Online Appendix of Deep Autoencoders for Nonlinear Factor Models: Theory and Applications

Zhouyu Shen*
Booth School of Business
University of Chicago

Dacheng Xiu†
Booth School of Business
University of Chicago and NBER

Abstract

This appendix presents supporting mathematical results for the proofs in the main text.

A Supplemental Mathematical Proofs

A.1 Approximation Error of Neural Networks

Proposition A1. *Assume $f \in \mathcal{H}^p([-a,a]^r, C)$ for some $p = q + s$, $q \in \mathbb{N}_0$ and $s \in (0,1]$, and $C > 0$. Suppose $a \geq 1$ and ζ is sufficiently large (depending only on fixed constants including p, a, C and $\|f\|_{C^q}$). If $d \asymp \lceil \log_4(\zeta^{2p}) \rceil \cdot (\lceil \log_2(\max\{q,r\}+1\}) \rceil + 1)$ and $w \asymp 2^r \cdot \binom{r+q}{r} \cdot r^2 \cdot (q+1) \cdot \zeta^r$, then there exists a neural network $\widehat{f}_{wide} \in \mathcal{F}_r^1(d, w, \zeta^{5p+5}, B)$ such that $\left\| f - \widehat{f}_{wide} \right\|_\infty \leq C\zeta^{-2p}$, for some constant C depending only on fixed parameters.*

Proof. The proof largely follows the argument in Kohler and Langer (2021). The key distinction lies in ensuring that the neural networks used in our construction have uniformly bounded weights, enabling us to establish an explicit bound on the estimation error of AEs. We begin by introducing the notations required for the proof.

Notations: Let $x = (x^{(1)}, \ldots, x^{(r)}) \in \mathbb{R}^r$ denote a generic vector. Let $C \subset \mathbb{R}^r$ be an r-dimensional half-open cube that takes the form $[\alpha, \beta) = \left[\alpha^{(1)}, \beta^{(1)}\right) \times \cdots \times \left[\alpha^{(r)}, \beta^{(r)}\right)$, with $\alpha, \beta \in \mathbb{R}^r$. We denote by C_{left} the bottom-left corner of C: $(\alpha^{(1)}, \alpha^{(2)}, \ldots, \alpha^{(r)})$, that

*Address: 5807 S Woodlawn Avenue, Chicago, IL 60637 USA. Email: zshen10@chicagobooth.edu.
†Address: 5807 S Woodlawn Avenue, Chicago, IL 60637, USA. Email: dacheng.xiu@chicagobooth.edu.

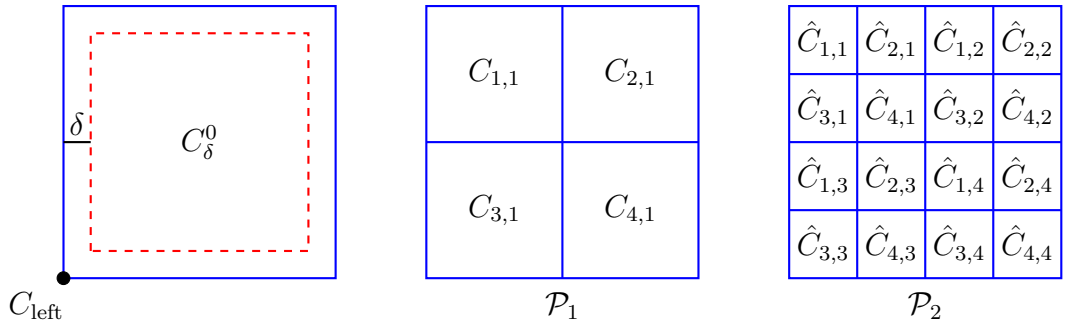

Figure 1: Illustration of Notations on Cubes

is, the point in $\mathbb{R}^r$ where all coordinate-wise lower bounds are attained; see the left panel of Figure 1 for illustration. Then each such cube C with side length s can be expressed as the polytope: $-x^{(j)} + C^{(j)}_{\text{left}} \leq 0$ and $x^{(j)} - C^{(j)}_{\text{left}} - s < 0$, for all $j \in \{1, \ldots, r\}$.

We define the interior-shrunken cube $C^0_\delta \subset C$ to be the subset of points that are at least δ away from the boundary of C, i.e., $-x^{(j)} + C^{(j)}_{\text{left}} \leq -\delta$ and $x^{(j)} - C^{(j)}_{\text{left}} - s < -\delta$, for all $j \in \{1, \ldots, r\}$.

Below we consider partitions and geometric constructions within subsets of $[-a, a)^r$. If $\mathcal{P}$ is a partition of $[-a, a)^r$ into cubes, and $x \in [-a, a)^r$, we denote by $C_{\mathcal{P}}(x)$ the unique cube $C \in \mathcal{P}$ such that $x \in C$.

We partition $[-a, a)^r$ into ζ^r and ζ^{2r} numbers of half-open, equi-volume cubes. Let

$$\mathcal{P}_1 = \{C_{k,1}\}_{k \in \{1,\ldots,\zeta^r\}} \text{ and } \mathcal{P}_2 = \{C_{j,2}\}_{j \in \{1,\ldots,\zeta^{2r}\}} \tag{A1}$$

denote the corresponding partitions of $[-a, a)^r$, respectively. For each $i \in \{1, \ldots, \zeta^r\}$, we denote by $\widehat{C}_{1,i}, \ldots, \widehat{C}_{\zeta^r,i}$ the cubes in $\mathcal{P}_2$ that are contained within $C_{i,1}$. We order these subcubes so that their bottom-left corners satisfy $\left(\widehat{C}_{k,i}\right)_{\text{left}} = (C_{i,1})_{\text{left}} + v_k$ for all $k \in \{1, \ldots, \zeta^r\}$ and $i \in \{1, \ldots, \zeta^r\}$, where v_k is a vector whose entries lie in the set $\{0, 2a/\zeta^2, \ldots, (\zeta - 1) \cdot 2a/\zeta^2\}$. Note that the collection of all $\widehat{C}_{k,i}$'s coincides with the elements of $\mathcal{P}_2$, but they are arranged in this specific order and thus given a new notation. See the middle and right panels of Figure 1 for an illustration. Each vector v_k specifies the relative position of the subcube $\left(\widehat{C}_{k,i}\right)_{\text{left}}$ within its parent cube $C_{i,1}$. Without loss of generality, the ordering is chosen such that these relative positions are the same for each i.

With these notations established, the proof proceeds by constructing neural networks with bounded weights, following the approach in the proof of Theorem 2(a) of Kohler and Langer (2021). The proof is divided into three steps, corresponding to steps 2–4 in their

construction.

A.1.1 Step I: An Initial Approximation

We begin by constructing a neural network, denoted by $\widehat{f}_{\mathcal{P}_2}(x)$, that approximates the target function $f(x)$ for all $x \in \bigcup_{j\in\{1,\ldots,\zeta^{2r}\}} (C_{j,2})^0_{1/\zeta^{2p+2}}$.

To establish this, we prove in Lemma 2 that there exists a weight-bounded version of the network $\widehat{f}_p$—originally defined in Lemma 5 of Kohler and Langer (2021)—that approximates monomials with the same error bounds as theirs, but using twenty times as many layers.

We define the identity function as implemented by a ReLU neural network, constructed in a way that respects our bounded-weight constraint. For any scalar $z \in \mathbb{R}$, let $\widehat{f}_{id}(z) = \sigma(z) - \sigma(-z) = z$, where σ denotes the ReLU activation $\max(z, 0)$. For a vector $x = \left(x^{(1)}, \ldots, x^{(r)}\right) \in \mathbb{R}^r$, we define $\widehat{f}_{id}(x) = \left(\widehat{f}_{id}\left(x^{(1)}\right), \ldots, \widehat{f}_{id}\left(x^{(r)}\right)\right) = x$. This function can be interpreted as a neural network realizing the identity map on $\mathbb{R}^r$. Its iterates are defined recursively by $\widehat{f}^{t+1}_{id}(x) = \widehat{f}_{id}\left(\widehat{f}^{t}_{id}(x)\right) = x$, for all $t \in \mathbb{N}_+, x \in \mathbb{R}^r$. The network uses ReLU activations with weights in $\{\pm 1\}$, which are trivially bounded by ζ^{5p+5} and thus satisfy the required weight constraint.

We also adopt the neural networks $\widehat{f}_{\text{ind},[a,b)}$ and $\widehat{f}_{\text{test}}(x, a, b, s)$ as introduced in Lemma 6 of their paper:

$$
\begin{aligned}
&\widehat{f}_{\text{ind},[a,b)}(x) = \sigma\left(1 - \zeta^{2p+2} \cdot \sum_{i=1}^{r} \left(\sigma\left(a^{(i)} + \zeta^{-2p-2} - x^{(i)}\right) + \sigma\left(x^{(i)} - b^{(i)} + \zeta^{-2p-2}\right)\right)\right) \\
&\widehat{f}_{\text{test}}(x, a, b, s) \qquad\qquad (\text{A2}) \\
&= \sigma\left(\widehat{f}_{id}(s) - \zeta^{4p+4} \cdot \sum_{i=1}^{r} \left(\sigma\left(a^{(i)} + \zeta^{-2p-2} - x^{(i)}\right) + \sigma\left(x^{(i)} - b^{(i)} + \zeta^{-2p-2}\right)\right)\right) \\
&\quad - \sigma\left(-\widehat{f}_{id}(s) - \zeta^{4p+4} \cdot \sum_{i=1}^{r} \left(\sigma\left(a^{(i)} + \zeta^{-2p-2} - x^{(i)}\right) + \sigma\left(x^{(i)} - b^{(i)} + \zeta^{-2p-2}\right)\right)\right).
\end{aligned}
$$

Note that the neural network $\widehat{f}_{\text{ind},[a,b)}(x)$, which belongs to the class $\mathcal{F}^1_r(2, 2d, \zeta^{5p+5})$, satisfies $\widehat{f}_{\text{ind},[a,b)}(x) = \mathbf{1}_{[a,b)}(x)$ for all x in the set $K_\zeta := \{x \in \mathbb{R}^r : x^{(i)} \notin \left[a^{(i)}, a^{(i)} + \zeta^{-2p-2}\right) \cup \left(b^{(i)} - \zeta^{-2p-2}, b^{(i)}\right), \forall i \in \{1, \ldots, r\}\}$, and further satisfies $|\widehat{f}_{\text{ind},[a,b)}(x) - \mathbf{1}_{[a,b)}(x)| \leq 1$ for all $x \in \mathbb{R}^r$. Similarly, the function $\widehat{f}_{\text{test}}(x, a, b, s)$, also in $\mathcal{F}^1_r(2, 2d, \zeta^{5p+5})$, satisfies $\widehat{f}_{\text{test}}(x, a, b, s) = s\mathbf{1}_{[a,b)}(x)$ for all $x \in K_{1/R}$ (defined above with ζ replaced by $1/R$), and $|\widehat{f}_{\text{test}}(x, a, b, s) - s\mathbf{1}_{[a,b)}(x)| \leq |s|$ for $x \in \mathbb{R}^r$.

Additionally, we consider the neural networks introduced in their proof of Lemma 3.

$$
\begin{aligned}
&\widehat{\phi}_{1,1} = \left(\widehat{\phi}_{1,1}^{(1)}, \ldots, \widehat{\phi}_{1,1}^{(r)}\right) = \widehat{f}_{id}^{2}(x), \\
&\widehat{\phi}_{2,1} = \left(\widehat{\phi}_{2,1}^{(1)}, \ldots, \widehat{\phi}_{2,1}^{(r)}\right) = \sum_{i\in\{1,\ldots,\zeta^r\}} (C_{i,1})_{left} \cdot \widehat{f}_{ind,C_{i,1}}(x), \\
&\widehat{\phi}_{3,1}^{(l,j)} = \sum_{i\in\{1,\ldots,\zeta^r\}} \left(D^l f\right)\left(\left(\tilde{C}_{j,i}\right)_{left}\right) \cdot \widehat{f}_{ind,C_{i,1}}(x), \\
&\widehat{\phi}_{1,2} = \left(\widehat{\phi}_{1,2}^{(1)}, \ldots, \widehat{\phi}_{1,2}^{(r)}\right) = \widehat{f}_{id}^{2}\left(\widehat{\phi}_{1,1}\right), \\
&\widehat{\phi}_{2,2}^{(i)} = \sum_{j=1}^{\zeta^r} \widehat{f}_{\text{test}}\left(\widehat{\phi}_{1,1}, \widehat{\phi}_{2,1} + v_j, \widehat{\phi}_{2,1} + v_j + \frac{2a}{\zeta^2}\cdot 1, \widehat{\phi}_{2,1}^{(i)} + v_j^{(i)}\right), \\
&\widehat{\phi}_{3,2}^{(l)} = \sum_{j=1}^{\zeta^r} \widehat{f}_{\text{test}}\left(\widehat{\phi}_{1,1}, \widehat{\phi}_{2,1} + v_j, \widehat{\phi}_{2,1} + v_j + \frac{2a}{\zeta^2}\cdot 1, \widehat{\phi}_{3,1}^{(1,j)}\right),
\end{aligned}
\tag{A3}
$$

for $j \in \{1, \ldots, \zeta^r\}$, $l \in \mathbb{N}_0^d$ with $\|l\|_1 \leq q$ and $i \in \{1, \ldots, r\}$. These functions play an essential role in constructing the neural network that approximates the target function. It is straightforward to verify that all weight parameters of the above networks are bounded by ζ^{5p+5}, assuming ζ is sufficiently large and that $\|f\|_{C^q}$ is finite.

Next, let $l_1, \ldots, l_{\binom{r+q}{r}}$ be the collection of all multi-indices $l_i = (s_1, \ldots, s_r) \in \mathbb{N}_0^r$ such that $s_1 + \cdots + s_r \leq q$. That is, $\left\{l_1, \ldots, l_{\binom{r+q}{r}}\right\} = \{(s_1, \ldots, s_r) \in \mathbb{N}_0^r : s_1 + \cdots + s_r \leq q\}$. For each i, we define the factorial of the multi-index $l_i = (s_1, \ldots, s_r)$ and the corresponding monomial as

$$
l_i! = s_1! \cdots s_r!, \quad m_i(z) = z^{l_i} = \left(z^{(1)}\right)^{s_1} \cdots \left(z^{(r)}\right)^{s_r}, \tag{A4}
$$

where $z = (z^{(1)}, \ldots, z^{(r)}) \in \mathbb{R}^r$. We then define the function $p\left(z, y_1, \ldots, y_{\binom{r+q}{r}}\right) = \sum_{i=1}^{\binom{r+q}{r}} c_i\, y_i\, m_i(z)$, where $c_i = \frac{1}{l_i!}$.

By Lemma 2, $\max_i |c_i| = 1$, and the fact that $(4\max\{2a, \|f\|_{C^q}\})^{2q+2} < \zeta^{5p+5}$ for sufficiently large ζ, writting $B_{\zeta,p} := \lceil \log_4(\zeta^{2p}) \rceil$, there exists a function

$$
\widehat{f}_p \in \mathcal{F}^1_{\binom{r+q}{r}+r}\left(20B_{\zeta,p} \cdot \lceil \log_2(\max\{q+1, 2\}) \rceil, 18\cdot(q+1)\cdot\binom{r+q}{r}, \zeta^{5p+5}\right)
$$

such that for all $|z^{(1)}|, \ldots, |z^{(r)}|, |y_1|, \ldots, |y_{\binom{r+q}{r}}| \leq \max\{2a, \|f\|_{C^q}\}$, it holds that

$$
\left|\widehat{f}_p\left(z, y_1, \ldots, y_{\binom{r+q}{r}}\right) - p\left(z, y_1, \ldots, y_{\binom{r+q}{r}}\right)\right| \leq C \cdot (\max\{2a, \|f\|_{C^q}\})^{4(q+1)} \cdot 4^{-B_{\zeta,p}}
$$

for some fixed constant C.

Finally, we define the network

$$\widehat{f}_{\mathcal{P}_2}(x) := \widehat{f}_p\left(\widehat{\phi}_{1,2}(x) - \widehat{\phi}_{2,2}(x), y_1, \ldots, y_{\binom{r+q}{r}}\right) \tag{A5}$$

where $y_v := \widehat{\phi}_{3,2}^{(l_v)}(x)$ for $v \in \{1, \ldots, \binom{r+q}{r}\}$. This function belongs to the class $\mathcal{F}_r^1(d, w, \zeta^{5p+5})$ with $d = 80 + 20B_{\zeta,p} \cdot \lceil \log_2(\max\{q+1, 2\}) \rceil$ and $w = \max(\binom{r+q}{r} \cdot \zeta^r \cdot 2 \cdot (2+2r) + 2r, 18 \cdot (q+1) \cdot \binom{r+q}{r})$. Following an argument analogous to Lemma 3 in Kohler and Langer (2021), it can be shown that $\left|\widehat{f}_{\mathcal{P}_2}(x) - f(x)\right| \leq C \cdot (\max\{2a, \|f\|_{C^q}\})^{4(q+1)} \cdot \zeta^{-2p}$ for all $x \in \bigcup_{j \in \{1,\ldots,\zeta^{2r}\}} (C_{j,2})^0_{1/\zeta^{2p+2}}$. Moreover, for all $x \in [-a, a)^r$, the network output remains uniformly bounded: $\left|\widehat{f}_{\mathcal{P}_2}(x)\right| \leq 2 \cdot e^{2ad} \cdot \max\{\|f\|_{C^q}, 1\}$. That is, $\widehat{f}_{\mathcal{P}_2}$ provides a good approximation of f over this region and maintains boundedness outside it.

A.1.2 Step II: Approximating $w_{\mathcal{P}_2}(x) \cdot f(x)$

Define

$$w_{\mathcal{P}_2}(x) = \prod_{j=1}^{r} \left(1 - \frac{\zeta^2}{a} \cdot \left|(C_{\mathcal{P}_2}(x))^{(j)}_{\text{left}} + \frac{a}{\zeta^2} - x^{(j)}\right|\right)_+, \tag{A6}$$

where $()_+$ denotes the positive part. This function is a linear tensor-product B-spline that attains its maximum at the center of the cube $C_{\mathcal{P}_2}(x)$.

In this section, we construct a neural network that approximates the product $w_{\mathcal{P}_2}(x) \cdot f(x)$. To this end, we introduce two weight-bounded networks, $\widehat{f}_{w_{\mathcal{P}_2}}$ and $\widehat{f}_{check,\mathcal{P}_2}$, as established in Lemmas 3 and 4, respectively. These are modified versions of the corresponding functions defined in Lemmas 9 and 10 of Kohler and Langer (2021), with parameters bounded in magnitude. Based on the previous section and Lemma 3, the networks $\widehat{f}_{\mathcal{P}_2}$ and $\widehat{f}_{w_{\mathcal{P}_2}}$ provide accurate approximations to f and $w_{\mathcal{P}_2}$, respectively, except over a small region. The network $\widehat{f}_{check,\mathcal{P}_2}$ is designed to detect whether a given input x lies within that exceptional region. Based on the previous construction, we define the network

$$\widehat{f}_{\mathcal{P}_2,\text{true}}(x) = \sigma\left(\widehat{f}_{\mathcal{P}_2}(x) - B_{\text{true}} \cdot \widehat{f}_{\text{check},\mathcal{P}_2}(x)\right) - \sigma\left(-\widehat{f}_{\mathcal{P}_2}(x) - B_{\text{true}} \cdot \widehat{f}_{\text{check},\mathcal{P}_2}(x)\right),$$

where $B_{\text{true}} = 2 \cdot e^{2ar} \cdot \max\{\|f\|_{C^q}, 1\}$. Following the same argument as Lemma 3 in Kohler and Langer (2021), it can be shown that $|\widehat{f}_{\mathcal{P}_2}(x)| \leq B_{\text{true}}$ for all $x \in [-a, a)^r$. Intuitively, $\widehat{f}_{\mathcal{P}_2,\text{true}}$ is designed to equal $\widehat{f}_{\mathcal{P}_2}$ in regions where $\widehat{f}_{\mathcal{P}_2}$ approximates the target function well,

and to be zero in regions where the approximation fails, as identified by the indicator $\widehat{f}_{\text{check},\mathcal{P}_2}$. Since B_{true} and the weights of both $\widehat{f}_{\mathcal{P}_2}$ and $\widehat{f}_{\text{check},\mathcal{P}_2}(x)$ are bounded by ζ^{5p+5} for sufficiently large ζ, it follows that the weights of $\widehat{f}_{\mathcal{P}_2,\text{true}}$ are also bounded by ζ^{5p+5}.

Let $\widehat{f}_{\text{mult}} \in \mathcal{F}_2^1(20\lceil \log_4(\zeta^{2p})\rceil, 18, \zeta^{5p+5})$ be the function defined in Eq. (A9), with c in that definition replaced by $2 \cdot \max\{\|f\|_{\infty,[-a,a]^r}, 1\}$. This function approximates the product of its two inputs. We then define

$$\widehat{f}(x) = \widehat{f}_{\text{mult}}\left(\widehat{f}_{w_{\mathcal{P}_2}}(x), \widehat{f}_{\mathcal{P}_2,\text{true}}(x)\right), \tag{A7}$$

which belongs to $\mathcal{F}_r^1(d, w, \zeta^{5p+5})$ with $d = 100 + 20\lceil \log_4(\zeta^{2p})\rceil \cdot (\lceil \log_2(\max\{q,r\}+1)\rceil + 1)$, and $w = 64 \cdot \binom{r+q}{r} \cdot r^2 \cdot (q+1) \cdot \zeta^r$. Following an argument analogous to Lemma 7 in Kohler and Langer (2021), it can be shown that $\left|\widehat{f}(x) - w_{\mathcal{P}_2}(x) \cdot f(x)\right| \leq C \cdot (\max\{2a, \|f\|_{C^q}\})^{4(q+1)} \cdot \zeta^{-2p}$ holds for $x \in [-a,a)^r$.

A.1.3 Step III: Applying $\widehat{f}$ to Slightly Shifted Partitions

Having constructed a network that accurately approximates the product $w_{\mathcal{P}_2}(x) \cdot f(x)$, we now turn to the task of recovering the function f itself. The weight function $w_{\mathcal{P}_2}(x)$ attains its maximum at the center of the cube $C_{\mathcal{P}_2}(x)$, and serves as a localized window that emphasizes the contribution of f near that center. The key idea is to define a collection of such weight functions whose supports together cover the entire domain and whose values sum to one at each point. By summing the corresponding approximations, such as $\widehat{f}$, over these shifted partitions, we obtain a neural network that effectively approximates the target function f over the entire domain. The remainder of the proof formalizes this construction.

Recall the definition of $\mathcal{P}_1$ and $\mathcal{P}_2$ from Eq. (A1). Let $\mathcal{P}_{1,1} = \mathcal{P}_1$ and $\mathcal{P}_{2,1} = \mathcal{P}_2$. For each $v \in \{2, 3, \ldots, 2^r\}$, define partitions $\mathcal{P}_{1,v}$ and $\mathcal{P}_{2,v}$ as shifted versions of $\mathcal{P}_{1,1}$ and $\mathcal{P}_{2,1}$, where at least one coordinate is shifted by $\zeta^{-2}a$. Specifically, each $\mathcal{P}_{1,v}$ takes the form $\mathcal{P}_1 + \sum_{k\in S} \zeta^{-2} a e_k$ for some subset $S \subset \{1, 2, \ldots, r\}$, where e_k denotes the k-th basis vector in $\mathbb{R}^r$. Corresponding to each pair of partitions $\mathcal{P}_{1,v}$ and $\mathcal{P}_{2,v}$, define the functions $w_{\mathcal{P}_{2,v}}$ and the neural networks $\widehat{f}_v$ using the same construction as in Eq. (A6) and Eq. (A7). That is, each $\widehat{f}_v$ approximates the product $w_{\mathcal{P}_{2,v}}(x) \cdot f(x)$. Importantly, as shown on page 32 of Supplement A of Kohler and Langer (2021), the weight functions satisfy the identity $w_{\mathcal{P}_{2,1}} + \cdots + w_{\mathcal{P}_{2,2^r}} = 1$ for all $x \in [-a/2, a/2]^d$. We now define a wide network as:

$$\widehat{f}_{\text{wide}}(x) = \sum_{v=1}^{2^r} \widehat{f}_v(x) \in \mathcal{F}_r^1(d, w, \zeta^{5p+5})$$

with $d = 100 + 20\lceil \log_4(\zeta^{2p}) \rceil \cdot (\lceil \log_2(\max\{q, r\} + 1) \rceil + 1)$ and $w = 2^r \cdot 64 \cdot \binom{r+q}{r} \cdot r^2 \cdot (q+1) \cdot \zeta^r$. Based on the previous section, we have

$$|\widehat{f}_{\text{wide}}(x) - f(x)| \leq \sum_{v=1}^{2^r} \left| \widehat{f}_v(x) - w_{\mathcal{P}_{2,v}}(x) \cdot f(x) \right| \lesssim \zeta^{-2p}, \tag{A8}$$

for $x \in [-a/2, a/2]^r$. Although the result holds for $x \in [-a/2, a/2]^r$, we can always increase a if necessary. Note that $\widehat{f}_{\text{wide}} \in \mathcal{F}_r^1(d, w, \zeta^{5p+5})$, but it may not be uniformly bounded by a constant B. To enforce boundedness, we append two additional layers to the network, using the function $\sigma(2B - \sigma(B - \widehat{f}_{\text{wide}}(\cdot))) - B$. This modification ensures that the final output belongs to $\mathcal{F}_r^1(d, w, \zeta^{5p+5}, 2B)$, while preserving the value of $\widehat{f}_{\text{wide}}(x)$ whenever $|\widehat{f}_{\text{wide}}(x)| \leq B$. Therefore, Eq. (A8) still holds given that $|f(x)| \leq B$. □

A.2 Proofs of Technical Lemmas

Lemma 1. *There exists* $\widehat{f}_{mult,r} \in \mathcal{F}_r^1(20R\lceil \log_2(r) \rceil, 18r, 4^{2r}a^{2r})$ *such that*

$$|\widehat{f}_{mult,r}(x) - \prod_{i=1}^{r} x^{(i)}| \leq 4^{4r+1} \cdot a^{4r} \cdot r \cdot 4^{-R}, \quad \forall x \in [-a, a]^r,$$

for any $a \geq 1$ *and any* $R \geq \log_4(2 \cdot 4^{2r} \cdot a^{2r})$.

Proof. For any $c > 0$, we construct a neural network $\widehat{f}_{mult} \in \mathcal{F}_2^1(10R, 18, 4c^2)$ satisfying

$$|\widehat{f}_{mult}(x, y) - xy| \leq 2c^2 4^{-R}, \quad \text{for all } x, y \in [-c, c]. \tag{A9}$$

By Lemma A.2 of Schmidt-Hieber (2020), there exists a neural network $\widehat{\phi} \in \mathcal{F}_2^1(2R+5, 6, 1)$, satisfying $\left|\widehat{\phi}(x, y) - xy\right| \leq 2^{-2R-1}$, for all $x, y \in [0, 1]$. Define $\widehat{f}_{mult}(x, y) = 4c^2\widehat{\phi}\left(\frac{x+c}{2c}, \frac{y+c}{2c}\right) - c(x + y) - c^2$. Then $\widehat{f}_{mult} \in \mathcal{F}_2^1(10R, 18, 4c^2)$. Observe that for all $x, y \in [-c, c]$, we have $(x + c)/2c, (y + c)/2c \in [0, 1]$. Therefore,

$$\begin{aligned} |\widehat{f}_{mult}(x, y) - xy| &= \left| 4c^2\widehat{\phi}\left(\frac{x+c}{2c}, \frac{y+c}{2c}\right) - c(x + y) - c^2 - xy \right| \\ &= 4c^2 \left| \widehat{\phi}\left(\frac{x+c}{2c}, \frac{y+c}{2c}\right) - \frac{(x+c)(y+c)}{4c^2} \right| \leq 4c^2 2^{-2R-1}, \end{aligned}$$

which implies Eq. (A9).

To construct $\widehat{f}_{\text{mult},r}$, we use the network $\widehat{f}_{\text{mult}}$ defined in Eq. (A9) with the choice $c = 4^r a^r$. This network satisfies $\widehat{f}_{\text{mult}} \in \mathcal{F}_2^1\left(20R, 18, 4^{2r}a^{2r}\right)$ and guarantees the approximation bound

$$\left|\widehat{f}_{\text{mult}}(x,y) - xy\right| \leq 2 \cdot 4^{2r} a^{2r} \cdot 4^{-R}, \tag{A10}$$

for all $x, y \in [-4^r a^r, 4^r a^r]$.

For $r > 2$, we set $q := \lceil \log_2(r) \rceil$ and define

$$(z_1, \ldots, z_{2^q}) = \left(x^{(1)}, x^{(2)}, \ldots, x^{(r)}, 1, \ldots, 1\right),$$

where the remaining entries are filled with ones to reach dimension 2^q.

We are now ready to construct $\widehat{f}_{\text{mult},r}$. Using the above setup, the first $20R$ layers of the network compute

$$\left(\widehat{f}_{\text{mult}}(z_1, z_2),\ \widehat{f}_{\text{mult}}(z_3, z_4),\ \ldots,\ \widehat{f}_{\text{mult}}(z_{2^q-1}, z_{2^q})\right),$$

which requires $20R$ layers and at most $18 \cdot 2^{q-1} \leq 18r$ neurons. The output at this stage is a vector of length 2^{q-1}.

We then recursively pair neighboring entries and apply $\widehat{f}_{\text{mult}}$ to each pair, halving the dimension at each step. This process continues until a single output remains. The resulting network $\widehat{f}_{\text{mult},r}$ thus belongs to the class $\mathcal{F}_r^1\left(20qR, 18r, 4^{2r}a^{2r}\right)$. The remainder of the proof proceeds along the same lines as Lemma 8 in Kohler and Langer (2021), and is omitted for brevity. □

Lemma 2. *Recall the notation introduced in (A4). Let $m_1, \ldots, m_{\binom{r+n}{r}}$ denote all monomials in $\mathcal{P}_n^r$ for some $n \in \mathbb{N}_+$. Given coefficients $c_1, \ldots, c_{\binom{r+n}{r}} \in \mathbb{R}$, we define the function*

$$p\left(x, y_1, \ldots, y_{\binom{r+n}{r}}\right) = \sum_{i=1}^{\binom{r+n}{r}} c_i \cdot y_i \cdot m_i(x), \quad x \in [-a,a]^r, \quad y_1, \ldots, y_{\binom{r+n}{r}} \in [-a,a].$$

Set $\overline{r}(p) = \max_{1 \leq i \leq \binom{r+n}{r}} |c_i|$. Then for any $a \geq 1$ and $R \geq \log_4(2 \cdot 4^{2\cdot(n+1)} \cdot a^{2\cdot(n+1)})$, there exists a neural network $\widehat{f}_p \in \mathcal{F}^1_{\binom{r+n}{r}+r}(d, w, 4^{2n+2}a^{2n+2} \vee \overline{r}_p)$ with $d = 20R \cdot \lceil \log_2(n+1) \rceil$ and $w = 18 \cdot (n+1) \cdot \binom{r+n}{r}$, such that

$$\left|\widehat{f}_p\left(x, y_1, \ldots, y_{\binom{r+n}{r}}\right) - p\left(x, y_1, \ldots, y_{\binom{r+n}{r}}\right)\right| \leq C \cdot \overline{r}(p) \cdot a^{4(n+1)} \cdot 4^{-R}$$

for $x \in [-a,a]^r, y_1, \ldots, y_{\binom{r+n}{r}} \in [-a,a]$, *where* C *is a constant depending only on* d *and* n.

Proof. By Lemma 1, for each monomial m_i, there exists a neural network $\widehat{f}_{m_i} : \mathbb{R}^{r+1} \to \mathbb{R}$, $\widehat{f}_{m_i} \in \mathcal{F}^1_{r+1}\left(20R \cdot \lceil \log_2(n+1) \rceil, 18 \cdot (n+1), 4^{2n+2}a^{2n+2}\right)$, such that

$$|\widehat{f}_{m_i}(x, y_i) - y_i m_i(x)| \leq 4 \cdot 4^{4(n+1)} \cdot a^{4(n+1)} \cdot (n+1) \cdot 4^{-R}.$$

Define

$$\widehat{f}_p := \sum_{i=1}^{\binom{r+n}{r}} c_i \cdot \widehat{f}_{m_i}(x, y_i).$$

It is straightforward to verify that $\widehat{f}_p \in \mathcal{F}^1_{\binom{r+n}{r}+r}(d, w, 4^{2n+2}a^{2n+2} \vee \overline{r}_p)$ and that

$$\begin{aligned}\left|\widehat{f}_p\left(x, y_1, \ldots, y_{\binom{r+n}{r}}\right) - p\left(x, y_1, \ldots, y_{\binom{r+n}{r}}\right)\right| &\leq \sum_{i=1}^{\binom{r+n}{r}} |c_i| \cdot \left|y_i \cdot m_i(x) - \widehat{f}_{m_i}(x, y_i)\right| \\ &\leq \binom{r+n}{r} \cdot \overline{r}(p) \cdot 4 \cdot 4^{4\cdot(n+1)} \cdot a^{4\cdot(n+1)} \cdot (n+1) \cdot 4^{-R}.\end{aligned}$$

□

Lemma 3. *Let* $1 \leq a < \infty$ *and* $\zeta \geq 4^{4r+1}r$. *For* $w_{\mathcal{P}_2}$ *defined in Eq. (A6), there exists a neural network* $\widehat{f}_{w_{\mathcal{P}_2}} \in \mathcal{F}^1_r\left(d, w, \zeta^{5p+5}\right)$, *with* $d = 100 + 20\lceil \log_4(\zeta^{2p}) \rceil \cdot \lceil \log_2(r) \rceil$ *and* $w = \max\{18r, 2r + r \cdot \zeta^r \cdot 2 \cdot (2+2r)\}$, *such that* $\left|\widehat{f}_{w_{\mathcal{P}_2}}(x) - w_{\mathcal{P}_2}(x)\right| \leq 4^{4r+1} \cdot r \cdot \zeta^{-2p}$ *for* $x \in \bigcup_{i \in \{1,\ldots,\zeta^{2r}\}} (C_{i,2})^0_{1/\zeta^{2p+2}}$ *and* $|\widehat{f}_{w_{\mathcal{P}_2}}(x)| \leq 2$ *for* $x \in [-a,a)^r$.

Proof. We make use of $\widehat{\phi}_{1,2}$ and $\widehat{\phi}_{2,2}$ defined in Eq. (A3). Let

$$\begin{aligned}\widehat{f}_{w_{\mathcal{P}_2,j}}(x) =& \sigma\left(\frac{\zeta^2}{a} \cdot \left(\widehat{\phi}^{(j)}_{1,2} - \widehat{\phi}^{(j)}_{2,2}\right)\right) - 2 \cdot \sigma\left(\frac{\zeta^2}{a} \cdot \left(\widehat{\phi}^{(j)}_{1,2} - \widehat{\phi}^{(j)}_{2,2} - \frac{a}{\zeta^2}\right)\right) \\ &+ \sigma\left(\frac{\zeta^2}{a} \cdot \left(\widehat{\phi}^{(j)}_{1,2} - \widehat{\phi}^{(j)}_{2,2} - \frac{2 \cdot a}{\zeta^2}\right)\right).\end{aligned}$$

Since the weights of $\widehat{\phi}_{1,2}$ and $\widehat{\phi}_{2,2}$ are bounded by ζ^{5p+5}, it follows that $\widehat{f}_{w_{\mathcal{P}_2,j}}(x)$ also has weights bounded by ζ^{5p+5}. We now define the full function $\widehat{f}_{w_{\mathcal{P}_2}}$ as

$$\widehat{f}_{w_{\mathcal{P}_2}}(x) = \widehat{f}_{mult,d}\left(\widehat{f}_{w_{\mathcal{P}_2,1}}(x), \ldots, \widehat{f}_{w_{\mathcal{P}_2,d}}(x)\right),$$

where we set $a = 1$ and $R = 20\lceil \log_4(\zeta^{2p}) \rceil$ as in Lemma 1. It is important to note that the construction of each $\widehat{f}_{w_{\mathcal{P}_2,j}}$ exactly mirrors that in Lemma 9 of Kohler and Langer (2021).

Furthermore, our network $\widehat{f}_{mult,d}$ achieves the same approximation error as theirs. Thus, the remainder of the proof follows identically to Lemma 9 of Kohler and Langer (2021) and is omitted. □

Lemma 4. *Let $1 \leq a < \infty$. There exists a neural network $\widehat{f}_{check,\mathcal{P}_2} \in \mathcal{F}_r^1(100, 2r + (4r^2+4r) \cdot \zeta^r, \zeta^{5p+5})$, satisfying $\widehat{f}_{check,\mathcal{P}_2}(x) = 1_{\bigcup_{i\in\{1,\dots,\zeta^{2r}\}} C_{i,2}\setminus(C_{i,2})^0_{1/\zeta^{2p+2}}}(x)$, for $x \notin \bigcup_{i\in\{1,\dots,\zeta^{2r}\}} (C_{i,2})^0_{1/\zeta^{2p+2}} \setminus (C_{i,2})^0_{2/\zeta^{2p+2}}$, and $\widehat{f}_{check,\mathcal{P}_2}(x) \in [0,1]$, for $x \in [-a,a)^r$.*

Proof. We adopt the definition of $\widehat{f}_{\text{check},\mathcal{P}_2}$ in Lemma 10 of Kohler and Langer (2021):

$$\begin{aligned}
&\widehat{f}_{\text{check},\mathcal{P}_2}(x) = 1 - \sigma\left(1 - \widehat{f}_2(x) - \widehat{f}_{id}^2\left(\widehat{f}_1(x)\right)\right), \text{ where}\\
&\widehat{f}_1(x) := 1 - \sum_{k=1}^{\zeta^r} \widehat{f}_{ind,\left(C_{k,1}\right)^0_{1/\zeta^{2p+2}}}(x) \quad \text{and} \qquad \text{(A11)}\\
&\widehat{f}_2(x) := 1 - \sum_{j=1}^{\zeta^r} \widehat{f}_{\text{test}}\left(\widehat{f}_{id}^2(x), \widehat{\phi}_{2,1} + v_j + \frac{1}{\zeta^{2p+2}} \cdot 1, \widehat{\phi}_{2,1} + v_j + \frac{2a}{\zeta^2} \cdot 1 - \frac{1}{\zeta^{2p+2}} \cdot 1, 1\right).
\end{aligned}$$

Here $\widehat{f}_{ind,[a,b)}$ and $\widehat{f}_{test}$ are defined in Eq. (A2) and the function $\widehat{\phi}_{2,1}$ is defined in Eq. (A3). Since all these functions are identical to those used in Kohler and Langer (2021), the only remaining task is to verify that the weights of $\widehat{f}_{\text{check},\mathcal{P}_2}$ lie within the interval $[-\zeta^{5p+5}, \zeta^{5p+5}]$. Given that the weights of $\widehat{f}_{ind,[a,b)}$, $\widehat{f}_{\text{test}}$, and $\widehat{\phi}_{2,1}$ are all bounded within this range, it follows directly from their definitions that the weights of $\widehat{f}_1$ and $\widehat{f}_2$ and consequently $\widehat{f}_{\text{check},\mathcal{P}_2}$, also lie within $[-\zeta^{5p+5}, \zeta^{5p+5}]$. □

Lemma 5. *Under Assumption 1, for any set of constants $\{C_{it}\}_{i=1,\dots,N,t=1,\dots,N}$, conditional on $\{F_t^\star\}_{t=1}^T$, $(\sum_{t=1}^T \sum_{i=1}^N C_{it}U_{it})/(\sum_{t=1}^T \sum_{i=1}^N C_{it}^2)^{1/2}$ is sub-Gaussian with its sub-Gaussian norm bounded by some fixed constant σ_u^2.*

Proof. Let $C = (C_{it})$ denote a matrix of fixed constants. Then we can write

$$\frac{\sum_{t=1}^T \sum_{i=1}^N C_{it}U_{it}}{\left(\sum_{t=1}^T \sum_{i=1}^N C_{it}^2\right)^{1/2}} = \frac{\operatorname{vec}(C)^\top \operatorname{vec}(U)}{\left(\sum_{t=1}^T \sum_{i=1}^N C_{it}^2\right)^{1/2}} = \frac{\operatorname{vec}(C)^\top \Sigma^{1/2} \operatorname{vec}(Z)}{\left(\sum_{t=1}^T \sum_{i=1}^N C_{it}^2\right)^{1/2}}.$$

Since $\operatorname{vec}(Z)$ is a vector of independent sub-Gaussian random variables with sub-Gaussian norm bounded by σ_z^2, the sub-Gaussian norm of the right hand side is bounded by

$$\frac{\sigma_z^2 \operatorname{vec}(C)^\top \Sigma \operatorname{vec}(C)}{\sum_{t=1}^T \sum_{i=1}^N C_{it}^2} \leq \frac{\sigma_z^2 \|\Sigma\| \|\operatorname{vec}(C)\|^2}{\sum_{t=1}^T \sum_{i=1}^N C_{it}^2} = \sigma_z^2 \|\Sigma\|. \qquad \square$$

Lemma 6. *Under Assumption 1, with probability at least* $1 - C\exp(-cT)$*, we have* $\sum_{t=1}^{T} \|U_t\|_1 \lesssim NT$*.*

Proof. Since $\|\Sigma\|$ is bounded, we have

$$\begin{aligned}\sum_{t=1}^{T} \|U_t\|_1 \leq & \sum_{t=1}^{T} N^{1/2}\|U_t\| \leq N^{1/2}T^{1/2}\left(\sum_{i=1}^{N}\sum_{t=1}^{T} U_{it}^2\right)^{1/2} \\ = & N^{1/2}T^{1/2}\left(\operatorname{vec}(Z)^\top \Sigma \operatorname{vec}(Z)\right)^{1/2} \lesssim N^{1/2}T^{1/2}\left(\operatorname{vec}(Z)^\top \operatorname{vec}(Z)\right)^{1/2}.\end{aligned}$$

Therefore, it is sufficient to bound the quadratic form $N^{-1}T^{-1}\operatorname{vec}(Z)^\top \operatorname{vec}(Z)$. Recall that each Z_{it} is sub-Gaussian with sub-Gaussian norm bounded by σ_z^2. By the Hansen-Wright inequality,

$$\mathrm{P}\left(\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(Z_{it}^2 - \mathbb{E}Z_{it}^2) \geq \sigma_z^2\right) \leq 2\exp(-cNT).$$

Moreover, from the properties of sub-Gaussian random variables, $\mathbb{E}Z_{it}^2 \leq 2\sigma_z^2$. Thus, with probability at least $1 - C\exp(-cT)$, it holds that $\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} Z_{it}^2 \leq 3\sigma_z^2$. □

Lemma 7. *Under Assumption 1 and Definition 2, and assuming that* $W^\star$ *satisfies Eq. (16), if* $\log T = o(\epsilon_N^{-1})$*, then with probability at least* $1 - C\exp(-c\epsilon_N^{-1})$*, it holds that* $\|W^\star U_t\|_\infty \leq 1$*, for all* $1 \leq t \leq T$*.*

Proof. Write $\Sigma_t \in \mathbb{R}^{T\times NT}$ as the submatrix of $\Sigma^{1/2}$ consisting of rows $(t-1)N+1$ to tN. As a result, we can express $U_t = \Sigma_t \operatorname{vec}(Z)$. For $i = 1, \ldots, K$, let $W_i^\star$ denote the i-th row of the matrix $W^\star$. Then, $\|W^\star U_t\|_\infty = \max_{i=1,\ldots,K}\|W_i^\star \Sigma_t \operatorname{vec}(Z)\|$. Observe that each $W_i^\star \Sigma_t \operatorname{vec}(Z)$ is a sub-Gaussian random variable. Its sub-Gaussian norm is bounded by $\sigma_z^2 W_i^\star \Sigma_t \Sigma_t^\top (W_i^\star)^\top \lesssim \|W_i^\star\|^2 \|\Sigma_t\Sigma_t^\top\| \lesssim \epsilon_N \|\Sigma_t\Sigma_t^\top\|$. Since $\Sigma_t\Sigma_t^\top$ is a submatrix of Σ and $\|\Sigma\| \lesssim 1$, it follows that $\|\Sigma_t\Sigma_t^\top\| \lesssim 1$. Therefore, the sub-Gaussian norm of $W_i^\star\Sigma_t \operatorname{vec}(Z)$ is bounded by a constant multiple of ϵ_N. By standard sub-Gaussian tail bounds, this implies $\mathrm{P}(|W_i^\star \Sigma_t \operatorname{vec}(Z)| \geq 1) \leq C\exp(-c\epsilon_N^{-1})$. Applying the union bound over all $t = 1, \ldots, T$ and $i = 1, \ldots, K$, and using the assumption $\log(T) = o(\epsilon_N^{-1})$, we obtain

$$\mathrm{P}(\max_{1\leq t\leq T}\max_{1\leq i\leq K}|W_i^\star \Sigma_t \operatorname{vec}(Z)| \geq 1) \leq CTK\exp(-c\epsilon_N^{-1}) \lesssim C\exp(-c\epsilon_N^{-1}/2).$$ □

Lemma 8. *For* $W^\star$ *satisfying Eq. (16), it holds that with probability at least* $1 - C\exp(-cT)$*,* $T^{-1}\sum_{t=1}^{T}\|W^\star U_t\|^2 \lesssim \epsilon_N$*.*

Proof. Define $Q := \mathbb{I}_T \otimes (W^\star)^\top W^\star$. Then,

$$T^{-1}\sum_{t=1}^{T}\|W^\star U_t\|^2 = T^{-1}\operatorname{vec}(U)^\top Q\operatorname{vec}(U) = T^{-1}\operatorname{vec}(Z)^\top\Sigma^{1/2}Q\Sigma^{1/2}\operatorname{vec}(Z).$$

Note that $\|Q\| \le \|(W^\star)^\top W^\star\| \lesssim \epsilon_N$, which implies $\|\Sigma^{1/2}Q\Sigma^{1/2}\| \lesssim \epsilon_N$. Also, the rank satisfies $\operatorname{rank}(\Sigma^{1/2}Q\Sigma^{1/2}) \le \operatorname{rank}(Q) \le T\operatorname{rank}(W^\star) = TK$. As a consequence, we have

$$\begin{aligned}
&T^{-1}\operatorname{Tr}(\Sigma^{1/2}Q\Sigma^{1/2}) \le \|T^{-1}\Sigma^{1/2}Q\Sigma^{1/2}\|\operatorname{rank}(\Sigma^{1/2}Q\Sigma^{1/2}) \lesssim \epsilon_N,\\
&\|T^{-1}\Sigma^{1/2}Q\Sigma^{1/2}\|_F^2 \lesssim \|T^{-1}\Sigma^{1/2}Q\Sigma^{1/2}\|^2\operatorname{rank}(\Sigma^{1/2}Q\Sigma^{1/2}) \lesssim T^{-1}\epsilon_N^2.
\end{aligned}$$

Applying the Hanson-Wright inequality, we conclude that

$$\mathrm{P}\left(|(1-\mathbb{E})T^{-1}\operatorname{vec}(Z)^\top\Sigma^{1/2}Q\Sigma^{1/2}\operatorname{vec}(Z)| \ge \epsilon_N\right) \le C\exp(-cT).$$

Finally, since $\mathbb{E}\left[T^{-1}\operatorname{vec}(Z)^\top\Sigma^{1/2}Q\Sigma^{1/2}\operatorname{vec}(Z)\right] \asymp T^{-1}\operatorname{Tr}(\Sigma^{1/2}Q\Sigma^{1/2}) \lesssim \epsilon_N$, the proof is complete. □

Lemma 9. *For any $\varphi \in \mathcal{F}_{n_0}^{n_{d+1}}(d, w, T^{5\beta+5})$ and $x, y \in \mathbb{R}^{n_0}$, it holds that*

$$|\varphi(x) - \varphi(y)| \le n_0 T^{(5\beta+5)(d+1)} w^d \|x - y\|_\infty.$$

Proof. Write $\varphi(x) = W_d\sigma_{v_d}W_{d-1}\sigma_{v_{d-1}}\dots W_1\sigma_{v_1}W_0x$, where each $W_i \in \mathbb{R}^{n_{i+1}\times n_i}$. By assumption, each weight matrix satisfies $\|W_i\|_\infty \le T^{5\beta+5}$, and hence for any vectors x, y, $\|W_ix - W_iy\|_\infty \le n_iT^{5\beta+5}\|x - y\|_\infty$. Furthermore, the activation functions σ_{v_i} are 1-Lipschitz, so $\|\sigma_{v_i}x - \sigma_{v_i}y\|_\infty \le \|x - y\|_\infty$. Now, consider composing two functions f_1 and f_2 with Lipschitz constants L_1 and L_2, respectively. Then the composition satisfies $\|f_2 \circ f_1(x) - f_2 \circ f_1(y)\|_\infty \le L_1L_2\|x - y\|_\infty$. Applying this repeatedly to the composition $W_d\sigma_{v_d}W_{d-1}\sigma_{v_{d-1}}\dots W_1\sigma_{v_1}W_0x$, and since $\prod_{i=0}^{d} n_i \le n_0w^d$, we obtain

$$\|\varphi(x) - \varphi(y)\| \le T^{(5\beta+5)(d+1)}\Big(\prod_{i=0}^{d} n_i\Big)\|x - y\|_\infty \le n_0T^{(5\beta+5)(d+1)}w^d\|x - y\|_\infty. \qquad \square$$

Lemma 10. *Consider the class of neural networks $\mathcal{F}_{n_0}^{n_{d+1}}(d, w, T^{5\beta+5}, B)$ with width vector $n = (n_0, \dots, n_{d+1})$. Assume that the total number of nonzero weights in the network is bounded by S. Then, there exists a subset $\mathcal{F}_{n_0,\delta}^{n_{d+1}} \subset \mathcal{F}_{n_0}^{n_{d+1}}(d, w, T^{5\beta+5})$ such that:*
(a) Its cardinality satisfies: $|\mathcal{F}_{n_0,\delta}^{n_{d+1}}| \le \left(8\delta^{-1}CT^{(5\beta+5)(d+2)}(1+w)^dn_0n_{d+1}d\right)^{2S}$;
(b) For any $\varphi \in \mathcal{F}_{n_0}^{n_{d+1}}(d, w, T^{5\beta+5}, B)$, there exists $\overline{\varphi} \in \mathcal{F}_{n_0,\delta}^{n_{d+1}}$ such that $\|\varphi(x) - \overline{\varphi}(x)\|_\infty \le \delta$ for all $\|x\|_\infty \le C$.

Proof. Let $\varphi(x) = W_d\sigma_{v_d}W_{d-1}\sigma_{v_{d-1}}\cdots W_1\sigma_{v_1}W_0x$. Define $A_k^+\varphi : [-C,C]^r \to \mathbb{R}^{n_k}$ and $A_k^-\varphi : \mathbb{R}^{n_{k-1}} \to \mathbb{R}$ as

$$A_k^+\varphi(x) = \sigma_{v_k}W_{k-1}\sigma_{v_{k-1}}\cdots W_1\sigma_{v_1}W_0x, \quad \text{for } k = 1,\ldots,d,$$
$$A_k^-\varphi(x) = W_d\sigma_{v_d}W_{d-1}\cdots W_k\sigma_{v_k}W_{k-1}x, \quad \text{for } k = 1,\ldots,d+1.$$

By convention, we set $A_0^+\varphi(x) = A_{d+2}^-\varphi(x) = x$. Given that all the parameters of φ are bounded by $T^{5\beta+5}$ and that $\|x\|_\infty \le C$, by induction, we have

$$\|A_k^+(x)\|_\infty \le T^{(5\beta+5)(k+1)}C\prod_{\ell=0}^{k-1}(n_\ell+1). \tag{A12}$$

Additionally, from Lemma 9, we derive

$$|A_k^-\varphi(x) - A_k^-\varphi(y)| \le T^{(5\beta+5)(d-k+2)}\left(\prod_{\ell=k-1}^{d} n_\ell\right)\|x-y\|_\infty. \tag{A13}$$

Now fix $\varepsilon > 0$. Consider two DNNs, $\varphi(x) = W_d\sigma_{v_d}W_{d-1}\sigma_{v_{d-1}}\cdots W_1\sigma_{v_1}W_0x$ and $\overline{\varphi}(x) = \overline{W}_d\sigma_{\overline{v}_d}\overline{W}_{d-1}\sigma_{\overline{v}_{d-1}}\cdots\overline{W}_1\sigma_{\overline{v}_1}\overline{W}_0x$, both belonging to the class $\mathcal{F}_{n_0}^{n_{d+1}}(d,w,T^{5\beta+5})$. Suppose that their corresponding parameters are ε-close in the sup-norm, i.e., $\|W_k-\overline{W}_k\|_\infty \le \varepsilon, \|v_k-\overline{v}_k\|_\infty \le \varepsilon$, for all k. By successively applying the triangle inequality, and replacing the weights and biases in $\varphi(x)$ with those in $\overline{\varphi}(x)$ one layer at a time, we obtain the following upper bound:

$$\|\varphi(x)-\overline{\varphi}(x)\|_\infty \le \sum_{k=1}^{d+1}\left\|A_{k+1}^-\varphi\circ\sigma_{v_k}W_{k-1}A_{k-1}^+\overline{\varphi}(x) - A_{k+1}^-\varphi\circ\sigma_{\overline{v}_k}\overline{W}_{k-1}A_{k-1}^+\overline{\varphi}(x)\right\|_\infty.$$

By applying the Lipschitz bound for $A_{k+1}^-\varphi$ from Eq. (A13), this is bounded by

$$\sum_{k=1}^{d+1}T^{(5\beta+5)(d-k+1)}\left(\prod_{\ell=k}^{d}n_\ell\right)\left\|\sigma_{v_k}W_{k-1}A_{k-1}^+\overline{\varphi}(x) - \sigma_{\overline{v}_k}\overline{W}_{k-1}A_{k-1}^+\overline{\varphi}(x)\right\|_\infty.$$

Next, using the 1-Lipschitz property of σ_v, we bound this difference by

$$\sum_{k=1}^{d+1}T^{(5\beta+5)(d-k+1)}\left(\prod_{\ell=k}^{d}n_\ell\right)\left(\|(W_{k-1}-\overline{W}_{k-1})A_{k-1}^+\overline{\varphi}(x)\|_\infty + \|v_k-\overline{v}_k\|_\infty\right).$$

Since the parameter difference is at most ε, and $W_{k-1} - \overline{W}_{k-1}$ has at most n_{k-1} rows, this yields

$$\varepsilon \sum_{k=1}^{d+1} T^{(5\beta+5)(d-k+1)} \left(\prod_{\ell=k}^{d} n_\ell \right) \left(n_{k-1} \|A_{k-1}^{+} \overline{\varphi}(x)\|_\infty + 1 \right).$$

Applying Eq. (A12) to bound $\|A_{k-1}^{+} \overline{\varphi}(x)\|_\infty$, we obtain

$$\|\varphi(x) - \overline{\varphi}(x)\|_\infty \leq 4\varepsilon T^{(5\beta+5)(d+1)} C(1+w)^d n_0 d.$$

Choosing $\varepsilon := \delta\big(4T^{(5\beta+5)(d+1)} C(1+w)^d n_0 d\big)^{-1}$ ensures that $\|\varphi(x) - \overline{\varphi}(x)\|_\infty \leq \delta$ uniformly over all $\|x\|_\infty \leq C$. Therefore, we conclude that for any two DNNs in the class $\mathcal{F}_{n_0}^{n_{d+1}}(d, w, T^{5\beta+5})$, if their corresponding parameters differ by at most ε, then their outputs differ by at most δ uniformly over $\|x\|_\infty \leq C$.

Consequently, we construct a finite subset $\mathcal{F}_{n_0,\delta}^{n_{d+1}} \subset \mathcal{F}_{n_0}^{n_{d+1}}(d, w, T^{5\beta+5})$ by discretizing the parameters of any $\varphi \in \mathcal{F}_{n_0}^{n_{d+1}}(d, w, T^{5\beta+5})$ on a uniform grid. Specifically, we form a grid of mesh width ε over the compact domain $[-T^{5\beta+5}, T^{5\beta+5}]$ for each parameter entry in the collection $\{W_0, v_1, \ldots, v_d, W_d\}$. Since each weight matrix and bias vector contains at most w^2 and w nonzero entries, respectively, and each such entry can be discretized into approximately $2T^{5\beta+5}/\varepsilon$ points, this procedure yields a finite class of networks.

By construction, for any φ in the original class, we can find $\overline{\varphi} \in \mathcal{F}_{n_0,\delta}^{n_{d+1}}$ whose parameters lie within ε (in ℓ^∞ norm) of those of φ. Owing to the Lipschitz continuity of the neural network with respect to its parameters, the output of $\overline{\varphi}$ deviates from that of φ by at most δ, provided ε is sufficiently small. Thus, condition (b) is satisfied.

To verify condition (a), note that the total number of parameters is bounded by

$$\sum_{\ell=0}^{d} (n_\ell + 1) n_{\ell+1} \leq (d+1) \cdot 2^{-d} \prod_{\ell=0}^{d+1} (n_\ell + 1) \leq 4 n_0 n_{d+1} (1+w)^d.$$

The number of ways to select S nonzero entries among these parameters is at most

$$\binom{4n_0 n_{d+1}(1+w)^d}{S} \leq \left(4n_0 n_{d+1}(1+w)^d\right)^S.$$

Each of the S nonzero parameters can take at most $2T^{5\beta+5}/\varepsilon$ values on a grid, so the cardinality satisfies

$$|\mathcal{F}_{n_0,\delta}^{n_{d+1}}| \le \sum_{S^\star \le S} \left(8\delta^{-1}CT^{(5\beta+5)(d+2)}(1+w)^d n_0 d\right)^{S^\star} \le \left(8\delta^{-1}CT^{(5\beta+5)(d+2)}(1+w)^d n_0 n_{d+1} d\right)^{2S}. \square$$

Lemma 11. *Under the same conditions of Theorem 1, assume* $(\varphi_1, \ldots, \varphi_N) \circ \rho \in \mathcal{F}_{AE}^{K_1}$ *and that* $\varphi_1, \ldots, \varphi_N$ *are deterministic, then with probability at least* $1 - C\exp(-cT)$,

$$\sum_{t=1}^{T}\sum_{i=1}^{N} U_{it}\big(X_{it}^\star - \varphi_i(\rho(X_t))\big) \lesssim T^{1/2}K_1^{1/2}\log(T)\Big(\sum_{t=1}^{T}\sum_{i=1}^{N}\big(X_{it}^\star - \varphi_i(\rho(X_t))\big)^2\Big)^{1/2} + K_1^{1/2}T^{1/2}N^{1/2}\log(T) + N.$$

Proof. Write $\varepsilon_T = K_1^{-1}T^{-1-(5\beta+5)(d_2+1)}w_2^{-d_2}$ and define the set

$$\mathcal{S} = \{x \in \mathbb{R}^{K_1} : \forall 1 \le i \le K_1,\ \exists 1 \le j(\in \mathbb{N}) \le 2B\varepsilon_T^{-1},\ \text{s.t. } x_i = -B + j\varepsilon_T\}. \tag{A14}$$

By definition, we see that for any $F \in [-B, B]^{K_1}$, there exists an $\bar{F} \in \mathcal{S}$ such that $\|F - \bar{F}\|_\infty \le \varepsilon_T$. It is straightforward to show that the cardinality of $\mathcal{S}$ satisfies that:

$$\log|\mathcal{S}| \le K_1\log(2BK_1T^{1+(5\beta+5)(d_2+1)}w_2^{d_2}) \lesssim K_1\log^2(T). \tag{A15}$$

Denote the elements of this set by $\bar{F}_1, \bar{F}_2, \ldots, \bar{F}_{|\mathcal{S}|}$. For $1 \le t \le T$, by definition, there exists a vector $\bar{F} \in \mathcal{S}$ such that $\|\rho(X_t) - \bar{F}\|_\infty \le \varepsilon_T$. We denote its subscript by $k_t^\star$, i.e., $\|\rho(X_t) - \bar{F}_{k_t^\star}\|_\infty \le \varepsilon_T$. As a consequence, by Lemma 9,

$$|\varphi_i(\rho(X_t)) - \varphi_i(\bar{F}_{k_t^\star})| \le T^{-1}. \tag{A16}$$

Together with Lemma 6, with probability at least $1 - C\exp(-cT)$,

$$\begin{aligned}
&\sum_{t=1}^{T}\sum_{i=1}^{N} U_{it}\big(X_{it}^\star - \varphi_i(\rho(X_t))\big) \lesssim \sum_{t=1}^{T}\sum_{i=1}^{N} U_{it}\big(X_{it}^\star - \varphi_i(\bar{F}_{k_t^\star})\big) + N \\
&\le \max_{\substack{k_t\in[|\mathcal{S}|]\\ 1\le t\le T}} \frac{\Big|\sum_{t=1}^{T}\sum_{i=1}^{N} U_{it}\big(X_{it}^\star - \varphi_i(\bar{F}_{k_t})\big)\Big|}{\Big(\sum_{t=1}^{T}\sum_{i=1}^{N}\big(X_{it}^\star - \varphi_i(\bar{F}_{k_t})\big)^2\Big)^{1/2}}\Big(\sum_{t=1}^{T}\sum_{i=1}^{N}\big(X_{it}^\star - \varphi_i(\bar{F}_{k_t^\star})\big)^2\Big)^{1/2} + N \qquad \text{(A17)} \\
&\lesssim \max_{\substack{k_t\in[|\mathcal{S}|]\\ 1\le t\le T}} \frac{\Big|\sum_{t=1}^{T}\sum_{i=1}^{N} U_{it}\big(X_{it}^\star - \varphi_i(\bar{F}_{k_t})\big)\Big|}{\Big(\sum_{t=1}^{T}\sum_{i=1}^{N}\big(X_{it}^\star - \varphi_i(\bar{F}_{k_t})\big)^2\Big)^{1/2}}\left(\Big(\sum_{t=1}^{T}\sum_{i=1}^{N}\big(X_{it}^\star - \varphi_i(\rho(X_t))\big)^2\Big)^{1/2} + N^{1/2}\right) + N,
\end{aligned}$$

where we use Eq. (A16) and Lemma 6 in the first inequality and Eq. (A16) in the last

inequality. For any $u \geq 0$, by Lemma 5 and the property of sub-Gaussian distribution, as well as the fact that $\varphi_1, \ldots, \varphi_N$, $\bar{F}_{k_1}, \ldots, \bar{F}_{k_T}$, and $\{X^\star_{it}\}_{i,t}$ are deterministic conditional on $\{F^\star_t\}_{t=1}^T$, it follows that

$$\mathrm{P}\left(\frac{\left|\sum_{t=1}^T \sum_{i=1}^N U_{it}\big(X^\star_{it} - \varphi_i(\bar{F}_{k_t})\big)\right|}{\left(\sum_{t=1}^T \sum_{i=1}^N \big(X^\star_{it} - \varphi_i(\bar{F}_{k_t})\big)^2\right)^{1/2}} > u \middle| \{F^\star_t\}_{t=1}^T \right) \leq 2\exp\left(-\frac{u^2}{2\sigma_u^2}\right).$$

Applying the union bound over all possible sequences $(k_1, \ldots, k_T)$, and integrating both sides with respect to the distribution of $\{F^\star_t\}_{t=1}^T$, we conclude

$$\mathrm{P}\left(\max_{\substack{k_t \in [|\mathcal{S}|] \\ 1 \leq t \leq T}} \frac{\left|\sum_{t=1}^T \sum_{i=1}^N U_{it}\big(X^\star_{it} - \varphi_i(\bar{F}_{k_t})\big)\right|}{\left(\sum_{t=1}^T \sum_{i=1}^N \big(X^\star_{it} - \varphi_i(\bar{F}_{k_t})\big)^2\right)^{1/2}} > u \right) \leq 2|\mathcal{S}|^T \exp\left(-\frac{u^2}{2\sigma_u^2}\right).$$

By Eq. (A15), Eq. (A17), and setting $u = 2\sigma_u(T\log|\mathcal{S}|)^{1/2}$, we complete the proof. □

Lemma 12. *Under the same conditions of Theorem 1, with probability at least* $1 - C\exp(-cT)$*, we have*

$$\begin{aligned}\left|\sum_{t=1}^T \sum_{i=1}^N \big(\widehat{\varphi}_i(\widehat{\rho}(X_t)) - X^\star_{it}\big)U_{it}\right| \lesssim & (NT^{\frac{K}{2\beta+K}} + TK_1)^{1/2}\log^2(T) \\ & \times \left(\left(\sum_{t=1}^T \sum_{i=1}^N \big(\widehat{\varphi}_i(\widehat{\rho}(X_t)) - X^\star_{it}\big)^2\right)^{1/2} + N^{1/2}\right) + N.\end{aligned}$$

Proof. We begin by invoking Lemma 10, which states that for $\delta = T^{-1}$, there exists a function class $\mathcal{F}_{2,\delta} \subset \mathcal{F}_2$ satisfying

$$\log|\mathcal{F}_{2,\delta}| \lesssim (w_2^2 d_2 + K_1 w_2)\log\left(8\delta^{-1}BT^{(5\beta+5)(d_2+2)}(1+w_2)^{d_2}K_1 d_2\right) \lesssim T^{\frac{K}{2\beta+K}}\log^4(T), \quad \text{(A18)}$$

such that for any $\varphi \in \mathcal{F}_2$, there exists $\overline{\varphi} \in \mathcal{F}_{2,\delta}$ with $|\overline{\varphi}(x) - \varphi(x)| \leq T^{-1}$, for all $\|x\|_\infty \leq B$. Since $\|\varphi\|_\infty \leq B$, it follows that $|\overline{\varphi}(x)| \leq T^{-1} + B$. For simplicity, we may thus assume all functions in $\mathcal{F}_{2,\delta}$ are bounded by $2B$, i.e., $\mathcal{F}_{2,\delta} \subset \mathcal{F}^1_{K_1}(d_2, w_2, T^{5\beta+5}, 2B)$. Write $\mathcal{F}_{2,\delta} = \{\overline{\varphi}_j, 1 \leq j \leq |\mathcal{F}_{2,\delta}|\}$. By construction, for each $i \leq N$, there exists an index $\ell_i^\star \in [|\mathcal{F}_{2,\delta}|]$ such that $|\widehat{\varphi}_i(x) - \overline{\varphi}_{\ell_i^\star}(x)| \leq T^{-1}$, for all $\|x\|_\infty \leq B$. Similarly, by the definition of $\mathcal{S}$ in Eq. (A14), for each $t \leq T$, there exists $k_t^\star \in [|\mathcal{S}|]$ such that $\|\overline{F}_{k_t^\star} - \widehat{\rho}(X_t)\|_\infty \leq \varepsilon_T$. Since each $\widehat{\varphi}_i \in \mathcal{F}^1_{K_1}(d_2, w_2, T^{5\beta+5}, B)$, we can apply Lemma 9 to obtain

$$|\widehat{\varphi}_i(\widehat{\rho}(X_t)) - \overline{\varphi}_{\ell_i^\star}(\overline{F}_{k_t^\star})| \le |\widehat{\varphi}_i(\widehat{\rho}(X_t)) - \widehat{\varphi}_i(\overline{F}_{k_t^\star})| + |\widehat{\varphi}_i(\overline{F}_{k_t^\star}) - \overline{\varphi}_{\ell_i^\star}(\overline{F}_{k_t^\star})| \le 2T^{-1}.$$

Together with Lemma 6, we obtain that, with probability at least $1 - C\exp(-cT)$,

$$\begin{aligned}
\left|\sum_{t=1}^{T}\sum_{i=1}^{N}\left(\widehat{\varphi}_i(\widehat{\rho}(X_t)) - X_{it}^\star\right)U_{it}\right| &\lesssim \left|\sum_{t=1}^{T}\sum_{i=1}^{N}\left(\overline{\varphi}_{\ell_i^\star}(\overline{F}_{k_t^\star}) - X_{it}^\star\right)U_{it}\right| + N \qquad \text{(A19)}\\
&\le \max_{\substack{k_t\in[|\mathcal{S}|],1\le t\le T\\ \ell_i\in[|\mathcal{F}_{2,\delta}|],1\le i\le N}} \frac{\left|\sum_{t=1}^{T}\sum_{i=1}^{N}\left(\overline{\varphi}_{\ell_i}(\overline{F}_{k_t}) - X_{it}^\star\right)U_{it}\right|}{\left(\sum_{t=1}^{T}\sum_{i=1}^{N}\left(\overline{\varphi}_{\ell_i}(\overline{F}_{k_t}) - X_{it}^\star\right)^2\right)^{1/2}}\left(\sum_{t=1}^{T}\sum_{i=1}^{N}\left(\overline{\varphi}_{\ell_i^\star}(\overline{F}_{k_t^\star}) - X_{it}^\star\right)^2\right)^{1/2} + N\\
&\lesssim \max_{\substack{k_t\in[|\mathcal{S}|],1\le t\le T\\ \ell_i\in[|\mathcal{F}_{2,\delta}|],1\le i\le N}} \frac{\left|\sum_{t=1}^{T}\sum_{i=1}^{N}\left(\overline{\varphi}_{\ell_i}(\overline{F}_{k_t}) - X_{it}^\star\right)U_{it}\right|}{\left(\sum_{t=1}^{T}\sum_{i=1}^{N}\left(\overline{\varphi}_{\ell_i}(\overline{F}_{k_t}) - X_{it}^\star\right)^2\right)^{1/2}}\\
&\qquad\times\left(\left(\sum_{t=1}^{T}\sum_{i=1}^{N}\left(\widehat{\varphi}_i(\widehat{\rho}(X_t)) - X_{it}^\star\right)^2\right)^{1/2} + N^{1/2}\right) + N.
\end{aligned}$$

Now, using Lemma 5 and the union bound inequality, we have

$$\begin{aligned}
&\mathrm{P}\left(\max_{\substack{k_t\in[|\mathcal{S}|],1\le t\le T\\ \ell_i\in[|\mathcal{F}_{2,\delta}|],1\le i\le N}} \frac{\left|\sum_{t=1}^{T}\sum_{i=1}^{N}\left(\overline{\varphi}_{\ell_i}(\overline{F}_{k_t}) - X_{it}^\star\right)U_{it}\right|}{\left(\sum_{t=1}^{T}\sum_{i=1}^{N}\left(\overline{\varphi}_{\ell_i}(\overline{F}_{k_t}) - X_{it}^\star\right)^2\right)^{1/2}} > u \,\middle|\, \{F_t^\star\}_{t=1}^T\right)\\
&\le 2|\mathcal{S}|^T\cdot|\mathcal{F}_{2,\delta}|^N\exp\left(-\frac{u^2}{2\sigma_u^2}\right).
\end{aligned}$$

Integrating both sides with respect to the distribution of $\{F_t^\star\}_{t=1}^T$, we conclude

$$\mathrm{P}\left(\max_{\substack{k_t\in[|\mathcal{S}|],1\le t\le T\\ \ell_i\in[|\mathcal{F}_{2,\delta}|],1\le i\le N}} \frac{\left|\sum_{t=1}^{T}\sum_{i=1}^{N}\left(\overline{\varphi}_{\ell_i}(\overline{F}_{k_t}) - X_{it}^\star\right)U_{it}\right|}{\left(\sum_{t=1}^{T}\sum_{i=1}^{N}\left(\overline{\varphi}_{\ell_i}(\overline{F}_{k_t}) - X_{it}^\star\right)^2\right)^{1/2}} > u\right) \le 2|\mathcal{S}|^T\cdot|\mathcal{F}_{2,\delta}|^N\exp\left(-\frac{u^2}{2\sigma_u^2}\right).$$

Setting $u = 2\sigma_u(T\log|\mathcal{S}| + N\log|\mathcal{F}_{2,\delta}|)^{1/2}$, and using bounds from Eqs. (A15), (A18), and (A19), the claim follows. □

Lemma 13. *Under the same conditions as Theorem 2, there exists a fixed constant C, independent of i, such that*

$$\left|\mathbb{E}\left(\widehat{\varphi}_i(\widehat{\rho}(X_{T+1})) - \varphi_i^\star(F_{T+1}^\star)\right)^2 - \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left(\widehat{\varphi}_i(\widehat{\rho}(X_t)) - \varphi_i^\star(F_t^\star)\right)^2\right|$$

$$\lesssim C(T^{-\frac{2\beta}{2\beta+K}} + T^{-1}L_N)\log^4(NT)$$
$$+ C(T^{-\frac{\beta}{2\beta+K}} + T^{-1/2}L_N^{1/2})\log^2(NT)\mathbb{E}^{1/2}[\big(\widehat{X}_{T+1,i} - \varphi_i^\star(F_{T+1}^\star)\big)^2].$$

Proof. Consider the function class $\mathcal{F}_N^1(d_1+d_2, \max(w_1,w_2), T^{5\beta+5}, B)$. We focus on its subset in which every DNN has a total number of nonzero weights bounded by $CL_N + CT^{\frac{K}{2\beta+K}}\log T$ for some fixed constant C. For notational convenience, we denote this subclass by $\widetilde{\mathcal{F}}$. By Lemma 10, there exists a discretized class of networks, denoted by $\mathcal{F}_{N,\delta}^1$ with $\delta = T^{-1}$, such that the logarithm of its cardinality satisfies $\log|\mathcal{F}_{N,\delta}^1| \lesssim (L_N + T^{\frac{K}{2\beta+K}})\log^4 NT$, and for any $f \in \widetilde{\mathcal{F}}$, there exists a function $\overline{f}(\cdot) \in \mathcal{F}_{N,\delta}^1$ such that $|f(x) - \overline{f}(x)| \le T^{-1}$ whenever $\|x\|_\infty \le 2NT$. For convenience, we denote the elements of $\mathcal{F}_{N,\delta}^1$ by $\overline{f}_1, \overline{f}_2, \ldots, \overline{f}_{|\mathcal{F}_{N,\delta}^1|}$. We further assume, without loss of generality, that all functions in $\mathcal{F}_{N,\delta}^1$ are uniformly bounded in sup-norm by B, i.e., $\|\overline{f}_j\|_\infty \le B$ for all j. Otherwise, one can append two additional layers to each network to enforce the bound using the transformation $\sigma(2B - \sigma(B - \overline{f}_j)) - B$. Note that $\widehat{\varphi}_i \circ \widehat{\rho}$ belongs to $\widetilde{\mathcal{F}}$. Therefore, there exists a function belonging to $\{\overline{f}_j\}_{j=1}^{|\mathcal{F}_{N,\delta}^1|}$, denoted by $\overline{f}_{i^\star}$, such that

$$|\widehat{\varphi}_i \circ \widehat{\rho}(x) - \overline{f}_{i^\star}(x)| \le \delta = T^{-1}, \tag{A20}$$

for $\|x\|_\infty \le 2NT$. Suppose $\{(\tilde{F}_t^\star, \tilde{U}_t, \tilde{X}_t)\}_{t=1}^{T+1}$ is an independent copy of $\{(F_t^\star, U_t, X_t)\}_{t=1}^{T+1}$. As a consequence,

$$\left|\mathbb{E}\big(\widehat{\varphi}_i(\widehat{\rho}(X_{T+1})) - \varphi_i^\star(F_{T+1}^\star)\big)^2 - \frac{1}{T}\sum_{t=1}^T \mathbb{E}\big(\widehat{\varphi}_i(\widehat{\rho}(X_t)) - \varphi_i^\star(F_t^\star)\big)^2\right|$$
$$= \left|\frac{1}{T}\sum_{t=1}^T \mathbb{E}\big(\widehat{\varphi}_i(\widehat{\rho}(\tilde{X}_t)) - \varphi_i^\star(\tilde{F}_t^\star)\big)^2 - \frac{1}{T}\sum_{t=1}^T \mathbb{E}\big(\widehat{\varphi}_i(\widehat{\rho}(X_t)) - \varphi_i^\star(F_t^\star)\big)^2\right|. \tag{A21}$$

Note that by Lemma 6 and the fact that $\|X_t^\star\|_\infty \lesssim 1, \|\tilde{X}_t^\star\|_\infty \lesssim 1$, it is straightforward to verify that with probability at least $1 - C\exp(-cT)$, $\max_{1\le t\le T+1}\{\|X_t\|_\infty, \|\tilde{X}_t\|_\infty\} \le 2NT$. Therefore, under this event, Eq. (A20) holds for $X_1, \tilde{X}_1, \ldots, X_{T+1}, \tilde{X}_{T+1}$. Let us define

$$r_{ij} := \max\left(\sqrt{T^{-1}\log|\mathcal{F}_{N,\delta}^1|}, \mathbb{E}^{1/2}\left[\left(\overline{f}_j(X_{T+1}) - \varphi_i^\star\left(F_{T+1}^\star\right)\right)^2\right]\right) \text{ and}$$
$$V_i := \max_{j\in[|\mathcal{F}_{N,\delta}^1|]}\left|\frac{1}{r_{ij}B}\sum_{t=1}^T\left[\big(\overline{f}_j(\tilde{X}_t) - \varphi_i^\star(\tilde{F}_t^\star)\big)^2 - \big(\overline{f}_j(X_t) - \varphi_i^\star(F_t^\star)\big)^2\right]\right|.$$

Additionally, we define $r_i^\star$ as r_{ij} for $j = i^\star$, which is the same as

$$
\begin{aligned}
r_i^\star &= \max\left(\sqrt{T^{-1}\log|\mathcal{F}^1_{N,\delta}|}, \mathbb{E}^{1/2}\left[\left(\overline{f}_{i^\star}(X_{T+1}) - \varphi_i^\star\left(F^\star_{T+1}\right)\right)^2 \middle| \{X_t\}_{t=1}^T\right]\right) \\
&\le \sqrt{T^{-1}\log|\mathcal{F}^1_{N,\delta}|} + \mathbb{E}^{1/2}\left[\left(\widehat{\varphi}_i(\widehat{\rho}(X_{T+1}) - \varphi_i^\star\left(F^\star_{T+1}\right)\right)^2 \middle| \{X_t\}_{t=1}^T\right] + T^{-1}, \qquad \text{(A22)}
\end{aligned}
$$

where the last part follows from triangle inequality and Eq. (A20). Moreover, by Eq. (A20), with probability at least $1 - C\exp(-cT)$,

$$
\begin{aligned}
&\left|\frac{1}{T}\sum_{t=1}^T \left(\widehat{\varphi}_i(\widehat{\rho}(\tilde{X}_t)) - \varphi_i^\star(\tilde{F}_t^\star)\right)^2 - \frac{1}{T}\sum_{t=1}^T \left(\widehat{\varphi}_i(\widehat{\rho}(X_t)) - \varphi_i^\star(F_t^\star)\right)^2\right| \\
\le &\left|\frac{1}{T}\sum_{t=1}^T \left(\overline{f}_{i^\star}(\tilde{X}_t) - \varphi_i^\star(\tilde{F}_t^\star)\right)^2 - \frac{1}{T}\sum_{t=1}^T \left(\overline{f}_{i^\star}(X_t) - \varphi_i^\star(F_t^\star)\right)^2\right| + 9T^{-1}B.
\end{aligned}
$$

With the above inequality and the fact that $\exp(-cT) = o(T^{-1})$, we conclude

$$
\begin{aligned}
&\left|\frac{1}{T}\sum_{t=1}^T \mathbb{E}\left(\widehat{\varphi}_i(\widehat{\rho}(\tilde{X}_t)) - \varphi_i^\star(\tilde{F}_t^\star)\right)^2 - \frac{1}{T}\sum_{t=1}^T \mathbb{E}\left(\widehat{\varphi}_i(\widehat{\rho}(X_t)) - \varphi_i^\star(F_t^\star)\right)^2\right| \\
\le &\left|\frac{1}{T}\sum_{t=1}^T \mathbb{E}\left(\overline{f}_{i^\star}(\tilde{X}_t) - \varphi_i^\star(\tilde{F}_t^\star)\right)^2 - \frac{1}{T}\sum_{t=1}^T \mathbb{E}\left(\overline{f}_{i^\star}(X_t) - \varphi_i^\star(F_t^\star)\right)^2\right| + 10T^{-1}B \\
\le &\frac{B}{T}\mathbb{E}(V_i r_i^\star) + 10T^{-1}B. \qquad \text{(A23)}
\end{aligned}
$$

By Cauchy-Schwarz inequality and Eq. (A22), we have

$$
\begin{aligned}
\mathbb{E}(V_i r_i^\star) &\le \left(\mathbb{E}\left(\mathbb{E}\left[\left(\widehat{\varphi}_i \circ \widehat{\rho}(X_{T+1}) - \varphi_i^\star(F^\star_{T+1})\right)^2 \middle| \{X_t\}_{t=1}^T\right]\right)\right)^{1/2} (\mathbb{E}V_i^2)^{1/2} \\
&\quad + \sqrt{T^{-1}\log|\mathcal{F}^1_{N,\delta}|}\mathbb{E}V_i + T^{-1}\mathbb{E}V_i \qquad \text{(A24)} \\
&= \left(\mathbb{E}\left(\widehat{\varphi}_i \circ \widehat{\rho}(X_{T+1}) - \varphi_i^\star(F^\star_{T+1})\right)^2\right)^{1/2} (\mathbb{E}V_i^2)^{1/2} + \sqrt{T^{-1}\log|\mathcal{F}^1_{N,\delta}|}\mathbb{E}V_i + T^{-1}\mathbb{E}V_i.
\end{aligned}
$$

Next we analyze the term V_i. For any fixed $j \in [|\mathcal{F}^1_{N,\delta}|]$, define $Y_{tj} := \left(\overline{f}_j(\tilde{X}_t) - \varphi_i^\star(\tilde{F}_t^\star)\right)^2 - \left(\overline{f}_j(X_t) - \varphi_i^\star(F_t^\star)\right)^2$. Then it is straightforward to see that $\{Y_{tj}\}_{t=1}^T$ is a series of i.i.d. random variables satisfying $\mathbb{E}Y_{tj} = 0$, $|Y_{tj}| \le 4B^2$, and

$$
\mathrm{Var}(Y_{tj}) = 2\,\mathrm{Var}\left(\left(\overline{f}_j(\tilde{X}_t) - \varphi_i^\star(\tilde{F}_t^\star)\right)^2\right) \le 2\mathbb{E}\left[\left(\overline{f}_j(\tilde{X}_t) - \varphi_i^\star(\tilde{F}_t^\star)\right)^4\right] \le 8B^2 r_{ij}^2.
$$

With the above results, by Bernstein's inequality and union bound inequality, we have

$$\mathrm{P}(V_i \geq x) = \mathrm{P}\left(\max_{j \in [|\mathcal{F}^1_{N,\delta}|]} \left| \frac{1}{r_{ij} B} \sum_{t=1}^{T} Y_{tj} \right| \geq x \right) \leq 2|\mathcal{F}^1_{N,\delta}| \max_{j \in [|\mathcal{F}^1_{N,\delta}|]} \exp\left(-\frac{x^2}{8x/(3r_{ij}) + 16T} \right).$$

Since $r_{ij} \geq \sqrt{T^{-1} \log |\mathcal{F}^1_{N,\delta}|}$, it holds that $\mathrm{P}(V_i \geq x) \leq 2\exp(-3x\sqrt{\log |\mathcal{F}^1_{N,\delta}|}/16\sqrt{T})$ whenever $x \geq 6\sqrt{T \log |\mathcal{F}^1_{N,\delta}|}$. Therefore, when T is large enough, $\mathbb{E}V_i$ equals

$$\begin{aligned} &\int_0^\infty \mathrm{P}(V_i \geq x) dx \leq 6\sqrt{T \log |\mathcal{F}^1_{N,\delta}|} + \int_{6\sqrt{T \log |\mathcal{F}^1_{N,\delta}|}}^\infty 2|\mathcal{F}^1_{N,\delta}| \exp\left(\frac{-3x\sqrt{\log |\mathcal{F}^1_{N,\delta}|}}{16\sqrt{T}} \right) dx \\ &\leq 6\sqrt{T \log |\mathcal{F}^1_{N,\delta}|} + \frac{32}{3}\sqrt{\frac{T}{\log |\mathcal{F}^1_{N,\delta}|}} \leq 7\sqrt{T \log |\mathcal{F}^1_{N,\delta}|}. \end{aligned} \tag{A25}$$

In a similar way, it can be shown that $\mathbb{E}V_i^2 \leq 40T \log |\mathcal{F}^1_{N,\delta}|$. Combining Eq. (A24) and bounds for the moment of V_i, $T^{-1}B\mathbb{E}(V_i r_{ij^*})$ is bounded by a constant multiple of

$$\frac{B}{T}\left(\mathbb{E}^{1/2}\left[\left(\widehat{X}_{T+1,i} - \varphi_i^\star(F^\star_{T+1}) \right)^2 \right] \sqrt{T \log |\mathcal{F}^1_{N,\delta}|} + \sqrt{T^{-1} \log |\mathcal{F}^1_{N,\delta}|} + \log |\mathcal{F}^1_{N,\delta}| \right).$$

Recall that $\log |\mathcal{F}^1_{N,\delta}| \lesssim (T^{\frac{K}{2\beta+K}} + L_N) \log^4(NT)$, implying $T^{-1}B\mathbb{E}(V_i r_{ij^*})$ is bounded by $C(T^{-\frac{2\beta}{2\beta+K}} + T^{-1}L_N) \log^4(NT) + C(T^{-\frac{\beta}{2\beta+K}} + T^{-1/2}L_N^{1/2}) \log^2(NT)\mathbb{E}^{1/2}[\left(\widehat{X}_{T+1,i} - \varphi_i^\star(F^\star_{T+1}) \right)^2]$. Together with Eqs. (A21) and (A23), we complete the proof. □

References

M. Kohler and S. Langer. On the rate of convergence of fully connected deep neural network regression estimates. *The Annals of Statistics*, 49(4):2231 – 2249, 2021.

J. Schmidt-Hieber. Nonparametric regression using deep neural networks with ReLU activation function. *The Annals of Statistics*, 48(4):1875 – 1897, 2020.