Supplementary Materials for "Embedding learning"

This supplementary materials include proofs in "Embedding learning".

Proof of Lemma 1. By the definition of sufficient embedding, we have

$$\begin{aligned}\min_f \mathbb{E}\big(L(f(\mathcal{X}(S)),Y)\big) &= \min_f \mathbb{E}\big(\mathbb{E}\big(L(f(\mathcal{X}(S)),Y)|S\big)\big) = \min_f \mathbb{E}\big(\mathbb{E}\big(L\big(f(\mathcal{X}(S)),Y\big)\big|\mathcal{X}(S)\big)\big)\\ &= \min_g \mathbb{E}\big(\mathbb{E}\big(L\big(g(S),Y\big)\big|S\big)\big) = \min_g \mathbb{E}\big(L\big(g(S),Y\big)\big),\end{aligned}$$

where first equality follows from the law of conditioning, the second equality follows from Definition 1, and the second last equality holds by the pointwise minimization of both sides. The desired result then follows. □

Proof of Lemma 2. Note that $\mathcal{X}^*$ is a sufficient embedding, then Lemma 1 yields that

$$\mathbb{E}\big(L\big(f^*(\mathcal{X}^*(S)),Y\big)\big) = \min_f \mathbb{E}\big(L\big(f(\mathcal{X}^*(S)),Y\big)\big) = \min_g \mathbb{E}\big(L\big(g(S),Y\big)\big),$$

Note further that we can choose $g(s) = f(\mathcal{X}(s))$, thus for any $\mathcal{X}$ and f, $\min_g \mathbb{E}\big(L\big(g(S),Y\big)\big) \le \mathbb{E}\big(L\big(f(\mathcal{X}(s)),Y\big)\big)$. Then for any $C \ge U(\mathcal{X}^*)$,

$$\begin{aligned}\min_g \ \mathbb{E}\big(L\big(g(S),Y\big)\big) &\le \min_{f,\mathcal{X}:U(\mathcal{X})\le C} \ \mathbb{E}\big(L\big(f(\mathcal{X}(s)),Y\big)\big)\\ &\le \mathbb{E}\big(L\big(f^*(\mathcal{X}^*(S)),Y\big)\big) = \min_g \ \mathbb{E}\big(L\big(g(S),Y\big)\big),\end{aligned}$$

where the second last inequality follows from $U(\mathcal{X}^*) \le C$. The desired result follows. □

Proof of Theorem 1. By Lemma 2,

$$\mathbb{P}\big(e(\widehat{\boldsymbol{\theta}}) \geq c_1\delta_n^{2\mu}\big) = \mathbb{P}\Big(\mathbb{E}\big(L\big(\widehat{f}(\widehat{\mathcal{X}}(S)), Y\big) - L\big(f^*(\mathcal{X}^*(S)), Y\big)\big) \geq c_1\delta_n^{2\mu}\Big).$$

Then the desired result follows from Theorem 1 of [3]. This completes the proof. □

Proof of Corollary 1. Without loss of generality, assume that the Lipschitz constant is 1 for $V(z)$, then it suffices to verify Assumption A3. Let $\mathcal{V}(r) = \big\{V\big(yf\big(\mathcal{X}(s)\big)\big) - V\big(y\bar{f}\big(\bar{\mathcal{X}}(s)\big)\big) : f \in \mathcal{F}, J(f) \leq \bar{J}r, U(\mathcal{X}) \leq C\big\}$ and $\mathcal{F}(r) = \{\boldsymbol{\beta}^\top\boldsymbol{x} : \|\boldsymbol{\beta}\|_2^2 \leq \bar{J}r\}$. Then we compute the cover number $\mathcal{N}\big(\mathcal{V}(r), \nu\big)$ of the functional space. For $\boldsymbol{\beta}$ and $\tilde{\boldsymbol{\beta}}$ in $\mathcal{F}(r)$, $\boldsymbol{X}$ and $\widetilde{\boldsymbol{X}}$ in $\mathcal{D}_C$,

$$\begin{aligned}
&\sup_{S,Y}\big|V\big(Y\widetilde{\boldsymbol{\beta}}^\top\widetilde{\mathcal{X}}(S)\big) - V\big(Y\boldsymbol{\beta}^\top\mathcal{X}(S)\big)\big| \leq \max_{u=1,\cdots,|\mathcal{S}|}\big|\widetilde{\boldsymbol{\beta}}^\top\widetilde{\boldsymbol{x}}_u - \boldsymbol{\beta}^\top\widetilde{\boldsymbol{x}}_u + \boldsymbol{\beta}^\top\widetilde{\boldsymbol{x}}_u - \boldsymbol{\beta}^\top\boldsymbol{x}_u\big| \\
\leq &\max_{u=1,\cdots,|\mathcal{S}|}\|\widetilde{\boldsymbol{x}}_u\|_2\big\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\big\|_2 + \|\boldsymbol{\beta}\|_2\max_{u=1,\cdots,|\mathcal{S}|}\big\|\widetilde{\boldsymbol{x}}_u - \boldsymbol{x}_u\big\|_2 \\
\leq &B\sqrt{p}\big\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\big\|_2 + \sqrt{\bar{J}pr}\big\|\mathrm{vec}(\widetilde{\boldsymbol{X}}) - \mathrm{vec}(\boldsymbol{X})\big\|_\infty,
\end{aligned}$$

where $\boldsymbol{x}_u = \mathcal{X}(s_u)$ is the u-th column vector of $\boldsymbol{X}$, the second inequality follows from the Cauchy-Schwarz inequality. Then, it suffices to compute the covering number of $\mathcal{F}(r)$ and $\mathcal{X}_C$ separately, i.e.,

$$\begin{aligned}
H_\infty\big(\nu, \mathcal{V}(r)\big) &\leq \log\mathcal{N}\big(\nu/(2B\sqrt{p})\big), \mathcal{F}(r)\big) + \log\mathcal{N}\big(\nu/(2\sqrt{\bar{J}rp}), \mathcal{X}_C\big) \\
&\leq p\Big(\log\big(\frac{6B\sqrt{\bar{J}rp}}{\nu}\big) + \sum_{u=1}^{|\mathcal{S}|}\max\big(\log(\frac{\sqrt{2\bar{J}pr}\|\boldsymbol{P}\|_\infty\min(\sqrt{C\sigma_u^{-1}}, B\|\boldsymbol{P}\|_1)}{\nu}), 0\big)\Big),
\end{aligned}$$

where $\mathcal{X}_C = \big\{\mathrm{vec}(\boldsymbol{X}) \in \mathbb{R}^{p|\mathcal{S}|} : \mathrm{Tr}(\boldsymbol{X}\boldsymbol{Q}\boldsymbol{X}^T) \leq 2C,\ \|\mathrm{vec}(\boldsymbol{X})\|_\infty \leq B\big\}$, and the last inequality from Lemma 7. To apply Theorem 1, we verify the required entropy condition for $\mathcal{V}(r)$, $\sup_{r\geq 2}\phi(\epsilon_n, r) \leq c_3 n^{1/2}$, where $\phi(\epsilon_n, r) = \int_{c_5\zeta}^{c_4^{1/2}\zeta^{\kappa/2}} H_\infty^{1/2}\big(u, \mathcal{V}(r)\big)du/\zeta$, $\zeta = \min(\epsilon_n^2 + \lambda\bar{J}(r/2-1), 1)$. Note that $\phi(\epsilon_n, r)$ is non-increasing in $r \geq 2$ for any fixed $\epsilon_n > 0$. Then

$\sup_{r\geq 2}\phi(\epsilon_n, r) = \phi(\epsilon_n, 2)$ is bounded by

$$
\begin{aligned}
&p^{1/2}\Big(\log\big(\frac{6B\sqrt{2\bar{J}p}}{\epsilon_n^2}\big) + \sum_{u=1}^{|\mathcal{S}|}\max\big(\log(\frac{2\sqrt{\bar{J}p}\|\boldsymbol{P}\|_\infty \min(\sqrt{C\sigma_u^{-1}}, B\|\boldsymbol{P}\|_1)}{\epsilon_n^2}), 0\big)\Big)^{1/2}\epsilon_n^{\kappa-2} \\
&\leq p^{1/2}\Big(\log\big(\frac{6B\sqrt{2\bar{J}p}}{\epsilon_n^2}\big) + \sum_{u=1}^{K^*}\log(\frac{2\sqrt{\bar{J}p}B\|\boldsymbol{P}\|_1\|\boldsymbol{P}\|_\infty}{\epsilon_n^2})\Big)^{1/2}\epsilon_n^{\kappa-2} \\
&\leq p^{1/2}(1+K^*)^{1/2}\Big(\log(\frac{6B\sqrt{2\bar{J}p}(1+\|\boldsymbol{P}\|_1\|\boldsymbol{P}\|_\infty)}{\epsilon_n^2})\Big)^{1/2}\epsilon_n^{\kappa-2} \leq c_3 n^{1/2}, \qquad (1)
\end{aligned}
$$

where the second inequality follows from the definition of K^*, and the last inequality implies from the fact that $\epsilon_n^2 \sim \Big(\frac{(K^*+1)p^*}{n}\log\Big(\frac{nB\sqrt{\bar{J}p^*}(1+\|\boldsymbol{P}\|_1\|\boldsymbol{P}\|_\infty)}{(K^*+1)p^*}\Big)\Big)^{\frac{1}{2-\kappa}}$. □

Proof of Corollary 2. Without loss of generality, we assume $V(z)$ is Lipschitz continuous with the Lipschitz constant equal to 1. We bound the approximation error and estimation error separately.

For the estimation error, let $\mathcal{I} = [-B, B]^{p^*}$. For any $f_M, \tilde{f}_M \in \mathcal{F}(r) = \{f \in \mathcal{F}, J(f) \leq r\bar{J}\}$, with $\mathcal{F}$ being a space of neural networks and $\boldsymbol{X}, \widetilde{\boldsymbol{X}} \in \mathcal{D}_C$,

$$
\begin{aligned}
&\sup_{S,Y}\big|V\big(Y\tilde{f}_M\big(\widetilde{\mathcal{X}}(S)\big)\big) - V\big(Yf_M\big(\mathcal{X}(S)\big)\big)\big| \\
&\leq \max_{u=1,\cdots,|\mathcal{S}|}\big|\tilde{f}_M\big(\widetilde{\boldsymbol{x}}_u\big) - f_M\big(\boldsymbol{x}_u\big)\big| \leq \max_{u=1,\cdots,|\mathcal{S}|}\big(\big|\tilde{f}_M\big(\widetilde{\boldsymbol{x}}_u\big) - f_M\big(\widetilde{\boldsymbol{x}}_u\big)\big| + \big|f_M\big(\widetilde{\boldsymbol{x}}_u\big) - f_M(\boldsymbol{x}_u)\big|\big) \\
&\leq \sup_{\boldsymbol{x}\in\mathcal{I}}|\tilde{f}_M(\boldsymbol{x}) - f_M(\boldsymbol{x})| + \max_{u=1,\cdots,|\mathcal{S}|}\big|f_M\big(\widetilde{\boldsymbol{x}}_u\big) - f_M(\boldsymbol{x}_u)\big| \\
&\leq \sup_{\boldsymbol{x}\in\mathcal{I}}|\tilde{f}_M(\boldsymbol{x}) - f_M(\boldsymbol{x})| + \|\boldsymbol{A}_M\|_{1,1}\cdots\|\boldsymbol{A}_1\|_{1,1}\max_{u=1,\cdots,|\mathcal{S}|}\|\widetilde{\boldsymbol{x}}_u - \boldsymbol{x}_u\|_\infty \\
&\leq \sup_{\boldsymbol{x}\in\mathcal{I}}|\tilde{f}_M(\boldsymbol{x}) - f_M(\boldsymbol{x})| + (r\bar{J}/M)^M\big\|\mathrm{vec}(\widetilde{\boldsymbol{X}}) - \mathrm{vec}(\boldsymbol{X})\big\|_\infty,
\end{aligned}
$$

where the last inequality follows from the means inequality, and the second last inequality

follows from the fact that

$$\begin{aligned}
&\max_{1\le j\le h_m}\big|\big(\mathbf{f}_m(\tilde{\boldsymbol{x}}_u)-\mathbf{f}_m(\boldsymbol{x}_u)\big)_j\big| = \max_{1\le j\le h_m}\big|\boldsymbol{\sigma}_j\big(\boldsymbol{A}_m\mathbf{f}_{m-1}(\widetilde{\boldsymbol{x}}_u)+\boldsymbol{b}_m\big)-\boldsymbol{\sigma}_j\big(\boldsymbol{A}_m\mathbf{f}_{m-1}(\boldsymbol{x}_u)+\boldsymbol{b}_m\big)\big|\\
&\le \max_{1\le j\le h_m}\big|(\boldsymbol{A}_m)_j\mathbf{f}_{m-1}(\widetilde{\boldsymbol{x}}_u)-(\boldsymbol{A}_m)_j\mathbf{f}_{m-1}(\boldsymbol{x}_u)\big|\\
&\le \max_{1\le j\le h_m}\big\|(\boldsymbol{A}_m)_j\big\|_1\max_{1\le j\le h_{m-1}}\big|\big(\mathbf{f}_{m-1}(\widetilde{\boldsymbol{x}}_u)-\mathbf{f}_{m-1}(\boldsymbol{x}_u)\big)_j\big|\\
&\le \|\boldsymbol{A}_m\|_{1,1}\max_{1\le j\le h_{m-1}}\big|\big(\mathbf{f}_{m-1}(\widetilde{\boldsymbol{x}}_u)-\mathbf{f}_{m-1}(\boldsymbol{x}_u)\big)_j\big|.
\end{aligned}$$

Then it suffices to bound the entropy of $\mathcal{F}(r)$ and $\mathcal{D}_C$. Since $\max_{s\in\mathcal{S}}\|\mathcal{X}(s)\|_\infty\le B$,

$$\begin{aligned}
&H_\infty(\nu,\mathcal{V}(r))\le \log\mathcal{N}(\nu/2,\mathcal{F}(r))+\log\mathcal{N}\big(\frac{\nu}{2(r\bar{J}/M)^M},\mathcal{D}_C\big)\\
&\le |\boldsymbol{\Theta}_N|\log\Big(\frac{6r\bar{J}}{\nu}\Big(MB(\frac{2r\bar{J}}{M-1})^{M-1}+(\frac{r\bar{J}(1-(r\bar{J})^M)}{1-r\bar{J}})^2\Big)\Big)\\
&\qquad\qquad +p^*\sum_{u=1}^{|\mathcal{S}|}\max\Big(\log\big(\frac{\sqrt{2}\|\boldsymbol{P}\|_\infty(r\bar{J}/M)^M\min(\sqrt{C\sigma_u^{-1}},B\|\boldsymbol{P}\|_1)}{\nu}\big),0\Big)\\
&\le |\boldsymbol{\Theta}_N|\log\Big(\frac{6(MB+1)(r\bar{J})^{2M+1}}{\nu}\Big)\\
&\qquad\qquad +p^*\sum_{u=1}^{|\mathcal{S}|}\max\Big(\log\big(\frac{\sqrt{2}\|\boldsymbol{P}\|_\infty(r\bar{J}/M)^M\min(\sqrt{C\sigma_u^{-1}},B\|\boldsymbol{P}\|_1)}{\nu}\big),0\Big),
\end{aligned}$$

where the last inequality follows from Lemma 6. To apply Theorem 1, we verify the required entropy condition for $\mathcal{V}(r)$, $\sup_{r\ge2}\phi(\epsilon_n,r)\le c_3n^{1/2}$. Note that $\phi(\epsilon_n,r)$ is non-increasing in r for $r\ge2$ for any fixed $\epsilon_n>0$,

$$\begin{aligned}
&\sup_{r\ge2}\phi(\epsilon_n,2)=\phi(\epsilon_n,2)\le\Big(|\boldsymbol{\Theta}_N|\log\Big(\frac{6MB(2\bar{J})^{2M+1}}{\epsilon_n^2}\Big)\\
&\qquad +p^*\sum_{u=1}^{|\mathcal{S}|}\max\big(\log(\frac{\sqrt{2}\|\boldsymbol{P}\|_\infty(r\bar{J}/M)^M\min(\sqrt{C\sigma_u^{-1}},B\|\boldsymbol{P}\|_1)}{\epsilon_n^2}),0\big)\Big)^{1/2}\epsilon_n^{\kappa-2}\\
&\le\Big(|\boldsymbol{\Theta}_N|\log\Big(\frac{6MB(2\bar{J})^{2M+1}}{\epsilon_n^2}\Big)+p^*K^*\log(\frac{\sqrt{2}B(2\bar{J}/M)^M\|\boldsymbol{P}\|_1\|\boldsymbol{P}\|_\infty}{\epsilon_n^2})\Big)^{1/2}\epsilon_n^{\kappa-2}\\
&\le(|\boldsymbol{\Theta}_N|+p^*K^*)^{1/2}\Big(\log(\frac{6B(2\bar{J})^{2M+1}(M+\|\boldsymbol{P}\|_1\|\boldsymbol{P}\|_\infty)}{\epsilon_n^2})\Big)^{1/2}\le c_3n^{1/2},
\end{aligned}$$

where $\epsilon_n^2 \sim \left(\frac{|\boldsymbol{\Theta}_N|+K^*p^*}{n}\log\left(\frac{nB(2\bar{J})^{2M}(M+\|\boldsymbol{P}\|_1\|\boldsymbol{P}\|_\infty)}{M(|\boldsymbol{\Theta}_N|+K^*p)}\right)\right)^{\frac{1}{2-\kappa}}$, the second inequality follows from the fact that $2\sqrt{C}\|\boldsymbol{P}\|_\infty(2\bar{J}_M/M)^M/\sqrt{\sigma_k} \leq \epsilon_n^2$, for any $k > K^*$.

Next, we bound the approximation error $\upsilon_n^2 = e_V(\bar{\boldsymbol{\theta}})$ based on Theorem 1 of [6]. First, we transform $\mathcal{X}^*(s)$ from $[-B^*, B^*]^{p^*}$ to $[0,1]^{p^*}$. To proceed, let $\mathcal{X}^*_{B^*}(s) = \frac{\mathcal{X}^*(s)+B^*}{2B^*\tau} \in [0, 1/\tau]^{p^*} \subset [0,1]^{p^*}$ and $\mathcal{X}^*_{B^*} = \left\{\mathcal{X}^*_{B^*}(s)\right\}_{s\in\mathcal{S}}$. Since $f^* \in \mathcal{W}^\alpha_\infty([-B^*, B^*]^{p^*})$, then $f^*_{B^*}(\boldsymbol{x}) = f^*(2\tau B^*\boldsymbol{x} - B^*) \in \mathcal{W}^\alpha_\infty([0,1/\tau]^{p^*})$, where $\tau = \max(1, \Gamma/U(\mathcal{X}^*))^{1/2} \geq 1$, and $\Gamma = p|\mathcal{S}|\mathrm{Tr}(\boldsymbol{Q})$. Then the desired result follows from the fact that

$$\begin{aligned}
\upsilon_n^2 = e_V(\bar{\boldsymbol{\theta}}) &= \mathbb{E}\Big(V\big(Y\bar{f}_M(\bar{\mathcal{X}}(S))\big) - V\big(Yf^*(\mathcal{X}^*(S))\big)\Big) \\
&= \mathbb{E}\Big(V\big(Y\bar{f}_M(\mathcal{X}^*_{B^*}(S))\big) - V\big(Yf^*_{B^*}(\mathcal{X}^*_{B^*}(S))\big)\Big) \leq \big\|\bar{f}_M - f^*_{B^*}\big\|_\infty \\
&\leq a_1 \max\big(\frac{1}{e^{M-1}}, |\boldsymbol{\Theta}_N|^{-\frac{\alpha}{p^*}}\log(\frac{n}{M(|\boldsymbol{\Theta}_N|+K^*p^*)})\big),
\end{aligned}$$

where $a_1 > 0$ is a constant, the second inequality follows from Theorem 1 of [6], and the first equality holds by taking $\bar{\mathcal{X}} = \mathcal{X}^*_{B^*}$, and $\mathcal{X}^*_{B^*} \in \mathcal{D}_C$, since

$$\begin{aligned}
U(\mathcal{X}^*_{B^*}) &\leq \frac{1}{\max(1, \Gamma/U(\mathcal{X}^*))} \sup_{\boldsymbol{X}\in[0,1]^{|\mathcal{S}|\times|\mathcal{S}|}} \mathrm{Tr}(\boldsymbol{X}\boldsymbol{Q}\boldsymbol{X}^\top) \\
&\leq \sup_{\boldsymbol{X}\in[0,1]^{|\mathcal{S}|\times|\mathcal{S}|}} \frac{1}{\max(1, \Gamma/U(\mathcal{X}^*))} \sum_i \sigma_i\sigma_i(\boldsymbol{X}^\top\boldsymbol{X}) \\
&\leq \frac{p^*|\mathcal{S}|\mathrm{Tr}(\boldsymbol{Q})}{\max(1, \Gamma/U(\mathcal{X}^*))} = \min(\Gamma, U(\mathcal{X}^*)) \leq C,
\end{aligned}$$

where $\sigma_i(\boldsymbol{X}^\top\boldsymbol{X})$ is the i-th eigenvalue of $\boldsymbol{X}^\top\boldsymbol{X}$. This completes the proof. □

Proof of Lemma 3. Note that (14) and (17) are convex minimization. Then convergence of Algorithm 1 follows from Theorem 4.7 of [1]. Moreover, local minimum of the solution follows from [5]. This completes the proof. □

Proof of Lemma 4. Note that the definition of a stationary point in Proposition 4 of [2] is in fact a local minimizer. By Proposition 4 there, we obtain the desired result. This completes the proof. □

Proof of Lemma 5. Let $\boldsymbol{g} = (\mathbb{E}(Y|S = s_1), \cdots, \mathbb{E}(Y|S = s_{|\mathcal{S}|}))^\top$ and $\mathcal{B}_{|\mathcal{S}|}$ is a unit L_2-ball in $\mathbb{R}^{|\mathcal{S}|}$. Then, (27) is equivalent to the existence of linear equations $\boldsymbol{g} = \widetilde{\boldsymbol{X}}^\top\boldsymbol{\beta}$ for any $\boldsymbol{g} \in \mathcal{B}_{|\mathcal{S}|}$ and a given embedding matrix $\widetilde{\boldsymbol{X}}$. By the Rouche-Capelli Theorem [4], it equivalent to $\text{Rank}(\widetilde{\boldsymbol{X}}) = \text{Rank}([\widetilde{\boldsymbol{X}}; \boldsymbol{g}])$ for any $\boldsymbol{g} \in \mathcal{B}_{|\mathcal{S}|}$, where $[\widetilde{\boldsymbol{X}}; \boldsymbol{g}]$ is the augmented matrix. Then, the sufficiency is established. Now, we prove the necessity by contradiction. Suppose $\text{Rank}(\widetilde{\boldsymbol{X}}) < |\mathcal{S}|$. Then, $\text{Span}(\widetilde{\boldsymbol{X}})^\perp \neq \emptyset$, and thus $\text{Rank}([\widetilde{\boldsymbol{X}}; \boldsymbol{g}]) > \text{Rank}(\widetilde{\boldsymbol{X}})$ when $\boldsymbol{g} = \boldsymbol{z}/\|\boldsymbol{z}\|_2$, $\boldsymbol{z} \in \text{Span}(\widetilde{\boldsymbol{X}})^\perp$. This contradicts to that (27). This completes the proof. □

Lemma 6. *Consider a function space of M-depth ReLU neural networks defined as*

$$\mathcal{F}(r) = \big\{\mathbf{f}_M(\boldsymbol{x}), \textit{with } (\mathbf{f}_m(\boldsymbol{x}) = \boldsymbol{\sigma}(\boldsymbol{A}_m\mathbf{f}_{m-1}(\boldsymbol{x}) + \boldsymbol{b}_m))_{m=1}^M : (\boldsymbol{A}, \boldsymbol{b}) \in \boldsymbol{\Theta}_N(r)\big\},$$

where $\boldsymbol{\sigma}(\boldsymbol{z}_m) = (\sigma(\boldsymbol{z}_{m,1}), \cdots, \sigma(\boldsymbol{z}_{m,h_m}))^\top$, $\sigma(\cdot)$ is a rectified linear unit (ReLU), $\boldsymbol{\Theta}_N(r) = \big\{(\boldsymbol{A}, \boldsymbol{b}) : \boldsymbol{A} = (\boldsymbol{A}_m)_{m=1}^M, \boldsymbol{b} = (\boldsymbol{b}_m)_{m=1}^M, \|\boldsymbol{A}\|_{1,1} + \|\boldsymbol{b}\|_{1,1} = \sum_{m=1}^M \big(\sum_{j=1}^{h_m} \|\boldsymbol{A}_{m,j}\|_1 + \|\boldsymbol{b}_m\|_1\big) \leq r\big\}$, and $\boldsymbol{x} \in \mathcal{I} \subset [-B, B]^p$, then for any $\tilde{\mathbf{f}}_M$ and $\mathbf{f}_M$ in $\mathcal{F}(r)$, we have

$$H_\infty(\mathcal{F}(r), \nu) \leq |\boldsymbol{\Theta}_N| \log\Big(\frac{6r}{\nu}\big(MB(\frac{2r}{M-1})^{M-1} + (\frac{r(1-r^M)}{1-r})^2\big)\Big),$$

where $|\boldsymbol{\Theta}_N|$ is the number of parameters in the neural network.

Proof of Lemma 6. Let $\Lambda_m = \|\boldsymbol{A}_m\|_{1,1}$ and $\widetilde{\Lambda}_m = \|\widetilde{\boldsymbol{A}}_m\|_{1,1}$, for any $\mathbf{f}_M$ and $\tilde{\mathbf{f}}_M$ in $\mathcal{F}(r)$, we have

$$\begin{aligned}
&\max_{1\leq j\leq h_m} |\big(\mathbf{f}_m(\boldsymbol{x}) - \tilde{\mathbf{f}}_m(\boldsymbol{x})\big)_j| = \max_{1\leq j\leq h_m} \big|\boldsymbol{\sigma}_j\big(\boldsymbol{A}_m^\top\mathbf{f}_{m-1}(\boldsymbol{x}) + \boldsymbol{b}_m\big) - \boldsymbol{\sigma}_j\big(\widetilde{\boldsymbol{A}}_m^\top\tilde{\mathbf{f}}_{m-1}(\boldsymbol{x}) + \tilde{\boldsymbol{b}}_m\big)\big| \\
&\leq \max_{1\leq j\leq h_m} \big|(\boldsymbol{A}_m)_j^\top\mathbf{f}_{m-1}(\boldsymbol{x}) - (\widetilde{\boldsymbol{A}}_m)_j^\top\tilde{\mathbf{f}}_{m-1}(\boldsymbol{x})\big| + \|\boldsymbol{b}_m - \tilde{\boldsymbol{b}}_m\|_\infty \\
&\leq \|\boldsymbol{A}_m - \widetilde{\boldsymbol{A}}_m\|_{1,1} \max_{1\leq j\leq h_m} |(\mathbf{f}_{m-1}(\boldsymbol{x}))_j| + \tilde{\Lambda}_m \max_{1\leq j\leq h_m} |(\mathbf{f}_{m-1}(\boldsymbol{x}) - \tilde{\mathbf{f}}_{m-1}(\boldsymbol{x}))_j| + \|\boldsymbol{b}_m - \tilde{\boldsymbol{b}}_m\|_\infty \\
&\leq \|\boldsymbol{A}_m - \widetilde{\boldsymbol{A}}_m\|_{1,1}\big(\Lambda_{m-1}\cdots\Lambda_1\|\boldsymbol{x}\|_\infty + \frac{r(1-r^m)}{1-r}\big) \\
&\qquad\qquad + \tilde{\Lambda}_m \max_{1\leq j\leq h_m} |(\mathbf{f}_{m-1}(\boldsymbol{x}) - \tilde{\mathbf{f}}_{m-1}(\boldsymbol{x}))_j| + \|\boldsymbol{b}_m - \tilde{\boldsymbol{b}}_m\|_\infty,
\end{aligned}$$

where the second last inequality follows from the Cauchy-Schwarz inequality, and the last inequality follows from the fact that

$$\begin{aligned}\max_{1\le j\le h_m}|(\mathbf{f}_m(\boldsymbol{x}))_j| &\le \max_{1\le j\le h_m}\big|\boldsymbol{\sigma}_j\big(\boldsymbol{A}_m^\top \mathbf{f}_{m-1}(\boldsymbol{x})+\boldsymbol{b}_m\big)\big| \le \max_{1\le j\le h_m}\big|(\boldsymbol{A}_m^\top \mathbf{f}_{m-1}(\boldsymbol{x})+\boldsymbol{b}_m)_j\big| \\ &\le \Lambda_m \max_{1\le j\le h_m}|\mathbf{f}_{m-1}(\boldsymbol{x})|+\|\boldsymbol{b}_m\|_\infty.\end{aligned}$$

Next, multiplying both sides by $\tilde{\Lambda}_M\cdots\tilde{\Lambda}_{m+1}$ and summing up from $m=1$ to $m=M$ yield that

$$\begin{aligned}\max_{1\le j\le h_m}\big|f_M(\boldsymbol{x})-\tilde{f}_M(\boldsymbol{x})\big| &\le \big(\Lambda_1\cdots\Lambda_{M-1}+\Lambda_1\cdots\Lambda_{M-2}\tilde{\Lambda}_M\cdots+\tilde{\Lambda}_2\cdots\tilde{\Lambda}_M\big)\|\boldsymbol{A}-\widetilde{\boldsymbol{A}}\|_{1,1}\|\boldsymbol{x}\|_\infty \\ &\quad+\big(1+\tilde{\Lambda}_M\cdots+\tilde{\Lambda}_2\cdots\tilde{\Lambda}_M\big)\big(\frac{r(1-r^M)}{1-r}\|\boldsymbol{A}-\widetilde{\boldsymbol{A}}\|_{1,1}+\|\boldsymbol{b}-\tilde{\boldsymbol{b}}\|_\infty\big) \\ &\le \Big(\big(\frac{\Lambda_1+\cdots+\Lambda_M}{M-1}\big)^{M-1}\cdots+\big(\frac{\tilde{\Lambda}_2+\cdots+\tilde{\Lambda}_M}{M-1}\big)^{M-1}\Big)\|\boldsymbol{A}-\widetilde{\boldsymbol{A}}\|_{1,1}\|\boldsymbol{x}\|_\infty \\ &\quad+\Big(1+r+\cdots+\big(\frac{r}{M-1}\big)^{M-1}\Big)\big(\frac{r(1-r^M)}{1-r}\|\boldsymbol{A}-\widetilde{\boldsymbol{A}}\|_{1,1}+\|\boldsymbol{b}-\tilde{\boldsymbol{b}}\|_\infty\big) \\ &\le \Big(M\big(\frac{2r}{M-1}\big)^{M-1}\|\boldsymbol{x}\|_\infty+\big(\frac{r(1-r^M)}{1-r}\big)^2\Big)\Big(\big\|\boldsymbol{A}-\widetilde{\boldsymbol{A}}\big\|_{1,1}+\|\boldsymbol{b}-\tilde{\boldsymbol{b}}\|_{1,1}\Big).\end{aligned}$$

Taking $\sup_{\boldsymbol{x}\in\mathcal{I}}$ on both sides yields that

$$\begin{aligned}H_\infty(\mathcal{F}(r),\nu) &\le \log\mathcal{N}\Big(\boldsymbol{\Theta}_N(r),\nu/\Big(2MB\big(\frac{2r}{M-1}\big)^{M-1}+2\big(\frac{r(1-r^M)}{1-r}\big)^2\Big)\Big) \\ &\le |\boldsymbol{\Theta}_N|\log\Big(\frac{6r}{\nu}\big(MB\big(\frac{2r}{M-1}\big)^{M-1}+\big(\frac{r(1-r^M)}{1-r}\big)^2\big)\Big).\end{aligned}$$

This completes the proof. □

Lemma 7. *Let* $\mathcal{X}_C=\big\{vec(\boldsymbol{X})\in\mathbb{R}^{p|\mathcal{S}|}:\ Tr(\boldsymbol{X}\boldsymbol{Q}\boldsymbol{X}^T)\le 2C,\ \|vec(\boldsymbol{X})\|_\infty\le B\big\}$*, then its covering entropy under the infinity norm is upper bounded by*

$$\log\mathcal{N}(\mathcal{X}_C,\nu)\le p\sum_{u=1}^{|\mathcal{S}|}\max\Big(\log\big(\frac{\|\boldsymbol{P}\|_\infty\min(\sqrt{C\sigma_u^{-1}},B\|\boldsymbol{P}\|_1)}{\sqrt{2}\nu}\big),0\Big),$$

where $\mathcal{N}(\mathcal{X}_C, \nu)$ is the covering number of $\mathcal{X}_C$ under $\|\cdot\|_\infty$, $\sigma_u = \boldsymbol{\Sigma}_{uu}$ is the u-th largest eigenvalue of $\boldsymbol{Q}$, $\boldsymbol{P}$ is an orthogonal matrix whose columns are the eigenvectors of $\boldsymbol{Q}$, $\boldsymbol{\Sigma}$ is a diagonal matrix whose entries are the eigenvalues of $\boldsymbol{Q}$, and $\boldsymbol{Q} = \boldsymbol{P}\boldsymbol{\Sigma}\boldsymbol{P}^T$.

Proof of Lemma 7. Let $\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{P}$, and $\boldsymbol{X} = \boldsymbol{Z}\boldsymbol{P}^T$, and $\mathcal{Z}_C = \{\mathrm{vec}(\boldsymbol{Z}) \in \mathbb{R}^{p|\mathcal{S}|} : \mathrm{Tr}(\boldsymbol{Z}\boldsymbol{\Sigma}\boldsymbol{Z}^T) \leq 2C,\ \|\mathrm{vec}(\boldsymbol{Z})\|_\infty \leq B\|\boldsymbol{P}\|_1\}$, and $(B_i)_{i=1}^N$ are balls centered at $\mathrm{vec}(\boldsymbol{C}_i)$ with radius ν to cover $\mathcal{Z}_C$, then we define $\tilde{B}_i = \big\{(\boldsymbol{P} \otimes \boldsymbol{I}_p)\mathrm{vec}(\boldsymbol{Z}) : \mathrm{vec}(\boldsymbol{Z}) \in B_i\big\}$ with center $\mathrm{vec}(\tilde{\boldsymbol{C}}_i) = (\boldsymbol{P} \otimes \boldsymbol{I}_p)\mathrm{vec}(\boldsymbol{C}_i)$, then for any $\mathrm{vec}(\boldsymbol{X}) \in \tilde{B}_i$, there exists $\mathrm{vec}(\boldsymbol{Z}) = (\boldsymbol{P} \otimes \boldsymbol{I}_p)^\top \mathrm{vec}(\boldsymbol{X})$, satisfying $\mathrm{Tr}(\boldsymbol{Z}\boldsymbol{\Sigma}\boldsymbol{Z}^\top) \leq 2C$, and $\|\mathrm{vec}(\boldsymbol{Z})\|_\infty = \|(\boldsymbol{P} \otimes \boldsymbol{I}_p)^\top \mathrm{vec}(\boldsymbol{X})\|_\infty \leq \|\boldsymbol{P}^\top\|_\infty \|\mathrm{vec}(\boldsymbol{X})\|_\infty \leq B\|\boldsymbol{P}\|_1$, thus there exists a ball $\in B_i$ centered at $\boldsymbol{C}_i$, such that $\mathrm{vec}(\boldsymbol{Z}) \in B_i$ and

$$\begin{aligned}\|\mathrm{vec}(\boldsymbol{X}) - \mathrm{vec}(\tilde{\boldsymbol{C}})\|_\infty &\leq \|(\boldsymbol{P} \otimes \boldsymbol{I}_p)\big(\mathrm{vec}(\boldsymbol{Z}) - \mathrm{vec}(\boldsymbol{C}_i)\big)\|_\infty \\ &= \|\boldsymbol{P} \otimes \boldsymbol{I}_p\|_\infty \|\mathrm{vec}(\boldsymbol{Z}) - \mathrm{vec}(\boldsymbol{C}_i)\|_\infty \leq \nu\|\boldsymbol{P}\|_\infty,\end{aligned}$$

where $\otimes$ is the Kronecker product. Thus, the covering entropy of $\mathcal{X}_C$ is upper bounded by the covering entropy of $\mathcal{Z}_C$, that is, $\log \mathcal{N}(\mathcal{X}_C, \nu) \leq \log \mathcal{N}(\mathcal{Z}_C, \nu/\|\boldsymbol{P}\|_\infty)$, and it suffices to compute the covering number of $\mathcal{Z}_C$. Note that

$$\frac{1}{2C}\mathrm{Tr}(\boldsymbol{Z}\boldsymbol{\Lambda}\boldsymbol{Z}^T) = \frac{1}{2}\sum_{u=1}^{|\mathcal{S}|}\sum_{j=1}^{p}\sigma_u z_{uj}^2/C = \sum_{u=1}^{|\mathcal{S}|}\sum_{j=1}^{p}\big(\frac{z_{uj}}{\sqrt{2C/\sigma_u}}\big)^2 \leq 1, \tag{2}$$

which stands for an ellipse in $\mathbb{R}^{p|\mathcal{S}|}$, with axles $\big\{\sqrt{2C/\sigma_u}\big\}_{u=1}^{|\mathcal{S}|}$. Note that $\|\mathrm{vec}(\boldsymbol{Z})\|_\infty \leq B\|\boldsymbol{P}\|_1$, by partitioning the axle in $\min(\sqrt{2C/\sigma_u}, B\|\boldsymbol{P}\|_1)/(2\nu)$ parts, then the radius for each segment will less or equal than ν. Thus,

$$\log \mathcal{N}(\mathcal{Z}_C, \nu) \leq p\sum_{u=1}^{|\mathcal{S}|}\max\Big(\log\big(\frac{\min(\sqrt{C\sigma_u^{-1}}, B\|\boldsymbol{P}\|_1)}{\sqrt{2}\nu}\big), 0\Big).$$

The desired result then follows. □

The Armijo search in Algorithm 2. Let $\boldsymbol{\Theta}^{(k+1)}(\gamma^{(k)})$ denote $(\boldsymbol{A}_m{}^{(k+1)}, \boldsymbol{b}_m{}^{(k+1)})_{m=1}^M$ in

(23). Let $\boldsymbol{X}^{(k+1)}(\gamma^{(k)})$ be the embedding in (22) with step size $\gamma^{(k)}$, $V_n^{(k)} = V_n(\boldsymbol{\Theta}^{(k)}, \boldsymbol{X}^{(k)})$ and $\gamma^{(k)} = 2^{-t(k)}$ is determined with the Armijo search with,

$$t(k) = \min\Big\{j \in \mathbb{Z}^+ : V_n\big(\boldsymbol{\Theta}^{(k+1)}(2^{-j}), \boldsymbol{X}^{(k+1)}(2^{-j})\big) \\ \leq V_n^{(k)} - \phi 2^{-j}\mathrm{vec}(\frac{\partial V_n^{(k)}}{\partial \boldsymbol{\Theta}^{(k)}})^\top \mathrm{vec}(\frac{\partial V_n^{(k)}}{\partial \boldsymbol{\Theta}^{(k)}}) - \phi 2^{-j}\mathrm{vec}(\frac{\partial V_n^{(k)}}{\boldsymbol{X}^{(k)}})^\top \mathrm{vec}(\boldsymbol{X}^{(k)} - \boldsymbol{Z}^{(k)})\Big\},$$

for some $\phi \in (0, 1)$, $\frac{\partial V_n^{(k)}}{\partial \boldsymbol{\Theta}^{(k)}}$ and $\frac{\partial V_n^{(k)}}{\partial \boldsymbol{X}^{(k)}}$ are given in (20) and (21).

References

[1] Jochen Gorski, Frank Pfeuffer, and Kathrin Klamroth. Biconvex sets and optimization with biconvex functions: a survey and extensions. *Mathematical Methods of Operations Research*, 66(3):373–407, 2007.

[2] A. N. Iusem. On the convergence properties of the projected gradient method for convex optimization. *Computational and Applied Mathematics*, 22(1):37–52, 2003.

[3] Xiaotong Shen, Lifeng Wang, et al. Generalization error for multi-class margin classification. *Electronic Journal of Statistics*, 1:307–330, 2007.

[4] PK Suetin, Alexandra I Kostrikin, and Yu I Manin. *Linear algebra and geometry*. CRC Press, 1989.

[5] Paul Tseng. Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of Optimization Theory and Applications*, 109(3):475–494, 2001.

[6] Dmitry Yarotsky. Error bounds for approximations with deep relu networks. *Neural Networks*, 94:103–114, 2017.