Supplementary files for "On strong convergence of the two-tower model for recommender system"

Anonymous authors
Paper under double-blind review

Proof of Theorem 1. Note that $R^*(\boldsymbol{x}_u, \tilde{\boldsymbol{x}}_i) = \langle \boldsymbol{f}^*(\boldsymbol{x}_u), \tilde{\boldsymbol{f}}^*(\tilde{\boldsymbol{x}}_i)\rangle$ with $f_j^* \in \mathcal{H}(\beta, [0,1]^{D_u}, M)$ and $\tilde{f}_j^* \in \mathcal{H}(\beta, [0,1]^{D_i}, M)$. It follows from Theorem 5 in Nakada and Imaizumi (2020) that there exist $\mathcal{F}_{D_u}(W, L, B, M)$ and $\mathcal{F}_{D_i}(\tilde{W}, \tilde{L}, \tilde{B}, M)$ with $W = O(\epsilon^{-d_u/\beta}), \tilde{W} = O(\epsilon^{-d_i/\beta}), B = O(\epsilon^{-s})$ and $\tilde{B} = O(\epsilon^{-s})$ such that for each j, we have

$$\begin{aligned} \inf_{f_j \in \mathcal{F}_{D_u}(W,L,B,M)} \|f_j - f_j^*\|_{L^\infty(\mu_u)} \le \epsilon, \\ \inf_{\tilde{f}_j \in \mathcal{F}_{D_i}(\tilde{W},\tilde{L},\tilde{B},M)} \|\tilde{f}_j - \tilde{f}_j^*\|_{L^\infty(\mu_i)} \le \epsilon. \end{aligned} \tag{1}$$

By the triangle inequality and the Cauchy-Schwartz inequality, we have that

$$\begin{aligned} &|R(\boldsymbol{x}_u, \tilde{\boldsymbol{x}}_i) - R^*(\boldsymbol{x}_u, \tilde{\boldsymbol{x}}_i)| = |\langle \boldsymbol{f}(\boldsymbol{x}_u), \tilde{\boldsymbol{f}}(\tilde{\boldsymbol{x}}_i)\rangle - \langle \boldsymbol{f}^*(\boldsymbol{x}_u), \tilde{\boldsymbol{f}}^*(\tilde{\boldsymbol{x}}_i)\rangle| \\ \le& |\langle \boldsymbol{f}(\boldsymbol{x}_u), \tilde{\boldsymbol{f}}(\tilde{\boldsymbol{x}}_i) - \tilde{\boldsymbol{f}}^*(\tilde{\boldsymbol{x}}_i)\rangle| + |\langle \boldsymbol{f}(\boldsymbol{x}_u) - \boldsymbol{f}^*(\boldsymbol{x}_u), \tilde{\boldsymbol{f}}^*(\tilde{\boldsymbol{x}}_i)\rangle| \\ \le& \|\boldsymbol{f}(\boldsymbol{x})\|_2 \|\tilde{\boldsymbol{f}}(\tilde{\boldsymbol{x}}_i) - \tilde{\boldsymbol{f}}^*(\tilde{\boldsymbol{x}}_i)\|_2 + \|\boldsymbol{f}(\boldsymbol{x}_u) - \boldsymbol{f}^*(\boldsymbol{x}_u)\|_2 \|\tilde{\boldsymbol{f}}^*(\tilde{\boldsymbol{x}}_i)\|_2. \end{aligned}$$

Since $\boldsymbol{f} \in \mathcal{F}_{D_u}(W, L, B, M)$ and $\tilde{\boldsymbol{f}}^* \in \mathcal{H}^p(\beta, [0,1]^{D_i}, M)$, we have $\|\boldsymbol{f}(\boldsymbol{x}_u)\|_2 \le 2\sqrt{p}M$ and $\|\tilde{\boldsymbol{f}}^*(\tilde{\boldsymbol{x}}_i)\|_2 \le \sqrt{p}M$, which further implies that

$$|R(\boldsymbol{x}_u, \tilde{\boldsymbol{x}}_i) - R^*(\boldsymbol{x}_u, \tilde{\boldsymbol{x}}_i)| \le M\Big(2\sum_{j=1}^p \|\tilde{f}_j - \tilde{f}_j^*\|_{L^\infty(\mu_i)} + \sum_{j=1}^p \|f_j - f_j^*\|_{L^\infty(\mu_u)}\Big).$$

Let $\Phi = (W, L, B, M, \tilde{W}, \tilde{L}, \tilde{B})$, it then follows from (1) that

$$\begin{aligned} &\inf_{\mathbb{R} \in \mathcal{R}^\Phi} |R(\boldsymbol{x}_u, \tilde{\boldsymbol{x}}_i) - R^*(\boldsymbol{x}_u, \tilde{\boldsymbol{x}}_i)| = \inf_{f_j \in \mathcal{F}_{D_u}(W,L,B,M), \tilde{f}_j \in \mathcal{F}_{D_i}(\tilde{W},\tilde{L},\tilde{B},M)} |R(\boldsymbol{x}_u, \tilde{\boldsymbol{x}}_i) - R^*(\boldsymbol{x}_u, \tilde{\boldsymbol{x}}_i)| \\ \le& M\Big(2\sum_{j=1}^p \inf_{f_j \in \mathcal{F}_{D_u}(W,L,B,M)} \|\tilde{f}_j - \tilde{f}_j^*\|_{L^\infty(\mu_i)} + \sum_{j=1}^p \inf_{\tilde{f}_j \in \mathcal{F}_{D_i}(\tilde{W},\tilde{L},\tilde{B},M)} \|f_j - f_j^*\|_{L^\infty(\mu_u)}\Big) \\ \le& 3pM\epsilon. \end{aligned}$$

This completes the proof of Theorem 1. ■

Proof of Lemma 1. For $\boldsymbol{f}(\boldsymbol{x}) \in \mathcal{F}_D(W, L, B, M)$ with $U(\boldsymbol{f}) \le L$, we let $\boldsymbol{y}_l$ denote the output of the l-th layer of $\boldsymbol{f}$ and $\Theta = ((\boldsymbol{A}_1, \boldsymbol{b}_1), (\boldsymbol{A}_2, \boldsymbol{b}_2), \dots, (\boldsymbol{A}_{U(\boldsymbol{f})}, \boldsymbol{b}_{U(\boldsymbol{f})}))$ the parameter of $\boldsymbol{f}$, where $\boldsymbol{A}_l \in [-B, B]^{p_l \times p_{l-1}}$, $\boldsymbol{b}_l \in [-B, B]^{p_l}$, $p_0 = D$ and $p_{U(\boldsymbol{f})} = p$. We then construct $\boldsymbol{f}' = Q(\boldsymbol{f})$ with $\Theta' = ((\boldsymbol{A}_1', \boldsymbol{b}_1'), (\boldsymbol{A}_2', \boldsymbol{b}_2'), \dots, (\boldsymbol{A}_L', \boldsymbol{b}_L'))$ as follows.

For $l = 1$, we let $\boldsymbol{A}_1' = (\boldsymbol{A}_1^T, \boldsymbol{0}_{D\times(2W-p_1)})^T$ and $\boldsymbol{b}_1' = (\boldsymbol{b}_1^T, \boldsymbol{0}_{2W-p_1}^T)^T$, and then the output of the first layer $\boldsymbol{y}_1'$ is given by

$$\boldsymbol{y}_1' = \sigma(\boldsymbol{A}_1'\boldsymbol{x} + \boldsymbol{b}_1') = \begin{pmatrix} \sigma(\boldsymbol{A}_1\boldsymbol{x} + \boldsymbol{b}_1) \\ \boldsymbol{0}_{2W-p_1} \end{pmatrix} = \begin{pmatrix} \boldsymbol{y}_1 \\ \boldsymbol{0}_{2W-p_1} \end{pmatrix},$$

where $\sigma(\cdot)$ is the element-wise ReLU function. For $l = 2, \dots, U(\boldsymbol{f}) - 1$, we let $\boldsymbol{A}_l' = \mathrm{diag}(\boldsymbol{A}_l, \boldsymbol{0}_{(2W-p_l)\times(2W-p_{l-1})})$ and $\boldsymbol{b}_l' = (\boldsymbol{b}_l^T, \boldsymbol{0}_{2W-p_l}^T)^T$, and then

$$\boldsymbol{y}_l' = \sigma(\boldsymbol{A}_l'\boldsymbol{y}_{l-1} + \boldsymbol{b}_l') = \begin{pmatrix} \sigma(\boldsymbol{A}_l\boldsymbol{y}_{l-1} + \boldsymbol{b}_l) \\ \boldsymbol{0}_{2W-p_l} \end{pmatrix} = \begin{pmatrix} \boldsymbol{y}_l \\ \boldsymbol{0}_{2W-p_l} \end{pmatrix}.$$

The remaining $(\boldsymbol{A}_l', \boldsymbol{b}_l')$'s for $l = U(\boldsymbol{f}), \ldots, L$ are constructed based on the value of $U(\boldsymbol{f})$. If $U(\boldsymbol{f}) = L$, as the last layer of $\boldsymbol{f}$ and $\boldsymbol{f}'$ are both linear, we set $\boldsymbol{A}_L' = (\boldsymbol{A}_L, \boldsymbol{0}_{p\times(2W-p_{L-1})})$ and $\boldsymbol{b}_L' = \boldsymbol{b}_L$, and then

$$\boldsymbol{y}_L' = \boldsymbol{A}_L'\boldsymbol{y}_{L-1}' + \boldsymbol{b}_L' = \boldsymbol{A}_L\boldsymbol{y}_{L-1} + \boldsymbol{b}_L = \boldsymbol{y}_{U(\boldsymbol{f})}.$$

If $U(\boldsymbol{f}) = L-1$, we set

$$\boldsymbol{A}_{L-1}' = \begin{pmatrix} \boldsymbol{A}_{L-1} & \boldsymbol{0}_{p_{L-1}\times(2W-p_{L-2})} \\ -\boldsymbol{A}_{L-1} & \boldsymbol{0}_{p_{L-1}\times(2W-p_{L-2})} \\ \boldsymbol{0}_{(2W-2p_{L-1})\times p_{L-2}} & \boldsymbol{0}_{(2W-2p_{L-1})\times(2W-2p_{L-2})} \end{pmatrix}, \boldsymbol{b}_{L-1}' = \begin{pmatrix} \boldsymbol{b}_{L-1} \\ -\boldsymbol{b}_{L-1} \\ \boldsymbol{0}_{2W-2p_{L-1}} \end{pmatrix}.$$

Then we have

$$\boldsymbol{y}_{L-1}' = \sigma(\boldsymbol{A}_{L-1}'\boldsymbol{y}_{L-2}' + \boldsymbol{b}_{L-1}') = \begin{pmatrix} \sigma(\boldsymbol{A}_{L-1}\boldsymbol{y}_{L-2} + \boldsymbol{b}_{L-1}) \\ \sigma(-\boldsymbol{A}_{L-1}\boldsymbol{y}_{L-2} - \boldsymbol{b}_{L-1}) \\ \boldsymbol{0}_{2W-2p} \end{pmatrix}.$$

We further let $\boldsymbol{A}_L' = \big(\boldsymbol{I}_p, -\boldsymbol{I}_p, \boldsymbol{0}_{p\times(2W-2p)}\big)$ and $\boldsymbol{b}_L = \boldsymbol{0}_p$, and then

$$\boldsymbol{y}_L' = \sigma(\boldsymbol{A}_{L-1}\boldsymbol{y}_{L-2} + \boldsymbol{b}_{L-1}) - \sigma(-\boldsymbol{A}_{L-1}\boldsymbol{y}_{L-2} - \boldsymbol{b}_{L-1}) = \boldsymbol{y}_{U(\boldsymbol{f})},$$

where the second equality follows from property of the ReLU function that $\sigma(x) - \sigma(-x) = x$.

If $U(\boldsymbol{f}) \leq L-2$, we first construct $(\boldsymbol{A}_l', \boldsymbol{b}_l'); l = U(\boldsymbol{f})+1, \ldots, L-1$ as

$$\boldsymbol{A}_l' = \begin{pmatrix} \boldsymbol{I}_p & -\boldsymbol{I}_p & \boldsymbol{0}_{p\times(2W-2p)} \\ -\boldsymbol{I}_p & \boldsymbol{I}_p & \boldsymbol{0}_{p\times(2W-2p)} \\ \boldsymbol{0}_{(2W-2p)\times p} & \boldsymbol{0}_{(2W-2p)\times p} & \boldsymbol{0}_{(2W-2p)\times(2W-2p)} \end{pmatrix} \text{ and } \boldsymbol{b}_l' = \boldsymbol{0}_{2W}.$$

Then we have

$$\boldsymbol{y}_l' = \sigma(\boldsymbol{A}_l'\boldsymbol{y}_{l-1}' + \boldsymbol{b}_l') = \begin{pmatrix} \sigma(\boldsymbol{A}_{U(\boldsymbol{f})}\boldsymbol{y}_{U(\boldsymbol{f})-1} + \boldsymbol{b}_{U(\boldsymbol{f})}) \\ \sigma(-\boldsymbol{A}_{U(\boldsymbol{f})}\boldsymbol{y}_{U(\boldsymbol{f})-1} - \boldsymbol{b}_{U(\boldsymbol{f})}) \\ \boldsymbol{0}_{2W-2p} \end{pmatrix}. \tag{2}$$

We further set $\boldsymbol{A}_L' = \big(\boldsymbol{I}_p, -\boldsymbol{I}_p, \boldsymbol{0}_{p\times(2W-2p)}\big)$ and $\boldsymbol{b}_L = \boldsymbol{0}_p$, then we have

$$\boldsymbol{y}_L' = \sigma(\boldsymbol{A}_{U(\boldsymbol{f})}\boldsymbol{y}_{U(\boldsymbol{f})-1} + \boldsymbol{b}_{U(\boldsymbol{f})}) - \sigma(-\boldsymbol{A}_{U(\boldsymbol{f})}\boldsymbol{y}_{U(\boldsymbol{f})-1} - \boldsymbol{b}_{U(\boldsymbol{f})}) = \boldsymbol{y}_{U(\boldsymbol{f})}.$$

By the definition of $\mathcal{F}_D(W, L, B, M)$, the non-zero elements of $\boldsymbol{A}_l$ is at most W, and hence the number of non-zero elements in $\boldsymbol{A}_l'$ is at most

$$4W + \sum_{s=1}^{2W}\big(\lfloor\frac{2W}{s}\rfloor + 1\big) \leq 8W + \sum_{s=2}^{2W}(\frac{2W}{s} \times 1) \leq 8W + \int_1^{2W}\frac{2W}{x}dx \leq 12W\log W,$$

where $\lfloor\cdot\rfloor$ is the floor function. Similarly, the number of non-zero elements in $\boldsymbol{b}_l'$ is less than $2W\log W$. The desired result then follows immediately. ■

Proof of Lemma 2: For an L-layer neural network $\boldsymbol{f}(\boldsymbol{x};\Theta) \in \mathcal{K}_D(W, L, B, M)$, its l-th layer can be formulated as

$$\boldsymbol{h}_l(\boldsymbol{x}) = \big(h_{l1}(\boldsymbol{x}), h_{l2}(\boldsymbol{x}), \ldots, h_{lp_l}(\boldsymbol{x})\big) = \boldsymbol{A}_l\boldsymbol{x} + b_l,$$

where $h_{li}(\boldsymbol{x}) = \sum_{j=1}^{p_{l-1}} A_{lij}x_j + b_{li}$, with $p_0 = D$ and $p_{l-1} = 2W$ for $2 \leq l \leq L$. It follows from the triangle inequality that

$$\begin{aligned} &\sup_{\|\boldsymbol{x}\|_\infty\leq 1}\|\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{f}'(\boldsymbol{x})\|_2 = \sup_{\|\boldsymbol{x}\|_\infty\leq 1}\|\boldsymbol{h}_L \circ \boldsymbol{h}_{L-1} \circ \cdots \circ \boldsymbol{h}_1(\boldsymbol{x}) - \boldsymbol{h}_L' \circ \boldsymbol{h}_{L-1}' \circ \cdots \circ \boldsymbol{h}_1'(\boldsymbol{x})\|_2 \\ \leq &\sup_{\|\boldsymbol{x}\|_\infty\leq 1}\|\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{g}_{L-1}(\boldsymbol{x}) + \boldsymbol{g}_{L-1}(\boldsymbol{x}) - \boldsymbol{g}_{L-2}(\boldsymbol{x}) + \cdots + \boldsymbol{g}_1(\boldsymbol{x}) - \boldsymbol{f}'(\boldsymbol{x})\|_2 \\ \leq &\sup_{\|\boldsymbol{x}\|_\infty\leq 1}\|\boldsymbol{g}_L(\boldsymbol{x}) - \boldsymbol{g}_{L-1}(\boldsymbol{x})\|_2 + \ldots + \sup_{\|\boldsymbol{x}\|_\infty\leq 1}\|\boldsymbol{g}_1(\boldsymbol{x}) - \boldsymbol{g}_0(\boldsymbol{x})\|_2, \end{aligned} \tag{3}$$

where $\boldsymbol{g}_l(\boldsymbol{x}) = \boldsymbol{h}'_L \circ \cdots \circ \boldsymbol{h}'_{l+1} \circ \boldsymbol{h}_l \circ \cdots \circ \boldsymbol{h}_1(\boldsymbol{x})$. It then suffices to bound $\sup_{\|\boldsymbol{x}\|_\infty \le 1} \|\boldsymbol{g}_l(\boldsymbol{x}) - \boldsymbol{g}_{l-1}(\boldsymbol{x})\|_2$ for $l = 1, \ldots, L$ separately.

Before we proceed, we first bound

$$\sup_{\|\boldsymbol{x}\|_\infty \le 1} \|\boldsymbol{h}_l \circ \cdots \circ \boldsymbol{h}_1(\boldsymbol{x})\|_\infty \le (WB)^l (1 + \frac{B}{WB-1}) - \frac{B}{WB-1} \triangleq E_l, \tag{4}$$

for any $l \ge 1$ by mathematical induction. When $l = 1$, note that the ReLU function is a Lipschitz-1 function, then we have

$$\sup_{\|\boldsymbol{x}\|_\infty \le 1} |h_{1i}(\boldsymbol{x})| \le \sup_{\|\boldsymbol{x}\|_\infty \le 1} \sum_{j=1}^{D} |A_{lij}| \cdot |x_j| + |b_{li}| \le WB + B = E_1,$$

for $i = 1, \ldots, p_1$. It then follows that $\sup_{\|\boldsymbol{x}\|_\infty \le 1} \|\boldsymbol{h}_1(\boldsymbol{x})\|_\infty \le E_1$. Next, suppose that equation 4 holds true for $l \le k-1$, then

$$\begin{aligned}\sup_{\|\boldsymbol{x}\|_\infty \le 1} |h_{ki} \circ \cdots \circ \boldsymbol{h}_1(\boldsymbol{x})| \le \sup_{\|\boldsymbol{x}\|_\infty \le E_{k-1}} |h_{ki}(\boldsymbol{x})| \le \sup_{\|\boldsymbol{x}\|_\infty \le E_{k-1}} \sum_{j=1}^{p_{k-1}} |A_{kij}| \cdot |x_j| + |b_{li}| \\ \le WBE_{k-1} + B = (WB)^k (1 + \frac{B}{WB-1}) - \frac{B}{WB-1} = E_k,\end{aligned}$$

for $i = 1, \ldots, p_k$. It then follows that $\sup_{\|\boldsymbol{x}\|_\infty \le 1} \|\boldsymbol{h}_k \circ \cdots \boldsymbol{h}_1(\boldsymbol{x})\|_\infty \le E_k$, and thus equation 4 holds true for any $l \ge 1$.

We now turn to bound $\sup_{\|\boldsymbol{x}\|_\infty \le 1} \|\boldsymbol{g}_l(\boldsymbol{x}) - \boldsymbol{g}_{l-1}(\boldsymbol{x})\|_2$. Note that

$$\begin{aligned}&\sup_{\|\boldsymbol{x}\|_\infty \le 1} \|\boldsymbol{g}_l(\boldsymbol{x}) - \boldsymbol{g}_{l-1}(\boldsymbol{x})\|_2 \le \sum_{i=1}^{p} \sup_{\|\boldsymbol{x}\|_\infty \le 1} \left| g_{li}(\boldsymbol{x}) - g_{l-1,i}(\boldsymbol{x}) \right| \\ =& \sum_{i=1}^{p} \sup_{\|\boldsymbol{x}\|_\infty \le 1} \left| h'_{Li} \circ \cdots \circ \boldsymbol{h}'_{l+1} \circ \boldsymbol{h}_l \circ \cdots \circ \boldsymbol{h}_1(\boldsymbol{x}) - h'_{Li} \circ \cdots \circ \boldsymbol{h}'_l \circ \boldsymbol{h}_{l-1} \circ \cdots \circ \boldsymbol{h}_1(\boldsymbol{x}) \right| \\ \le& \sum_{i=1}^{p} \sup_{\|\boldsymbol{x}\|_\infty \le E_{l-1}} \left| h'_{Li} \circ \cdots \circ \boldsymbol{h}'_{l+1} \circ \boldsymbol{h}_l(\boldsymbol{x}) - h'_{Li} \circ \cdots \circ \boldsymbol{h}'_{l+1} \circ \boldsymbol{h}'_l(\boldsymbol{x}) \right| \\ \le& \sum_{i=1}^{p} \sup_{\|\boldsymbol{x} - \boldsymbol{x}'\|_\infty \le \epsilon(WE_{l-1}+1)} \left| h'_{Li} \circ \cdots \circ \boldsymbol{h}'_{l+1}(\boldsymbol{x}) - h'_{Li} \circ \cdots \circ \boldsymbol{h}'_{l+1}(\boldsymbol{x}') \right| \\ \le& p\epsilon (WB)^{L-l} (WE_{l-1} + 1),\end{aligned}$$

where $\boldsymbol{g} = (g_{l1}, \ldots, g_{lp})$, the second inequality follows from the fact that

$$\sup_{\|\boldsymbol{x}\|_\infty \le E_{l-1}} |h_{li}(\boldsymbol{x}) - h'_{li}(\boldsymbol{x})| \le \sup_{\|\boldsymbol{x}\|_\infty \le E_{l-1}} \sum_{j=1}^{p_{l-1}} |A_{lij} - A'_{lij}| \cdot |x_j| + |b_{li} - b'_{li}| \le \epsilon(WE_{l-1} + 1),$$

and the last inequality is derived by repeatedly using the fact that $\sup_{\|\boldsymbol{x} - \boldsymbol{x}'\|_\infty \le E} |h_{li}(\boldsymbol{x}) - h_{li}(\boldsymbol{x}')| \le WBE$ for any $E \ge 0$ and $l \ge 1$.

Therefore, after plugging the definition of E_l in equation 4, we have

$$\begin{aligned}\sup_{\|\boldsymbol{x}\| \le 1} \|\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{f}'(\boldsymbol{x})\|_2 \le& \sum_{l=1}^{L} \sup_{\|\boldsymbol{x}\| \le 1} \|\boldsymbol{g}_l(\boldsymbol{x}) - \boldsymbol{g}_{l-1}(\boldsymbol{x})\|_2 \\ \le& \sum_{l=1}^{L} p\epsilon \Big((WB)^L (\frac{1}{B} + \frac{1}{WB-1}) - \frac{(WB)^{L-l}}{WB-1} \Big) \\ =& p\epsilon \Big((WB)^L (\frac{L}{B} + \frac{L}{WB-1}) - \frac{(WB)^L - 1}{(WB-1)^2} \Big).\end{aligned}$$

This completes the proof of Lemma 2. ■

Proof of Lemma 3: For any $R \in \mathcal{R}^{\Phi}$, we have $R(\boldsymbol{x}_u, \tilde{\boldsymbol{x}}_i) = \langle \boldsymbol{f}(\boldsymbol{x}_u), \tilde{\boldsymbol{f}}(\tilde{\boldsymbol{x}}_i)\rangle$, where $\boldsymbol{f}(\boldsymbol{x}_u) \in \mathcal{F}_{D_u}(W, L, B, M)$ and $\tilde{\boldsymbol{f}}(\tilde{\boldsymbol{x}}_i) \in \mathcal{F}_{D_i}(\tilde{W}, \tilde{L}, \tilde{B}, M)$. It follows from Lemma 2 that there exists mappings $Q_u : \mathcal{F}_{D_u}(W, L, B, M) \to \mathcal{K}_{D_u}(W, L, B, M)$ and $Q_i : \mathcal{F}_{D_i}(\tilde{W}, \tilde{L}, \tilde{B}, M) \to \mathcal{K}_{D_i}(\tilde{W}, \tilde{L}, \tilde{B}, M)$ such that

$$R(\boldsymbol{x}_u, \tilde{\boldsymbol{x}}_i) = \langle \boldsymbol{f}(\boldsymbol{x}_u), \tilde{\boldsymbol{f}}(\tilde{\boldsymbol{x}}_i)\rangle = \langle Q_u(\boldsymbol{f})(\boldsymbol{x}_u), Q_i(\tilde{\boldsymbol{f}})(\tilde{\boldsymbol{x}}_i)\rangle,$$

for any $(\boldsymbol{x}_u, \tilde{\boldsymbol{x}}_i) \in \mathrm{Supp}(\mu_{ui})$.

Let Θ_Q and $\tilde{\Theta}_Q$ denote the effective parameters of $Q_u(\boldsymbol{f})$ and $Q_i(\tilde{\boldsymbol{f}})$, then R can be parametrized by $\Lambda_Q = (\Theta_Q, \tilde{\Theta}_Q)$. Let $\mathcal{Q} = \{\Lambda_Q : R(\cdot; \Lambda_Q) \in \mathcal{R}^{\Phi}\}$ and $\mathcal{G} = \{\Lambda_Q^{(1)}, \ldots, \Lambda_Q^{(N)}\}$ be an $\epsilon/2-$covering set of $\mathcal{Q}$ under the $\|\cdot\|_\infty$ metric. For any $R(\cdot; \Lambda_Q) \in \mathcal{R}^{\Phi}$, there exists $\Lambda'_Q \in \mathcal{G}$ such that $\|\Lambda_Q - \Lambda'_Q\|_\infty < \epsilon/2$, and thus

$$\begin{aligned}
&\sup_{\|(\boldsymbol{x}_u, \tilde{\boldsymbol{x}}_i)\|_\infty \le 1} |R(\boldsymbol{x}_u, \tilde{\boldsymbol{x}}_i) - R'(\boldsymbol{x}_u, \tilde{\boldsymbol{x}}_i)| = \sup_{\|(\boldsymbol{x}_u, \tilde{\boldsymbol{x}}_i)\|_\infty \le 1} |\langle \boldsymbol{f}(\boldsymbol{x}_u), \tilde{\boldsymbol{f}}(\tilde{\boldsymbol{x}}_i)\rangle - \langle \boldsymbol{f}'(\boldsymbol{x}_u), \tilde{\boldsymbol{f}}'(\tilde{\boldsymbol{x}}_i)\rangle| \\
\le &\sup_{\|(\boldsymbol{x}_u, \tilde{\boldsymbol{x}}_i)\|_\infty \le 1} |\langle \boldsymbol{f}(\boldsymbol{x}_u), \tilde{\boldsymbol{f}}(\tilde{\boldsymbol{x}}_i) - \tilde{\boldsymbol{f}}'(\tilde{\boldsymbol{x}}_i)\rangle| + \sup_{\|(\boldsymbol{x}_u, \tilde{\boldsymbol{x}}_i)\|_\infty \le 1} |\langle \boldsymbol{f}(\boldsymbol{x}_u) - \boldsymbol{f}'(\boldsymbol{x}_u), \tilde{\boldsymbol{f}}'(\tilde{\boldsymbol{x}}_i)\rangle| \\
\le &\sup_{\|(\boldsymbol{x}_u, \tilde{\boldsymbol{x}}_i)\|_\infty \le 1} \|\boldsymbol{f}(\boldsymbol{x}_u)\|_2 \|\tilde{\boldsymbol{f}}(\tilde{\boldsymbol{x}}_i) - \tilde{\boldsymbol{f}}'(\tilde{\boldsymbol{x}}_i)\|_2 + \sup_{\|(\boldsymbol{x}_u, \tilde{\boldsymbol{x}}_i)\|_\infty \le 1} \|\boldsymbol{f}(\boldsymbol{x}_u) - \boldsymbol{f}'(\boldsymbol{x}_u)\|_2 \|\tilde{\boldsymbol{f}}'(\tilde{\boldsymbol{x}}_i)\|_2 \\
\le &2Mp^{1/2}\Big(\sup_{\|\tilde{\boldsymbol{x}}_i\|_\infty \le 1} \|Q_i(\tilde{\boldsymbol{f}})(\tilde{\boldsymbol{x}}_i) - Q_i(\tilde{\boldsymbol{f}}')(\tilde{\boldsymbol{x}}_i)\|_2 + \sup_{\|\boldsymbol{x}_u\|_\infty \le 1} \|Q_u(\boldsymbol{f})(\boldsymbol{x}_u) - Q_u(\boldsymbol{f}')(\boldsymbol{x}_u)\|_2\Big) \\
\le &\epsilon M p^{3/2}(C(W, L, B) + C(\tilde{W}, \tilde{L}, \tilde{B})) \triangleq C_4\epsilon,
\end{aligned} \tag{5}$$

where the last inequality follows from Lemma 2.

For each $\Lambda_Q^{(n)} \in \mathcal{G}$, we define a $C_4\epsilon$-bracket as

$$g_n^U(\boldsymbol{x}_u, \tilde{\boldsymbol{x}}_i) = R(\boldsymbol{x}_u, \tilde{\boldsymbol{x}}_i; \Lambda_Q^{(n)}) + \frac{C_4\epsilon}{2}, g_n^L(\boldsymbol{x}_u, \tilde{\boldsymbol{x}}_i) = R(\boldsymbol{x}_u, \tilde{\boldsymbol{x}}_i; \Lambda_Q^{(n)}) - \frac{C_4\epsilon}{2}.$$

Combined with (5), it follows that for any $\Lambda_Q \in \mathcal{Q}$, there exists $1 \le k \le N$ such that

$$\begin{aligned}
g_k^U(\boldsymbol{x}_u, \tilde{\boldsymbol{x}}_i) - R(\boldsymbol{x}_u, \tilde{\boldsymbol{x}}_i, \Lambda_Q) &\ge \frac{C_4\epsilon}{2} - |R(\boldsymbol{x}_u, \tilde{\boldsymbol{x}}_i; \Lambda_Q) - R(\boldsymbol{x}_u, \tilde{\boldsymbol{x}}_i; \Lambda_Q^{(k)})| \ge 0, \\
g_k^L(\boldsymbol{x}_u, \tilde{\boldsymbol{x}}_i) - R(\boldsymbol{x}_u, \tilde{\boldsymbol{x}}_i, \Lambda_Q) &\le |R(\boldsymbol{x}_u, \tilde{\boldsymbol{x}}_i; \Lambda_Q) - R(\boldsymbol{x}_u, \tilde{\boldsymbol{x}}_i; \Lambda_Q^{(k)})| - \frac{C_4\epsilon}{2} \le 0,
\end{aligned}$$

for any $(\boldsymbol{x}_u, \tilde{\boldsymbol{x}}_i) \in \mathrm{Supp}(\mu_{ui})$. Therefore, $\mathcal{B} = \{[g_1^L, g_1^U], [g_2^L, g_2^U], \ldots, [g_N^L, g_N^U]\}$ forms a $C_4\epsilon$-bracketing set of $\mathcal{R}^{\Phi}$ under the $\|\cdot\|_{L^2(\mu_{ui})}$ metric.

By Lemma 2, the size of Λ_Q is at most $14LW\log W + 14\tilde{L}\tilde{W}\log\tilde{W}$. Combined with the definition of $\mathcal{G}$, this yields that

$$\log N \le (14LW\log W + 14\tilde{L}\tilde{W}\log\tilde{W})\log\Big(\epsilon^{-1}2\max\{B, \tilde{B}\}\Big).$$

Substituting ϵ by $\tilde{\epsilon}/C_4$ leads to the desired upper bound immediately. ■

Proof of Theorem 2: Let $L_{ui} = \max\{L, \tilde{L}\}$, $\eta_{|\Omega|}^2 = L_{ui}|\Omega|^{-2\beta/(2\beta+d_{ui})}\log^2|\Omega|$, $\mathcal{M} = \{R \in \mathcal{R}^{\Phi} : \|R - R^*\|_{L^2(\mu_{ui})}^2 > \eta_{|\Omega|}^2\}$, and let $R_0 \in \mathcal{R}^{\Phi}$ satisfy $\|R_0 - R^*\|_{L^\infty(\mu_{ui})}^2 \le \eta_{|\Omega|}^2/4$. Further, we denote $\|R - K\|_\Omega^2 = \frac{1}{|\Omega|}\sum_{(u,i)\in\Omega}(R(\boldsymbol{x}_u, \tilde{\boldsymbol{x}}_i) - K_{ui})^2$, and then it follows from the definition of $\hat{R}$ that

$$\begin{aligned}
&P\big(\|\hat{R} - R^*\|_{L^2(\mu_{ui})}^2 > \eta_{|\Omega|}^2\big) \\
\le &P\Big(\sup_{R\in\mathcal{M}}\big(\|R_0 - K\|_\Omega^2 + \lambda_{|\Omega|}J_0 - \|R - K\|_\Omega^2 - \lambda_{|\Omega|}J(R)\big) \ge 0\Big) \equiv I,
\end{aligned}$$

where $J_0 = J(R_0)$. We further decompose $\mathcal{M}$ into small subsets. Specifically, we let $\mathcal{M}_{ij} = \{R \in \mathcal{R}^{\Phi} : 2^{i-1}\eta_{|\Omega|}^2 < \|R - R^*\|_{L^2(\boldsymbol{\mu}_{ui})}^2 \le 2^i\eta_{|\Omega|}^2, 2^{j-1}J_0 < J(R) \le 2^j J_0\}$ for $i, j \ge 1$, and $\mathcal{M}_{i0} = \{R \in \mathcal{R}^{\Phi} : 2^{i-1}\eta_{|\Omega|}^2 < \|R -$

$R^*\|^2_{L^2(\boldsymbol{\mu}_{ui})} \leq 2^i\eta^2_{|\Omega|}, J(R) \leq J_0\}$ for $i \geq 1$. Then we have

$$
\begin{aligned}
I \leq & \sum_{i=1}^{\infty}\sum_{j=0}^{\infty} P\Big(\sup_{R\in\mathcal{M}_{ij}} \big(\|R_0 - K\|^2_\Omega + \lambda_{|\Omega|}J_0 - \|R-K\|^2_\Omega - \lambda_{|\Omega|}J(R)\big) \geq 0\Big) \\
= & \sum_{i,j=1}^{\infty} P\Big(\sup_{R\in\mathcal{M}_{ij}} \big(\|R_0 - K\|^2_\Omega + \lambda_{|\Omega|}J_0 - \|R-K\|^2_\Omega - \lambda_{|\Omega|}J(R)\big) \geq 0\Big) \\
& + \sum_{i=1}^{\infty} P\Big(\sup_{R\in\mathcal{M}_{i0}} \big(\|R_0 - K\|^2_\Omega + \lambda_{|\Omega|}J_0 - \|R-K\|^2_\Omega - \lambda_{|\Omega|}J(R)\big) \geq 0\Big) \equiv I_1 + I_2.
\end{aligned}
$$

It thus suffices to bound I_1 and I_2 separately.

Let $\epsilon = K - R^*$, then we have

$$
\|R - K\|^2_\Omega = \|R - R^*\|^2_\Omega + \|\epsilon\|^2_\Omega - \frac{2}{|\Omega|}\sum_{(u,i)\in\Omega} \epsilon_{ui}\big(R(\boldsymbol{x}_u, \tilde{\boldsymbol{x}}_i) - R^*(\boldsymbol{x}_u, \tilde{\boldsymbol{x}}_i)\big).
$$

Therefore, $\mathbb{E}\|R - K\|^2_\Omega = \|R - R^*\|^2_{L^2(\mu_{ui})} + \mathbb{E}\|\epsilon\|^2_\Omega$, and thus

$$
\begin{aligned}
\mathbb{E}\big(\|R-K\|^2_\Omega - \|R_0 - K\|^2_\Omega\ \big) &= \|R - R^*\|^2_{L^2(\mu_{ui})} - \|R_0 - R^*\|^2_{L^2(\mu_{ui})} \\
&\geq \|R - R^*\|^2_{L^2(\mu_{ui})} - \eta^2_{|\Omega|}/4.
\end{aligned}
$$

Let $E_\Omega(R) = \|R - K\|^2_\Omega - \mathbb{E}\big(\|R-K\|^2_\Omega\big)$, then we have

$$
\begin{aligned}
& P\Big(\sup_{R\in\mathcal{M}_{ij}} \big(\|R_0 - K\|^2_\Omega + \lambda_{|\Omega|}J(R_0) - \|R-K\|^2_\Omega - \lambda_{|\Omega|}J(R)\big) \geq 0\Big) \\
=& P\Big(\sup_{R\in\mathcal{M}_{ij}} \big(E_\Omega(R_0) - E_\Omega(R)\big) \geq \inf_{R\in\mathcal{M}_{ij}} \lambda_{|\Omega|}(J(R) - J(R_0)) + \inf_{R\in\mathcal{M}_{ij}} \mathbb{E}\big(\|R-K\|^2_\Omega - \|R_0 - K\|^2_\Omega\big)\Big) \\
\leq& P\Big(\sup_{R\in\mathcal{M}_{ij}} \big(E_\Omega(R_0) - E_\Omega(R)\big) \geq \inf_{R\in\mathcal{M}_{ij}} \lambda_{|\Omega|}(J(R) - J(R_0)) + \inf_{R\in\mathcal{M}_{ij}} \|R - R^*\|^2_{L^2(\mu_{ui})} - \eta^2_{|\Omega|}/4\Big) \\
\leq& P\Big(\sup_{R\in\mathcal{M}_{ij}} \big(E_\Omega(R_0) - E_\Omega(R)\big) \geq (2^{j-1} - 1)\lambda_{|\Omega|}J_0 + (2^{i-1} - 1/4)\eta^2_{|\Omega|}\Big) \\
=& P\Big(\sup_{R\in\mathcal{M}_{ij}} \big(E_\Omega(R_0) - E_\Omega(R)\big) \geq M(i,j)\Big),
\end{aligned}
$$

where $M(i,j) = (2^{j-1} - 1)\lambda_{|\Omega|}J_0 + (2^{i-1} - 1/4)\eta^2_{|\Omega|}$.

Next, it follows from the assumption $\lambda_{|\Omega|}J_0 \leq 1/4\eta^2_{|\Omega|}$ that

$$
\begin{aligned}
&\sup_{R\in\mathcal{M}_{ij}} \mathrm{Var}\Big((R(\boldsymbol{x}_u,\tilde{\boldsymbol{x}}_i)-K_{ui})^2-(R_0(\boldsymbol{x}_u,\tilde{\boldsymbol{x}}_i)-K_{ui})^2\Big)\\
=&\sup_{R\in\mathcal{M}_{ij}} \mathrm{Var}\Big((R(\boldsymbol{x}_u,\tilde{\boldsymbol{x}}_i)-R^*(\boldsymbol{x}_u,\tilde{\boldsymbol{x}}_i))^2-(R_0(\boldsymbol{x}_u,\tilde{\boldsymbol{x}}_i)-R^*(\boldsymbol{x}_u,\tilde{\boldsymbol{x}}_i))^2\Big)\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad+\mathrm{Var}\Big(2\epsilon_{ui}(R(\boldsymbol{x}_u,\tilde{\boldsymbol{x}}_i)-R_0(\boldsymbol{x}_u,\tilde{\boldsymbol{x}}_i))\Big)\\
\leq&\sup_{R\in\mathcal{M}_{ij}} 2\mathrm{Var}\Big((R(\boldsymbol{x}_u,\tilde{\boldsymbol{x}}_i)-R^*(\boldsymbol{x}_u,\tilde{\boldsymbol{x}}_i))^2\Big)+2\mathrm{Var}\Big((R_0(\boldsymbol{x}_u,\tilde{\boldsymbol{x}}_i)-R^*(\boldsymbol{x}_u,\tilde{\boldsymbol{x}}_i))^2\Big)\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad+4\mathbb{E}\epsilon_{ui}^2\sup_{R\in\mathcal{M}_{ij}}\mathbb{E}(R(\boldsymbol{x}_u,\tilde{\boldsymbol{x}}_i)-R_0(\boldsymbol{x}_u,\tilde{\boldsymbol{x}}_i))^2\\
\leq&2\sup_{R\in\mathcal{M}_{ij}}\mathbb{E}(R(\boldsymbol{x}_u,\tilde{\boldsymbol{x}}_i)-R^*(\boldsymbol{x}_u,\tilde{\boldsymbol{x}}_i))^4+2\mathbb{E}((R_0(\boldsymbol{x}_u,\tilde{\boldsymbol{x}}_i)-R^*(\boldsymbol{x}_u,\tilde{\boldsymbol{x}}_i))^2\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad+4\sigma^2\sup_{R\in\mathcal{M}_{ij}}\mathbb{E}(R(\boldsymbol{x}_u,\tilde{\boldsymbol{x}}_i)-R_0(\boldsymbol{x}_u,\tilde{\boldsymbol{x}}_i))^2\\
\leq&\sup_{R\in\mathcal{M}_{ij}}(50p^2M^4+4\sigma^2)\big(\|R-R^*\|^2_{L^2(\mu_{ui})}+\|R_0-R^*\|^2_{L^2(\mu_{ui})}\big)\\
\leq&(50p^2M^4+4\sigma^2)\big(2^i\eta^2_{|\Omega|}+\frac{1}{4}\eta^2_{|\Omega|}\big)\leq C_5M(i,j)\equiv v(i,j),
\end{aligned}
\tag{6}
$$

where $C_5 = 16\max\{(50p^2M^4+4\sigma^2),1\}(25p^2M^4+B_e^2)$.

In the following, we proceed to verify conditions (4.5)-(4.7) in Shen and Wong (1994). First, the relation between $M(i,j)$ and $v(i,j)$ in (6) directly implies (4.6) with $T = 2(25p^2M^4+B_e^2)$ and $\epsilon = 1/2$. Second, we let $\mathcal{R}^\Phi(\tau) = \{R\in\mathcal{R}^\Phi : J(R)\leq\tau J_0\}$, and note that $J(R)\leq\tau J_0$ implies that $\max\{B,\tilde{B}\}\leq\sqrt{\tau J_0}$. Then it follows from Lemma 3 that

$$
\log\mathcal{N}_{[\cdot]}(\epsilon,\mathcal{R}^\Phi(\tau),\|\cdot\|_{L^2(\mu_{ui})})\leq C_2(W\log W+\tilde{W}\log\tilde{W})\log\Big(C_6\epsilon^{-1}\Big),
$$

where $C_6 = C_3\big(C(W,L,\sqrt{\tau J_0})+C(\tilde{W},\tilde{L},\sqrt{\tau J_0})\big)$, and C_2 and C_3 are defined as in Lemma 3. It then follows that

$$
\begin{aligned}
&\int_{\frac{\epsilon}{32}M(i,j)}^{v^{1/2}(i,j)}\sqrt{\log\mathcal{N}_{[\cdot]}(u,\mathcal{R}^\Phi(\tau),\|\cdot\|_{L^2(\mu_{ui})})}du/M(i,j)\\
&\qquad\leq\int_{\frac{\epsilon}{32}M(i,j)}^{v^{1/2}(i,j)}\sqrt{C_2(W\log W+\tilde{W}\log\tilde{W})\log\Big(C_6u^{-1}\Big)}du/M(i,j).
\end{aligned}
\tag{7}
$$

Notice that the right-hand side of (7) is non-increasing in i and $M(i,j)$, it then follows that

$$
\begin{aligned}
&\int_{\frac{\epsilon}{32}M(i,j)}^{v^{1/2}(i,j)}\sqrt{C_2(W\log W+\tilde{W}\log\tilde{W})\log\Big(C_6u^{-1}\Big)}du/M(i,j)\\
&\qquad\leq\int_{\frac{\epsilon}{32}M(1,j)}^{v^{1/2}(1,j)}\sqrt{C_2(W\log W+\tilde{W}\log\tilde{W})\log\Big(C_6u^{-1}\Big)}du/M(1,j).
\end{aligned}
\tag{8}
$$

Note that W and $\tilde{W}$ are adaptive parameters governing the rate of approximation error $\|R_0-R^*\|_{L^\infty(\mu_{ui})}$, which must satisfy $\|R_0-R^*\|_{L^\infty(\mu_{ui})}\leq 1/2\eta_{|\Omega|}$. Thus, (4.7) holds by setting $W = O(|\Omega|^{d_{ui}/(2\beta+d_{ui})}\log|\Omega|)$ and $\tilde{W} = O(|\Omega|^{d_{ui}/(2\beta+d_{ui})}\log|\Omega|)$, and (4.7) directly implies (4.5). By Theorem 3 in Shen and Wong (1994) with $M =$

$|\Omega|^{1/2}M(i,j)$ and $v=v(i,j)$, we have

$$
\begin{aligned}
I_1 &\leq \sum_{j=1}^{\infty}\sum_{i=1}^{\infty} 3\exp\Big(-\frac{(1-\epsilon)|\Omega|M(i,j)^2}{2(4C_5M(i,j)+M(i,j)T/3)}\Big)\\
&\leq 3\sum_{j=1}^{\infty}\sum_{i=1}^{\infty}\exp\Big(-C_7(1-\epsilon)|\Omega|\big((2^{j-1}-1)\lambda_{|\Omega|}J_0+(2^{i-1}-1/4)\eta^2_{|\Omega|}\big)\Big)\\
&\leq 3\sum_{i=1}^{n}\exp(-C_7(1-\epsilon)|\Omega|(i-1/4)\eta^2_{|\Omega|})\sum_{j=1}^{n}\exp(-C_7(1-\epsilon)|\Omega|(j-1)\lambda_{|\Omega|}J_0)\\
&\leq 3\frac{\exp(-C_7(1-\epsilon)|\Omega|\eta^2_{|\Omega|}/4)}{1-\exp(-C_7(1-\epsilon)|\Omega|\eta^2_{|\Omega|})}\frac{1}{1-\exp(-C_7(1-\epsilon)|\Omega|\lambda_{|\Omega|}J_0)}\\
&\leq 3\frac{\exp(-C_7(1-\epsilon)|\Omega|\eta^2_{|\Omega|}/4)}{(1-\exp(-C_7(1-\epsilon)|\Omega|\eta^2_{|\Omega|}/4))^2},
\end{aligned}
\tag{9}
$$

where $C_7=3/(26C_5)$ and the last inequality follows from the fact that $\lambda_{|\Omega|}J_0\leq 1/4\eta^2_{|\Omega|}$.

Similarly, I_2 can be bounded by

$$
\begin{aligned}
I_2 &\leq \sum_{i=1}^{n}3\exp(-\frac{(1-\epsilon)|\Omega|M^2(i,0)}{2(4v(i,0)+M(i,0)T/3)})\leq\sum_{i=1}^{n}3\exp(-C_7(1-\epsilon)|\Omega|M(i,0))\\
&\leq\sum_{i=1}^{\infty}3\exp(-C_7(1-\epsilon)|\Omega|(2^{i-1}-1/2)\eta^2_{|\Omega|})\leq 3\frac{\exp(-C_7(1-\epsilon)|\Omega|\eta^2_{|\Omega|}/2)}{1-\exp(-C_7(1-\epsilon)|\Omega|\eta^2_{|\Omega|})}.
\end{aligned}
\tag{10}
$$

Combining (9) and (10), we have

$$
I\leq I_1+I_2\leq 3\frac{\exp(-C_7(1-\epsilon)|\Omega|\eta^2_{|\Omega|}/4)}{(1-\exp(-C_7(1-\epsilon)|\Omega|\eta^2_{|\Omega|}/4)^2}+3\frac{\exp(-C_7(1-\epsilon)|\Omega|\eta^2_{|\Omega|}/2)}{1-\exp(-C_7(1-\epsilon)|\Omega|\eta^2_{|\Omega|})}.
$$

Let $s=\exp(-C_7(1-\epsilon)|\Omega|\eta^2_{|\Omega|}/4)$, then

$$
I\leq\frac{3s^2}{(1-s)^2}+\frac{3s^2}{1-s^4}\leq\frac{3s^2}{(1-s)^2}+\frac{3s^2}{1-s}=\frac{6s^2-3s^3}{(1-s)^2}\leq 24s^2,
$$

as $s\leq 1/2$. The desired result then follows immediately. ■

References

Nakada, R. and Imaizumi, M. (2020). Adaptive approximation and generalization of deep neural network with intrinsic dimensionality. *Journal of Machine Learning Research*, 21(174):1–38.

Shen, X. and Wong, W. H. (1994). Convergence rate of sieve estimates. *The Annals of Statistics*, pages 580–615.