Supplementary Materials of “Recurrent Neural Networks for Nonlinear Time Series”

Xiao Chen[*] Yu Chen[†] Zhouyu Shen[‡] Dacheng Xiu[§]

This Version: March 1, 2026

1 RNN variants: LSTM and GRU

This section introduces the architecture of LSTM and GRU (see Figure 1 for an illustration). To fix ideas, we briefly describe the gating mechanisms and the corresponding update equations used by these architectures.

A common innovation in the recurrent structure of LSTM and GRU is the use of “gates.”[1] Intuitively, gates are element-wise nonlinear filters taking values in $[0,1]$ (typically via a sigmoid) that act as soft, multiplicative selectors of information. By modulating how much prior state and new input are written, retained, or exposed at each step, LSTM and GRU mitigate vanishing gradients (a potential issue when training RNNs on data with long-range temporal dependencies). GRUs achieve similar control with a simpler parameterization, which can ease training and reduce variance in small samples.

[*] School of Management, University of Science and Technology of China, cx990621@mail.ustc.edu.cn.

[†] School of Public Affairs, University of Science and Technology of China, cyu@ustc.edu.cn.

[‡] Guanghua School of Management, Peking University, shenzhouyu@gsm.pku.edu.cn.

[§] Booth School of Business, University of Chicago and NBER, dacheng.xiu@chicagobooth.edu.

[1] More specifically, with input I_t, hidden state H_t, and cell state C_t, the LSTM updates are

$$i_t = \sigma_1(W_{i1}H_{t-1} + W_{i2}\rho(I_t) + b_i),\ f_t = \sigma_1(W_{f1}H_{t-1} + W_{f2}\rho(I_t) + b_f),\ o_t = \sigma_1(W_{o1}H_{t-1} + W_{o2}\rho(I_t) + b_o),$$
$$\tilde{C}_t = \sigma_2(W_{c1}H_{t-1} + W_{c2}\rho(I_t) + b_c),\ C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t,\ H_t = o_t \odot \sigma_2(C_t).$$

Here $\sigma_1(\cdot)$ denotes the logistic sigmoid and $\sigma_2(\cdot)$ is typically tanh; $\odot$ is the Hadamard (element-wise) product. Similarly, with input I_t and hidden state H_t, the GRU updates are

$$r_t = \sigma_1(W_{r1}H_{t-1} + W_{r2}\rho(I_t) + b_r),\ z_t = \sigma_1(W_{z1}H_{t-1} + W_{z2}\rho(I_t) + b_z),$$
$$\tilde{H}_t = \sigma_2\big(W_{h1}(r_t \odot H_{t-1}) + W_{h2}\rho(I_t) + b_h\big), H_t = (1 - z_t) \odot H_{t-1} + z_t \odot \tilde{H}_t.$$

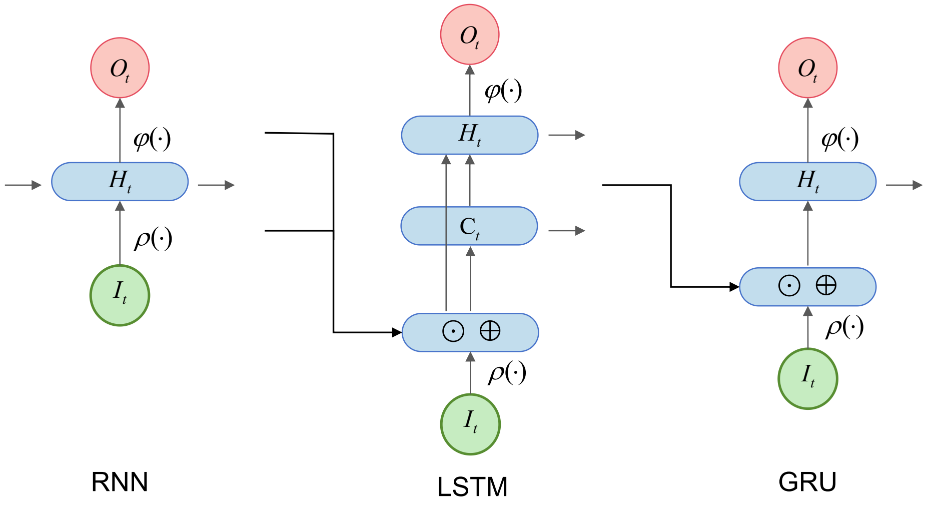

Figure 1: Plain and Gated RNN Architectures

Note: Three panels labeled RNN, LSTM, and GRU. Each shows nodes I_t (input), H_t (hidden), and O_t (output). The recurrent update takes $(\rho(I_t), H_{t-1})$ to H_t via gated operations when applicable, and $\varphi(\cdot)$ maps H_t to O_t. The LSTM includes an additional cell state C_t drawn as a horizontal path alongside H_t; elementwise operations are indicated by $\odot$ (multiplication) and $\oplus$ (addition) on the connections into and out of C_t. The GRU shows only the hidden state pathway (no separate cell state), with $\odot$ and $\oplus$ marking gated elementwise operations on input-hidden and (past) hidden-hidden connections. The RNN panel shows a single hidden-state recurrence.

2 Technical Lemmas and Their Proofs

Throughout the proof we use the following shorthand. Let

$$\Lambda_T := \max\bigl(w^2 w_h \log T,\ w_h^3\bigr) \log^3 T,\ \Delta_T := \Bigl(\max\bigl(w^2 w_h \log T,\ w_h^3\bigr)\Bigr)^{1/2}(T w_h^{-1})^{-1/2} \log^3 T.$$

In nonparametric statistics, econometrics, and learning theory, many convergence and generalization results depend not only on sample size but also on how complex the function class is. For this reason, we introduce:

Definition 1 (Covering Number and Metric Entropy)**.** *Let $\mathcal{F}$ be a class of functions equipped with a norm $\|\cdot\|$. A subset $\mathcal{F}' \subseteq \mathcal{F}$ is called a* δ-cover *of $\mathcal{F}$ if, for every $\varphi \in \mathcal{F}$, there exists $\varphi' \in \mathcal{F}'$ such that $\|\varphi - \varphi'\| \leq \delta$. The* δ-covering number *of $\mathcal{F}$, denoted $\mathcal{N}(\delta, \|\cdot\|, \mathcal{F})$, is the cardinality of the smallest δ-cover of $\mathcal{F}$. The* metric entropy *of $\mathcal{F}$ is defined as the logarithm of its covering number, i.e., $\log \mathcal{N}(\delta, \|\cdot\|, \mathcal{F})$.*

At the outset we work with the range-bounded class $\mathcal{F}_{n_0}^{n_d+1}(d, w, C, B)$ from Section 2.2. For convenience we also introduce a larger class that strictly contains it; by monotonicity of covering numbers under set inclusion, any upper bound for the larger class immediately bounds the covering number of the constrained subclass.

Definition 2. *Denote the function class $\mathcal{F}_{n_0}^{n_{d+1}}(d, w, C)$ as follows:*

$$\mathcal{F}_{n_0}^{n_{d+1}}(d, w, C) := \left\{ f \text{ of the form (2)} : \max_{j=0,\ldots,d} \left(\max(\|W_j\|_\infty, |v_j|_\infty) \right) \le C \right\}.$$

To control temporal dependence, we approximate the hidden state H_t by truncated states $\{H_t^\ell\}_{\ell \ge 0}$ that depend only on the most recent ℓ lags of (Y, X). These finite-memory proxies are used to establish mixing and uniform bounds, and relate to H_t as ℓ-step truncations of its recursion.

Definition 3. *For $s \ge 2$ and $h \in \mathbb{R}^{w_h}$, define the update map*

$$\Phi_s(h) := \sigma_{v_h}\big(\rho(Y_{s-1}, X_{s-1}) + W_h h\big).$$

For $(\rho, W_h, v_h, \varphi) \in \mathcal{F}_{rnn}^{w_h}$ and $2 \le t \le T$, define the truncated states by

$$H_t^0 := \Phi_t(0), \qquad H_t^\ell := \begin{cases} 0, & t \le \ell, \\ \Phi_t\big(H_{t-1}^{\ell-1}\big), & t > \ell, \ \ell \ge 1. \end{cases} \tag{1}$$

Equivalently, for $t > \ell$, $H_t^\ell = \Phi_t \circ \Phi_{t-1} \circ \cdots \circ \Phi_{t-\ell}(0)$. In contrast, the true hidden state follows $H_t = \Phi_t(H_{t-1}) = \Phi_t \circ \Phi_{t-1} \circ \cdots \circ \Phi_2(H_1)$.

We begin with Lemmas 1-3; detailed proofs are provided in the Online Appendix of Shen and Xiu (2024).

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

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

Lemma 2. *Consider the class of neural networks $\mathcal{F}_{n_0}^{n_{d+1}}(d, w, T^{5\beta+5})$ with width vector $n = (n_0, \ldots, n_{d+1})$. Assume that the total number of nonzero weights in the network is bounded by S and the input $\|x\|_\infty \le C$ for some $C \ge 1$. 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:*

(i) Its cardinality satisfies: $|\mathcal{F}_{n_0,\delta}^{n_{d+1}}| \le \left(8\delta^{-1} C T^{(5\beta+5)(d+2)} (1+w)^d n_0 n_{d+1} d\right)^{2S}$;

(ii) 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 $\|x\|_\infty \le C$.

Lemma 3. *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.*

These lemmas supply the regularity, complexity, and approximation ingredients needed for our main result. Lemma 1 provides a uniform Lipschitz bound for networks in $\mathcal{F}_{n_0}^{n_{d+1}}(d, w, T^{5\beta+5})$, ensuring stability of outputs with respect to inputs. Lemma 2 furnishes a covering-number bound that discretizes the hypothesis class with explicit cardinality control, thereby quantifying its capacity. Lemma 3 establishes a uniform approximation guarantee for smooth target functions by networks of controlled depth, width, and parameter magnitude. Together, these results jointly control smoothness, statistical complexity, and approximation error, and they underwrite the bounds on estimation and approximation errors required to derive our convergence rates.

Next, we prove several lemmas specific to RNNs that serve as key ingredients for the subsequent theoretical analysis.

Lemma 4. *Let H_t^ℓ be defined in (1) and let ϵ be as in (4). Then,*

(i) for all integers $t \geq 1$ and $\ell \geq 0$, $\|H_t\|_\infty$, $\|H_t^\ell\|_\infty \leq 4\big(\epsilon^{-1} T^{5\beta+5}\big)^{w_h}$;

(ii) for all $t \geq \ell + 2$ and $\ell \geq 1$, $\|H_t^\ell - H_t\|_\infty \leq 8\,(1-\epsilon)^{\ell - w_h}\,\big(2\ell\,\epsilon^{-1} T^{5\beta+5}\big)^{w_h}$.

Proof. Set $a := T^{5\beta+5}\epsilon^{-1}$. Throughout the proof, absolute values and inequalities are understood elementwise. Using (5), we have

$$\begin{aligned} |H_t| = \big|\sigma_{v_h}\big(\rho(Y_{t-1}, X_{t-1}) + W_h H_{t-1}\big)\big| &\leq |v_h| + |\rho(Y_{t-1}, X_{t-1})| + |W_h H_{t-1}| \\ &\leq 2T^{5\beta+5}\, 1_{w_h} + |W_h H_{t-1}|. \end{aligned}$$

Let $H_{k,t}$ denote the k-th entry of H_t. By (4), W_h is upper triangular with diagonal entries bounded by $1-\epsilon$ and upper-diagonal magnitudes bounded by $T^{5\beta+5}$, hence

$$\begin{aligned} |H_{w_h,t}| &\leq 2T^{5\beta+5} + (1-\epsilon)|H_{w_h,t-1}|, \\ |H_{w_h-1,t}| &\leq 2T^{5\beta+5} + (1-\epsilon)|H_{w_h-1,t-1}| + T^{5\beta+5}|H_{w_h,t-1}|, \\ &\vdots \\ |H_{1,t}| &\leq 2T^{5\beta+5} + (1-\epsilon)|H_{1,t-1}| + T^{5\beta+5}\sum_{i=2}^{w_h} |H_{i,t-1}|. \end{aligned}$$

(i) Uniform bound. With $H_1 = 0_{w_h}$, the first inequality yields $|H_{w_h,t}| \leq 2a$ for all $t \geq 1$. Consequently, from the second inequality, we obtain $|H_{w_h-1,t}| \leq 2a + 2a^2$. Propagating upward gives

$$|H_{i,t}| \leq 2a\big(1 + a + \cdots + a^{w_h - i}\big) = 2a\,\frac{a^{w_h-i+1} - 1}{a-1}.$$

Using the elementary bound $\frac{a^m-1}{a-1} \leq 2a^{m-1}$ (and hence $2a\frac{a^m-1}{a-1} \leq 4a^m$), we obtain $|H_{i,t}| \leq 4\,a^{w_h-i+1}$. Maximizing over i yields $\|H_t\|_\infty \leq 4a^{w_h}$. The same recursion with H_{t-1} replaced by $H_{t-1}^{\ell-1}$ shows $|H_{i,t}^{l}| \leq 4\,a^{w_h-i+1}$ and $\|H_t^\ell\|_\infty \leq 4a^{w_h}$ as well.

(ii) Truncation error. Since σ_{v_h} is 1-Lipschitz elementwise,

$$|H_t^\ell - H_t| = \big|\sigma_{v_h}(\rho + W_h H_{t-1}^{\ell-1}) - \sigma_{v_h}(\rho + W_h H_{t-1})\big| \leq |W_h(H_{t-1}^{\ell-1} - H_{t-1})|.$$

Therefore, by (4),

$$\begin{aligned}
|H_{w_h,t}^\ell - H_{w_h,t}| &\leq (1-\epsilon)|H_{w_h,t-1}^{\ell-1} - H_{w_h,t-1}|,\\
|H_{w_h-1,t}^\ell - H_{w_h-1,t}| &\leq (1-\epsilon)|H_{w_h-1,t-1}^{\ell-1} - H_{w_h-1,t-1}| + T^{5\beta+5}|H_{w_h,t-1}^{\ell-1} - H_{w_h,t-1}|,\\
&\;\vdots\\
|H_{1,t}^\ell - H_{1,t}| &\leq (1-\epsilon)|H_{1,t-1}^{\ell-1} - H_{1,t-1}| + T^{5\beta+5}\sum_{i=2}^{w_h}|H_{i,t-1}^{\ell-1} - H_{i,t-1}|.
\end{aligned} \tag{2}$$

We claim that, for all integers $\ell \geq 1$, $t \geq \ell + 2$, and $1 \leq i \leq w_h$,

$$|H_{i,t}^\ell - H_{i,t}| \leq 8\,(1-\epsilon)^{\ell - w_h + i}\,\big(2\ell\,\epsilon^{-1}T^{5\beta+5}\big)^{\,w_h-i+1}. \tag{3}$$

This immediately implies the desired bound.

Base case $\ell = 1$. From the first line in (2) and the bound in (i),

$$|H_{w_h,t}^1 - H_{w_h,t}| \leq (1-\epsilon)|H_{w_h,t-1}^0 - H_{w_h,t-1}| \leq 8(1-\epsilon)\,a \leq 8(1-\epsilon)\,\big(2\,\epsilon^{-1}T^{5\beta+5}\big).$$

For $k \leq w_h - 1$, (2) gives

$$\begin{aligned}
|H_{k,t}^1 - H_{k,t}| &\leq (1-\epsilon)|H_{k,t-1} - H_{k,t-1}^0| + \sum_{i=k+1}^{w_h} T^{5\beta+5}|H_{i,t-1} - H_{i,t-1}^0|\\
&\leq (1-\epsilon)\cdot 8\,a^{w_h-k+1} + T^{5\beta+5}\sum_{i=k+1}^{w_h} 8\,a^{\,w_h-i+1}\\
&\leq 8(1-\epsilon)a^{w_h-k+1} + 16a^{w_h-k+1} \leq 8(1-\epsilon)^{1-w_h+k}\big(2\,\epsilon^{-1}T^{5\beta+5}\big)^{w_h-k+1},
\end{aligned}$$

which is (3) with $\ell = 1$.

Inductive step. Assume (3) holds for all $1 \le \ell \le \ell_1 - 1$. For $i = w_h$, iterating the first line of (2) yields, for $t \ge \ell_1 + 2$,

$$|H^{\ell_1}_{w_h,t} - H_{w_h,t}| \le (1-\epsilon)^{\ell_1} |H^0_{w_h,t-\ell_1} - H_{w_h,t-\ell_1}| \le 8(1-\epsilon)^{\ell_1}\, a \le 8(1-\epsilon)^{\ell_1} \big(2\ell_1\, \epsilon^{-1} T^{5\beta+5}\big).$$

Now fix $k \le w_h - 1$ and suppose (3) holds for $i = k+1, \dots, w_h$ at level $\ell = \ell_1$. From (2),

$$\begin{aligned}
|H^{\ell_1}_{k,t} - H_{k,t}| &\le (1-\epsilon)|H^{\ell_1-1}_{k,t-1} - H_{k,t-1}| + T^{5\beta+5} \sum_{i=k+1}^{w_h} |H^{\ell_1-1}_{i,t-1} - H_{i,t-1}| \\
&\le (1-\epsilon)^{\ell_1-1} |H^1_{k,t-\ell_1+1} - H_{k,t-\ell_1+1}| \\
&\quad + (\ell_1 - 1) T^{5\beta+5} \sum_{i=k+1}^{w_h} 8\,(1-\epsilon)^{\ell_1 - w_h + i - 1} \big(2\ell_1\, \epsilon^{-1} T^{5\beta+5}\big)^{w_h - i + 1} \\
&=: Q_1 + Q_2, \quad \forall t \ge \ell_1 + 2.
\end{aligned}$$

Using the base case bound at $\ell = 1$ and that $\ell_1 \ge 2$, $w_h - k + 1 \ge 2$,

$$Q_1 \le 2\,(1-\epsilon)^{\ell_1 - w_h + k} \left(2\ell_1\, \epsilon^{-1} T^{5\beta+5}\right)^{w_h - k + 1}.$$

For Q_2, factor the leading term and bound the finite geometric series:

$$\begin{aligned}
Q_2 &\le 4\,(1-\epsilon)^{\ell_1 - w_h + k} \left(2\ell_1\, \epsilon^{-1} T^{5\beta+5}\right)^{w_h - k + 1} \left(1 - \frac{1-\epsilon}{2\ell_1\, \epsilon^{-1} T^{5\beta+5}}\right)^{-1} \\
&\le \frac{16}{3}\,(1-\epsilon)^{\ell_1 - w_h + k} \left(2\ell_1\, \epsilon^{-1} T^{5\beta+5}\right)^{w_h - k + 1},
\end{aligned}$$

since $\frac{1-\epsilon}{2\ell_1\, \epsilon^{-1} T^{5\beta+5}} \le \frac{1}{4}$ when $\ell_1 \ge 2$. Combining Q_1 and Q_2 and loosening constants gives

$$|H^{\ell_1}_{k,t} - H_{k,t}| \le 8\,(1-\epsilon)^{\ell_1 - w_h + k} \left(2\ell_1\, \epsilon^{-1} T^{5\beta+5}\right)^{w_h - k + 1}.$$

This proves (3) for $i = k$, and the induction on i and ℓ completes the proof. □

Lemma 5. *For $\mathcal{F}^{w_h}_{rnn}$ defined in (5) and satisfying the conditions of Theorem 1, there exists a function class $\mathcal{F}^{w_h}_{rnn,\delta} \subset \mathcal{F}^{w_h}_{rnn}$ with $\delta = T^{-1}$ such that:*

(i) Its cardinality satisfies $\log |\mathcal{F}^{w_h}_{rnn,\delta}| \lesssim w^2 w_h \log^4 T + w_h^3 \log^3 T$.

(ii) For any $(\rho, W_h, v_h, \varphi) \in \mathcal{F}^{w_h}_{rnn}$, there exists $(\overline{\rho}, \overline{W}_h, \overline{v}_h, \overline{\varphi}) \in \mathcal{F}^{w_h}_{rnn,\delta}$ such that

$$\big|\varphi(H_t) - \overline{\varphi}(\overline{H}_t)\big| \le \delta = T^{-1} \quad \textit{for all } t \ge C w_h \log^2 T,$$

where $C > 0$ is a fixed constant independent of $(\rho, W_h, v_h, \varphi)$ and T.

Proof. Recall H_t^ℓ from (1). By Lemma 4 (ii), for $t \ge \ell + 2$,

$$\|H_t - H_t^\ell\|_\infty \;\le\; 8\,(1-\epsilon)^{\ell - w_h}\left(2\ell\,\epsilon^{-1}T^{5\beta+5}\right)^{w_h}.$$

Let $(\overline{\rho}, \overline{W}_h, \overline{v}_h, \overline{\varphi}) \in \mathcal{F}_{rnn,\delta}^{w_h}$ satisfy,

$$\|W_h - \overline{W}_h\|_\infty\,, \quad \|v_h - \overline{v}_h\|_\infty\,, \quad \|\rho - \overline{\rho}\|_\infty \;\le\; \xi. \tag{4}$$

Then for $t \ge \ell + 2$,

$$\begin{aligned} \|H_t - \overline{H}_t\|_\infty &\le \|H_t^\ell - \overline{H}_t^\ell\|_\infty + \|H_t - H_t^\ell\|_\infty + \|\overline{H}_t - \overline{H}_t^\ell\|_\infty \\ &\le \|H_t^\ell - \overline{H}_t^\ell\|_\infty + 16\,(1-\epsilon)^{\ell - w_h}\left(2\ell\,\epsilon^{-1}T^{5\beta+5}\right)^{w_h}. \end{aligned} \tag{5}$$

Using the elementwise inequality $|\sigma_v(x) - \sigma_u(y)| \le |v-u| + |x-y|$ and Lemma 4 (i) (applied to bound $\|\overline{H}_{t-1}^{\ell-1}\|_\infty$), we have

$$\begin{aligned} \|H_t^\ell - \overline{H}_t^\ell\|_\infty &= \left\|\sigma_{v_h}\big(\rho(Y_{t-1}, X_{t-1}) + W_h H_{t-1}^{\ell-1}\big) - \sigma_{\overline{v}_h}\big(\overline{\rho}(Y_{t-1}, X_{t-1}) + \overline{W}_h \overline{H}_{t-1}^{\ell-1}\big)\right\|_\infty \\ &\le \|v_h - \overline{v}_h\|_\infty + \|\rho - \overline{\rho}\|_\infty + \|W_h H_{t-1}^{\ell-1} - \overline{W}_h \overline{H}_{t-1}^{\ell-1}\|_\infty \\ &\le 2\xi + w_h\|W_h - \overline{W}_h\|_\infty\,\|\overline{H}_{t-1}^{\ell-1}\|_\infty + w_h\|W_h\|_\infty\,\|H_{t-1}^{\ell-1} - \overline{H}_{t-1}^{\ell-1}\|_\infty \\ &\le 6\,\xi\,w_h(\epsilon^{-1}T^{5\beta+5})^{w_h} + w_h T^{5\beta+5}\,\|H_{t-1}^{\ell-1} - \overline{H}_{t-1}^{\ell-1}\|_\infty, \end{aligned}$$

where we used $2\xi \le 2\xi w_h(\epsilon^{-1}T^{5\beta+5})^{w_h}$, $\|W_h - \overline{W}_h\|_\infty \le \xi$, $\|\overline{H}_{t-1}^{\ell-1}\|_\infty \le 4(\epsilon^{-1}T^{5\beta+5})^{w_h}$ from Lemma 4 (i), and $\|W_h\|_\infty \le T^{5\beta+5}$.

Iterating this inequality yields

$$\begin{aligned} \|H_t^\ell - \overline{H}_t^\ell\|_\infty &\le 6\,\xi\,w_h\Big(\epsilon^{-1}T^{5\beta+5}\Big)^{w_h}\sum_{k=0}^{\ell-1}(w_h T^{5\beta+5})^k + (w_h T^{5\beta+5})^\ell\,\|H_{t-\ell}^0 - \overline{H}_{t-\ell}^0\|_\infty \\ &\le 6\,\xi\,w_h\Big(\epsilon^{-1}T^{5\beta+5}\Big)^{w_h}(w_h T^{5\beta+5})^\ell + (w_h T^{5\beta+5})^\ell\,\|\sigma_{v_h}(\rho(\cdot)) - \sigma_{\overline{v}_h}(\overline{\rho}(\cdot))\|_\infty \\ &\le 6\,\xi\,w_h\Big(\epsilon^{-1}T^{5\beta+5}\Big)^{w_h}(w_h T^{5\beta+5})^\ell + 2\xi\,(w_h T^{5\beta+5})^\ell \\ &\le 8\,\xi\,w_h\Big(\epsilon^{-1}T^{5\beta+5}\Big)^{w_h}(w_h T^{5\beta+5})^\ell. \end{aligned} \tag{6}$$

Combining (5) and (6), for $t \ge \ell + 2$ we obtain

$$\|H_t - \overline{H}_t\|_\infty \le 8\,\xi\,w_h\Big(\epsilon^{-1}T^{5\beta+5}\Big)^{w_h}(w_h T^{5\beta+5})^\ell + 16\,(1-\epsilon)^{\ell - w_h}\left(2\ell\,\epsilon^{-1}T^{5\beta+5}\right)^{w_h}$$

$$\leq 24\left(2\ell\,\epsilon^{-1}T^{5\beta+5}\right)^{w_h}\Big(\xi\, w_h\,(w_h T^{5\beta+5})^{\ell} + (1-\epsilon)^{\ell-w_h}\Big).$$

Choose $\ell = C_1 w_h \log^2 T$ and $\xi = \exp(-C_2 w_h \log^3 T)$. Since $w_h \asymp T^{\alpha_2}\log T$ with $\alpha_2 < 1$, we have $w_h \leq T$ for large T. Then

$$\begin{aligned}\xi\, w_h\,(w_h T^{5\beta+5})^{\ell} &= \exp\Big(-C_2 w_h \log^3 T + C_1 w_h \log^2 T \log(w_h T^{5\beta+5}) + \log w_h\Big)\\ &\leq \exp\Big(-C_2 w_h \log^3 T + C_1(5\beta{+}6) w_h \log^3 T + \log T\Big),\end{aligned}$$

and

$$(1-\epsilon)^{\ell-w_h} = \exp\Big((\ell - w_h)\log(1-\epsilon)\Big) = \exp\Big(C_1 w_h \log^2 T\, \log(1-\epsilon) - w_h \log(1-\epsilon)\Big).$$

Hence, for any fixed $C_0 > 0$, by taking C_1, C_2 sufficiently large (independent of T and the parameters) we can ensure $\max(\xi\, w_h\,(w_h T^{5\beta+5})^{\ell}, (1-\epsilon)^{\ell-w_h}) \leq \exp\big(-C_0 w_h \log^2 T\big)$. Therefore,

$$\|H_t - \overline{H}_t\|_\infty \leq 48\left(2\ell\,\epsilon^{-1}T^{5\beta+5}\right)^{w_h}\ \exp\big(-C_0 w_h \log^2 T\big).$$

By Lemma 1, for $\varphi, \overline{\varphi} \in \mathcal{F}^r_{w_h}(d, w, T^{5\beta+5})$,

$$\|\varphi(H_t) - \overline{\varphi}(\overline{H}_t)\|_\infty \leq w_h T^{(5\beta+5)(d+1)} w^d \|H_t - \overline{H}_t\|_\infty + \|\varphi - \overline{\varphi}\|_\infty.$$

Since $d \lesssim \log T$, $w \leq T$, $w_h \leq T$, and $\ell = C_1 w_h \log^2 T$, we can choose C_0 large enough so that

$$48\, w_h T^{(5\beta+5)(d+1)} w^d \left(2\ell\,\epsilon^{-1}T^{5\beta+5}\right)^{w_h}\ \exp\big(-C_0 w_h \log^2 T\big)\ \leq\ (2T)^{-1}.$$

Thus, whenever $\|\varphi - \overline{\varphi}\|_\infty \leq (2T)^{-1}$ and (4) holds, we have

$$\|\varphi(H_t) - \overline{\varphi}(\overline{H}_t)\|_\infty \leq T^{-1} \quad \text{for all } t \geq \ell + 2 = C_1 w_h \log^2 T + 2,$$

which proves (ii).

For (i), combine coverings of φ at accuracy $(2T)^{-1}$ with coverings of W_h, v_h, ρ at accuracy $\xi = \exp(-C_2 w_h \log^3 T)$. The total number of functions is at most $\mathcal{N}_1\mathcal{N}_2\mathcal{N}_3\mathcal{N}_4$, where

$$\begin{aligned}\mathcal{N}_1 &:= \mathcal{N}\Big((2T)^{-1},\ \|\cdot\|_{\infty,\ [-4(\epsilon^{-1}T^{5\beta+5})^{w_h},\,4(\epsilon^{-1}T^{5\beta+5})^{w_h}]^{w_h}},\ \mathcal{F}^r_{w_h}(d, w, T^{5\beta+5}, B)\Big),\\ \mathcal{N}_2 &:= \mathcal{N}\Big(\xi,\ \|\cdot\|_{\infty,\mathcal{B}^{r+k}},\ \mathcal{F}^{w_h}_{r+k}(d, w, T^{5\beta+5}, T^{5\beta+5})\Big),\\ \mathcal{N}_3 &:= \mathcal{N}\Big(\xi,\ \|\cdot\|_\infty,\ \{v_h \in \mathbb{R}^{w_h} : \|v_h\|_\infty \leq T^{5\beta+5}\}\Big),\\ \mathcal{N}_4 &:= \mathcal{N}\Big(\xi,\ \|\cdot\|_\infty,\ \{W_h :\ W_h \text{ of the form in (4)}\}\Big),\end{aligned}$$

where $\|f\|_{\infty,\Omega}$ denotes the sup norm on the domain Ω for function $f(\cdot)$. By Lemma 2(i) and $\xi = \exp(-C_2 w_h \log^3 T)$,

$$\begin{aligned} &\log \mathcal{N}_1 \lesssim (w^2 w_h \log T + w w_h^2) \log^2 T, \qquad \log \mathcal{N}_2 \lesssim (w^2 w_h \log T + w w_h^2) \log^3 T, \\ &\text{and directly,} \qquad \log \mathcal{N}_3 \lesssim w_h^2 \log^3 T, \qquad \log \mathcal{N}_4 \lesssim w_h^3 \log^3 T. \end{aligned}$$

Using $xy \leq x^2 + y^2$ to absorb mixed terms (in w and w_h), we obtain

$$\log |\mathcal{F}_{rnn,\delta}^{w_h}| \;=\; \sum_{i=1}^{4} \log \mathcal{N}_i \;\lesssim\; w^2 w_h \log^4 T \;+\; w_h^3 \log^3 T,$$

which proves (i). □

Lemma 6. *Let $\{Y_t\}_{t\in\mathbb{Z}}$ be a stationary α-mixing process with coefficients $\alpha(j) \leq e^{-\kappa j}$ for some $\kappa > 0$. Let $f : \mathbb{R} \to \mathbb{R}$ be Borel measurable with $|f(Y_t)| \leq B$ a.s. and $\mathbb{E}[f(Y_t)] = 0$. Then there exists a constant $C_\kappa > 0$ depending only on κ such that, for all $T \geq 2$ and $x \geq 0$,*

$$\mathrm{P}\Big(\big|S_T\big| \geq x\Big) \;\leq\; \exp\left(-\frac{C_\kappa\, x^2}{2T \log T\;\; \mathrm{Var}(f(Y_1)) \;+\; B^2 \;+\; xB(\log T)^2}\right), \quad S_T := \sum_{t=1}^{T} f(Y_t).$$

Proof. Set $v^2 := \mathrm{Var}(f(Y_1)) + 2\sum_{t>1}|\operatorname{Cov}(f(Y_t), f(Y_1))|$. By Theorem 2 of Merlevède et al. (2009), there exists $C > 0$, which depends only on κ, such that, for all $T \geq 2$,

$$\mathrm{P}\Big(\big|S_T\big| \geq x\Big) \;\leq\; \exp\left(-\frac{C\, x^2}{v^2 T + B^2 + xB(\log T)^2}\right), \quad \forall x \geq 0. \tag{7}$$

For $t \geq 2$, the covariance inequality for α-mixing and bounded functions (Lemma 3 of Doukhan (1994)) gives $\big|\operatorname{Cov}(f(Y_t), f(Y_1))\big| \leq 4B^2\, \alpha(t-1)$. Hence, for any integer $k \geq 1$,

$$\begin{aligned} v^2 &\leq \mathrm{Var}(f(Y_1)) + 2\sum_{t=2}^{k} \mathrm{Var}(f(Y_1)) \;+\; 2\sum_{t>k} \big|\operatorname{Cov}(f(Y_t), f(Y_1))\big| \\ &\leq (2k-1)\; \mathrm{Var}(f(Y_1)) \;+\; 8B^2 \sum_{m\geq k} \alpha(m) \leq (2k-1)\; \mathrm{Var}(f(Y_1)) \;+\; \frac{8}{1-e^{-\kappa}}\, B^2 e^{-\kappa k}. \end{aligned}$$

With $k := \lceil \kappa^{-1} \log T \rceil$ we have $e^{-\kappa k} \leq T^{-1}$, so

$$v^2 \;\leq\; 2\kappa^{-1} \log T\;\; \mathrm{Var}(f(Y_1)) \;+\; \frac{8}{1-e^{-\kappa}}\, B^2\, T^{-1}.$$

Plugging this into (7) and absorbing the factors κ^{-1} and $8(1-e^{-\kappa})^{-1}$ into the universal constant yields the claim with $C_\kappa > 0$ depending only on κ. $\square$

Lemma 7. *For H_t^ℓ defined in (1), there exists a fixed constant $C^\star > 0$, independent of $(\rho, W_h, v_h, \varphi)$ and T, such that with $\ell^\star = C^\star w_h \log^2 T$,*

$$\sup_{t \geq \ell^\star + 2} \left\| \varphi(H_t) - \varphi(H_t^{\ell^\star}) \right\|_\infty \;\leq\; T^{-1}. \tag{8}$$

Proof. By Lemma 4 (ii), for every $t \geq \ell + 2$,

$$\|H_t^\ell - H_t\|_\infty \;\leq\; 8\,(1-\epsilon)^{\ell - w_h} \left(2\ell\,\epsilon^{-1} T^{5\beta+5}\right)^{w_h}.$$

Since $\varphi \in \mathcal{F}_{w_h}^r(d, w, T^{5\beta+5}, B)$, Lemma 1 yields the (coordinatewise) Lipschitz bound

$$\|\varphi(z) - \varphi(z')\|_\infty \;\leq\; L_\varphi\,\|z - z'\|_\infty, \qquad L_\varphi \;\leq\; w_h\,T^{(5\beta+5)(d+1)}\,w^d.$$

Combining the last two displays gives, for every $t \geq \ell + 2$,

$$\|\varphi(H_t) - \varphi(H_t^\ell)\|_\infty \;\leq\; 8\,w_h\,T^{(5\beta+5)(d+1)} w^d\,(1-\epsilon)^{\ell - w_h} \left(2\ell\,\epsilon^{-1} T^{5\beta+5}\right)^{w_h}.$$

The right-hand side does not depend on t, hence the same bound holds with $\sup_{t \geq \ell+2}$ on the left.

It remains to choose ℓ. Using the structural relations $d \lesssim \log T$ and $\max(w, w_h) \leq T$, set $\ell = \ell^\star := C^\star w_h \log^2 T$. Since $\log(1-\epsilon) \leq -\epsilon$, the factor

$$(1-\epsilon)^{\ell^\star - w_h} = \exp\!\big((\ell^\star - w_h)\log(1-\epsilon)\big) \leq \exp\!\big(-\epsilon\,C^\star w_h \log^2 T + \epsilon\,w_h\big)$$

decays as $\exp\!\big(-\epsilon\,C^\star w_h \log^2 T\big)$, which dominates the polynomial growth coming from $w_h T^{(5\beta+5)(d+1)} w^d \big(2\ell^\star \epsilon^{-1} T^{5\beta+5}\big)^{w_h}$. Hence, by taking $C^\star > 0$ sufficiently large (independent of $(\rho, W_h, v_h, \varphi)$ and T), we ensure

$$\|\varphi(H_t) - \varphi(H_t^\ell)\|_\infty \leq 8\,w_h\,T^{(5\beta+5)(d+1)} w^d\,(1-\epsilon)^{\ell^\star - w_h} \left(2\ell^\star\,\epsilon^{-1} T^{5\beta+5}\right)^{w_h} \;\leq\; T^{-1}. \qquad \square$$

Lemma 8. *Under the conditions in Theorem 1, for any fixed $(\rho, W_h, v_h, \varphi) \in \mathcal{F}_{rnn}^{w_h}$ and fixed $c > 0$, there exists $C < \infty$ independent of $(\rho, W_h, v_h, \varphi)$ and T such that, with probability at least $1 - (\ell^\star + 2)\exp(-c\Lambda_T)$,*

$$\left| \frac{1}{T} \sum_{t=\ell^\star}^{T} \varepsilon_t^\top \big(\mathbb{E}_{t-1} Y_t - \varphi(H_t)\big) \right| \;\leq\; C \left(\frac{1}{T} \sum_{t=\ell^\star}^{T} \mathbb{E}\big\| \mathbb{E}_{t-1} Y_t - \varphi(H_t) \big\|^2 \right)^{1/2} \Delta_T \;+\; C\,\Delta_T^2 \log T,$$

for all T large enough.

Proof. For integers $i \le j$, let $\mathfrak{F}_i^j$ be the sigma-algebra generated by $\{(Y_s, \varepsilon_s, X_s) : i \le s \le j\}$. By unrolling (1) for ℓ steps, H_t^ℓ depends only on $\{(Y_s, \varepsilon_s, X_s) : t-1-\ell \le s \le t-1\}$, hence H_t^ℓ is $\mathfrak{F}_{t-1-\ell}^{t-1}$-measurable for $t \ge \ell + 2$.

Fix $\ell = \ell^\star$ as in (8) and define, for $t \ge \ell^\star + 2$,

$$D_t := \varepsilon_t^\top \big(\mathbb{E}_{t-1} Y_t - \varphi(H_t^{\ell^\star})\big).$$

Then $\mathbb{E}(D_t) = 0$ since $H_t^{\ell^\star}$ is $\mathfrak{F}_{t-1-\ell^\star}^{t-1}$-measurable and $\mathbb{E}_{t-1}\varepsilon_t = 0$. Moreover,

$$|D_t| \le \|\varepsilon_t\|_1 \big\|\mathbb{E}_{t-1} Y_t - \varphi(H_t^{\ell^\star})\big\|_\infty \le 2rB^2 =: M, \tag{9}$$

$$\mathbb{E}|D_t|^2 \le \mathbb{E}\Big(\|\varepsilon_t\|^2 \big\|\mathbb{E}_{t-1} Y_t - \varphi(H_t^{\ell^\star})\big\|^2\Big) \le rB^2\, \mathbb{E}\big\|\mathbb{E}_{t-1} Y_t - \varphi(H_t^{\ell^\star})\big\|^2, \tag{10}$$

where the last inequality uses $\|\varepsilon_t\|^2 \le rB^2$.

Let $L := \ell^\star + 2$ and form the L interlaced subsequences $\{D_{L+j+kL}\}_{k\ge 0}$, $j = 0, \dots, L-1$ (truncated when indices exceed T). Each subsequence is α-mixing with coefficients bounded by $\alpha_L(u) \le e^{-\kappa u}$. Denote $n := \lfloor T/L \rfloor$. By Lemma 6 (a Craig-Bernstein inequality for α-mixing arrays), for all $x > 0$,

$$\mathrm{P}\left(\Big|\sum_{k=0}^{n-1} D_{L+j+kL}\Big| \ge x\right) \le \exp\left(-\frac{C\,x^2}{2n(\log n)\,\mathbb{E}|D_t|^2 + M^2 + Mx(\log n)^2}\right).$$

By the union bound, writing $A := 2n(\log n)\mathbb{E}|D_t|^2 + M^2$ and $B := M(\log n)^2$,

$$\mathrm{P}\left(\Big|\sum_{t=L}^{T} D_t\Big| \ge Tz\right) \le L \exp\left(-\frac{C\,(Tz/L)^2}{A + B(Tz/L)}\right), \quad \forall z > 0. \tag{11}$$

Set $y := Tz/L$. Note that $\frac{t^2}{a+bt} \ge \frac{1}{2}\min\Big(\frac{t^2}{a}, \frac{t}{b}\Big)$ holds for all $a, b, t > 0$. Applying this in (11) yields

$$\mathrm{P}\left(\Big|\sum_{t=L}^{T} D_t\Big| \ge Tz\right) \le L \exp\left(-\frac{C}{2}\min\left(\frac{y^2}{A}, \frac{y}{B}\right)\right).$$

Hence, for any $\vartheta > 0$, choosing

$$y \;\ge\; \max\left(\sqrt{\frac{2A\vartheta}{C}}, \frac{2B\vartheta}{C}\right) \qquad \text{(e.g. take } y = \sqrt{\tfrac{2A\vartheta}{C}} + \tfrac{2B\vartheta}{C}\text{)} \tag{12}$$

makes the exponent at least ϑ. Equivalently, with probability at least $1 - Le^{-\vartheta}$,

$$\frac{1}{T}\Big|\sum_{t=L}^{T} D_t\Big| \le \frac{L}{T}\left(\sqrt{\frac{2A\,\vartheta}{C}} + \frac{2B\,\vartheta}{C}\right).$$

Using $\sqrt{A} \le \sqrt{2n(\log n)\mathbb{E}|D_t|^2} + M$ and $Ln \le T$ gives

$$\frac{1}{T}\Big|\sum_{t=L}^{T} D_t\Big| \le C\left(\sqrt{\frac{L}{T}\,(\log n)\,\mathbb{E}|D_t|^2\,\vartheta} + \frac{L}{T}\,M(\log n)^2\,\vartheta + \frac{L}{T}\,M\sqrt{\vartheta}\right). \tag{13}$$

Take $\vartheta = c\Lambda_T$ and use $L = \ell^\star + 2 \asymp w_h \log^2 T$, $n \asymp T/L$, and $\log n \le \log T$. With probability at least $1 - Le^{-c\Lambda_T}$, combining (13) with (10) yields

$$\begin{aligned}\frac{1}{T}\Big|\sum_{t=\ell^\star+2}^{T} D_t\Big| \le C\sqrt{\frac{\ell^\star}{T}\,\frac{1}{T-\ell^\star-1} rB^2 \sum_{t=\ell^\star+2}^{T} \mathbb{E}\big\|\mathbb{E}_{t-1}Y_t - \varphi(H_t^{\ell^\star})\big\|^2}\,\sqrt{\Lambda_T \log n} \\ +\; C\frac{\ell^\star}{T}\,M(\log n)^2\Lambda_T \;+\; C\frac{\ell^\star}{T}\,M\sqrt{\Lambda_T}.\end{aligned} \tag{14}$$

Since $T - \ell^\star - 1 \ge T/2$ for large T and $\sqrt{\Lambda_T \log n} \le \sqrt{\max(w^2 w_h \log T,\ w_h^3)}\,\log^2 T$, the first term in (14) is bounded by

$$C\left(\frac{2rB^2}{T}\sum_{t=\ell^\star+2}^{T} \mathbb{E}\|\mathbb{E}_{t-1}Y_t - \varphi(H_t^{\ell^\star})\|^2\right)^{1/2} \underbrace{\left(\sqrt{\tfrac{\ell^\star}{T}}\sqrt{\max(w^2 w_h \log T,\ w_h^3)}\,\log^2 T\right)}_{\le C\,\Delta_T},$$

and, using (9) together with $L \asymp w_h \log^2 T$ and $\log n \le \log T$,

$$\frac{\ell^\star}{T}\,M(\log n)^2\Lambda_T \;\lesssim\; \Delta_T^2 \log T, \qquad \frac{\ell^\star}{T}\,M\sqrt{\Lambda_T} \;\lesssim\; \Delta_T^2 \log T.$$

Therefore,

$$\frac{1}{T}\Big|\sum_{t=\ell^\star+2}^{T} D_t\Big| \le C\left(\frac{2rB^2}{T}\sum_{t=\ell^\star+2}^{T} \mathbb{E}\big\|\mathbb{E}_{t-1}Y_t - \varphi(H_t^{\ell^\star})\big\|^2\right)^{1/2} \Delta_T + C\,\Delta_T^2 \log T.$$

Finally, write

$$\begin{aligned}\frac{1}{T}\sum_{t=\ell^\star}^{T} \varepsilon_t^\top\big(\mathbb{E}_{t-1}Y_t - \varphi(H_t)\big) = \frac{1}{T}\sum_{t=\ell^\star+2}^{T} D_t + \frac{1}{T}\sum_{t=\ell^\star}^{\ell^\star+1} \varepsilon_t^\top\big(\mathbb{E}_{t-1}Y_t - \varphi(H_t)\big) \\ + \frac{1}{T}\sum_{t=\ell^\star+2}^{T} \varepsilon_t^\top\big(\varphi(H_t^{\ell^\star}) - \varphi(H_t)\big).\end{aligned}$$

By (8) and $\|\varepsilon_t\|_1 \le rB$, the last two terms contribute at most $10T^{-1}rB^2$. Combining the pieces and applying $\sqrt{x+y} \le \sqrt{x} + \sqrt{y}$ to absorb negligible terms, we conclude that, for all T large enough,

$$\left| \frac{1}{T} \sum_{t=\ell^\star}^{T} \varepsilon_t^\top \big(\mathbb{E}_{t-1} Y_t - \varphi(H_t)\big) \right| \le \tilde{C} \left(\frac{1}{T} \sum_{t=\ell^\star}^{T} \mathbb{E}\big\|\mathbb{E}_{t-1} Y_t - \varphi(H_t)\big\|^2 \right)^{1/2} \Delta_T + \tilde{C}\, \Delta_T^2 \log T,$$

where $C, \tilde{C}$ are constants independent of $(\rho, W_h, v_h, \varphi)$ and T. □

Lemma 9. *Under the conditions of Theorem 1, for any fixed $(\rho, W_h, v_h, \varphi) \in \mathcal{F}_{rnn}^{w_h}$ and any fixed $c > 0$, there exists a constant $C > 0$, independent of $(\rho, W_h, v_h, \varphi)$ and T, such that, for the iterated prediction procedure and any fixed $h \in \mathbb{N}^+$, the following holds with probability at least $1 - (\ell^\star + 2)\exp(-c\Lambda_T)$:*

$$\left| \frac{1}{T} \sum_{t=2}^{T} \big\|\mathbb{E}_{t-1} Y_t - \varphi(H_t)\big\|^2 - \mathbb{E}\big\|\mathbb{E}_{T+h-1} Y_{T+h} - \varphi(H_{T+h})\big\|^2 \right|$$
$$\le C \left(\frac{1}{T} \sum_{t=\ell^\star}^{T} \mathbb{E}\big\|\mathbb{E}_{t-1} Y_t - \varphi(H_t)\big\|^2 \right)^{1/2} \Delta_T + C\, \Delta_T^2 \log T,$$

for all T large enough. Here $\ell^\star$ is defined in (8).

Proof. For any $\ell \ge 1$, define

$$\tilde{D}_t := \big\|\mathbb{E}_{t-1} Y_t - \varphi(H_t^\ell)\big\|^2 - \mathbb{E}\big\|\mathbb{E}_{T+h-1} Y_{T+h} - \varphi(H_{T+h}^\ell)\big\|^2, \qquad t \ge \ell + 2.$$

Since H_t^ℓ depends only on $\{(Y_s, \varepsilon_s, X_s) : t-1-\ell \le s \le t-1\}$, we have $\tilde{D}_t \in \mathfrak{F}_{t-1-\ell}^{t}$. By stationarity (time-homogeneity) of the iterated procedure, $\mathbb{E}\tilde{D}_t = 0$. Moreover, using $\|\mathbb{E}_{t-1} Y_t - \varphi(\cdot)\| \le 2\sqrt{r}\, B$, we obtain

$$|\tilde{D}_t| \le 4rB^2 \quad \text{and} \quad \mathbb{E}\tilde{D}_t^2 \le (4rB^2)\, \mathbb{E}\big\|\mathbb{E}_{t-1} Y_t - \varphi(H_t^\ell)\big\|^2,$$

where the second bound follows from $0 \le X \le 4rB^2 \Rightarrow X^2 \le (4rB^2)X$ with $X = \|\mathbb{E}_{t-1} Y_t - \varphi(H_t^\ell)\|^2$.

From here, apply exactly the same interlacing/blocking and Craig-Bernstein concentration argument as in Lemma 8, with the random array $\{\tilde{D}_t\}$ in place of the variables used there. This yields the desired probability level $1 - (\ell^\star + 2)\exp(-c\Lambda_T)$ after the union bound over the $L = \ell^\star + 2$ interlaced subsequences. Finally, use (8) to replace $H_t^{\ell^\star}$ by H_t (and similarly at time $T + h$), and absorb boundary terms for $t < \ell^\star + 2$ into $C\, \Delta_T^2 \log T$, exactly as in Lemma 8. This gives the stated inequality. □

The following lemma strengthens Lemma 8 by allowing $(\rho, W_h, v_h, \varphi)$ to be chosen in a data-dependent way. The proof covers $\mathcal{F}_{rnn}^{w_h}$ by a finite δ-net $\mathcal{F}_{rnn,\delta}^{w_h}$ and applies a union bound to obtain a uniform high-probability inequality; hence the bound holds for any (random) selector as well.

Lemma 10. *Under the conditions of Theorem 1, for any $(\rho^*, W_h^*, v_h^*, \varphi^*) \in \mathcal{F}_{rnn}^{w_h}$, possibly depending on the in-sample data, the following bounds hold with probability at least $1 - C\exp(-\Lambda_T)$ for some fixed constant C independent of $(\rho^*, W_h^*, v_h^*, \varphi^*)$ and T:*

$$\left|\frac{1}{T}\sum_{t=2}^{T} \varepsilon_t^\top\big(\mathbb{E}_{t-1}Y_t - \varphi^*(H_t^*)\big)\right| \lesssim \Delta_T^2 \log T \;+\; \left(\frac{1}{T}\sum_{t=2}^{T}\mathbb{E}\big\|\mathbb{E}_{t-1}Y_t - \varphi^*(H_t^*)\big\|^2\right)^{1/2}\Delta_T, \tag{15}$$

$$\left|\frac{1}{T}\sum_{t=2}^{T}\big[\|\mathbb{E}_{t-1}Y_t - \varphi^*(H_t^*)\|^2 - \mathbb{E}\|\mathbb{E}_T Y_{T+1} - \varphi^*(H_{T+1}^*)\|^2\big]\right|$$
$$\lesssim \frac{1}{T}\sum_{t=2}^{T}\mathbb{E}\big\|\mathbb{E}_{t-1}Y_t - \varphi^*(H_t^*)\big\|^2 \;+\; \Delta_T^2 \log T. \tag{16}$$

Here $H_t^* := \sigma_{v_h^*}\big(\rho^*(Y_{t-1}, X_{t-1}) + W_h^* H_{t-1}^*\big)$ and $H_1^* := 0$.

Proof. By Lemmas 5 and 7, there exists a finite subclass $\mathcal{F}_{rnn,\delta}^{w_h} \subset \mathcal{F}_{rnn}^{w_h}$ such that for any $(\rho, W_h, v_h, \varphi) \in \mathcal{F}_{rnn}^{w_h}$, there is a corresponding $(\overline{\rho}, \overline{W}_h, \overline{v}_h, \overline{\varphi}) \in \mathcal{F}_{rnn,\delta}^{w_h}$ with

$$\big\|\varphi(H_t) - \overline{\varphi}(\overline{H}_t)\big\|_\infty \le T^{-1} \qquad \text{for all } t \ge C_1 w_h \log^2 T, \tag{17}$$

where $C_1 > 0$ is universal and

$$\log|\mathcal{F}_{rnn,\delta}^{w_h}| \;\lesssim\; w^2 w_h \log^4 T \;+\; w_h^3 \log^3 T.$$

By construction, there exists $(\overline{\rho}^*, \overline{W}_h^*, \overline{v}_h^*, \overline{\varphi}^*) \in \mathcal{F}_{rnn,\delta}^{w_h}$ such that

$$\big\|\varphi^*(H_t^*) - \overline{\varphi}^*(\overline{H}_t^*)\big\|_\infty \le T^{-1} \qquad \text{for all } t \ge C_1 w_h \log^2 T, \tag{18}$$

where $\overline{H}_t^* := \sigma_{\overline{v}_h^*}\Big(\overline{\rho}^*(Y_{t-1}, X_{t-1}) + \overline{W}_h^*\overline{H}_{t-1}^*\Big)$ and $\overline{H}_1^* := 0$.

Step 1. Fix $f = (\rho, W_h, v_h, \varphi) \in \mathcal{F}_{rnn,\delta}^{w_h}$ and define the event

$$\mathcal{E}_f := \left\{\left|\frac{1}{T}\sum_{t=\ell^\star}^{T}\varepsilon_t^\top\big(\mathbb{E}_{t-1}Y_t - \varphi(H_t)\big)\right| \le C_2\left(\frac{1}{T}\sum_{t=\ell^\star}^{T}\mathbb{E}\big\|\mathbb{E}_{t-1}Y_t - \varphi(H_t)\big\|^2\right)^{1/2}\Delta_T + C_2\,\Delta_T^2\log T\right\}.$$

By Lemma 8, for each fixed f,

$$\mathrm{P}(\mathcal{E}_f^c) \;\leq\; L\, e^{-c\Lambda_T}.$$

Now apply the union bound over the finite class $\mathcal{F}^{w_h}_{rnn,\delta}$:

$$\mathrm{P}\Big(\bigcap\nolimits_{f\in\mathcal{F}^{w_h}_{rnn,\delta}} \mathcal{E}_f\Big) \;\geq\; 1-\sum\nolimits_{f\in\mathcal{F}^{w_h}_{rnn,\delta}} \mathrm{P}(\mathcal{E}_f^c) \;\geq\; 1-|\mathcal{F}^{w_h}_{rnn,\delta}|\, L\, e^{-c\Lambda_T}.$$

Using $\log|\mathcal{F}^{w_h}_{rnn,\delta}| \lesssim w^2 w_h \log^4 T + w_h^3\log^3 T$ and $L \asymp w_h \log^2 T$, we have

$$\log\big(|\mathcal{F}^{w_h}_{rnn,\delta}|\, L\big) = O\big(w^2 w_h\log^4 T + w_h^3 \log^3 T + \log\log T\big) = o(\Lambda_T)$$

by the definition of Λ_T. Hence there exists a fixed $c>1$ (e.g. $c=2$) such that

$$|\mathcal{F}^{w_h}_{rnn,\delta}|\, L\, e^{-c\Lambda_T} \;\leq\; e^{-\Lambda_T} \quad \text{for all } T \text{ large enough.}$$

Therefore,

$$\mathrm{P}\Big(\bigcap\nolimits_{f\in\mathcal{F}^{w_h}_{rnn,\delta}} \mathcal{E}_f\Big) \;\geq\; 1-e^{-\Lambda_T}.$$

That is,

$$\left|\frac{1}{T}\sum_{t=\ell^\star}^{T}\varepsilon_t^\top\big(\mathbb{E}_{t-1}Y_t-\varphi(H_t)\big)\right| \leq C_2\left(\frac{1}{T}\sum_{t=\ell^\star}^{T}\mathbb{E}\big\|\mathbb{E}_{t-1}Y_t-\varphi(H_t)\big\|^2\right)^{1/2}\Delta_T + C_2\,\Delta_T^2\log T,$$

holds *uniformly* for all $(\rho, W_h, v_h, \varphi)\in\mathcal{F}^{w_h}_{rnn,\delta}$ with probability at least $1-e^{-\Lambda_T}$.

Step 2: Proof of (15). Using Step 1, $\ell^\star \asymp w_h\log^2 T$, and $\|\varepsilon_t\|_\infty \leq B$, we have, with probability at least $1-\exp(-\Lambda_T)$,

$$\begin{aligned}
\left|\frac{1}{T}\sum_{t=2}^{T}\varepsilon_t^\top\big(\mathbb{E}_{t-1}Y_t-\varphi^*(H_t^*)\big)\right| &\lesssim \left|\frac{1}{T}\sum_{t=\ell^\star}^{T}\varepsilon_t^\top\big(\mathbb{E}_{t-1}Y_t-\overline{\varphi}^*(\overline{H}_t^*)\big)\right| + \frac{\ell^\star}{T} \\
&\lesssim \left(\frac{1}{T}\sum_{t=\ell^\star}^{T}\mathbb{E}\big\|\mathbb{E}_{t-1}Y_t-\overline{\varphi}^*(\overline{H}_t^*)\big\|^2\right)^{1/2}\Delta_T \;+\; \Delta_T^2\log T \\
&\lesssim \left(\frac{1}{T}\sum_{t=\ell^\star}^{T}\mathbb{E}\big\|\mathbb{E}_{t-1}Y_t-\varphi^*(H_t^*)\big\|^2\right)^{1/2}\Delta_T \;+\; \Delta_T^2\log T \\
&\lesssim \left(\frac{1}{T}\sum_{t=2}^{T}\mathbb{E}\big\|\mathbb{E}_{t-1}Y_t-\varphi^*(H_t^*)\big\|^2\right)^{1/2}\Delta_T \;+\; \Delta_T^2\log T.
\end{aligned} \tag{19}$$

In the first and the third line we used $\|\varphi^*(H_t^*)-\overline{\varphi}^*(\overline{H}_t^*)\|_\infty \leq T^{-1}$ from (18). This proves

(15).

Step 3: Proof of (16). By Lemma 9, for any fixed $(\rho, W_h, v_h, \varphi) \in \mathcal{F}_{rnn}^{w_h}$ and $c > 0$, there exists $C_3 > 0$, independent of $(\rho, W_h, v_h, \varphi)$ and T, such that with probability at least $1 - (\ell^\star + 2)\exp(-c\Lambda_T)$,

$$\begin{aligned}
&\left|\frac{1}{T}\sum_{t=2}^{T}\big\|\mathbb{E}_{t-1}Y_t - \varphi(H_t)\big\|^2 - \mathbb{E}\big\|\mathbb{E}_T Y_{T+1} - \varphi(H_{T+1})\big\|^2\right| \\
&\leq C_3\left(\frac{1}{T}\sum_{t=\ell^\star}^{T}\mathbb{E}\big\|\mathbb{E}_{t-1}Y_t - \varphi(H_t)\big\|^2\right)^{1/2}\Delta_T + C_3\,\Delta_T^2\log T. \qquad (20)
\end{aligned}$$

As above, choosing c large enough and applying a union bound over $\mathcal{F}_{rnn,\delta}^{w_h}$ yields a uniform version of (20). Repeating the argument in (19) (approximating $\varphi^*(H_t^*)$ by $\overline{\varphi}^*(\overline{H}_t^*)$ and handling the boundary $t < \ell^\star$) gives, with probability at least $1 - \exp(-\Lambda_T)$,

$$\begin{aligned}
&\left|\frac{1}{T}\sum_{t=2}^{T}\big[\|\mathbb{E}_{t-1}Y_t - \varphi^*(H_t^*)\|^2 - \mathbb{E}\|\mathbb{E}_T Y_{T+1} - \varphi^*(H_{T+1}^*)\|^2\big]\right| \\
&\lesssim \left(\frac{1}{T}\sum_{t=\ell^\star}^{T}\mathbb{E}\big\|\mathbb{E}_{t-1}Y_t - \varphi^*(H_t^*)\big\|^2\right)^{1/2}\Delta_T \;+\; \Delta_T^2\log T.
\end{aligned}$$

Finally, by Cauchy-Schwarz inequality,

$$\left(\frac{1}{T}\sum_{t=\ell^\star}^{T}\mathbb{E}\big\|\mathbb{E}_{t-1}Y_t - \varphi^*(H_t^*)\big\|^2\right)^{1/2}\Delta_T \;\lesssim\; \frac{1}{T}\sum_{t=2}^{T}\mathbb{E}\big\|\mathbb{E}_{t-1}Y_t - \varphi^*(H_t^*)\big\|^2 \;+\; \Delta_T^2\log T,$$

which yields (16). This completes the proof. □

Lemma 11. *Under the conditions of Theorem 1, for any $(\rho^*, W_h^*, v_h^*, \varphi^*) \in \mathcal{F}_{rnn}^{w_h}$, possibly depending on the in-sample data, there exists a constant $C > 0$, independent of $(\rho^*, W_h^*, v_h^*, \varphi^*)$ and T, such that for each fixed $h \in \mathbb{N}^+$ the following holds with probability at least $1 - C\exp(-\Lambda_T)$:*

$$\begin{aligned}
&\left|\frac{1}{T}\sum_{t=2}^{T}\Big[\big\|\mathbb{E}_{t-1}Y_t - \varphi^*(H_t^*)\big\|^2 - \mathbb{E}\,\big\|\mathbb{E}_{T+h-1}Y_{T+h} - \varphi^*(H_{T+h}^*)\big\|^2\Big]\right| \\
&\lesssim \frac{1}{T}\sum_{t=2}^{T}\mathbb{E}\big\|\mathbb{E}_{t-1}Y_t - \varphi^*(H_t^*)\big\|^2 \;+\; \Delta_T^2\log T.
\end{aligned}$$

Here $H_t^* := \sigma_{v_h^*}\big(\rho^*(Y_{t-1}, X_{t-1}) + W_h^* H_{t-1}^*\big)$ with $H_1^* := 0$.

Proof. By Lemma 9 (applied to a fixed h) and the same covering/approximation and union-bound argument used in Lemma 10, the stated bound follows directly; the truncation and boundary terms are handled identically. We omit the details. □

Lemma 12. *For any function $\varphi^\star \in \mathcal{H}^\beta([-a,a]^r, B)$, there exists a single-hidden-layer neural network $f^\dagger \in \mathcal{F}_r^1(1, \tau, T^{2\beta+2})$ with $\tau < T$ such that*

$$\iota_\varphi = \|\varphi^\star - f^\dagger\|_\infty \lesssim \begin{cases} (\log T)^{\frac{1}{2}}\, \tau^{-\frac{\beta}{r}}, & \text{if } \beta < \frac{r+3}{2}, \\ (\log T)^{\frac{1}{2}}\, \tau^{-\frac{1}{2}-\frac{3}{2r}}, & \text{if } \beta \geq \frac{r+3}{2}, \end{cases} \tag{21}$$

where the implicit constant depends only on (a, r, B, β) (and not on T).

Proof. Define the class

$$\mathcal{G}_r^1(\tau, M) := \Big\{ f : \mathbb{R}^r \to \mathbb{R} :\ f(x) = \sum_{i=1}^{\tau} a_i\, \sigma\big((x^\top, 1) v_i\big),\ \|v_i\| = 1,\ \sum_{i=1}^{\tau} |a_i| \leq M \Big\}.$$

By definition, any $f \in \mathcal{G}_r^1(\tau, M)$ is a one-hidden-layer ReLU network with width τ and output-layer ℓ_1-norm bounded by M. Hence, for $M \geq 1$, $\mathcal{G}_r^1(\tau, M) \subset \mathcal{F}_r^1(1, \tau, M)$.

By Corollary 2.4 of Yang and Zhou (2025), for any $\tau \in \mathbb{N}$ there exists $f^\dagger \in \mathcal{G}_r^1(\tau, M_\tau)$ such that

$$\|\varphi^\star - f^\dagger\|_\infty \lesssim \begin{cases} (\log T)^{1/2}\, \tau^{-\beta/r}, & \text{if } \beta < \frac{r+3}{2} \text{ and } M_\tau \asymp \tau^{\frac{r+3-2\beta}{2r}}, \\ (\log T)^{1/2}\, \tau^{-\frac{1}{2}-\frac{3}{2r}}, & \text{if } \beta \geq \frac{r+3}{2} \text{ and } M_\tau \asymp (\log \tau)^{1/2}. \end{cases}$$

It remains to embed this $f^\dagger$ into the target class $\mathcal{F}_r^1(1, \tau, T^{2\beta+2})$. Since $\tau < T$, we have:

- If $\beta < \frac{r+3}{2}$, then $0 < \frac{r+3-2\beta}{2r} \leq 2\beta + 2$, so

$$M_\tau \asymp \tau^{\frac{r+3-2\beta}{2r}} \leq T^{\frac{r+3-2\beta}{2r}} \leq T^{2\beta+2}.$$

- If $\beta \geq \frac{r+3}{2}$, then

$$M_\tau \asymp (\log \tau)^{1/2} \leq (\log T)^{1/2} \leq T^{2\beta+2}.$$

Therefore, in both cases $M_\tau \leq T^{2\beta+2}$ for all sufficiently large T, and thus $f^\dagger \in \mathcal{G}_r^1(\tau, M_\tau) \subset \mathcal{F}_r^1(1, \tau, T^{2\beta+2})$. This proves (21). □

Lemma 13. *Recall that $\bar\theta$ is defined in (19) and that ι_φ introduced in the proof of Theorem 2. For $1 \leq i \leq q$ and $n \geq 1$, define*

$$I_{n,i} := \mathbb{I}\Big(n - \Big\lfloor \frac{n-2}{q} \Big\rfloor q - i > 0\Big)\,. \tag{22}$$

If $0 < \bar\theta < 1$, *then*

$$\Bigg(\sum_{j=0}^{n-i-1} \bar\theta^j\Bigg)\,\iota_\varphi \;+\; \bar\theta^{\,\lfloor (n-2)/q \rfloor + I_{n,i}}\,B \;\le\; B,$$

for all $1 \le i \le q$, $1 \le n \le \lceil \log_{1/\bar\theta} T \rceil q$, *and all sufficiently large* T. *(By convention, the sum is zero when* $n - i - 1 < 0$*.)*

Proof. Throughout, adopt the convention that an empty sum equals zero. Set

$$S_{n,i} := \sum_{j=0}^{n-i-1} \bar\theta^j, \qquad k_{n,i} := \Big\lfloor \frac{n-2}{q} \Big\rfloor + I_{n,i}.$$

We split into two cases.

Case 1: $n \le i$ *(equivalently,* $n < i+1$*).* Then $S_{n,i} = 0$. It suffices to show $k_{n,i} \ge 0$. If $n = 1$, then $\lfloor (n-2)/q \rfloor = -1$ and, since $1 \le i \le q$, $I_{n,i} = \mathbb{I}(1 + q - i > 0) = 1$, so $k_{n,i} = -1 + 1 = 0$. If $n \ge 2$, then $\lfloor (n-2)/q \rfloor \ge 0$ and $I_{n,i} \ge 0$, hence $k_{n,i} \ge 0$. Thus $\bar\theta^{k_{n,i}} B \le B$, proving the claim in this case.

Case 2: $n \ge i + 1$ *(hence* $n \ge 2$*).* Now $S_{n,i} > 0$. For sufficiently large T, we have $\iota_\varphi \le (1-\bar\theta)^2 B$, by (21). Using the geometric sum formula,

$$S_{n,i}\,\iota_\varphi = \frac{1 - \bar\theta^{\,n-i}}{1 - \bar\theta}\,\iota_\varphi \le (1 - \bar\theta^{\,n-i})(1-\bar\theta)\,B \le (1-\bar\theta)\,B.$$

Therefore it is enough to prove $k_{n,i} \ge 1$, since then $\bar\theta^{k_{n,i}} B \le \bar\theta B$ and

$$S_{n,i}\,\iota_\varphi + \bar\theta^{k_{n,i}} B \le (1-\bar\theta)B + \bar\theta B = B.$$

To verify $k_{n,i} \ge 1$, observe:

- If $2 \le n < q + 2$, then $\lfloor (n-2)/q \rfloor = 0$ and

$$k_{n,i} = I_{n,i} = \mathbb{I}(n - i > 0) = 1 \quad \text{(since } n \ge i+1\text{)}.$$

- If $n \ge q + 2$, then $\lfloor (n-2)/q \rfloor \ge 1$, so $k_{n,i} \ge 1$.

Combining the two cases yields

$$\Big(\sum_{j=0}^{n-i-1} \bar\theta^j\Big)\,\iota_\varphi \;+\; \bar\theta^{\,\lfloor (n-2)/q \rfloor + I_{n,i}}\,B \le B. \qquad \square$$

Lemma 14. *Let $\theta_{s,i}$, $1 \le s \le r$, $1 \le i \le qr$, be as in Assumption 4, let $\bar{\theta}$ be as in (19), and let $I_{n,i} := \mathbb{I}\Big(n - \Big\lfloor \frac{n-2}{q} \Big\rfloor q - i > 0\Big)$ (as in (22)). Then, for all $1 \le n \le \lceil \log_{1/\bar{\theta}} T \rceil q$ and $1 \le j \le q$,*

$$\sum_{i=1}^{q} \big(\theta_{s,(i-1)r+1} + \cdots + \theta_{s,ir}\big)\, \bar{\theta}^{\lfloor (n-2)/q \rfloor + I_{n,i}} \le \bar{\theta}^{\lfloor (n-1)/q \rfloor + I_{n+1,1}}, \tag{23}$$

$$\Big\lfloor \frac{n-2}{q} \Big\rfloor + I_{n,j} = \Big\lfloor \frac{n-1}{q} \Big\rfloor + I_{n+1,j+1}. \tag{24}$$

Proof. Set $l := \lfloor (n-2)/q \rfloor$, $l' := \lfloor (n-1)/q \rfloor$, $r_n := n - lq \in \{2, \dots, q+1\}$, and $r_{n+1} := (n+1) - l'q \in \{2, \dots, q+1\}$. A direct check gives

$$r_{n+1} = \begin{cases} r_n + 1, & r_n \in \{2, \dots, q\}, \\ 2, & r_n = q+1, \end{cases} \qquad l' = \begin{cases} l, & r_n \in \{2, \dots, q\}, \\ l+1, & r_n = q+1. \end{cases}$$

For any $1 \le j \le q$, note that

$$I_{n,j} = \mathbb{I}\big(r_n - j > 0\big) = \mathbb{I}(j < r_n), \qquad I_{n+1,j+1} = \mathbb{I}\big(r_{n+1} - (j+1) > 0\big) = \mathbb{I}(j < r_{n+1} - 1).$$

If $r_n \in \{2, \dots, q\}$ then $r_{n+1} - 1 = r_n$ and $l' = l$, so $I_{n+1,j+1} = \mathbb{I}(j < r_n) = I_{n,j}$ and hence $l + I_{n,j} = l' + I_{n+1,j+1}$. If $r_n = q+1$ then $I_{n,j} = \mathbb{I}(j < q+1) = 1$, $I_{n+1,j+1} = \mathbb{I}(j < 1) = 0$, and $l' = l+1$, so again $l + I_{n,j} = l' + I_{n+1,j+1}$. This proves (24).

Next, observe that $i \mapsto n - \lfloor (n-2)/q \rfloor q - i$ is strictly decreasing, hence $I_{n,1} \ge \cdots \ge I_{n,q}$. Since $0 < \bar{\theta} < 1$, for all $1 \le i \le q$, $\bar{\theta}^{\lfloor (n-2)/q \rfloor + I_{n,i}} \le \bar{\theta}^{\lfloor (n-2)/q \rfloor + I_{n,q}}$. Therefore,

$$\sum_{i=1}^{q} \big(\theta_{s,(i-1)r+1} + \cdots + \theta_{s,ir}\big)\, \bar{\theta}^{\lfloor (n-2)/q \rfloor + I_{n,i}} \le \Big(\sum_{i=1}^{qr} \theta_{s,i}\Big)\, \bar{\theta}^{\lfloor (n-2)/q \rfloor + I_{n,q}}.$$

By (19), $\sum_{i=1}^{qr} \theta_{s,i} \le \bar{\theta}$. Using (24) with $j = q$ yields

$$\Big\lfloor \frac{n-2}{q} \Big\rfloor + I_{n,q} = \Big\lfloor \frac{n-1}{q} \Big\rfloor + I_{n+1,q+1} = \Big\lfloor \frac{n-1}{q} \Big\rfloor,$$

since $I_{n+1,q+1} = \mathbb{I}(r_{n+1} - (q+1) > 0) = 0$ (because $r_{n+1} \le q+1$). Moreover $I_{n+1,1} = \mathbb{I}(r_{n+1} - 1 > 0) = 1$ as $r_{n+1} \ge 2$. Hence

$$\Big(\sum_{i=1}^{qr} \theta_{s,i}\Big)\, \bar{\theta}^{\lfloor (n-2)/q \rfloor + I_{n,q}} \le \bar{\theta} \cdot \bar{\theta}^{\lfloor (n-1)/q \rfloor} = \bar{\theta}^{\lfloor (n-1)/q \rfloor + I_{n+1,1}},$$

which proves (23). □

Lemma 15. *Let $I_{n,i}$ be as in (22), $(\rho^\dagger, W_h^\dagger, v_h^\dagger, \varphi^\dagger)$ as in (21), and $H_t^{\dagger,(i)}$ as in (23). Set $N := \lceil \log_{1/\bar\theta} T \rceil q$ and $i_n := N - n + 1$. If $1 \le n \le N$ and $t \ge \max\{p,q\} + n$, then*

$$\begin{aligned}
\left\| \sum_{i=1}^{\tau} a_i \, H_{i,t}^{\dagger,(i_n)} - \mathbb{E}_{t-1} Y_t \right\|_\infty \; &\le \; \Big(\sum_{j=0}^{n-2} \bar\theta^j \Big) \, \iota_\varphi \; + \; \bar\theta^{\lfloor (n-2)/q \rfloor + I_{n,1}} \, B, \\
\left\| H_{\tau+1:\tau+r,t}^{\dagger,(i_n)} - \sigma_{-B1_r}(\varepsilon_{t-1}) \right\|_\infty \; &\le \; \Big(\sum_{j=0}^{n-3} \bar\theta^j \Big) \, \iota_\varphi \; + \; \bar\theta^{\lfloor (n-2)/q \rfloor + I_{n,2}} \, B, \\
&\vdots \\
\left\| H_{\tilde\tau-r+1:\tilde\tau,t}^{\dagger,(i_n)} - \sigma_{-B1_r}(\varepsilon_{t-q+1}) \right\|_\infty \; &\le \; \Big(\sum_{j=0}^{n-q-1} \bar\theta^j \Big) \, \iota_\varphi \; + \; \bar\theta^{\lfloor (n-2)/q \rfloor + I_{n,q}} \, B,
\end{aligned} \tag{25}$$

where $H_t^{\dagger,(i)} = (H_{1,t}^{\dagger,(i)}, \dots, H_{\tilde\tau,t}^{\dagger,(i)})^\top$. By convention, $\sum_{j=0}^{g} \bar\theta^j = 0$ when $g < 0$.

Proof. Recall

$$H_t^\dagger \; = \; \sigma_{v_h^\dagger}\Big(W_h^\dagger H_{t-1}^\dagger + \rho^\dagger(Y_{t-1}, X_{t-1}) \Big) \, .$$

Using that $\sigma_{-B}(x) = x + B$ for $x \ge -B$ and the definitions in (21), for $t > p$ we have

$$H_t^{\dagger,(N)} \; = \; V_4, \qquad H_t^{\dagger,(N+1)} \; = \; \Big((Y_{t-p} + B1_r)^\top, \dots, (Y_{t-1} + B1_r)^\top \Big)^\top . \tag{26}$$

Let $\tilde b_i := (b_{i,pr+qr+1}, \dots, b_{i,pr+qr+k})^\top \in \mathbb{R}^k$. For any $1 \le n \le N-1$ and $t > p+1$, from

$$H_t^{\dagger,(n)} = \sigma_{V_1}\Big(W_1 H_{t-1}^{\dagger,(n+1)} + W_3 H_{t-1}^{\dagger,(N+1)} + V_2(Y_{t-1}^\top, X_{t-1}^\top) \Big)$$

we obtain, for $1 \le i \le \tau$,

$$\begin{aligned}
H_{i,t}^{\dagger,(n)} &= \sigma_{v_i}\Big(- \sum_{j=1}^{\tau} b_{i,\,pr+1:pr+r}^\top \, a_j \, H_{j,t-1}^{\dagger,(n+1)} + \sum_{j=1}^{(q-1)r} b_{i,\,pr+r+j} \, H_{\tau+j,t-1}^{\dagger,(n+1)} + b_{i,\,pr+1:pr+r}^\top Y_{t-1} \\
&\qquad + \sum_{j=1}^{p} b_{i,\,jr-r+1:jr}^\top Y_{t-j} + \tilde b_i^\top X_{t-1} - \sum_{j=r+1}^{qr} b_{i,\,pr+j} B \Big) \\
&= \sigma_{v_i}\left(b_{i,\,pr+1:pr+r}^\top \left(Y_{t-1} - \sum_{j=1}^{\tau} a_j H_{j,t-1}^{\dagger,(n+1)} \right) + b_{i,\,pr+r+1:pr+2r}^\top \big(H_{\tau+1:\tau+r,t-1}^{\dagger,(n+1)} - B1_r \big) + \cdots \right. \\
&\qquad \left. + \, b_{i,\,pr+qr-r+1:pr+qr}^\top \big(H_{\tilde\tau-r+1:\tilde\tau,t-1}^{\dagger,(n+1)} - B1_r \big) + \sum_{j=1}^{p} b_{i,\,jr-r+1:jr}^\top Y_{t-j} + \tilde b_i^\top X_{t-1} \right).
\end{aligned} \tag{27}$$

Moreover,

$$H^{\dagger,(n)}_{\tau+1:\tau+r,t} = \sigma_{-B1_r}\left(-\sum_{j=1}^{\tau} a_j H^{\dagger,(n+1)}_{j,t-1} + Y_{t-1}\right), \tag{28}$$

$$H^{\dagger,(n)}_{\tau+ir+1:\tau+(i+1)r,t} = \sigma_{-B1_r}\left(H^{\dagger,(n+1)}_{\tau+(i-1)r+1:\tau+ir,t-1} - B1_r\right), \qquad 1 \le i \le q-2. \tag{29}$$

We prove (25) by induction on n.

Base case $n = 1$. First note that, when $n = 1$, $\sum_{j=0}^{n-2} \bar{\theta}^j = \sum_{j=0}^{-1} \bar{\theta}^j = 0, \sum_{j=0}^{n-3} \bar{\theta}^j = \sum_{j=0}^{-2} \bar{\theta}^j = 0$, and $\left\lfloor \frac{n-2}{q} \right\rfloor = \left\lfloor -\frac{1}{q} \right\rfloor = -1, I_{1,i} = \mathbb{I}\big(1 - \lfloor -1/q \rfloor q - i > 0\big) = \mathbb{I}(1 + q - i > 0) = 1 \ \ (1 \le i \le q)$. Hence $\lfloor (n-2)/q \rfloor + I_{1,i} = 0$ for all $i \le q$, and each right-hand side in (25) reduces to B.

Next, by (26) (for $t > p$), $H^{\dagger,(N)}_{1:\tau,t} = 0$. For the first line of (25),

$$\left\| \sum_{i=1}^{\tau} a_i \, H^{\dagger,(N)}_{i,t} - \mathbb{E}_{t-1} Y_t \right\|_\infty = \|\mathbb{E}_{t-1} Y_t\|_\infty \ \le \ B.$$

For the remaining lines, by (26) (for $t > p$), $H^{\dagger,(N)}_{\tau+1:\tilde{\tau},t} = B1_{\tilde{\tau}-\tau}$. Using that $\sigma_{-B}(x) = x + B$ for $x \ge -B$, we have

$$\left\| H^{\dagger,(N)}_{\tau+1:\tau+r,t} - \sigma_{-B1_r}(\varepsilon_{t-1}) \right\|_\infty = \|B1_r - (\varepsilon_{t-1} + B1_r)\|_\infty = \left\|\varepsilon_{t-1}\right\|_\infty \le \ B,$$

and similarly, for $j = 2, \ldots, q-1$,

$$\left\| H^{\dagger,(N)}_{\tau+jr+1:\tau+(j+1)r,t} - \sigma_{-B1_r}(\varepsilon_{t-j}) \right\|_\infty \le B.$$

Thus all inequalities in (25) hold for $n = 1$, completing the base case.

Inductive step. Assume (25) holds for some $1 \le n_0 < N$; we prove it for $n_0 + 1$. Define $\widetilde{\varepsilon}_{(t-1:t-q)} := (\widetilde{\varepsilon}^\top_{t-1}, \ldots, \widetilde{\varepsilon}^\top_{t-q})^\top$ by

$$\widetilde{\varepsilon}_{t-1} := \varphi^\star(Y_{(t-2:t-p-1)}, \varepsilon_{(t-2:t-q-1)}, X_{t-2}) + \varepsilon_{t-1} - \sum_{j=1}^{\tau} a_j H^{\dagger,(i_{n_0})}_{j,t-1},$$
$$\widetilde{\varepsilon}_{t-2} := H^{\dagger,(i_{n_0})}_{\tau+1:\tau+r,t-1} - B1_r, \quad \ldots, \quad \widetilde{\varepsilon}_{t-q} := H^{\dagger,(i_{n_0})}_{\tilde{\tau}-r+1:\tilde{\tau},t-1} - B1_r.$$

Using (27) and the definition $f^\dagger(x) = \sum_{i=1}^{\tau} a_i \sigma_{v_i}(b_i^\top x)$ (see (20)),

$$\begin{aligned} &\sum_{i=1}^{\tau} a_i \, H^{\dagger,(i_{n_0}-1)}_{i,t} - \varphi^\star(Y_{(t-1:t-p)}, \varepsilon_{(t-1:t-q)}, X_{t-1}) \\ &= f^\dagger\big(Y_{(t-1:t-p)}, \widetilde{\varepsilon}_{(t-1:t-q)}, X_{t-1}\big) - \varphi^\star(Y_{(t-1:t-p)}, \varepsilon_{(t-1:t-q)}, X_{t-1}). \end{aligned} \tag{30}$$

By Assumption 1, $\|\varepsilon_t\|_\infty, \|Y_{(t-1:t-p)}\|_\infty, \|X_t\|_\infty \leq B$. By the inductive hypothesis ((25) with n_0) and Lemma 13,

$$\|\widetilde{\varepsilon}_{t-1}\|_\infty \leq B + \Big(\sum_{j=0}^{n_0-2} \bar{\theta}^j\Big)\iota_\varphi + \bar{\theta}^{\lfloor (n_0-2)/q \rfloor + I_{n_0,1}} B \leq 2B,$$

and similarly, for $2 \leq j \leq q$,

$$\|\widetilde{\varepsilon}_{t-j}\|_\infty \leq \Big(\sum_{u=0}^{n_0-j-1} \bar{\theta}^u\Big)\iota_\varphi + \bar{\theta}^{\lfloor (n_0-2)/q \rfloor + I_{n_0,j}} B + B \leq 2B.$$

Let $|x|_s := |x_s|$ for $x \in \mathbb{R}^r$. For $1 \leq s \leq r$, using (20), the inductive hypothesis, Assumption 4, and the triangle inequality on (30),

$$\begin{aligned}
&\left|\sum_{i=1}^{\tau} a_i H_{i,t}^{\dagger,(i_{n_0}-1)} - \varphi^\star(Y_{(t-1:t-p)}, \varepsilon_{(t-1:t-q)}, X_{t-1})\right|_s \\
&\leq \big|(f^\dagger - \varphi^\star)\big(Y_{(t-1:t-p)}, \widetilde{\varepsilon}_{(t-1:t-q)}, X_{t-1}\big)\big|_s \\
&\quad + \big|\varphi^\star\big(Y_{(t-1:t-p)}, \widetilde{\varepsilon}_{(t-1:t-q)}, X_{t-1}\big) - \varphi^\star(Y_{(t-1:t-p)}, \varepsilon_{(t-1:t-q)}, X_{t-1})\big|_s \\
&\leq \iota_\varphi + \sum_{i=1}^{r} \theta_{s,i}\Big(\sum_{j=0}^{n_0-2} \bar{\theta}^j \iota_\varphi + \bar{\theta}^{\lfloor (n_0-2)/q \rfloor + I_{n_0,1}} B\Big) \\
&+ \sum_{i=2}^{q} (\theta_{s,(i-1)r+1} + \cdots + \theta_{s,ir})\left(\sum_{j=0}^{n_0-i-1} \bar{\theta}^j \iota_\varphi + \bar{\theta}^{\lfloor (n_0-2)/q \rfloor + I_{n_0,i}} B\right) \\
&\leq \sum_{j=0}^{n_0-1} \bar{\theta}^j \iota_\varphi + \bar{\theta}^{\lfloor (n_0-1)/q \rfloor + I_{n_0+1,1}} B,
\end{aligned} \tag{31}$$

where the last line uses $\theta_{s,1} + \cdots + \theta_{s,qr} \leq \bar{\theta} < 1$ and Lemma 14.

Next, by (28) and the 1-Lipschitz property of σ_{-B1_r} (coordinatewise),

$$\begin{aligned}
&\left\|H_{\tau+1:\tau+r,t}^{\dagger,(i_{n_0}-1)} - \sigma_{-B1_r}(\varepsilon_{t-1})\right\|_\infty = \left\|\sigma_{-B1_r}\left(Y_{t-1} - \sum_{j=1}^{\tau} a_j H_{j,t-1}^{\dagger,(i_{n_0})}\right) - \sigma_{-B1_r}(\varepsilon_{t-1})\right\|_\infty \\
&\leq \left\|\varphi^\star(Y_{(t-2:t-p-1)}, \varepsilon_{(t-2:t-q-1)}, X_{t-2}) - \sum_{j=1}^{\tau} a_j H_{j,t-1}^{\dagger,(i_{n_0})}\right\|_\infty \\
&\leq \Big(\sum_{j=0}^{n_0-2} \bar{\theta}^j\Big)\iota_\varphi + \bar{\theta}^{\lfloor (n_0-2)/q \rfloor + I_{n_0,1}} B = \Big(\sum_{j=0}^{n_0-2} \bar{\theta}^j\Big)\iota_\varphi + \bar{\theta}^{\lfloor (n_0-1)/q \rfloor + I_{n_0+1,2}} B,
\end{aligned} \tag{32}$$

where the last equality uses Lemma 14.

Finally, by (29), for $1 \leq j \leq q-2$,

$$\begin{aligned}
&\Big\| H^{\dagger,(i_{n_0}-1)}_{\tau+jr+1:\tau+(j+1)r,t} - \sigma_{-B1_r}(\varepsilon_{t-j-1}) \Big\|_\infty \\
=&\Big\| \sigma_{-B1_r}\big(H^{\dagger,(i_{n_0})}_{\tau+(j-1)r+1:\tau+jr,t-1} - B1_r\big) - \sigma_{-B1_r}(\varepsilon_{t-j-1}) \Big\|_\infty \\
\leq&\big\| H^{\dagger,(i_{n_0})}_{\tau+(j-1)r+1:\tau+jr,t-1} - B1_r - \varepsilon_{t-j-1} \big\|_\infty = \big\| H^{\dagger,(i_{n_0})}_{\tau+(j-1)r+1:\tau+jr,t-1} - \sigma_{-B1_r}(\varepsilon_{t-j-1}) \big\|_\infty \\
\leq&\Big(\sum_{u=0}^{n_0-j-2} \bar{\theta}^u \Big) \iota_\varphi + \bar{\theta}^{\lfloor (n_0-2)/q \rfloor + I_{n_0,j+1}} B = \Big(\sum_{u=0}^{n_0-j-2} \bar{\theta}^u \Big) \iota_\varphi + \bar{\theta}^{\lfloor (n_0-1)/q \rfloor + I_{n_0+1,j+2}} B,
\end{aligned}$$

using the inductive hypothesis and Lemma 14 again.

Combining (31), (32), and the above inequality completes the induction step $n_0 \mapsto n_0+1$, hence the proof of Lemma 15. □

References

Doukhan, P. (1994). Mixing: Properties and examples. In *New York: Springer-Verlag*.

Merlevède, F., M. Peligrad, and E. Rio (2009). Bernstein inequality and moderate deviations under strong mixing conditions. In *High dimensional probability V: the Luminy volume*, Volume 5, pp. 273–293. Institute of Mathematical Statistics.

Shen, Z. and D. Xiu (2024). Deep autoencoders for nonlinear factor models: Theory and applications. *Available at SSRN*.

Yang, Y. and D.-X. Zhou (2025). Optimal rates of approximation by shallow ReLU^k neural networks and applications to nonparametric regression. *Constructive Approximation 62*, 329–360.