-q_1}\right)^2. \quad (31)
\end{aligned}$$

Combining (28), (29), (30), (31) together and defining

$$\Upsilon_1 \coloneqq d_1\|\mathbf{V}\|_F^{2q_2}r^{1-q_2} + d_2\|\mathbf{U}\|_F^{2q_1}r^{1-q_1}, \quad \Upsilon_2 \coloneqq \|\mathbf{V}\|_F^{2q_2}r^{1-q_2} + \|\mathbf{U}\|_F^{2q_1}r^{1-q_1}, \quad (32)$$

we know conditions in Lemma 24 hold for $\mathcal{J}_{11}$ with parameters (up to constants)

$$\begin{aligned}
\mu_1(\mathcal{J}_{11}) &\coloneqq (K_3^{(1,1)})^2\left(d_1(K_2^{(1,1)})^{q_2}\|\mathbf{V}\|_F^{2q_2}r^{1-q_2} + d_2(K_1^{(1,1)})^{q_1}\|\mathbf{U}\|_F^{2q_1}r^{1-q_1}\right)^2,\\
\nu_1(\mathcal{J}_{11}) &\coloneqq \exp\left(-(d_1\wedge d_2)K_3^{(1,1)}\right) + q_2\exp\left(-\frac{\|\mathbf{V}\|_F^2K_2^{(1,1)}}{\|\mathbf{V}\|_2^2}\right) + q_1\exp\left(-\frac{\|\mathbf{U}\|_F^2K_1^{(1,1)}}{\|\mathbf{U}\|_2^2}\right),\\
\nu_2(\mathcal{J}_{11}) &\coloneqq \Upsilon_1^3\Upsilon_2, \qquad \nu_3(\mathcal{J}_{11}) \coloneqq \Upsilon_1\Upsilon_2, \qquad \|\mathbb{E}[\mathcal{J}_{11}]\| \lesssim \Upsilon_1\Upsilon_2.
\end{aligned}$$

Thus, $\forall t > 0$

$$\begin{aligned}
&P\big(\big\|\mathcal{J}_{11}-\mathbb{E}[\mathcal{J}_{11}]\big\|_2 > t + \Upsilon_1\Upsilon_2\sqrt{\nu_1(\mathcal{J}_{11})}\big)\\
&\leq n_1n_2\nu_1(\mathcal{J}_{11}) + 2r(d_1+d_2)\exp\left(-\frac{(n_1\wedge n_2)t^2}{(2\Upsilon_1^3\Upsilon_2 + 4\Upsilon_1^2\Upsilon_2^2 + 4\Upsilon_1^2\Upsilon_3^2\nu_1(\mathcal{J}_{11})) + 4\mu_1(\mathcal{J}_{11})t}\right)\\
&\leq n_1n_2\nu_1(\mathcal{J}_{11}) + 2r(d_1+d_2)\exp\left(-\frac{(n_1\wedge n_2)t^2}{10\Upsilon_1^3\Upsilon_2 + 4\mu_1(\mathcal{J}_{11})t}\right).
\end{aligned}$$

In the above inequality, for any constant $s \geq 1$ we let

$$K_1^{(1,1)} = K_2^{(1,1)} = \log(n_1n_2) + s\log(d_1+d_2), \qquad K_3^{(1,1)} = 1.$$

By simple calculation, we can let

$$\epsilon_1 \asymp \sqrt{\frac{s(d_1+d_2)\log\left(r(d_1+d_2)\right)}{n_1 \wedge n_2}} \vee \frac{s(d_1+d_2)\left\{\log\left(r(d_1+d_2)\right)\right\}^{1+2(q_1\vee q_2)}}{n_1\wedge n_2}$$

and further have

$$P\left(\left\|\mathcal{J}_{11}-\mathbb{E}[\mathcal{J}_{11}]\right\|_2 > \epsilon_1\Upsilon_1\Upsilon_2\right) \lesssim \frac{1}{(d_1+d_2)^s}.$$

Under the conditions of Lemma 15, we combine the above inequality with (28) and have $P(\|\mathcal{J}_{11}\|_2 \gtrsim \Upsilon_1\Upsilon_2) \lesssim 1/(d_1+d_2)^s, \forall s \geq 1$. Dealing with $\mathcal{J}_{12}$ in (27) similarly, one can show (28) and the above result hold for $\mathcal{J}_{12}$ as well. So $P(\|\mathcal{J}_{12}\|_2 \gtrsim \Upsilon_1\Upsilon_2) \lesssim 1/(d_1+d_2)^s$. Plugging back into (27), we can define $\nu_2(\mathcal{J}_1) = \beta^4\Upsilon_1\Upsilon_2$ and then conditions of Lemma 22 hold for $\mathcal{J}_1$ with parameters $\nu_1(\mathcal{J}_1)$ (defined in (26)) and $\nu_2(\mathcal{J}_1)$. Therefore, we have $\forall t > 0$

$$P\left(\mathcal{J}_1 > t\right) \lesssim 2r(d_1+d_2)\exp\left(-\frac{mt^2}{4\nu_2(\mathcal{J}_1)+4\nu_1(\mathcal{J}_1)t}\right).$$

For any $s \geq 1$, we let

$$\epsilon_2 \asymp \sqrt{\frac{s(d_1+d_2)\log\left(r(d_1+d_2)\right)}{m}} \vee \frac{s(d_1+d_2)\left\{\log\left(r(d_1+d_2)\right)\right\}^{1+q}}{m}$$

and have

$$P\left(\mathcal{J}_1 > \beta^2\epsilon_2\Upsilon_2\right) \lesssim \frac{1}{(d_1+d_2)^s}.$$

The result follows by the definition of Υ_2 in (32) and noting that the first term in ϵ_2 is the dominant term.

Proof of $\mathcal{J}_2$. We apply Lemma 23 to bound $\mathcal{J}_2$. We check all conditions of Lemma 23. Some of steps are similar as above. By definition of $\mathbf{H}_1$,

$$\nabla^2\bar{\mathcal{L}}_1(\mathbf{U},\mathbf{V}) = \frac{1}{n_1^2 n_2^2}\sum_{(\mathbf{x},\mathbf{z})\in\mathcal{D}}\sum_{(\mathbf{x}',\mathbf{z}')\in\mathcal{D}'}\mathbf{H}_1\big((\mathbf{x},\mathbf{z}),(\mathbf{x}',\mathbf{z}')\big).$$

We first bound $\|\mathbb{E}[\mathbf{H}_1]\|_2$. We have

$$\|\mathbb{E}[\mathbf{H}_1]\|_2 \leq 2\beta^2 \max_{\|\boldsymbol{a}\|_F^2+\|\boldsymbol{b}\|_F^2=1}\mathbb{E}\left[(\boldsymbol{a}^T\boldsymbol{d}+\boldsymbol{b}^T\boldsymbol{p})^2\right] \lesssim \beta^2\Upsilon_2, \tag{33}$$

where the last inequality is derived similarly to (30). For the condition (a) in Lemma 23, we apply (25) and Lemma 20 (similar to (29)),

$$P\left(\|\mathbf{H}_1\|_2 \gtrsim \beta^2 K_3^{(1,2)}\left(d_1(K_2^{(1,2)})^{q_2}\|\mathbf{V}\|_F^{2q_2}r^{1-q_2} + d_2(K_1^{(1,2)})^{q_1}\|\mathbf{U}\|_F^{2q_1}r^{1-q_1}\right)\right)$$
$$\leq 2\exp\left(-(d_1\wedge d_2)K_3^{(1,2)}\right) + q_2\exp\left(-\frac{\|\mathbf{V}\|_F^2K_2^{(1,2)}}{\|\mathbf{V}\|_2^2}\right) + q_1\exp\left(-\frac{\|\mathbf{U}\|_F^2K_1^{(1,2)}}{\|\mathbf{U}\|_2^2}\right).$$

For the condition (b) in Lemma 23,

$$\|\mathbb{E}[\mathbf{H}_1\mathbf{H}_1^T]\|_2 \lesssim \beta^4\|\mathbb{E}[\mathcal{J}_{11}]\|_2 \overset{(28)}{\lesssim} \beta^4\Upsilon_1\Upsilon_2.$$

For the condition (c) in Lemma 23,

$$\max_{\|\boldsymbol{a}\|_F^2+\|\boldsymbol{b}\|_F^2=1}\mathbb{E}\Big[\Big(\begin{pmatrix}\boldsymbol{a}^T & \boldsymbol{b}^T\end{pmatrix}\mathbf{H}_1\begin{pmatrix}\boldsymbol{a}\\ \boldsymbol{b}\end{pmatrix}\Big)^2\Big] \lesssim \beta^4 \max_{\|\boldsymbol{a}\|_F^2+\|\boldsymbol{b}\|_F^2=1}\mathbb{E}\Big[\big(\boldsymbol{a}^T(\boldsymbol{d}-\boldsymbol{d}')+\boldsymbol{b}^T(\boldsymbol{p}-\boldsymbol{p}')\big)^4\Big] \overset{(31)}{\lesssim} \beta^4\Upsilon_2^2.$$

Thus, conditions of Lemma 23 hold with parameters (up to constants)

$$\begin{aligned}
\mu_1(\mathcal{J}_2) \coloneqq& \beta^2 K_3^{(1,2)} \left(d_1 (K_2^{(1,2)})^{q_2} \|\mathbf{V}\|_F^{2q_2} r^{1-q_2} + d_2 (K_1^{(1,2)})^{q_1} \|\mathbf{U}\|_F^{2q_1} r^{1-q_1} \right), \\
\nu_1(\mathcal{J}_2) \coloneqq& \exp\left(-(d_1 \wedge d_2) K_3^{(1,2)}\right) + q_2 \exp\left(-\frac{\|\mathbf{V}\|_F^2 K_2^{(1,2)}}{\|\mathbf{V}\|_2^2}\right) + q_1 \exp\left(-\frac{\|\mathbf{U}\|_F^2 K_1^{(1,2)}}{\|\mathbf{U}\|_2^2}\right), \\
\nu_2(\mathcal{J}_2) \coloneqq& \beta^4 \Upsilon_1 \Upsilon_2, \qquad \nu_3(\mathcal{J}_2) \coloneqq \beta^2 \Upsilon_2, \qquad \|\mathbb{E}[\mathbf{H}_1]\| \lesssim \beta^2 \Upsilon_2.
\end{aligned}$$

Similar to the proof of $\mathcal{J}_1$, for any $s \geq 1$, we let $K_1^{(1,2)} = K_2^{(1,2)} = 2\log n_1 n_2 + s \log(d_1 + d_2)$, $K_3^{(1,2)} = 1$, and

$$\epsilon_3 \asymp \sqrt{\frac{s(d_1+d_2)\log\left(r(d_1+d_2)\right)}{n_1 \wedge n_2}} \vee \frac{s(d_1+d_2)\left\{\log\left(r(d_1+d_2)\right)\right\}^{1+q}}{n_1 \wedge n_2},$$

and then have

$$P\left(\mathcal{J}_2 \gtrsim \beta^2 \epsilon_3 \Upsilon_2\right) \lesssim \frac{1}{(d_1+d_2)^s}.$$

Noting that the first term in ϵ_3 is the dominant term, we complete the proof.

G.5. Proof of Lemma 16

Proof of $\mathcal{T}_1$. For any two samples $(y, \mathbf{x}, \mathbf{z}) \in \mathcal{D}$ and $(y', \mathbf{x}', \mathbf{z}') \in \mathcal{D}'$, we define

$$\mathbf{H}_2\left((\mathbf{x},\mathbf{z}),(\mathbf{x}',\mathbf{z}')\right) = \frac{y-y'}{1+\exp\left((y-y')(\mathbf{\Theta}-\mathbf{\Theta}')\right)} \cdot \begin{pmatrix} \boldsymbol{Q}-\boldsymbol{Q}' & \boldsymbol{S}-\boldsymbol{S}' \\ \boldsymbol{S}^T-\boldsymbol{S}'^T & \boldsymbol{R}-\boldsymbol{R}' \end{pmatrix}, \tag{34}$$

where $\mathbf{\Theta} = \langle \phi_1(\mathbf{U}^T\mathbf{x}), \phi_2(\mathbf{V}^T\mathbf{z})\rangle$. We follow the same proof sketch as Lemma 15. We apply Lemma 22 to bound $\mathcal{T}_1$. We first check all conditions of Lemma 22. By the boundedness assumption of y, y' in Assumption 2,

$$\begin{aligned}
\|\mathbf{H}_2\|_2 \leq& 4\beta \left\| \begin{pmatrix} \boldsymbol{Q} & \boldsymbol{S} \\ \boldsymbol{S}^T & \boldsymbol{R} \end{pmatrix} \right\|_2 \\
\leq& 4\beta \max_{\|\boldsymbol{a}\|_F^2 + \|\boldsymbol{b}\|_F^2 = 1} \left| \sum_{p=1}^r \phi_1''(\boldsymbol{u}_p^T\mathbf{x})\phi_2(\boldsymbol{v}_p^T\mathbf{z})(\boldsymbol{a}_p^T\mathbf{x})^2 + 2\sum_{p=1}^r \phi_1'(\boldsymbol{u}_p^T\mathbf{x})\phi_2'(\boldsymbol{v}_p^T\mathbf{z})\mathbf{x}^T\boldsymbol{a}_p\boldsymbol{b}_p^T\mathbf{z} \right. \\
&\left. + \sum_{p=1}^r \phi_1(\boldsymbol{u}_p^T\mathbf{x})\phi_2''(\boldsymbol{v}_p^T\mathbf{z})(\boldsymbol{b}_p^T\mathbf{z})^2 \right| \\
\lesssim& \beta \max_{\|\boldsymbol{a}\|_F^2 + \|\boldsymbol{b}\|_F^2 = 1} \left| \sum_{p=1}^r \mathbf{1}_{q_1=0} \cdot |\boldsymbol{v}_p^T\mathbf{z}|^{q_2}(\boldsymbol{a}_p^T\mathbf{x})^2 + 2\sum_{p=1}^r |\mathbf{x}^T\boldsymbol{a}_p| \cdot |\boldsymbol{b}_p^T\mathbf{z}| + \sum_{p=1}^r \mathbf{1}_{q_2=0} \cdot |\boldsymbol{u}_p^T\mathbf{x}|^{q_1}(\boldsymbol{b}_p^T\mathbf{z})^2 \right| \\
\lesssim& \beta \left((1-q_1)\mathbf{x}^T\mathbf{x} \max_{p\in[r]} |\mathbf{z}^T\boldsymbol{v}_p|^{q_2} + (1-q_2)\mathbf{z}^T\mathbf{z} \max_{p\in[r]} |\mathbf{x}^T\boldsymbol{u}_p|^{q_1} + \|\mathbf{x}\|_2\|\mathbf{z}\|_2 \right).
\end{aligned} \tag{35}$$

Here, the third inequality is due to the fact that $|\phi_i''| \leq 2$ if $\phi_i \in \{\text{sigmoid}, \tanh\}$ and $\phi_i'' = 0$ if ϕ_i is ReLU. Taking union bound over $\mathcal{D} \cup \mathcal{D}'$, noting that $\log(r(n_1+n_2)(d_1+d_2)) \asymp \log\left(r(d_1+d_2)\right)$, and applying Lemma 20, for any $s \geq 1$, we define

$$\begin{aligned}
\Upsilon_3 =& (1-q_1)d_1\left(\log(r(d_1+d_2))\right)^{q_2/2}\|\mathbf{V}\|_2^{q_2} + \sqrt{d_1 d_2} + (1-q_2)d_2\left(\log(r(d_1+d_2))\right)^{q_1/2}\|\mathbf{U}\|_2^{q_1} \\
\asymp& d_1^{\frac{2-q_1}{2}} d_2^{\frac{q_1}{2}} \left(\log\left(r(d_1+d_1)\right)\right)^{\frac{q_2(1-q_1)}{2}} \|\mathbf{V}\|_2^{q_2(1-q_1)} + d_2^{\frac{2-q_2}{2}} d_1^{\frac{q_2}{2}} \left(\log\left(r(d_1+d_1)\right)\right)^{\frac{q_1(1-q_2)}{2}} \|\mathbf{U}\|_2^{q_1(1-q_2)}
\end{aligned} \tag{36}$$

and have

$$P\left(\max_{\mathcal{D}\cup\mathcal{D}'} \|\mathbf{H}_2\|_2 \gtrsim \beta\Upsilon_3\right) \lesssim \frac{1}{(d_1+d_2)^s}. \tag{37}$$

Next, we bound the following quantity

$$
\begin{aligned}
&\frac{1}{n_1^2 n_2^2}\sum_{(\mathbf{x},\mathbf{z})\in\mathcal{D}}\sum_{(\mathbf{x}',\mathbf{z}')\in\mathcal{D}'}\mathbf{H}_2\big((\mathbf{x},\mathbf{z}),(\mathbf{x}',\mathbf{z}')\big)\mathbf{H}_2\big((\mathbf{x},\mathbf{z}),(\mathbf{x}',\mathbf{z}')\big)^T\\
&=\frac{1}{n_1^2 n_2^2}\sum_{(\mathbf{x},\mathbf{z})\in\mathcal{D}}\sum_{(\mathbf{x}',\mathbf{z}')\in\mathcal{D}'}\frac{(y-y')^2}{\big(1+\exp\big((y-y')(\mathbf{\Theta}-\mathbf{\Theta}')\big)\big)^2}\cdot\begin{pmatrix}\boldsymbol{Q}-\boldsymbol{Q}' & \boldsymbol{S}-\boldsymbol{S}'\\ \boldsymbol{S}^T-\boldsymbol{S}'^T & \boldsymbol{R}-\boldsymbol{R}'\end{pmatrix}^2\\
&\preceq\frac{16\beta^2}{n_1 n_2}\sum_{(\mathbf{x},\mathbf{z})\in\mathcal{D}}\begin{pmatrix}\boldsymbol{Q} & \boldsymbol{S}\\ \boldsymbol{S}^T & \boldsymbol{R}\end{pmatrix}^2=\frac{16\beta^2}{n_1 n_2}\sum_{(\mathbf{x},\mathbf{z})\in\mathcal{D}}\begin{pmatrix}\boldsymbol{Q}^2+\boldsymbol{S}\boldsymbol{S}^T & \boldsymbol{Q}\boldsymbol{S}+\boldsymbol{S}\boldsymbol{R}\\ \boldsymbol{S}^T\boldsymbol{Q}+\boldsymbol{R}\boldsymbol{S}^T & \boldsymbol{R}^2+\boldsymbol{S}^T\boldsymbol{S}\end{pmatrix}\coloneqq 16\beta^2\mathcal{T}_{11}. \qquad (38)
\end{aligned}
$$

Similarly to Lemma 15, we have two steps.

Step 1. Bound $\|\mathbb{E}[\mathcal{T}_{11}]\|_2$. For any vectors $\boldsymbol{a}=(\boldsymbol{a}_1;\ldots;\boldsymbol{a}_r)$ and $\boldsymbol{b}=(\boldsymbol{b}_1;\ldots;\boldsymbol{b}_r)$ such that $\boldsymbol{a}_p\in\mathbb{R}^{d_1}$, $\boldsymbol{b}_p\in\mathbb{R}^{d_2}$ for $p\in[r]$ and $\|\boldsymbol{a}\|_2^2+\|\boldsymbol{b}\|_2^2=1$,

$$
\begin{aligned}
&\left|\begin{pmatrix}\boldsymbol{a} & \boldsymbol{b}\end{pmatrix}\mathbb{E}[\mathcal{T}_{11}]\begin{pmatrix}\boldsymbol{a}\\ \boldsymbol{b}\end{pmatrix}\right|\\
=&\mathbb{E}\Big[\sum_{p=1}^r\Big(\big(\phi_1''(\boldsymbol{u}_p^T\mathbf{x})\phi_2(\boldsymbol{v}_p^T\mathbf{z})\big)^2\mathbf{x}^T\mathbf{x}+\big(\phi_1'(\boldsymbol{u}_p^T\mathbf{x})\phi_2'(\boldsymbol{v}_p^T\mathbf{z})\big)^2\mathbf{z}^T\mathbf{z}\Big)(\boldsymbol{a}_p^T\mathbf{x})^2\\
&+2\sum_{p=1}^r\Big(\phi_1''(\boldsymbol{u}_p^T\mathbf{x})\phi_1'(\boldsymbol{u}_p^T\mathbf{x})\phi_2(\boldsymbol{v}_p^T\mathbf{z})\phi_2'(\boldsymbol{v}_p^T\mathbf{z})\mathbf{x}^T\mathbf{x}+\phi_1(\boldsymbol{u}_p^T\mathbf{x})\phi_1'(\boldsymbol{u}_p^T\mathbf{x})\phi_2'(\boldsymbol{v}_p^T\mathbf{z})\phi_2''(\boldsymbol{v}_p^T\mathbf{z})\mathbf{z}^T\mathbf{z}\Big)\mathbf{x}^T\boldsymbol{a}_p\boldsymbol{b}_p^T\mathbf{z}\\
&+\sum_{p=1}^r\Big(\big(\phi_1(\boldsymbol{u}_p^T\mathbf{x})\phi_2''(\boldsymbol{v}_p^T\mathbf{z})\big)^2\mathbf{z}^T\mathbf{z}+\big(\phi_1'(\boldsymbol{u}_p^T\mathbf{x})\phi_2'(\boldsymbol{v}_p^T\mathbf{z})\big)^2\mathbf{x}^T\mathbf{x}\Big)(\boldsymbol{b}_p^T\mathbf{z})^2\Big]\\
\lesssim&\mathbb{E}\Big[\sum_{p=1}^r\big((1-q_1)(\mathbf{z}^T\boldsymbol{v}_p\boldsymbol{v}_p^T\mathbf{z})^{q_2}\mathbf{x}^T\mathbf{x}+\mathbf{z}^T\mathbf{z}\big)\cdot\mathbf{x}^T\boldsymbol{a}_p\boldsymbol{a}_p^T\mathbf{x}+\sum_{p=1}^r\big((1-q_2)(\mathbf{x}^T\boldsymbol{u}_p\boldsymbol{u}_p^T\mathbf{x})^{q_1}\mathbf{z}^T\mathbf{z}\\
&+\mathbf{x}^T\mathbf{x}\big)\cdot\mathbf{z}^T\boldsymbol{b}_p\boldsymbol{b}_p^T\mathbf{z}+\sum_{p=1}^r\big((1-q_1)|\boldsymbol{v}_p^T\mathbf{z}|^{q_2}\mathbf{x}^T\mathbf{x}+(1-q_2)|\boldsymbol{u}_p^T\mathbf{x}|^{q_1}\mathbf{z}^T\mathbf{z}\big)|\mathbf{x}^T\boldsymbol{a}_p\boldsymbol{b}_p^T\mathbf{z}|\Big]\\
\lesssim&\mathbb{E}\big[(1-q_1)\mathbf{x}^T\mathbf{x}\sum_{p=1}^r(\mathbf{z}^T\boldsymbol{v}_p\boldsymbol{v}_p^T\mathbf{z})^{q_2}\mathbf{x}^T\boldsymbol{a}_p\boldsymbol{a}_p^T\mathbf{x}+(1-q_2)\mathbf{z}^T\mathbf{z}\sum_{p=1}^r(\mathbf{x}^T\boldsymbol{u}_p\boldsymbol{u}_p^T\mathbf{x})^{q_1}\mathbf{z}^T\boldsymbol{b}_p\boldsymbol{b}_p^T\mathbf{z}\\
&+(1-q_1)\mathbf{x}^T\mathbf{x}\sum_{p=1}^r|\mathbf{x}^T\boldsymbol{a}_p|\cdot|\mathbf{z}^T\boldsymbol{b}_p|\cdot|\mathbf{z}^T\boldsymbol{v}_p|^{q_2}+(1-q_2)\mathbf{z}^T\mathbf{z}\sum_{p=1}^r|\mathbf{z}^T\boldsymbol{b}_p|\cdot|\mathbf{x}^T\boldsymbol{a}_p|\cdot|\mathbf{x}^T\boldsymbol{u}_p|^{q_1}\\
&+\mathbf{z}^T\mathbf{z}\cdot\mathbf{x}^T(\sum_{p=1}^r\boldsymbol{a}_p\boldsymbol{a}_p^T)\mathbf{x}+\mathbf{x}^T\mathbf{x}\cdot\mathbf{z}^T(\sum_{p=1}^r\boldsymbol{b}_p\boldsymbol{b}_p^T)\mathbf{z}\big].
\end{aligned}
$$

Using Lemma 21 and maximizing over set $\{(\boldsymbol{a},\boldsymbol{b}):\|\boldsymbol{a}\|_2^2+\|\boldsymbol{b}\|_2^2=1\}$, we get

$$
\|\mathbb{E}[\mathcal{T}_{11}]\|_2\lesssim(1-q_1)d_1\|\mathbf{V}\|_2^{2q_2}+(1-q_2)d_2\|\mathbf{U}\|_2^{2q_1}+d_1+d_2\lesssim d_1\|\mathbf{V}\|_2^{2q_2(1-q_1)}+d_2\|\mathbf{U}\|_2^{2q_1(1-q_2)}. \qquad (39)
$$

Step 2. Bound $\|\mathcal{T}_{11}-\mathbb{E}[\mathcal{T}_{11}]\|_2$. We still apply Lemma 24. Define the following random matrix

$$
\mathbf{T}_{11}(\mathbf{x},\mathbf{z})\coloneqq\begin{pmatrix}\boldsymbol{Q}^2+\boldsymbol{S}\boldsymbol{S}^T & \boldsymbol{Q}\boldsymbol{S}+\boldsymbol{S}\boldsymbol{R}\\ \boldsymbol{S}^T\boldsymbol{Q}+\boldsymbol{R}\boldsymbol{S}^T & \boldsymbol{R}^2+\boldsymbol{S}^T\boldsymbol{S}\end{pmatrix}.
$$

For the condition (a) in Lemma 24, we note that

$$
\|\mathbf{T}_{11}(\mathbf{x},\mathbf{z})\|_2=\max_{p\in[r]}\left\|\begin{pmatrix}\phi_1''(\boldsymbol{u}_p^T\mathbf{x})\phi_2(\boldsymbol{v}_p^T\mathbf{z})\cdot\mathbf{x}\mathbf{x}^T & \phi_1'(\boldsymbol{u}_p^T\mathbf{x})\phi_2'(\boldsymbol{v}_p^T\mathbf{z})\cdot\mathbf{x}\mathbf{z}^T\\ \phi_1'(\boldsymbol{u}_p^T\mathbf{x})\phi_2'(\boldsymbol{v}_p^T\mathbf{z})\cdot\mathbf{z}\mathbf{x}^T & \phi_1(\boldsymbol{u}_p^T\mathbf{x})\phi_2''(\boldsymbol{v}_p^T\mathbf{z})\cdot\mathbf{z}\mathbf{z}^T\end{pmatrix}\right\|_2^2
$$

$$
\begin{aligned}
&= \max_{p\in[r]} \bigg(\max_{\|\boldsymbol{a}_p\|_2^2+\|\boldsymbol{b}_p\|_2^2=1} \phi_1''(\boldsymbol{u}_p^T\mathbf{x})\phi_2(\boldsymbol{v}_p^T\mathbf{z})(\boldsymbol{a}_p^T\mathbf{x})^2 + 2\phi_1'(\boldsymbol{u}_p^T\mathbf{x})\phi_2'(\boldsymbol{v}_p^T\mathbf{z})\cdot(\boldsymbol{a}_p^T\boldsymbol{x})(\boldsymbol{b}_p^T\mathbf{z}) \\
&\quad + \phi_1(\boldsymbol{u}_p^T\mathbf{x})\phi_2''(\boldsymbol{v}_p^T\mathbf{z})(\boldsymbol{b}_p^T\mathbf{z})^2 \bigg)^2 \\
&\lesssim \max_{p\in[r]} \bigg(\max_{\|\boldsymbol{a}_p\|_2^2+\|\boldsymbol{b}_p\|_2^2=1} (1-q_1)|\boldsymbol{v}_p^T\mathbf{z}|^{q_2}\mathbf{x}^T\boldsymbol{a}_p\boldsymbol{a}_p^T\mathbf{x} + |\mathbf{x}^T\boldsymbol{a}_p\boldsymbol{b}_p^T\mathbf{z}| + (1-q_2)|\boldsymbol{u}_p^T\mathbf{x}|^{q_1}\mathbf{z}^T\boldsymbol{b}_p\boldsymbol{b}_p^T\mathbf{z} \bigg)^2 \\
&\lesssim \max_{p\in[r]} \Big((1-q_1)(\mathbf{z}^T\boldsymbol{v}_p\boldsymbol{v}_p^T\mathbf{z})^{q_2}(\mathbf{x}^T\mathbf{x})^2 + (\mathbf{x}^T\mathbf{x})(\mathbf{z}^T\mathbf{z}) + (1-q_2)(\mathbf{x}^T\boldsymbol{u}_p\boldsymbol{u}_p^T\mathbf{x})^{q_1}(\mathbf{z}^T\mathbf{z})^2 \Big).
\end{aligned}
$$

By Lemma 20, for any $K_1^{(2,1)} \wedge K_2^{(2,1)} \wedge K_3^{(2,1)} \geq 1$, defining

$$
\Upsilon_4 = d_1(K_2^{(2,1)})^{\frac{q_2(1-q_1)}{2}}\|\mathbf{V}\|_2^{q_2(1-q_1)} + d_2(K_1^{(2,1)})^{\frac{q_1(1-q_2)}{2}}\|\mathbf{U}\|_2^{q_1(1-q_2)} \tag{40}
$$

and we have

$$
\begin{aligned}
&P\Big(\|\mathbf{T}_{11}(\mathbf{x},\mathbf{z})\|_2 \gtrsim (K_3^{(2,1)})^2\Upsilon_4^2\Big) \\
&\qquad \leq 2\exp\Big(-(d_1\wedge d_2)K_3^{(2,1)}\Big) + (1-q_1)q_2 r\exp(-K_2^{(2,1)}) + (1-q_2)q_1 r\exp(-K_1^{(2,1)}).
\end{aligned} \tag{41}
$$

For the condition (b) in Lemma 24, let us define

$$
\begin{aligned}
\mathbf{T}_{11}^{(1)} &:= (1-q_1)(\boldsymbol{v}_p^T\mathbf{z})^{2q_2}\mathbf{x}^T\mathbf{x} + \mathbf{z}^T\mathbf{z}, \quad \mathbf{T}_{11}^{(3)} := (1-q_2)(\boldsymbol{u}_p^T\mathbf{x})^{2q_1}\mathbf{z}^T\mathbf{z} + \mathbf{x}^T\mathbf{x}, \\
\mathbf{T}_{11}^{(2)} &:= (1-q_1)|\boldsymbol{v}_p^T\mathbf{z}|^{q_2}\mathbf{x}^T\mathbf{x} + (1-q_2)|\boldsymbol{u}_p^T\mathbf{x}|^{q_1}\mathbf{z}^T\mathbf{z}.
\end{aligned}
$$

Then,

$$
\begin{aligned}
&\|\mathbb{E}[\mathbf{T}_{11}(\mathbf{x},\mathbf{z})\mathbf{T}_{11}(\mathbf{x},\mathbf{z})^T]\|_2 \lesssim \max_{\|\boldsymbol{a}\|_F^2+\|\boldsymbol{b}\|_F^2=1} \mathbb{E}\bigg[\bigg(\sum_{p=1}^r \big((\mathbf{T}_{11}^{(1)})^2\mathbf{x}^T\mathbf{x} + (\mathbf{T}_{11}^{(2)})^2\mathbf{z}^T\mathbf{z}\big)(\boldsymbol{a}_p^T\mathbf{x})^2\bigg) \\
&\quad + 2\bigg(\sum_{p=1}^r \big(\mathbf{T}_{11}^{(1)}\mathbf{T}_{11}^{(2)}\mathbf{x}^T\mathbf{x} + \mathbf{T}_{11}^{(3)}\mathbf{T}_{11}^{(2)}\mathbf{z}^T\mathbf{z}\big)|\mathbf{x}^T\boldsymbol{a}_p\boldsymbol{b}_p^T\mathbf{z}|\bigg) + \bigg(\sum_{p=1}^r \big((\mathbf{T}_{11}^{(3)})^2\mathbf{z}^T\mathbf{z} + (\mathbf{T}_{11}^{(2)})^2\mathbf{x}^T\mathbf{x}\big)(\boldsymbol{b}_p^T\mathbf{z})^2\bigg)\bigg].
\end{aligned}
$$

By simple calculations based on Lemma 21,

$$
\begin{aligned}
&\mathbb{E}\big[(\mathbf{T}_{11}^{(1)})^2\mathbf{x}^T\mathbf{x}\cdot\mathbf{x}^T\boldsymbol{a}_p\boldsymbol{a}_p^T\mathbf{x}\big] \lesssim \big((1-q_1)\|\boldsymbol{v}_p\|_2^{4q_2}d_1^2 + d_2^2\big)d_1\|\boldsymbol{a}_p\|_2^2, \\
&\mathbb{E}\big[(\mathbf{T}_{11}^{(3)})^2\mathbf{z}^T\mathbf{z}\cdot\mathbf{z}^T\boldsymbol{b}_p\boldsymbol{b}_p^T\mathbf{z}\big] \lesssim \big((1-q_2)\|\boldsymbol{u}_p\|_2^{4q_1}d_2^2 + d_1^2\big)d_2\|\boldsymbol{b}_p\|_2^2, \\
&\mathbb{E}\big[(\mathbf{T}_{11}^{(2)})^2\mathbf{z}^T\mathbf{z}\cdot\mathbf{x}^T\boldsymbol{a}_p\boldsymbol{a}_p^T\mathbf{x}\big] \lesssim \big((1-q_1)\|\boldsymbol{v}_p\|_2^{2q_2}d_1^2 + (1-q_2)\|\boldsymbol{u}_p\|_2^{2q_1}d_2^2\big)d_2\|\boldsymbol{a}_p\|_2^2, \\
&\mathbb{E}\big[(\mathbf{T}_{11}^{(2)})^2\mathbf{x}^T\mathbf{x}\cdot\mathbf{z}^T\boldsymbol{b}_p\boldsymbol{b}_p^T\mathbf{z}\big] \lesssim \big((1-q_1)\|\boldsymbol{v}_p\|_2^{2q_2}d_1^2 + (1-q_2)\|\boldsymbol{u}_p\|_2^{2q_1}d_2^2\big)d_1\|\boldsymbol{b}_p\|_2^2, \\
&\mathbb{E}\big[\mathbf{T}_{11}^{(1)}\mathbf{T}_{11}^{(2)}\mathbf{x}^T\mathbf{x}|\mathbf{x}^T\boldsymbol{a}_p\boldsymbol{b}_p^T\mathbf{z}|\big] \lesssim \big((1-q_1)\|\boldsymbol{v}_p\|_2^{2q_2}d_1 + d_2\big)\big((1-q_1)\|\boldsymbol{v}_p\|_2^{q_2}d_1 + (1-q_2)\|\boldsymbol{u}_p\|_2^{q_1}d_2\big)d_1\|\boldsymbol{a}_p\|_2\|\boldsymbol{b}_p\|_2, \\
&\mathbb{E}\big[\mathbf{T}_{11}^{(3)}\mathbf{T}_{11}^{(2)}\mathbf{z}^T\mathbf{z}|\mathbf{x}^T\boldsymbol{a}_p\boldsymbol{b}_p^T\mathbf{z}|\Big] \lesssim \big((1-q_1)\|\boldsymbol{u}_p\|_2^{2q_1}d_2 + d_1\big)\big((1-q_1)\|\boldsymbol{v}_p\|_2^{q_2}d_1 + (1-q_2)\|\boldsymbol{u}_p\|_2^{q_1}d_2\big)d_2\|\boldsymbol{a}_p\|_2\|\boldsymbol{b}_p\|_2.
\end{aligned}
$$

Combining the above two displays and maximizing over $\{(\boldsymbol{a},\boldsymbol{b}) : \|\boldsymbol{a}\|_2^2 + \|\boldsymbol{b}\_2^2 = 1\}$,

$$
\|\mathbb{E}[\mathbf{T}_{11}(\mathbf{x},\mathbf{z})\mathbf{T}_{11}(\mathbf{x},\mathbf{z})^T]\|_2 \lesssim d_1^{3-q_1}d_2^{q_1}\|\mathbf{V}\|_2^{4q_2(1-q_1)} + d_2^{3-q_2}d_1^{q_2}\|\mathbf{U}\|_2^{4q_1(1-q_2)}. \tag{42}
$$

For condition (c) of Lemma 24,

$$
\mathbb{E}[\,((\boldsymbol{a};\boldsymbol{b})^T\mathbf{T}_{11}(\mathbf{x},\mathbf{z})(\boldsymbol{a};\boldsymbol{b}))^2] \lesssim \mathbb{E}\bigg[\bigg(\sum_{p=1}^r \mathbf{T}_{11}^{(1)}\mathbf{x}^T\boldsymbol{a}_p\boldsymbol{a}_p^T\mathbf{x} + 2\sum_{p=1}^r \mathbf{T}_{11}^{(2)}|\mathbf{x}^T\boldsymbol{a}_p\boldsymbol{b}_p^T\mathbf{z}| + \sum_{p=1}^r \mathbf{T}_{11}^{(3)}\mathbf{z}^T\boldsymbol{b}_p\boldsymbol{b}_p^T\mathbf{z}\bigg)^2\bigg].
$$

Applying Lemma 21,

$$\mathbb{E}\big[\big(\sum_{p=1}^{r}\mathbf{T}_{11}^{(1)}\mathbf{x}^T\boldsymbol{a}_p\boldsymbol{a}_p^T\mathbf{x}\big)^2\big] \lesssim \big((1-q_1)\|\mathbf{V}\|_2^{4q_2}d_1^2+d_2^2\big)\bigg(\sum_{p=1}^{r}\|\boldsymbol{a}_p\|_2^2\bigg)^2,$$
$$\mathbb{E}\big[\big(\sum_{p=1}^{r}\mathbf{T}_{11}^{(2)}|\mathbf{x}^T\boldsymbol{a}_p\boldsymbol{b}_p^T\mathbf{z}|\big)^2\big] \lesssim \big((1-q_1)\|\mathbf{V}\|_2^{2q_2}d_1^2+(1-q_2)\|\mathbf{U}\|_2^{2q_1}d_2^2\big)\bigg(\sum_{p=1}^{r}\|\boldsymbol{a}_p\|_2^2\bigg)\bigg(\sum_{p=1}^{r}\|\boldsymbol{b}_p\|_2^2\bigg),$$
$$\mathbb{E}\big[\big(\sum_{p=1}^{r}\mathbf{T}_{11}^{(3)}\mathbf{z}^T\boldsymbol{b}_p\boldsymbol{b}_p^T\mathbf{z}\big)^2\big] \lesssim \big((1-q_2)\|\mathbf{U}\|_2^{4q_1}d_2^2+d_1^2\big)\bigg(\sum_{p=1}^{r}\|\boldsymbol{b}_p\|_2^2\bigg)^2.$$

Thus,

$$\max_{\|\boldsymbol{a}\|_F^2+\|\boldsymbol{b}\|_F^2=1}\Big(\mathbb{E}\big[\big((\boldsymbol{a};\boldsymbol{b})^T\mathbf{T}_{11}(\mathbf{x},\mathbf{z})(\boldsymbol{a};\boldsymbol{b})\big)^2\big]\Big)^{1/2} \lesssim d_1\|\mathbf{V}\|_2^{2q_2(1-q_1)}+d_2\|\mathbf{U}\|_2^{2q_1(1-q_2)}. \tag{43}$$

Combining (39), (41), (42), (43), and defining

$$\Upsilon_5 = d_1^{3-q_1}d_2^{q_1}\|\mathbf{V}\|_2^{4q_2(1-q_1)}+d_2^{3-q_2}d_1^{q_2}\|\mathbf{U}\|_2^{4q_1(1-q_2)}, \quad \Upsilon_6 = d_1\|\mathbf{V}\|_2^{2q_2(1-q_1)}+d_2\|\mathbf{U}\|_2^{2q_1(1-q_2)}, \tag{44}$$

then conditions in Lemma 24 hold for $\mathcal{T}_{11}$ with parameters

$$\nu_1(\mathcal{T}_{11}) := \exp\Big(-(d_1\wedge d_2)K_3^{(2,1)}\Big)+q_2(1-q_1)r\exp\Big(-K_2^{(2,1)}\Big)+q_1(1-q_2)r\exp\Big(-K_1^{(2,1)}\Big),$$
$$\mu_1(\mathcal{T}_{11}) := (K_3^{(2,1)})^2\Upsilon_4^2, \qquad \nu_2(\mathcal{T}_{11}) := \Upsilon_5 \qquad \nu_3(\mathcal{T}_{11}) := \Upsilon_6, \qquad \|\mathbb{E}[\mathcal{T}_{11}]\| \lesssim \Upsilon_6.$$

Here, Υ_4, Υ_5, Υ_6 are defined in (40), (44), and $\{K_i^{(2,1)}\}_{i=1,2,3}$ are any constant. So, $\forall t>0$,

$$\begin{aligned}
&P\big(\big\|\mathcal{T}_{11}-\mathbb{E}[\mathcal{T}_{11}]\big\|_2 > t+\Upsilon_6\sqrt{\nu_1(\mathcal{T}_{11})}\big)\\
&\quad\le n_1n_2\nu_1(\mathcal{T}_{11})+2r(d_1+d_2)\exp\bigg(-\frac{(n_1\wedge n_2)t^2}{\big(2\Upsilon_5+4\Upsilon_6^2+4\Upsilon_6^2\nu_1(\mathcal{T}_{11})\big)+4\mu_1(\mathcal{T}_{11})t}\bigg)\\
&\quad\le n_1n_2\nu_1(\mathcal{I}_{21})+2r(d_1+d_2)\exp\bigg(-\frac{(n_1\wedge n_2)t^2}{10\Upsilon_5+4\mu_1(\mathcal{T}_{11})t}\bigg).
\end{aligned}$$

For any $s\ge 1$, we let

$$K_1^{2,1}=K_2^{2,1}=\log(n_1n_2r)+s\log(d_1+d_2), \quad K_3^{2,1}=1.$$

Then, $\Upsilon_4 \asymp \Upsilon_3$. Noting that $q=q_1\vee q_2$ and $q'=q_1q_2$, we can let

$$\epsilon_4 \asymp \sqrt{\frac{s(d_1+d_2)\log\left(r(d_1+d_2)\right)}{n_1\wedge n_2}} \vee \frac{s(d_1+d_2)\left\{\log\left(r(d_1+d_2)\right)\right\}^{1+q-q'}}{n_1\wedge n_2},$$

and then have

$$P\left(\|\mathcal{T}_{11}-\mathbb{E}[\mathcal{T}_{11}]\|_2 \ge \epsilon_4\Upsilon_6\right) \lesssim \frac{1}{(d_1+d_2)^s}.$$

Combining the above inequality with (39), $P(\|\mathcal{T}_{11}\|_2 \gtrsim \Upsilon_6) \lesssim 1/(d_1+d_2)^s$. We plug back into (38), combine with (37), and know Lemma 22 holds for $\mathcal{T}_1$ with parameters $\nu_1(\mathcal{T}_1)=\beta\Upsilon_3$ and $\nu_2(\mathcal{T}_1)=\beta^2\Upsilon_6$. Finally we apply Lemma 22 and obtain that $\forall t>0$

$$P\left(\mathcal{T}_1>t\right) \lesssim 2r(d_1+d_2)\exp\left(-\frac{mt^2}{4\nu_2(\mathcal{T}_1)+4\nu_1(\mathcal{T}_1)t}\right).$$

For any $s\ge 1$, we let

$$\epsilon_5 \asymp \sqrt{\frac{s(d_1+d_2)\log\left(r(d_1+d_2)\right)}{m}} \vee \frac{s(d_1+d_2)\left\{\log\left(r(d_1+d_2)\right)\right\}^{1+\frac{q-q'}{2}}}{m},$$

$$\Upsilon_7 = \|\mathbf{V}\|_2^{q_2(1-q_1)} + \|\mathbf{U}\|_2^{q_1(1-q_2)},$$

and have

$$P\left(\mathcal{T}_1 > \beta\epsilon_5\Upsilon_7\right) \lesssim \frac{1}{(d_1+d_2)^s}.$$

This completes the proof for the first part.

Proof of $\mathcal{T}_2$. We apply Lemma 23 to bound $\mathcal{T}_2$. We check all conditions of Lemma 23. By definition of $\mathbf{H}_2$ in (34),

$$\nabla^2\bar{\mathcal{L}}_2(\mathbf{U},\mathbf{V}) = \frac{1}{n_1^2 n_2^2}\sum_{(\mathbf{x},\mathbf{z})\in\mathcal{D}}\sum_{(\mathbf{x}',\mathbf{z}')\in\mathcal{D}'}\mathbf{H}_2\big((\mathbf{x},\mathbf{z}),(\mathbf{x}',\mathbf{z}')\big).$$

We first bound $\|\mathbb{E}[\mathbf{H}_2]\|_2$ as follows:

$$\begin{aligned}
\|\mathbb{E}[\mathbf{H}_2]\|_2 \lesssim & \beta\|\mathbb{E}\left[\begin{pmatrix}\boldsymbol{Q} & \boldsymbol{S}\\ \boldsymbol{S}^T & \boldsymbol{R}\end{pmatrix}\right]\|_2\\
\lesssim & \beta\max_{\|\boldsymbol{a}\|_F^2+\|\boldsymbol{b}\|_F^2=1}\Bigg|\sum_{p=1}^r\mathbb{E}\big[\phi_1''(\boldsymbol{u}_p^T\mathbf{x})\phi_2(\boldsymbol{v}_p^T\mathbf{z})(\boldsymbol{a}_p^T\mathbf{x})^2\big] + 2\sum_{p=1}^r\mathbb{E}\big[\phi_1'(\boldsymbol{u}_p^T\mathbf{x})\phi_2'(\boldsymbol{v}_p^T\mathbf{z})\mathbf{x}^T\boldsymbol{a}_p\boldsymbol{b}_p^T\mathbf{z}\big]\\
& + \sum_{p=1}^r\mathbb{E}\big[\phi_1(\boldsymbol{u}_p^T\mathbf{x})\phi_2''(\boldsymbol{v}_p^T\mathbf{z})(\boldsymbol{v}_p^T\mathbf{z})^2\big]\Bigg|\\
\lesssim & \beta\max_{\|\boldsymbol{a}\|_F^2+\|\boldsymbol{b}\|_F^2=1}\Bigg|(1-q_1)\sum_{p=1}^r\mathbb{E}\big[|\boldsymbol{v}_p^T\mathbf{z}|^{q_2}\mathbf{x}^T\boldsymbol{a}_p\boldsymbol{a}_p^T\mathbf{x}\big] + \sum_{p=1}^r\mathbb{E}\big[|\mathbf{x}^T\boldsymbol{a}_p\boldsymbol{b}_p^T\mathbf{z}|\big]\\
& + (1-q_2)\sum_{p=1}^r\mathbb{E}\big[|\boldsymbol{u}_p^T\mathbf{x}|^{q_1}\mathbf{z}^T\boldsymbol{b}_p\boldsymbol{b}_p^T\mathbf{z}\big]\Bigg|\\
\leq & \beta\max_{\|\boldsymbol{a}\|_F^2+\|\boldsymbol{b}\|_F^2=1}\left((1-q_1)\sum_{p=1}^r\|\boldsymbol{v}_p\|_2^{q_2}\|\boldsymbol{a}_p\|_2^2 + \sum_{p=1}^r\|\boldsymbol{a}_p\|_2\|\boldsymbol{b}_p\|_2 + (1-q_2)\sum_{p=1}^r\|\boldsymbol{u}_p\|_2^{q_1}\|\boldsymbol{b}_p\|_2^2\right)\\
\leq & \beta\big((1-q_1)\|\mathbf{V}\|_2^{q_2} + 1 + (1-q_2)\|\mathbf{U}\|_2^{q_1}\big)\\
\leq & \beta\Upsilon_7.
\end{aligned} \tag{45}$$

For the condition (a) in Lemma 23, we have shown in (35) that

$$\|\mathbf{H}_2\|_2 \lesssim \beta\big((1-q_1)\mathbf{x}^T\mathbf{x}\max_{p\in[r]}|\mathbf{z}^T\boldsymbol{v}_p|^{q_2} + (1-q_2)\mathbf{z}^T\mathbf{z}\max_{p\in[r]}|\mathbf{x}^T\boldsymbol{u}_p|^{q_1} + \|\mathbf{x}\|_2\|\mathbf{z}\|_2\big).$$

Thus, similar to (41),

$$\begin{aligned}
& P\left(\|\mathbf{H}_2\|_2 \gtrsim \beta K_3^{(2,2)}\left(d_1(K_2^{(2,2)})^{\frac{q_2(1-q_1)}{2}}\|\mathbf{V}\|_2^{q_2(1-q_1)} + d_2(K_1^{(2,2)})^{\frac{q_1(1-q_2)}{2}}\|\mathbf{U}\|_2^{q_1(1-q_2)}\right)\right)\\
& \qquad \leq 2\exp\left(-(d_1\wedge d_2)K_3^{(2,2)}\right) + (1-q_1)q_2 r\exp(-K_2^{(2,2)}) + (1-q_2)q_1\exp(-K_1^{(2,2)}).
\end{aligned}$$

For the condition (b) in Lemma 23,

$$\|\mathbb{E}[\mathbf{H}_2\mathbf{H}_2^T]\|_2 \lesssim \beta^2\|\mathbb{E}[\mathcal{T}_{11}]\|_2 \overset{(39)}{\lesssim} \beta^2\Upsilon_6.$$

For the condition (c) in Lemma 23, we use Lemma 21 and obtain

$$\begin{aligned}
& \max_{\|\boldsymbol{a}\|_F^2+\|\boldsymbol{b}\|_F^2=1}\mathbb{E}\big[\big(\begin{pmatrix}\boldsymbol{a}^T & \boldsymbol{b}^T\end{pmatrix}\mathbf{H}_2\begin{pmatrix}\boldsymbol{a}\\ \boldsymbol{b}\end{pmatrix}\big)^2\big]\\
& \qquad \lesssim \beta^2\max_{\|\boldsymbol{a}\|_F^2+\|\boldsymbol{b}\|_F^2=1}\mathbb{E}\big[\big(\begin{pmatrix}\boldsymbol{a}^T & \boldsymbol{b}^T\end{pmatrix}\begin{pmatrix}\boldsymbol{Q} & \boldsymbol{S}\\ \boldsymbol{S}^T & \boldsymbol{R}\end{pmatrix}\begin{pmatrix}\boldsymbol{a}\\ \boldsymbol{b}\end{pmatrix}\big)^2\big]
\end{aligned}$$

$$
\begin{aligned}
&\lesssim \beta^2 \max_{\|\boldsymbol{a}\|_F^2+\|\boldsymbol{b}\|_F^2=1} \mathbb{E}\Bigg[\Bigg(\sum_{p=1}^r (1-q_1)|\mathbf{z}^T\boldsymbol{v}_p|^{q_2}\mathbf{x}^T\boldsymbol{a}_p\boldsymbol{a}_p^T\mathbf{x} + \sum_{p=1}^r |\mathbf{x}^T\boldsymbol{a}_p\boldsymbol{b}_p^T\mathbf{z}| \\
&\qquad + \sum_{p=1}^r (1-q_2)|\mathbf{x}^T\boldsymbol{u}_p|^{q_1}\mathbf{z}^T\boldsymbol{b}_p\boldsymbol{b}_p^T\mathbf{z}\Bigg)^2\Bigg] \\
&\lesssim \beta^2\big((1-q_1)\|\mathbf{V}\|_2^{2q_2} + 1 + (1-q_2)\|\mathbf{U}\|_2^{2q_1}\big) \\
&\lesssim \beta^2\Upsilon_7^2.
\end{aligned}
$$

Thus, conditions of Lemma 23 hold for $\mathcal{T}_2$ with parameters (up to constants)

$$
\begin{aligned}
\mu_1(\mathcal{T}_2) &:= \beta K_3^{(2,2)}\left(d_1(K_2^{(2,2)})^{\frac{q_2(1-q_1)}{2}}\|\mathbf{V}\|_2^{q_2(1-q_1)} + d_2(K_1^{(2,2)})^{\frac{q_1(1-q_2)}{2}}\|\mathbf{U}\|_2^{q_1(1-q_2)}\right), \\
\nu_1(\mathcal{T}_2) &:= \exp\left(-(d_1\wedge d_2)K_3^{(2,2)}\right) + (1-q_1)q_2 r\exp(-K_2^{(2,2)}) + (1-q_2)q_1\exp(-K_1^{(2,2)}), \\
\nu_2(\mathcal{T}_2) &:= \beta^2\Upsilon_6, \qquad \nu_3(\mathcal{T}_2) := \beta\Upsilon_7, \qquad \|\mathbb{E}[\mathbf{H}_2]\| \lesssim \beta\Upsilon_7.
\end{aligned}
$$

For any $s \geq 1$, we let $K_1^{(2,2)} = K_2^{(2,2)} = 2\log n_1 n_2 r + s\log(d_1+d_2)$, $K_3^{(2,2)} = 1$, and

$$
\epsilon_6 \asymp \sqrt{\frac{s(d_1+d_2)\log(r(d_1+d_2))}{n_1\wedge n_2}} \vee \frac{s(d_1+d_2)\left\{\log(r(d_1+d_2))\right\}^{1+\frac{q-q'}{2}}}{n_1\wedge n_2},
$$

and then have

$$
P\big(\mathcal{T}_2 \gtrsim \beta\epsilon_6\Upsilon_7\big) \lesssim \frac{1}{(d_1+d_2)^s}.
$$

We finish the proof by noting that the first term of ϵ_6 is the dominant.

G.6. Proof of Lemma 17

By definition of $\mathcal{J}_3$,

$$
\begin{aligned}
&\|\mathbb{E}[\nabla^2\mathcal{L}_1(\mathbf{U},\mathbf{V})] - \mathbb{E}[\nabla^2\mathcal{L}_1(\mathbf{U}^\star,\mathbf{V}^\star)]\|_2 \\
&\qquad = \left\|\mathbb{E}\left[A\begin{pmatrix}\boldsymbol{d}-\boldsymbol{d}'\\ \boldsymbol{p}-\boldsymbol{p}'\end{pmatrix}\begin{pmatrix}\boldsymbol{d}-\boldsymbol{d}'\\ \boldsymbol{p}-\boldsymbol{p}'\end{pmatrix}^T\right] - \mathbb{E}\left[A^\star\begin{pmatrix}\boldsymbol{d}^\star-\boldsymbol{d}'^\star\\ \boldsymbol{p}^\star-\boldsymbol{p}'^\star\end{pmatrix}\begin{pmatrix}\boldsymbol{d}^\star-\boldsymbol{d}'^\star\\ \boldsymbol{p}^\star-\boldsymbol{p}'^\star\end{pmatrix}^T\right]\right\|_2 \\
&\qquad \leq \left\|\mathbb{E}\left[A\left(\begin{pmatrix}\boldsymbol{d}-\boldsymbol{d}'\\ \boldsymbol{p}-\boldsymbol{p}'\end{pmatrix}\begin{pmatrix}\boldsymbol{d}-\boldsymbol{d}'\\ \boldsymbol{p}-\boldsymbol{p}'\end{pmatrix}^T - \begin{pmatrix}\boldsymbol{d}^\star-\boldsymbol{d}'^\star\\ \boldsymbol{p}^\star-\boldsymbol{p}'^\star\end{pmatrix}\begin{pmatrix}\boldsymbol{d}^\star-\boldsymbol{d}'^\star\\ \boldsymbol{p}^\star-\boldsymbol{p}'^\star\end{pmatrix}^T\right)\right]\right\|_2 \\
&\qquad\quad + \left\|\mathbb{E}\left[(A-A^\star)\begin{pmatrix}\boldsymbol{d}^\star-\boldsymbol{d}'^\star\\ \boldsymbol{p}^\star-\boldsymbol{p}'^\star\end{pmatrix}\begin{pmatrix}\boldsymbol{d}^\star-\boldsymbol{d}'^\star\\ \boldsymbol{p}^\star-\boldsymbol{p}'^\star\end{pmatrix}^T\right]\right\|_2 := \|\mathcal{J}_{31}\|_2 + \|\mathcal{J}_{32}\|_2.
\end{aligned} \tag{46}
$$

For $\mathcal{J}_{31}$,

$$
\|\mathcal{J}_{31}\|_2 \lesssim \beta^2\left(\left\|\mathbb{E}\left[\begin{pmatrix}\boldsymbol{d}\\ \boldsymbol{p}\end{pmatrix}\begin{pmatrix}\boldsymbol{d}\\ \boldsymbol{p}\end{pmatrix}^T - \begin{pmatrix}\boldsymbol{d}^\star\\ \boldsymbol{p}^\star\end{pmatrix}\begin{pmatrix}\boldsymbol{d}^\star\\ \boldsymbol{p}^\star\end{pmatrix}^T\right]\right\|_2 + \left\|\mathbb{E}\begin{pmatrix}\boldsymbol{d}\\ \boldsymbol{p}\end{pmatrix}\mathbb{E}\begin{pmatrix}\boldsymbol{d}\\ \boldsymbol{p}\end{pmatrix}^T - \mathbb{E}\begin{pmatrix}\boldsymbol{d}^\star\\ \boldsymbol{p}^\star\end{pmatrix}\mathbb{E}\begin{pmatrix}\boldsymbol{d}^\star\\ \boldsymbol{p}^\star\end{pmatrix}^T\right\|_2\right).
$$

We only bound the first term. The second term has the same bound using the equation $\mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{x}]^T = \mathbb{E}[\mathbf{x}\mathbf{x}'^T]$ for any variable $\mathbf{x}'$ independent from $\mathbf{x}$. Note that

$$
\begin{aligned}
&\left\|\mathbb{E}\left[\begin{pmatrix}\boldsymbol{d}\\ \boldsymbol{p}\end{pmatrix}\begin{pmatrix}\boldsymbol{d}\\ \boldsymbol{p}\end{pmatrix}^T - \begin{pmatrix}\boldsymbol{d}^\star\\ \boldsymbol{p}^\star\end{pmatrix}\begin{pmatrix}\boldsymbol{d}^\star\\ \boldsymbol{p}^\star\end{pmatrix}^T\right]\right\|_2 \\
&= \max_{\|\boldsymbol{a}\|_F^2+\|\boldsymbol{b}\|_F^2=1}\Bigg|\sum_{i,j=1}^r \mathbb{E}\big[\big(\phi_1'(\boldsymbol{u}_i^T\mathbf{x})\phi_2(\boldsymbol{v}_i^T\mathbf{z})\phi_1'(\boldsymbol{u}_j^T\mathbf{x})\phi_2(\boldsymbol{v}_j^T\mathbf{z}) - \phi_1'(\boldsymbol{u}_i^{\star T}\mathbf{x})\phi_2(\boldsymbol{v}_i^{\star T}\mathbf{z})\phi_1'(\boldsymbol{u}_j^{\star T}\mathbf{x})\phi_2(\boldsymbol{v}_j^{\star T}\mathbf{z})\big)\cdot
\end{aligned}
$$

$$(\mathbf{x}^T \boldsymbol{a}_i \boldsymbol{a}_j^T \mathbf{x})\big] + 2\sum_{i,j=1}^{r} \mathbb{E}\big[\big(\phi_1'(\boldsymbol{u}_i^T\mathbf{x})\phi_2(\boldsymbol{v}_i^T\mathbf{z})\phi_1(\boldsymbol{u}_j^T\mathbf{x})\phi_2'(\boldsymbol{v}_j^T\mathbf{z}) - \phi_1'(\boldsymbol{u}_i^{\star T}\mathbf{x})\phi_2(\boldsymbol{v}_i^{\star T}\mathbf{z})\phi_1(\boldsymbol{u}_j^{\star T}\mathbf{x})\phi_2'(\boldsymbol{v}_j^{\star T}\mathbf{z})\big)$$
$$\cdot (\mathbf{x}^T \boldsymbol{a}_i \boldsymbol{b}_j^T \mathbf{x})\big] + \sum_{i,j=1}^{r} \mathbb{E}\big[\big(\phi_1(\boldsymbol{u}_i^T\mathbf{x})\phi_2'(\boldsymbol{v}_i^T\mathbf{z})\phi_1(\boldsymbol{u}_j^T\mathbf{x})\phi_2'(\boldsymbol{v}_j^T\mathbf{z}) - \phi_1(\boldsymbol{u}_i^{\star T}\mathbf{x})\phi_2'(\boldsymbol{v}_i^{\star T}\mathbf{z})\phi_1(\boldsymbol{u}_j^{\star T}\mathbf{x})\phi_2'(\boldsymbol{v}_j^{\star T}\mathbf{z})\big)$$
$$\cdot (\mathbf{z}^T \boldsymbol{b}_i \boldsymbol{b}_j^T \mathbf{z})\big]\bigg|. \tag{47}$$

We focus on the first term in the above equality. By simple calculations using the boundedness and Lipschitz continuity of ϕ_i, ϕ_i',

$$\begin{aligned}
&\big|\phi_1'(\boldsymbol{u}_i^T\mathbf{x})\phi_2(\boldsymbol{v}_i^T\mathbf{z})\phi_1'(\boldsymbol{u}_j^T\mathbf{x})\phi_2(\boldsymbol{v}_j^T\mathbf{z}) - \phi_1'(\boldsymbol{u}_i^{\star T}\mathbf{x})\phi_2(\boldsymbol{v}_i^{\star T}\mathbf{z})\phi_1'(\boldsymbol{u}_j^{\star T}\mathbf{x})\phi_2(\boldsymbol{v}_j^{\star T}\mathbf{z})\big| \\
&\qquad\le |\phi_1'(\boldsymbol{u}_i^T\mathbf{x}) - \phi_1'(\boldsymbol{u}_i^{\star T}\mathbf{x})| \cdot |\mathbf{z}^T\boldsymbol{v}_i^\star \boldsymbol{v}_j^{\star T}\mathbf{z}|^{q_2} + |\mathbf{z}^T(\boldsymbol{v}_i - \boldsymbol{v}_i^\star)| \cdot |\boldsymbol{v}_j^{\star T}\mathbf{z}|^{q_2} \\
&\qquad\quad + |\phi_1'(\boldsymbol{u}_j^T\mathbf{x}) - \phi_1'(\boldsymbol{u}_j^{\star T}\mathbf{x})| \cdot |\mathbf{z}^T\boldsymbol{v}_i^\star \boldsymbol{v}_j^{\star T}\mathbf{z}|^{q_2} + |\mathbf{z}^T(\boldsymbol{v}_j - \boldsymbol{v}_j^{\star T})| \cdot |\boldsymbol{v}_i^{\star T}\mathbf{z}|^{q_2}.
\end{aligned}$$

Plugging the above inequality back into (47), dealing with other terms similarly, and applying Lemma 25 by noting $\sigma_r(\mathbf{U}^\star) \wedge \sigma_r(\mathbf{V}^\star) \ge 1$,

$$\begin{aligned}
&\left\|\mathbb{E}\left[\begin{pmatrix}\boldsymbol{d}\\ \boldsymbol{p}\end{pmatrix}\begin{pmatrix}\boldsymbol{d}\\ \boldsymbol{p}\end{pmatrix}^T - \begin{pmatrix}\boldsymbol{d}^\star\\ \boldsymbol{p}^\star\end{pmatrix}\begin{pmatrix}\boldsymbol{d}^\star\\ \boldsymbol{p}^\star\end{pmatrix}^T\right]\right\|_2 \\
&\lesssim \max_{\|\boldsymbol{a}\|_F^2 + \|\boldsymbol{b}\|_F^2 = 1} \sum_{i,j=1}^r \|\boldsymbol{u}_i - \boldsymbol{u}_i^\star\|_2^{1-\frac{q_1}{2}} \|\boldsymbol{v}_i^\star\|_2^{q_2} \|\boldsymbol{v}_j^\star\|_2^{q_2} \|\boldsymbol{a}_i\|_2 \|\boldsymbol{a}_j\|_2 + \sum_{i,j=1}^r \|\boldsymbol{v}_i - \boldsymbol{v}_i^\star\|_2 \|\boldsymbol{v}_j^\star\|_2^{q_2} \|\boldsymbol{a}_i\|_2 \|\boldsymbol{a}_j\|_2 \\
&\quad + \sum_{i,j=1}^r \|\boldsymbol{u}_i - \boldsymbol{u}_i^\star\|_2^{1-\frac{q_1}{2}} \|\boldsymbol{u}_j^\star\|_2^{q_1} \|\boldsymbol{v}_i^\star\|^{q_2} \|\boldsymbol{a}_i\|_2 \|\boldsymbol{b}_j\|_2 + \sum_{i,j=1}^r \|\boldsymbol{v}_i - \boldsymbol{v}_i^\star\|_2 \|\boldsymbol{u}_j^\star\|^{q_1} \|\boldsymbol{a}_i\|_2 \|\boldsymbol{b}_j\|_2 \\
&\quad + \sum_{i,j=1}^r \|\boldsymbol{v}_i - \boldsymbol{v}_i^\star\|_2^{1-\frac{q_2}{2}} \|\boldsymbol{v}_j^\star\|_2^{q_1} \|\boldsymbol{u}_i^\star\|_2^{q_1} \|\boldsymbol{a}_j\|_2 \|\boldsymbol{b}_i\|_2 + \sum_{i,j=1}^r \|\boldsymbol{u}_i - \boldsymbol{u}_i^\star\|_2 \|\boldsymbol{v}_j^\star\|_2 \|\boldsymbol{a}_j\|_2 \|\boldsymbol{b}_i\|_2 \\
&\quad + \sum_{i,j=1}^r \|\boldsymbol{v}_i - \boldsymbol{v}_i^\star\|_2^{1-\frac{q_2}{2}} \|\boldsymbol{u}_i^\star\|_2^{q_1} \|\boldsymbol{u}_j^\star\|_2^{q_1} \|\boldsymbol{b}_i\|_2 \|\boldsymbol{b}_j\|_2 + \sum_{i,j=1}^r \|\boldsymbol{u}_i - \boldsymbol{u}_i^\star\|_2 \|\boldsymbol{u}_j^\star\|_2^{q_1} \|\boldsymbol{b}_i\|_2 \|\boldsymbol{b}_j\|_2 \\
&= \max_{\|\boldsymbol{a}\|_F^2 + \|\boldsymbol{b}\|_F^2 = 1} \Bigg(\sum_{i=1}^r \big(\|\boldsymbol{u}_i - \boldsymbol{u}_i^\star\|_2^{1-\frac{q_1}{2}} \|\boldsymbol{v}_i^\star\|_2^{q_2} + \|\boldsymbol{v}_i - \boldsymbol{v}_i^\star\|_2\big)\|\boldsymbol{a}_i\| + \sum_{j=1}^r \big(\|\boldsymbol{v}_j - \boldsymbol{v}_j^\star\|_2^{1-\frac{q_2}{2}} \|\boldsymbol{u}_j^\star\|_2^{q_1} \\
&\quad + \|\boldsymbol{u}_j - \boldsymbol{u}_j^\star\|_2\big)\|\boldsymbol{b}_j\|_2\Bigg) \cdot \Bigg(\sum_{i=1}^r \|\boldsymbol{v}_i^\star\|_2^{q_2} \|\boldsymbol{a}_i\|_2 + \sum_{j=1}^r \|\boldsymbol{u}_j^\star\|_2^{q_1} \|\boldsymbol{b}_j\|_2\Bigg) \\
&\le \sqrt{\|\mathbf{U} - \mathbf{U}^\star\|_F^2 + \|\mathbf{V} - \mathbf{V}^\star\|_F^2 + \sum_{i=1}^r \|\boldsymbol{u}_i - \boldsymbol{u}_i^\star\|_2^{2-q_1} \|\boldsymbol{v}_i^\star\|_2^{2q_2} + \|\boldsymbol{v}_i - \boldsymbol{v}_i^\star\|_2^{2-q_2} \|\boldsymbol{u}_i^\star\|_2^{2q_1}} \\
&\quad \cdot \sqrt{\sum_{i=1}^r \|\boldsymbol{v}_i^\star\|_2^{2q_2} + \|\boldsymbol{u}_i^\star\|_2^{2q_1}} \\
&\le \big(\|\mathbf{U} - \mathbf{U}^\star\|_F + \|\mathbf{V} - \mathbf{V}^\star\|_F + \|\mathbf{U} - \mathbf{U}^\star\|_2^{1-\frac{q_1}{2}} + \|\mathbf{V} - \mathbf{V}^\star\|_2^{1-\frac{q_2}{2}}\big)\Upsilon_2^\star,
\end{aligned} \tag{48}$$

where $\Upsilon_2^\star$ is defined in the same way as Υ_2 in (32) but calculated using $\mathbf{U}^\star, \mathbf{V}^\star$. Next, we bound $\mathcal{J}_{32}$. Since ψ is Lipschitz continuous,

$$\begin{aligned}
|A - A^\star| \lesssim \beta^3 &|\phi_1(\mathbf{U}^T\mathbf{x})^T\phi_2(\mathbf{V}^T\mathbf{z}) - \phi_1(\mathbf{U}^{\star T}\mathbf{x})^T\phi_2(\mathbf{V}^{\star T}\mathbf{z})| \\
&+ |\phi_1(\mathbf{U}^T\mathbf{x}')^T\phi_2(\mathbf{V}^T\mathbf{z}') - \phi_1(\mathbf{U}^{\star T}\mathbf{x}')^T\phi_2(\mathbf{V}^{\star T}\mathbf{z}')|.
\end{aligned}$$

Thus,

$$\|\mathcal{J}_{32}\|_2 \lesssim \beta^3 \left\|\mathbb{E}\left[\big|\phi_1(\mathbf{U}^T\mathbf{x})^T\phi_2(\mathbf{V}^T\mathbf{z}) - \phi_1(\mathbf{U}^{\star T}\mathbf{x})^T\phi_2(\mathbf{V}^{\star T}\mathbf{z})\big| \begin{pmatrix}\boldsymbol{d}^\star - \boldsymbol{d}'^\star\\ \boldsymbol{p}^\star - \boldsymbol{p}'^\star\end{pmatrix}\begin{pmatrix}\boldsymbol{d}^\star - \boldsymbol{d}'^\star\\ \boldsymbol{p}^\star - \boldsymbol{p}'^\star\end{pmatrix}^T\right]\right\|_2$$

$$\lesssim\beta^3\sqrt{\mathbb{E}\big[\,\Big(\phi_1(\mathbf{U}^T\mathbf{x})^T\phi_2(\mathbf{V}^T\mathbf{z})-\phi_1(\mathbf{U}^{\star T}\mathbf{x})^T\phi_2(\mathbf{V}^{\star T}\mathbf{z})\Big)^2\,\big]}\max_{\|\boldsymbol{a}\|_F^2+\|\boldsymbol{b}\|_F^2=1}\sqrt{\mathbb{E}[(\boldsymbol{d}^{\star T}\boldsymbol{a}+\boldsymbol{p}^{\star T}\boldsymbol{b})^4]}.$$

For the first term,

$$\begin{aligned}
&\mathbb{E}\big[\big(\phi_1(\mathbf{U}^T\mathbf{x})^T\phi_2(\mathbf{V}^T\mathbf{z})-\phi_1(\mathbf{U}^{\star T}\mathbf{x})^T\phi_2(\mathbf{V}^{\star T}\mathbf{z})\big)^2\big]\\
&\quad\lesssim\mathbb{E}\big[\big|\big(\phi_1(\mathbf{U}^T\mathbf{x})-\phi_1(\mathbf{U}^{\star T}\mathbf{x})\big)^T\phi_2(\mathbf{V}^{\star T}\mathbf{z})\big|^2\big]+\mathbb{E}\big[\big|\big(\phi_2(\mathbf{V}^T\mathbf{z})-\phi_2(\mathbf{V}^{\star T}\mathbf{z})\big)^T\phi_1(\mathbf{U}^{\star T}\mathbf{x})\big|^2\big]\\
&\quad\lesssim\mathbb{E}\big[\big(\sum_{p=1}^r|(\boldsymbol{u}_p-\boldsymbol{u}_p^\star)^T\mathbf{x}|\cdot|\boldsymbol{v}_p^{\star T}\mathbf{z}|^{q_2}\big)^2\big]+\mathbb{E}\big[\big(\sum_{p=1}^r|(\boldsymbol{v}_p-\boldsymbol{v}_p^\star)^T\mathbf{z}|\cdot|\boldsymbol{u}_p^{\star T}\mathbf{x}|^{q_1}\big)^2\big]\\
&\quad\lesssim\sum_{p=1}^r\|\boldsymbol{u}_p-\boldsymbol{u}_p^\star\|_2^2\sum_{p=1}^r\|\boldsymbol{v}_p^\star\|_2^{2q_2}+\sum_{p=1}^r\|\boldsymbol{v}_p-\boldsymbol{v}_p^\star\|_2^2\sum_{p=1}^r\|\boldsymbol{u}_p^\star\|_2^2\\
&\quad\lesssim\|\mathbf{U}-\mathbf{U}^\star\|_F^2\|\mathbf{V}^\star\|_F^{2q_2}r^{1-q_2}+\|\mathbf{V}-\mathbf{V}^\star\|_F^2\|\mathbf{U}^\star\|_F^{2q_1}r^{1-q_1}.
\end{aligned}$$

For the second term, from (31) we see $\max_{\|\boldsymbol{a}\|_F^2+\|\boldsymbol{b}\|_F^2=1}\sqrt{\mathbb{E}[(\boldsymbol{d}^{\star T}\boldsymbol{a}+\boldsymbol{p}^{\star T}\boldsymbol{b})^4]}\lesssim\Upsilon_2^\star$. Combining with the above two displays, and (48) and (46),

$$\begin{aligned}
&\|\mathbb{E}[\nabla^2\mathcal{L}_1(\mathbf{U},\mathbf{V})]-\mathbb{E}[\nabla^2\mathcal{L}_1(\mathbf{U}^\star,\mathbf{V}^\star)]\|_2\\
&\lesssim\beta^3\Upsilon_2^\star\big(\|\mathbf{U}-\mathbf{U}^\star\|_2^{1-\frac{q_1}{2}}+\|\mathbf{V}-\mathbf{V}^\star\|_2^{1-\frac{q_2}{2}}+\|\mathbf{U}-\mathbf{U}^\star\|_F\|\mathbf{V}^\star\|_F^{q_2}r^{\frac{1-q_2}{2}}+\|\mathbf{V}-\mathbf{V}^\star\|_F\|\mathbf{U}^\star\|_F^{q_1}r^{\frac{1-q_1}{2}}\big)\\
&\lesssim\beta^3(\Upsilon_2^\star)^{3/2}\big(\|\mathbf{U}-\mathbf{U}^\star\|_F^{1-\frac{q_1}{2}}+\|\mathbf{V}-\mathbf{V}^\star\|_F^{1-\frac{q_2}{2}}\big).
\end{aligned}$$

This completes the proof.

G.7. Proof of Lemma 18

We follow the same proof sketch as Lemma 17. By definition of $\mathcal{T}_3$,

$$\begin{aligned}
\|\mathbb{E}[\nabla^2\mathcal{L}_2(\mathbf{U},\mathbf{V})]-&\mathbb{E}[\nabla^2\mathcal{L}_2(\mathbf{U}^\star,\mathbf{V}^\star)]\|_2\\
&=\left\|\mathbb{E}\left[B\begin{pmatrix}\boldsymbol{Q}-\boldsymbol{Q}' & \boldsymbol{S}-\boldsymbol{S}'\\ \boldsymbol{S}^T-\boldsymbol{S}'^T & \boldsymbol{R}-\boldsymbol{R}'\end{pmatrix}\right]-\mathbb{E}\left[B^\star\begin{pmatrix}\boldsymbol{Q}^\star-\boldsymbol{Q}^{\star\prime} & \boldsymbol{S}^\star-\boldsymbol{S}^{\star\prime}\\ \boldsymbol{S}^{\star T}-\boldsymbol{S}^{\star\prime T} & \boldsymbol{R}^\star-\boldsymbol{R}^{\star\prime}\end{pmatrix}\right]\right\|_2\\
&\le\left\|\mathbb{E}\left[B\left(\begin{pmatrix}\boldsymbol{Q}-\boldsymbol{Q}' & \boldsymbol{S}-\boldsymbol{S}'\\ \boldsymbol{S}^T-\boldsymbol{S}'^T & \boldsymbol{R}-\boldsymbol{R}'\end{pmatrix}-\begin{pmatrix}\boldsymbol{Q}^\star-\boldsymbol{Q}^{\star\prime} & \boldsymbol{S}^\star-\boldsymbol{S}^{\star\prime}\\ \boldsymbol{S}^{\star T}-\boldsymbol{S}^{\star\prime T} & \boldsymbol{R}^\star-\boldsymbol{R}^{\star\prime}\end{pmatrix}\right)\right]\right\|_2\\
&\quad+\left\|\mathbb{E}\left[(B-B^\star)\begin{pmatrix}\boldsymbol{Q}^\star-\boldsymbol{Q}^{\star\prime} & \boldsymbol{S}^\star-\boldsymbol{S}^{\star\prime}\\ \boldsymbol{S}^{\star T}-\boldsymbol{S}^{\star\prime T} & \boldsymbol{R}^\star-\boldsymbol{R}^{\star\prime}\end{pmatrix}\right]\right\|_2\coloneqq\|\mathcal{T}_{31}\|_2+\|\mathcal{T}_{32}\|_2.
\end{aligned}$$

For $\mathcal{T}_{31}$,

$$\begin{aligned}
\mathcal{T}_{31}&\lesssim\beta\left\|\mathbb{E}\left[\begin{pmatrix}\boldsymbol{Q}-\boldsymbol{Q}^\star & \boldsymbol{S}-\boldsymbol{S}^\star\\ \boldsymbol{S}^T-\boldsymbol{S}^{\star T} & \boldsymbol{R}-\boldsymbol{R}^\star\end{pmatrix}\right]\right\|_2\\
&\lesssim\beta\max_{\|\boldsymbol{a}\|_F^2+\|\boldsymbol{b}\|_F^2=1}\bigg|\sum_{p=1}^r\mathbb{E}\big[\big(\phi_1''(\boldsymbol{u}_p^T\mathbf{x})\phi_2(\boldsymbol{v}_p^T\mathbf{z})-\phi_1''(\boldsymbol{u}_p^{\star T}\mathbf{x})\phi_2(\boldsymbol{v}_p^{\star T}\mathbf{z})\big)(\boldsymbol{a}_p^T\mathbf{x})^2\big]\\
&\qquad+2\sum_{p=1}^r\mathbb{E}\big[\big(\phi_1'(\boldsymbol{u}_p^T\mathbf{x})\phi_2'(\boldsymbol{v}_p^T\mathbf{z})-\phi_1'(\boldsymbol{u}_p^{\star T}\mathbf{x})\phi_2'(\boldsymbol{v}_p^{\star T}\mathbf{z})\big)(\mathbf{x}^T\boldsymbol{a}_p\boldsymbol{b}_p^T\mathbf{z})\big]\\
&\qquad+\sum_{p=1}^r\mathbb{E}\big[\big(\phi_1(\boldsymbol{u}_p^T\mathbf{x})\phi_2''(\boldsymbol{v}_p^T\mathbf{z})-\phi_1(\boldsymbol{u}_p^{\star T}\mathbf{x})\phi_2''(\boldsymbol{v}_p^{\star T}\mathbf{z})\big)(\boldsymbol{b}_p^T\mathbf{z})^2\big]\bigg|\\
&\lesssim\beta\max_{\|\boldsymbol{a}\|_F^2+\|\boldsymbol{b}\|_F^2=1}\bigg((1-q_1)\sum_{p=1}^r\mathbb{E}\big[\big(|(\boldsymbol{u}_p-\boldsymbol{u}_p^\star)^T\mathbf{x}|\cdot|\boldsymbol{v}_p^{\star T}\mathbf{z}|^{q_2}+|(\boldsymbol{v}_p-\boldsymbol{v}_p^\star)^T\mathbf{z}|\big)\mathbf{x}^T\boldsymbol{a}_p\boldsymbol{a}_p^T\mathbf{x}\big]\\
&\qquad+(1-q_2)\sum_{p=1}^r\mathbb{E}\big[\big(|(\boldsymbol{v}_p-\boldsymbol{v}_p^\star)^T\mathbf{z}|\cdot|\boldsymbol{u}_p^{\star T}\mathbf{x}|^{q_1}+|(\boldsymbol{u}_p-\boldsymbol{u}_p^\star)^T\mathbf{x}|\big)\mathbf{z}^T\boldsymbol{b}_p\boldsymbol{b}_p^T\mathbf{z}\big]
\end{aligned}$$

$$
\begin{aligned}
&+\sum_{p=1}^{r}\mathbb{E}\big[\big(|\phi_1'(\boldsymbol{u}_p^T\mathbf{x})-\phi_1'(\boldsymbol{u}_p^{\star T}\mathbf{x})|+|\phi_2'(\boldsymbol{v}_p^T\mathbf{z})-\phi_2'(\boldsymbol{v}_p^{\star T}\mathbf{z})|\big)\cdot|\mathbf{x}^T\boldsymbol{a}_p\boldsymbol{b}_p^T\mathbf{z}|\big]\Bigg)\\
\lesssim&\beta\max_{\|\boldsymbol{a}\|_F^2+\|\boldsymbol{b}\|_F^2=1}\Bigg((1-q_1)\sum_{p=1}^{r}\big(\|\boldsymbol{u}_p-\boldsymbol{u}_p^\star\|_2\|\boldsymbol{v}_p^\star\|_2^{q_2}+\|\boldsymbol{v}_p-\boldsymbol{v}_p^\star\|_2\big)\|\boldsymbol{a}_p\|_2^2+(1-q_2)\sum_{p=1}^{r}\big(\|\boldsymbol{v}_p-\boldsymbol{v}_p^\star\|_2\|\boldsymbol{u}_p^\star\|_2^{q_1}\\
&+\|\boldsymbol{u}_p-\boldsymbol{u}_p^\star\|_2\big)\|\boldsymbol{b}_p\|_2^2+\sum_{p=1}^{r}\|\boldsymbol{u}_p-\boldsymbol{u}_p^\star\|_2^{1-\frac{q_1}{2}}\|\boldsymbol{a}_p\|_2\|\boldsymbol{b}_p\|_2+\sum_{p=1}^{r}\|\boldsymbol{v}_p-\boldsymbol{v}_p^\star\|_2^{1-\frac{q_2}{2}}\|\boldsymbol{a}_p\|_2\|\boldsymbol{b}_p\|_2\Bigg)\\
\lesssim&\beta\Bigg((1-q_1)(\|\mathbf{U}-\mathbf{U}^\star\|_2\|\mathbf{V}^\star\|_2^{q_2}+\|\mathbf{V}-\mathbf{V}^\star\|_2)+(1-q_2)(\|\mathbf{V}-\mathbf{V}^\star\|_2\|\mathbf{U}^\star\|_2^{q_1}+\|\mathbf{U}-\mathbf{U}^\star\|_2)\\
&+\|\mathbf{U}-\mathbf{U}^\star\|_2^{1-\frac{q_1}{2}}+\|\mathbf{V}-\mathbf{V}^\star\|_2^{1-\frac{q_2}{2}}\Bigg)\\
\lesssim&\beta(\|\mathbf{U}-\mathbf{U}^\star\|_2^{1-\frac{q_1}{2}}\|\mathbf{V}^\star\|_2^{q_2(1-q_1)}+\|\mathbf{V}-\mathbf{V}^\star\|_2^{1-\frac{q_2}{2}}\|\mathbf{U}^\star\|_2^{q_1(1-q_2)})\\
\lesssim&\beta(\|\mathbf{U}-\mathbf{U}^\star\|_2^{1-\frac{q_1}{2}}+\|\mathbf{V}-\mathbf{V}^\star\|_2^{1-\frac{q_2}{2}})\Upsilon_7^\star,
\end{aligned}
$$

where $\Upsilon_7^\star$ has the same form as Υ_7 but is calculated using $\mathbf{U}^\star,\mathbf{V}^\star$. For $\mathcal{T}_{32}$, we use the Lipschitz continuity of $1/(1+\exp(x))$, and simplify analogously to $\mathcal{J}_{32}$. We obtain

$$
\|\mathcal{T}_{32}\|_2\lesssim\beta^2(\|\mathbf{U}-\mathbf{U}^\star\|_F\|\mathbf{V}^\star\|_F^{q_2}r^{\frac{1-q_2}{2}}+\|\mathbf{V}-\mathbf{V}^\star\|_F\|\mathbf{U}^\star\|_F^{q_1}r^{\frac{1-q_1}{2}})\Upsilon_7^\star.
$$

Combining the above three displays,

$$
\begin{aligned}
&\|\mathbb{E}[\nabla^2\mathcal{L}_2(\mathbf{U},\mathbf{V})]-\mathbb{E}[\nabla^2\mathcal{L}_2(\mathbf{U}^\star,\mathbf{V}^\star)]\|_2\\
&\lesssim\beta^2\Upsilon_7^\star\big(\|\mathbf{U}-\mathbf{U}^\star\|_2^{1-\frac{q_1}{2}}+\|\mathbf{V}-\mathbf{V}^\star\|_2^{1-\frac{q_2}{2}}+\|\mathbf{U}-\mathbf{U}^\star\|_F\|\mathbf{V}^\star\|_F^{q_2}r^{\frac{1-q_2}{2}}+\|\mathbf{V}-\mathbf{V}^\star\|_F\|\mathbf{U}^\star\|_F^{q_1}r^{\frac{1-q_1}{2}}\big)\\
&\lesssim\beta^2\Upsilon_7^\star\sqrt{\Upsilon_2^\star}\big(\|\mathbf{U}-\mathbf{U}^\star\|_F^{1-\frac{q_1}{2}}+\|\mathbf{V}-\mathbf{V}^\star\|_F^{1-\frac{q_2}{2}}\big).
\end{aligned}
$$

We complete the proof.

H. Auxiliary Results

Lemma 19 (Lemma D.4 in Zhong et al. (2018))**.** *Let* $\mathbf{U}\in\mathbb{R}^{d\times r}$ *be a full-column rank matrix. Let* $g:\mathbb{R}^k\to[0,\infty)$*. Define* $\bar\kappa(\mathbf{U})=\prod_{p=1}^{r}\frac{\sigma_p(\mathbf{U})}{\sigma_r(\mathbf{U})}$*, then we have*

$$
\mathbb{E}_{\mathbf{x}\in\mathcal{N}(0,I_d)}g(\mathbf{U}^T\mathbf{x})\geq\frac{1}{\bar\kappa(\mathbf{U})}\cdot\mathbb{E}_{\mathbf{z}\sim\mathcal{N}(0,I_r)}g(\sigma_r(\mathbf{U})\mathbf{z}).
$$

Lemma 20 (Concentration of quadratic form and norm)**.** *Suppose* $\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n\overset{iid}{\sim}\mathcal{N}(0,I_d)$ *and* $\mathbf{U}\in\mathbb{R}^{d\times r}$*, then* $\forall t>0$

(a) $P\left(\left|\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i^T\mathbf{U}\mathbf{U}^T\mathbf{x}_i-\|\mathbf{U}\|_F^2\right|>t\right)\leq2\exp(-\frac{nt^2}{4\|\mathbf{U}\mathbf{U}^T\|_F^2+4\|\mathbf{U}\|_2^2t})$.

(b) $P\left(\max_{i\in[n]}|\mathbf{x}_i^T\mathbf{U}\mathbf{U}^T\mathbf{x}_i-\|\mathbf{U}\|_F^2|>t\right)\leq2n\exp(-\frac{t^2}{4\|\mathbf{U}\mathbf{U}^T\|_F^2+4\|\mathbf{U}\|_2^2t})$.

(c) $P\left(\left|\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i^T\mathbf{U}\mathbf{U}^T\mathbf{x}_i-\|\mathbf{U}\|_F^2\right|>5\sqrt{\frac{s\log d}{n}}\|\mathbf{U}\|_F^2\right)\leq\frac{2}{d^s}$, $\forall s>0$.

(d) $P\left(\max_{i\in[n]}\mathbf{x}_i^T\mathbf{U}\mathbf{U}^T\mathbf{x}_i>(\|\mathbf{U}\|_F+2\sqrt{s\log n}\|\mathbf{U}\|_2)^2\right)\leq\frac{1}{n^{s-1}}$, $\forall s>0$.

(e) $P(\mathbf{x}^T\mathbf{U}\mathbf{U}^T\mathbf{x}\geq6K\|\mathbf{U}\|_F^2)\leq\exp(-\frac{\|\mathbf{U}\|_F^2K}{\|\mathbf{U}\|_2^2})$, $\forall K\geq1$.

(f) $P(\max_{i\in[n]}\left|\|\mathbf{x}_i\|_2-\sqrt{d}\right|>t)\leq2n\exp(-t^2/2)$.

(g) $P(\max_{i\in[n]}\left||\mathbf{x}_i^T\boldsymbol{u}|-\sqrt{\frac{2}{\pi}}\|\boldsymbol{u}\|_2\right|>t)\leq2n\exp(-\frac{t^2}{4\|\boldsymbol{u}\|_2^2})$, $\forall\boldsymbol{u}\in\mathbb{R}^d$.

Proof. Result in (a) directly comes from the Chernoff bound and Remark 2.3 in Hsu et al. (2012). We use union bound and (a) to prove (b). (c), (d) and (e) are directly from (a) and (b). (f) is from the Chapter 3 in Vershynin (2018). (g) is due to the fact that $|\mathbf{x}^T\boldsymbol{u}|$ is sub-Gaussian variable. $\square$

Lemma 21 (Expectation of product of quadratic form)**.** *Suppose* $\mathbf{x}\sim\mathcal{N}(0,I_d)$*,* $\mathbf{U}\in\mathbb{R}^{d\times r}$*,* $\boldsymbol{a},\boldsymbol{b}\in\mathbb{R}^d$*, then*

(a) $\mathbb{E}[\mathbf{x}^T\mathbf{U}\mathbf{U}^T\mathbf{x}\cdot|\mathbf{x}^T\boldsymbol{a}|]\lesssim\|\mathbf{U}\|_F^2\|\boldsymbol{a}\|_2$.

(b) $\mathbb{E}[\mathbf{x}^T\mathbf{U}\mathbf{U}^T\mathbf{x}\cdot|\mathbf{x}^T\boldsymbol{a}\boldsymbol{b}^T\mathbf{x}|]\lesssim\|\mathbf{U}\|_F^2\|\boldsymbol{a}\|_2\|\boldsymbol{b}\|_2$.

(c) suppose $\mathbf{U}_i\in\mathbb{R}^{d\times r_i}$ *for* $i\in[4]$*,* $\mathbb{E}\big[\prod_{i=1}^4\mathbf{x}^T\mathbf{U}_i\mathbf{U}_i^T\mathbf{x}\big]\lesssim\prod_{i=1}^4\|\mathbf{U}_i\|_F^2$.

Proof. Note that

$$\begin{aligned}\mathbb{E}[\mathbf{x}^T\mathbf{U}\mathbf{U}^T\mathbf{x}\cdot|\mathbf{x}^T\boldsymbol{a}|]\leq&\sqrt{\mathbb{E}[(\mathbf{x}^T\mathbf{U}\mathbf{U}^T\mathbf{x})^2]}\sqrt{\mathbb{E}[\mathbf{x}^T\boldsymbol{a}\boldsymbol{a}^T\mathbf{x}]}\\ =&\sqrt{2\mathrm{Trace}(\mathbf{U}\mathbf{U}^T\mathbf{U}\mathbf{U}^T)+\mathrm{Trace}(\mathbf{U}\mathbf{U}^T)^2}\cdot\|\boldsymbol{a}\|\lesssim\|\mathbf{U}\|_F^2\|\boldsymbol{a}\|.\end{aligned}$$

This shows the part (a). (b) can be showed similarly using the Hölder's inequality twice. For (c),

$$\begin{aligned}&\mathbb{E}\big[\prod_{i=1}^4\mathbf{x}^T\mathbf{U}_i\mathbf{U}_i^T\mathbf{x}\big]\\ &\leq\prod_{i=1}^4\sqrt[4]{\mathbb{E}\big[(\mathbf{x}^T\mathbf{U}_i\mathbf{U}_i^T\mathbf{x})^4\big]}\\ &=\prod_{i=1}^4\sqrt[4]{\|\mathbf{U}_i\|_F^8+32\|\mathbf{U}_i\|_F^2\|\mathbf{U}_i\mathbf{U}_i^T\mathbf{U}_i\|_F^2+12\|\mathbf{U}_i\mathbf{U}_i^T\|_F^4+12\|\mathbf{U}_i\|_F^4\|\mathbf{U}_i\mathbf{U}_i^T\|_F^2+48\|\mathbf{U}_i\mathbf{U}_i^T\mathbf{U}_i\mathbf{U}_i^T\|_F^2}\\ &\lesssim\prod_{i=1}^4\|\mathbf{U}_i\|_F^2.\end{aligned}$$

Here the first inequality is due to the Hölder's inequality and the second equality is from Lemma 2.2 in Magnus (1978). $\square$

Lemma 22 (Extension of Lemma E.13 in Zhong et al. (2018))**.** *Let* $\mathcal{D}=\{(\mathbf{x},\mathbf{z})\}$ *be a sample set, and let* $\Omega=\{(\mathbf{x}_k,\mathbf{z}_k)\}_{k=1}^m$ *be a collection of samples of* $\mathcal{D}$*, where each* $(\mathbf{x}_k,\mathbf{z}_k)$ *is sampled with replacement from* $\mathcal{D}$ *uniformly. Independently, we have another sets* $\mathcal{D}'=\{(\mathbf{x}',\mathbf{z}')\}$ *and* $\Omega'=\{(\mathbf{x}_k',\mathbf{z}_k')\}_{k=1}^m$*. For any pair* $(\mathbf{x},\mathbf{z})$ *and* $(\mathbf{x}',\mathbf{z}')$*, we have a matrix* $\mathbf{A}\big((\mathbf{x},\mathbf{z}),(\mathbf{x}',\mathbf{z}')\big)\in\mathbb{R}^{d_1\times d_2}$*. Define* $\mathbf{H}=\frac{1}{m^2}\sum_{k=1}^m\sum_{l=1}^m\mathbf{A}\left((\mathbf{x}_k,\mathbf{z}_k),(\mathbf{x}_l',\mathbf{z}_l')\right)$*. If the following conditions hold with* ν_1,ν_2 *not depending on* $\mathcal{D}$*,* $\mathcal{D}'$*:*

(a) $\|\mathbf{A}\big((\mathbf{x},\mathbf{z}),(\mathbf{x}',\mathbf{z}')\big)\|_2\leq\nu_1$, $\forall(\mathbf{x},\mathbf{z})\in\mathcal{D},(\mathbf{x}',\mathbf{z}')\in\mathcal{D}'$,

(b)
$$\begin{aligned}&\left\|\frac{1}{|\mathcal{D}||\mathcal{D}'|}\sum_{(\mathbf{x},\mathbf{z})\in\mathcal{D}}\sum_{(\mathbf{x}',\mathbf{z}')\in\mathcal{D}'}\mathbf{A}\big((\mathbf{x},\mathbf{z}),(\mathbf{x}',\mathbf{z}')\big)\mathbf{A}\big((\mathbf{x},\mathbf{z}),(\mathbf{x}',\mathbf{z}')\big)^T\right\|_2\\ &\qquad\vee\left\|\frac{1}{|\mathcal{D}||\mathcal{D}'|}\sum_{(\mathbf{x},\mathbf{z})\in\mathcal{D}}\sum_{(\mathbf{x}',\mathbf{z}')\in\mathcal{D}'}\mathbf{A}\big((\mathbf{x},\mathbf{z}),(\mathbf{x}',\mathbf{z}')\big)^T\mathbf{A}\big((\mathbf{x},\mathbf{z}),(\mathbf{x}',\mathbf{z}')\big)\right\|_2\leq\nu_2,\end{aligned}$$

then $\forall t>0$*,*

$$P\left(\left\|\mathbf{H}-\frac{1}{|\mathcal{D}||\mathcal{D}'|}\sum_{(\mathbf{x},\mathbf{z})\in\mathcal{D}}\sum_{(\mathbf{x}',\mathbf{z}')\in\mathcal{D}'}\mathbf{A}\big((\mathbf{x},\mathbf{z}),(\mathbf{x}',\mathbf{z}')\big)\right\|_2\geq t\right)\leq(d_1+d_2)\exp(-\frac{mt^2}{4\nu_2+4\nu_1t}).$$

Proof. For any integer k, we define $\bar{k}$ to be the remainder of k/m such that $1\leq\bar{k}\leq m$ (i.e. $\bar{m}=m$). Then we can express $\mathbf{H}$ as

$$\mathbf{H}=\frac{1}{m}\sum_{k=0}^{m-1}\left(\frac{1}{m}\sum_{l=1}^m\mathbf{A}\left((\mathbf{x}_l,\mathbf{z}_l),(\mathbf{x}'_{\overline{l+k}},\mathbf{z}'_{\overline{l+k}})\right)\right)=:\frac{1}{m}\sum_{k=0}^{m-1}\mathbf{H}_k.$$

Note that $\mathbf{H}_k$ is the sum of m independent samples, and for any $k = 0, 1, ..., m-1$, they have the same distribution with conditional expectation

$$\mathbb{E}[\mathbf{H}_k \mid \mathcal{D}, \mathcal{D}'] = \frac{1}{|\mathcal{D}||\mathcal{D}'|} \sum_{(\mathbf{x},\mathbf{z})\in\mathcal{D}} \sum_{(\mathbf{x}',\mathbf{z}')\in\mathcal{D}'} \mathbf{A}\left((\mathbf{x},\mathbf{z}),(\mathbf{x}',\mathbf{z}')\right).$$

Therefore,

$$\begin{aligned} P(\|\mathbf{H}-\mathbb{E}[\mathbf{H}]\|_2 > t \mid \mathcal{D},\mathcal{D}') \leq & P(\frac{1}{m}\sum_{k=0}^{m-1} \|\mathbf{H}_k - \mathbb{E}[\mathbf{H}_k]\|_2 > t \mid \mathcal{D},\mathcal{D}') \\ \leq & \inf_{s>0} e^{-st}\mathbb{E}[\exp(\frac{s}{m}\sum_{k=0}^{m-1}\|\mathbf{H}_k - \mathbb{E}[\mathbf{H}_k]\|_2) \mid \mathcal{D},\mathcal{D}'] \\ \leq & \inf_{s>0} e^{-st}\frac{1}{m}\sum_{k=0}^{m-1}\mathbb{E}[\exp(s\|\mathbf{H}_k - \mathbb{E}[\mathbf{H}_k]\|_2) \mid \mathcal{D},\mathcal{D}'] \\ = & \inf_{s>0} e^{-st}\mathbb{E}[\exp(s\|\mathbf{H}_0 - \mathbb{E}[\mathbf{H}_0]\|_2) \mid \mathcal{D},\mathcal{D}']. \end{aligned}$$

By the proof of Corollary 6.1.2 in Tropp et al. (2015), the right hand side satisfies

$$\inf_{s>0} e^{-st}\mathbb{E}[\exp(s\|\mathbf{H}_0 - \mathbb{E}[\mathbf{H}_0]\|_2) \mid \mathcal{D},\mathcal{D}'] \leq (d_1+d_2)\exp(-\frac{mt^2}{4\nu_2 + 4\nu_1 t}).$$

Combining the above two displays and using the equality that $P(\mathcal{A}) = \mathbb{E}\left[\mathbf{1}_{\mathcal{A}}\right] = \mathbb{E}\left[\mathbb{E}\left[\mathbf{1}_{\mathcal{A}} \mid \mathcal{D},\mathcal{D}'\right]\right]$ for any event $\mathcal{A}$, we finish the proof. □

Lemma 23 (Extension of Lemma E.10 in Zhong et al. (2018))**.** *Let $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{z}_j) : i \in [n_1], j \in [n_2], (\mathbf{x}_i, \mathbf{z}_j) \sim \mathcal{F}\}$ be a sample set with size $n_1 n_2$ and each pair $(\mathbf{x}, \mathbf{z})$ follows the same distribution $\mathcal{F}$; similarly but independently, let $\mathcal{D}' = \{(\mathbf{x}'_i, \mathbf{z}'_j) : i \in [n_1], j \in [n_2], (\mathbf{x}'_i, \mathbf{z}'_j) \sim \mathcal{F}'\}$ be another sample set. Let $\mathbf{A}\left((\mathbf{x},\mathbf{z}),(\mathbf{x}',\mathbf{z}')\right) \in \mathbb{R}^{d_1\times d_2}$ be a random matrix corresponding to $(\mathbf{x},\mathbf{z}) \in \mathcal{D}$, $(\mathbf{x}',\mathbf{z}') \in \mathcal{D}'$, and let $\mathbf{H} = \frac{1}{n_1^2 n_2^2}\sum_{(\mathbf{x},\mathbf{z})\in\mathcal{D}}\sum_{(\mathbf{x}',\mathbf{z}')\in\mathcal{D}'}\mathbf{A}\left((\mathbf{x},\mathbf{z}),(\mathbf{x}',\mathbf{z}')\right)$. Suppose the following conditions hold with parameters $\mu_1, \nu_1, \nu_2, \nu_3$ (when $\mathbf{A}$ is symmetric, one can let $\boldsymbol{v} = \boldsymbol{u}$ in condition (c)),*

(a) $P\left(\|\mathbf{A}\left((\mathbf{x},\mathbf{z}),(\mathbf{x}',\mathbf{z}')\right)\|_2 \geq \mu_1\right) \leq \nu_1$,

(b) $\left\|\mathbb{E}[\mathbf{A}\big((\mathbf{x},\mathbf{z}),(\mathbf{x}',\mathbf{z}')\big)\mathbf{A}\big((\mathbf{x},\mathbf{z}),(\mathbf{x}',\mathbf{z}')\big)^T]\right\|_2 \vee \left\|\mathbb{E}[\mathbf{A}\big((\mathbf{x},\mathbf{z}),(\mathbf{x}',\mathbf{z}')\big)^T\mathbf{A}\big((\mathbf{x},\mathbf{z}),(\mathbf{x}',\mathbf{z}')\big)]\right\|_2 \leq \nu_2$,

(c) $\max_{\|\boldsymbol{u}\|_2=\|\boldsymbol{v}\|_2=1}\left(\mathbb{E}\left[\left(\boldsymbol{u}^T\mathbf{A}\big((\mathbf{x},\mathbf{z}),(\mathbf{x}',\mathbf{z}')\big)\boldsymbol{v}\right)^2\right]\right)^{1/2} \leq \nu_3$,

then $\forall t > 0$

$$P\left(\|\mathbf{H}-\mathbb{E}[\mathbf{H}]\|_2 > t + \nu_3\sqrt{\nu_1}\right) \leq n_1^2 n_2^2 \nu_1 + (d_1+d_2)\exp\left(-\frac{(n_1\wedge n_2)t^2}{\left(2\nu_2 + 4\|\mathbb{E}[\mathbf{H}]\|_2^2 + 4\nu_3^2\nu_1\right) + 4\mu_1 t}\right).$$

Proof. We suppress the evaluation point of $\mathbf{A}$ for simplicity. Let $\bar{\mathbf{A}} = \mathbf{A}\cdot\mathbf{1}_{\|\mathbf{A}\|_2\leq\mu_1}$ and $\bar{\mathbf{H}} = \frac{1}{n_1^2 n_2^2}\sum_{(\mathbf{x},\mathbf{z})\in\mathcal{D}}\sum_{(\mathbf{x}',\mathbf{z}')\in\mathcal{D}'}\bar{\mathbf{A}}$. Then,

$$\|\mathbf{H}-\mathbb{E}[\mathbf{H}]\|_2 \leq \|\mathbf{H}-\bar{\mathbf{H}}\|_2 + \|\bar{\mathbf{H}}-\mathbb{E}[\bar{\mathbf{H}}]\|_2 + \|\mathbb{E}[\bar{\mathbf{H}}]-\mathbb{E}[\mathbf{H}]\|_2.$$

For the first term,

$$P(\|\mathbf{H}-\bar{\mathbf{H}}\|_2 = 0) \geq P(\mathbf{A} = \bar{\mathbf{A}}, \forall(\mathbf{x},\mathbf{z})\in\mathcal{D}, (\mathbf{x}',\mathbf{z}')\in\mathcal{D}') \geq 1 - n_1^2 n_2^2 \nu_1.$$

For the third term,

$$\|\mathbb{E}[\bar{\mathbf{H}}]-\mathbb{E}[\mathbf{H}]\|_2 = \|\mathbb{E}[\mathbf{A}\cdot\mathbf{1}_{\|\mathbf{A}\|_2>\mu_1}]\|_2 = \max_{\|\boldsymbol{u}\|_2=\|\boldsymbol{v}\|_2=1}\mathbb{E}\left[\boldsymbol{u}^T\mathbf{A}\boldsymbol{v}\cdot\mathbf{1}_{\|\mathbf{A}\|_2>\mu_1}\right]$$

$$\leq \max_{\|\boldsymbol{u}\|_2=\|\boldsymbol{v}\|_2=1} \sqrt{\mathbb{E}[(\boldsymbol{u}^T\mathbf{A}\boldsymbol{v})^2]}\sqrt{P(\|\mathbf{A}\|_2 > \mu_1)} \leq \nu_3\sqrt{\nu_1}.$$

For the second term, without loss of generality, we assume $n_1 \leq n_2$. For any integer k, we let $k = s_1 n_1 + \bar{k}$ where integer $s_1 \geq 0$ and remainder $\bar{k}$ satisfies $1 \leq \bar{k} \leq n_1$. Also, we let $k = s_2 n_2 + \tilde{k}$ where integer $s_2 \geq 0$ and $\tilde{k}$ satisfies $1 \leq \tilde{k} \leq n_2$. Then we can express $\bar{\mathbf{H}}$ as

$$\bar{\mathbf{H}} = \frac{1}{n_2^2}\sum_{k=0}^{n_2-1}\sum_{l=0}^{n_2-1}\frac{1}{n_1}\sum_{j=0}^{n_1-1}\Bigg(\underbrace{\frac{1}{n_1}\sum_{i=1}^{n_1}\bar{\mathbf{A}}\big((\mathbf{x}_i,\mathbf{z}_{\widetilde{i+k}}),(\mathbf{x}'_{\overline{i+j}},\mathbf{z}'_{\widetilde{\overline{i+j}+l}})\big)}_{\bar{\mathbf{H}}_{k,l,j}}\Bigg).$$

Based on this decomposition, we see $\bar{\mathbf{H}}_{k,l,j}$ is a sum of n_1 $i.i.d$ random matrices, and also $\{\bar{\mathbf{H}}_{k,l,j}\}$ have the same distribution. Similar to the proof of Lemma 22, we have

$$P\left(\|\bar{\mathbf{H}} - \mathbb{E}[\bar{\mathbf{H}}]\|_2 > t\right) \leq \inf_{s>0} e^{-st}\mathbb{E}[\exp(s\|\bar{\mathbf{H}}_{0,0,0} - \mathbb{E}[\bar{\mathbf{H}}_{0,0,0}]\|_2)].$$

We apply Corollary 6.1.2 in Tropp et al. (2015). Note that $\|\bar{\mathbf{A}} - \mathbb{E}[\bar{\mathbf{A}}]\|_2 \leq 2\mu_1$ and

$$\begin{aligned}\|\mathbb{E}[\bar{\mathbf{A}}\bar{\mathbf{A}}^T] - \mathbb{E}[\bar{\mathbf{A}}]\mathbb{E}[\bar{\mathbf{A}}^T]\|_2 \leq& \|\mathbb{E}[\mathbf{A}\mathbf{A}^T]\|_2 + \|\mathbb{E}[\bar{\mathbf{A}}]\|_2^2 \leq \nu_2 + (\|\mathbb{E}[\mathbf{H}]\|_2 + \nu_3\sqrt{\nu_1})^2 \\ \leq& \nu_2 + 2\|\mathbb{E}[\mathbf{H}]\|_2^2 + 2\nu_3^2\nu_1.\end{aligned}$$

We also have similar bound for $\|\mathbb{E}[\bar{\mathbf{A}}^T\bar{\mathbf{A}}] - \mathbb{E}[\bar{\mathbf{A}}^T]\mathbb{E}[\bar{\mathbf{A}}]\|_2$. Thus, we have

$$\inf_{s>0} e^{-st}\mathbb{E}[\exp(s\|\bar{\mathbf{H}}_{k,l,j} - \mathbb{E}[\bar{\mathbf{H}}_{k,l,j}]\|_2)] \leq (d_1+d_2)\exp\left(-\frac{n_1 t^2}{\left(2\nu_2 + 4\|\mathbb{E}[\mathbf{H}]\|_2^2 + 4\nu_3^2\nu_1\right) + 4\mu_1 t}\right).$$

Putting everything together finishes the proof. □

Lemma 24. *Let $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{z}_j) : i \in [n_1], j \in [n_2], (\mathbf{x}_i, \mathbf{z}_j) \sim \mathcal{F}\}$. Let $\mathbf{A}(\mathbf{x},\mathbf{z}) \in \mathbb{R}^{d_1\times d_2}$ be a random matrix corresponding to $(\mathbf{x},\mathbf{z}) \in \mathcal{D}$, and let $\mathbf{H} = \frac{1}{n_1 n_2}\sum_{(\mathbf{x},\mathbf{z})\in\mathcal{D}} \mathbf{A}(\mathbf{x},\mathbf{z})$. Suppose the following conditions hold with parameters $\mu_1, \nu_1, \nu_2, \nu_3$,*

(a) $P\left(\|\mathbf{A}(\mathbf{x},\mathbf{z})\|_2 \geq \mu_1\right) \leq \nu_1$,

(b) $\left\|\mathbb{E}\left[\mathbf{A}(\mathbf{x},\mathbf{z})\mathbf{A}(\mathbf{x},\mathbf{z})^T\right]\right\|_2 \vee \left\|\mathbb{E}\left[\mathbf{A}(\mathbf{x},\mathbf{z})^T\mathbf{A}(\mathbf{x},\mathbf{z})\right]\right\|_2 \leq \nu_2$,

(c) $\max_{\|\boldsymbol{u}\|_2=\|\boldsymbol{v}\|_2=1}\left(\mathbb{E}\left[\left(\boldsymbol{u}^T\mathbf{A}(\mathbf{x},\mathbf{z})\boldsymbol{v}\right)^2\right]\right)^{1/2} \leq \nu_3$,

then $\forall t > 0$

$$P\left(\|\mathbf{H} - \mathbb{E}[\mathbf{H}]\|_2 > t + \nu_3\sqrt{\nu_1}\right) \leq n_1 n_2 \nu_1 + (d_1+d_2)\exp\left(-\frac{(n_1\wedge n_2)t^2}{\left(2\nu_2 + 4\|\mathbb{E}[\mathbf{H}]\|_2^2 + 4\nu_3^2\nu_1\right) + 4\mu_1 t}\right).$$

Proof. The result is directly from Lemma 23. □

Lemma 25. *Suppose $\mathbf{x} \sim \mathcal{N}(0, I_d)$, $\phi \in \{sigmoid, tanh, ReLU\}$. For any vectors $\boldsymbol{u}, \boldsymbol{u}^\star, \boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^d$,*

$$\mathbb{E}[|\phi'(\boldsymbol{u}^T\mathbf{x}) - \phi'(\boldsymbol{u}^{\star T}\mathbf{x})| \cdot |\mathbf{x}^T\boldsymbol{a}\boldsymbol{b}^T\mathbf{x}|] \leq \left(\sqrt{\frac{\|\boldsymbol{u}-\boldsymbol{u}^\star\|_2}{\|\boldsymbol{u}^\star\|_2}}\right)^q \|\boldsymbol{u}-\boldsymbol{u}^\star\|_2^{1-q}\|\boldsymbol{a}\|_2\|\boldsymbol{b}\|_2,$$

where $q = 1$ if ϕ is ReLU and $q = 0$ otherwise.

Proof. By Hölder's inequality,

$$\mathbb{E}[|\phi'(\boldsymbol{u}^T\mathbf{x}) - \phi'(\boldsymbol{u}^{\star T}\mathbf{x})| \cdot |\mathbf{x}^T\boldsymbol{a}\boldsymbol{b}^T\mathbf{x}|] \leq \sqrt{\mathbb{E}[(\phi'(\boldsymbol{u}^T\mathbf{x}) - \phi'(\boldsymbol{u}^{\star T}\mathbf{x}))^2\mathbf{x}^T\boldsymbol{a}\boldsymbol{a}^T\mathbf{x}]}\sqrt{\mathbb{E}[\mathbf{x}^T\boldsymbol{b}\boldsymbol{b}^T\mathbf{x}]}.$$

If $\phi \in \{\text{sigmoid}, \tanh\}$, we finish the proof by using the Lipschitz continuity of ϕ' and Lemma 21. If ϕ is ReLU, we apply Lemma E.17 in Zhong et al. (2018) to complete the proof. □