Supplementary material for: Theoretical analysis of deep neural networks for temporally dependent observations

Mingliang Ma
Department of Statistics
University of Florida
Gainesville, FL 32611
`maminglian@ufl.edu`

Abolfazl Safikhani
Department of Statistics
George Mason University
Fairfax, VA 22030
`asafikha@gmu.edu`

We provide proofs of main theorems in Section A.2 while some useful Lemmas are stated and proved in Section A.1. In Section B, additional details related to the numerical experiments are provided. We also provide a numerical comparison between feed-forward deep neural networks (DNN) and Long Short-Term Memory networks (LSTM) in Section C.

A Proofs in Section 3

A.1 Useful Lemmas and their proofs

lemma 1. *Assume that $\{\mathbf{X}_t\}_{t\in\mathbb{Z}}$ satisfies Assumption 2, that is, $\{\mathbf{X}_t\}_{t\in\mathbb{Z}}$ is a stationary and exponential $\alpha-$mixing process. Further, assume f is a measurable function satisfying $\|f(\cdot)\|_\infty < M$. Then*

$$|\mathrm{Cov}(f(\mathbf{X}_i), f(\mathbf{X}_j))| \le 11M^2\exp(-c|i-j|/3),$$

for some positive constant c depending on $\tilde{c}$.

Proof. Consider the set of grid points

$$D(m) := \{(a_k, b_l),\ k, l = 1, 2, \cdots, 2m+1\}$$
$$a_k = M(k-m-1)/m,\ b_l = M(l-m-1)/m,$$

where m is some positive integer. Then we have

$$\begin{aligned}&\left|\mathbb{E}(f(\mathbf{X}_i)f(\mathbf{X}_j)) - \sum_{k=1}^{2m}\sum_{l=1}^{2m}(\frac{a_k+a_{k+1}}{2})(\frac{b_l+b_{l+1}}{2})\mathbb{P}(a_k \le f(\mathbf{X}_i) < a_{k+1}, b_l \le f(\mathbf{X}_j) < b_{l+1})\right|\\ &\le \sum_{k=1}^{2m}\sum_{l=1}^{2m}\mathbb{P}(a_k \le f(\mathbf{X}_i) < a_{k+1}, b_l \le f(\mathbf{X}_j) < b_{l+1})\frac{M^2}{m}\\ &= \frac{M^2}{m}.\end{aligned} \tag{16}$$

Similar to (16) we can prove that

$$\left|\mathbb{E}[f(\mathbf{X}_i)] - \sum_{k=1}^{2m}\frac{a_k+a_{k+1}}{2}\mathbb{P}(a_k \le f(\mathbf{X}_i) < a_{k+1})\right| \le \frac{M}{2m}.$$

To simplify the notation, we let $A_i := \sum_{k=1}^{2m}\frac{a_k+a_{k+1}}{2}\mathbb{P}(a_k \le f(\mathbf{X}_i) < a_{k+1})$. Then, we have

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

$$\begin{aligned}\left|\mathbb{E}[f(\mathbf{X}_i)]\mathbb{E}[f(\mathbf{X}_j)] - A_iA_j\right| &\leq \left|\mathbb{E}[f(\mathbf{X}_i)](\mathbb{E}[f(\mathbf{X}_j)] - A_j)\right| + \left|(\mathbb{E}[f(\mathbf{X}_j)] - A_j)A_i\right| \\ &\leq \frac{M^2}{2m} + \frac{M}{2m}(M + \frac{M}{2m}) = \frac{M^2}{m} + \frac{M^2}{4m^2}.\end{aligned} \tag{17}$$

Since $\{\mathbf{X}_t\}_{t\in\mathbb{Z}}$ is an exponential $\alpha-$mixing sequence, we know that there exists some positive constant c such that

$$\begin{aligned}&\mathbb{P}(a_k \leq f(\mathbf{X}_i) < a_{k+1}, b_l \leq f(\mathbf{X}_j) < b_{l+1}) \\ &\leq \exp(-c|i-j|) + \mathbb{P}(a_k \leq f(\mathbf{X}_i) < a_{k+1})\mathbb{P}(b_l \leq f(\mathbf{X}_j) < b_{l+1}).\end{aligned}$$

Therefore

$$\begin{aligned}&\left|\sum_{k=1}^{2m}\sum_{l=1}^{2m}(\frac{a_k + a_{k+1}}{2})(\frac{b_l + b_{l+1}}{2})\mathbb{P}(a_k \leq f(\mathbf{X}_i) < a_{k+1}, b_l \leq f(\mathbf{X}_j) < b_{l+1}) - A_iA_j\right| \\ &\leq \sum_{k=1}^{2m}\sum_{l=1}^{2m}|(\frac{a_k + a_{k+1}}{2})(\frac{b_l + b_{l+1}}{2})|\exp(-c|i-j|) \\ &\leq 4m^2M^2\exp(-c|i-j|).\end{aligned} \tag{18}$$

From (16),(17) and (18), we have that

$$\left|\mathbb{E}[f(\mathbf{X}_i)f(\mathbf{X}_j)] - \mathbb{E}[f(\mathbf{X}_i)]\mathbb{E}[f(\mathbf{X}_j)]\right| \leq \frac{2M^2}{m} + \frac{M^2}{4m^2} + 4m^2M^2\exp(-c|i-j|).$$

With the choice of $m = \lfloor \exp(c|i-j|/3) \rfloor$,

$$\begin{aligned}\left|\mathbb{E}[f(\mathbf{X}_i)f(\mathbf{X}_j)] - \mathbb{E}[f(\mathbf{X}_i)]\mathbb{E}[f(\mathbf{X}_j)]\right| &\leq 10M^2\exp(-c|i-j|/3) + \frac{M^2}{4}\exp(-2c|i-j|/3) \\ &\leq 11M^2\exp(-c|i-j|/3).\end{aligned}$$

□

lemma 2. *Let $\{\mathbf{X}_t\}_{t\in\mathbb{Z}}$ and f be as in lemma 1 while $\{a_n\}_{n\in\mathbb{Z}}$ is a sequence of real numbers such that $a_n \leq n^\alpha$ for some positive α. Let $Y_{ni} = a_nf(\mathbf{X}_i)$, then*

$$\mathrm{Var}(Y_{n0}) + 2\sum_{i>0}|\mathrm{Cov}(Y_{n0}, Y_{ni})| \leq ([24\alpha\log(n)/c] + 3)\mathrm{Var}(Y_{n0}) + 22M^2\frac{1}{n^{2\alpha}(\exp(\frac{c}{3}) - 1)},$$

where c is the same constant as in lemma 1 depending on $\tilde{c}$ (the α-mixing exponent).

Proof. Let $k = [\frac{12\alpha}{c}\log(n)] + 2$. Notice that $Y_{ni} < n^\alpha M$ for all $i \in \mathbb{Z}$. From lemma 1, we have that

$$\begin{aligned}2\sum_{i\geq k}|\mathrm{Cov}(Y_{n0}, Y_{nk})| \leq 22\sum_{i\geq k}(n^\alpha M)^2\exp(-ci/3) &= 22n^{2\alpha}M^2\frac{\exp(-\frac{c}{3}(k-1))}{\exp\frac{c}{3} - 1} \\ &\leq 22M^2\frac{1}{n^{2\alpha}\exp(\frac{c}{3} - 1)}.\end{aligned}$$

Therefore

$$\mathrm{Var}(Y_{n0}) + 2\sum_{i>0}|\mathrm{Cov}(Y_{n0}, Y_{nk})| \leq (2k-1)\mathrm{Var}(Y_{n0}) + 22M^2\frac{1}{n^{2\alpha}(\exp(\frac{c}{3}) - 1)}.$$

□

Let $\mathcal{F}$ be a class of function. We define $\mathcal{N}(\delta, \mathcal{F}, \|\cdot\|_\infty)$ to be the covering number, that is, the minimal number of $\|\cdot\|_\infty$-balls with radius δ that covers $\mathcal{F}$.

lemma 3. *Consider the d-variate nonparametric regression model with unknown regression function f_0, $Y_i = f_0(X_i) + \epsilon_i$, satisfying Assumptions 1-3. Let $\widehat{f}$ be any estimator taking values in $\mathcal{F}$. Define*

$$\Delta_n := \Delta_n(\widehat{f}, f_0, \mathcal{F}) := \mathbb{E}_{f_0}\left[\frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{f}(\mathbf{X}_i))^2 - \inf_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}(Y_i - f(\mathbf{X}_i))^2\right]$$

and assume $\{f_0\} \cup \mathcal{F} \subset \{f : [0,1]^d \to [-F, F]\}$ for some $F \geq 1$. If $N_n := \mathcal{N}(\delta, \mathcal{F}, \|\cdot\|_\infty) \geq 3$, then,

$$\begin{aligned}
&(1-\epsilon)^2\Delta_n - C(F,\sigma^2,c)\frac{\log^4 n \log N_n}{n\epsilon} - \delta C(F,\sigma^2,c)\log^2 n \\
&\leq R(\widehat{f}, f_0) \leq \\
&(1+\epsilon)^2\left(\inf_{f^*\in\mathcal{F}}\|f^* - f_0\|_\infty^2 + \Delta_n(\widehat{f}, f) + C(F,\sigma^2,c)\delta\log^2 n\right) \\
&+ \frac{(1+\epsilon)^3}{\epsilon}C(F,\sigma^2,c)\frac{\log^4 n\log N_n}{n},
\end{aligned}$$

where $C(F,\sigma^2,c)$ is defined as a constant depending only on F,σ^2, and c (the same constant in lemma 1).

Proof. Throughout the proof we write $\mathbb{E} = \mathbb{E}_{f_0}$. Define $\|g\|_n^2 := \frac{1}{n}\sum_{i=1}^n g(\mathbf{X}_i)^2$. For any estimator $\tilde{f}$, we introduce $\widehat{R}_n(\tilde{f}, f_0) := \mathbb{E}[\|\tilde{f} - f_0\|_n^2]$ for the empirical risk. In the first step, we show that we can restrict ourselves to the case $\log N_n \leq n$. Since $R(\widehat{f}, f_0) \leq 4F^2$, the upper bound trivially holds if $\log N_n \geq n$. To see that also the lower bound is trivial in this case, let $\tilde{f} \in \mathrm{argmin}_{f\in\mathcal{F}}\sum_{i=1}^n (Y_i - f(\mathbf{X}_i))^2$ be a (global) empirical risk minimizer. Observe that

$$\widehat{R}_n(\widehat{f}, f_0) - \widehat{R}_n(\tilde{f}, f_0) = \Delta_n + \mathbb{E}\left[\frac{2}{n}\sum_{i=1}^n \epsilon_i \widehat{f}(\mathbf{X}_i)\right] - \mathbb{E}\left[\frac{2}{n}\sum_{i=1}^n \epsilon_i \tilde{f}(\mathbf{X}_i)\right].$$

From this equation, it follows that $\Delta_n \leq 8F^2$ and this implies the lower bound in the statement of the lemma for $\log N_n \geq n$. We may therefore assume $\log N_n \leq n$. The proof is divided into four parts which are denoted by (I∼IV)

(I): We relate the risk $R(\widehat{f}, f_0) = \mathbb{E}[(\widehat{f}(\mathbf{X}) - f_0(\mathbf{X}))^2]$ to its empirical counterpart $\widehat{R}_n(\widehat{f}, f_0)$ via the inequalities

$$\begin{aligned}
&(1-\epsilon)\widehat{R}_n(\widehat{f}, f_0) - C_F\frac{\log^2 n\log N_n}{n} - C_F\delta\log^2 n - \frac{1}{\epsilon}\frac{C_F F^2\log^4 n\log N_n}{n} \\
&\leq R(\widehat{f}, f_0) \leq \\
&(1+\epsilon)\widehat{R}_n(\widehat{f}, f_0) + (1+\epsilon)\left(C_F\frac{\log^2 n\log N_n}{n} + C_F\delta\log^2 n\right) \\
&+ \frac{(1+\epsilon)^2}{\epsilon}\frac{C_F F^2\log^4 n\log N_n}{n},
\end{aligned}$$

where C_F is some constant which depends on F and c.

(II): For any estimator $\tilde{f}$ taking values in $\mathcal{F}$,

$$\begin{aligned}
\mathbb{E}\left[\left|\frac{2}{n}\sum_{i=1}^n \epsilon_i\tilde{f}(\mathbf{X}_i)\right|\right] &\leq \delta C(F,\sigma^2,c)\log^2 n + C(F,\sigma^2,c)\log^4 n\frac{\log N_n}{n} \\
&\quad + C(F,\sigma^2,c)\sqrt{\frac{\log^4 n\log N_n}{n}}\widehat{R}_n^{1/2}(\tilde{f}, f_0).
\end{aligned}$$

(III): We have

$$\widehat{R}_n(\widehat{f}, f_0) \leq (1+\epsilon)\Big[\inf_{f\in\mathcal{F}}\mathbb{E}[(f(\mathbf{X}) - f_0(\mathbf{X}))^2] + \Delta_n + \delta C(F,\sigma^2,c)\log^2 n \\ + C(F,\sigma^2,c)\log^4 n\frac{\log N_n}{n}\Big] + \frac{(1+\epsilon)^2}{\epsilon}C^2(F,\sigma^2,c)\log^4 n\frac{\log N_n}{n}.$$

(IV): We have

$$\widehat{R}_n(\widehat{f}, f_0) \geq (1-\epsilon)(\Delta_n - C(F,\sigma^2,c)\frac{\log^4 n\log N_n}{n\epsilon} - 2\delta C(F,\sigma^2,c)\log^2 n).$$

Combining (I) and (IV) gives the lower bound of the assertion. The upper bound follows from (I) and (III).

(I): Given a minimal δ-covering of $\mathcal{F}$, denote the centers of the balls by f_j . By construction there exists a (random) j^* such that $\|\widehat{f} - f_{j^*}\|_\infty \leq \delta$. Without loss of generality, we can assume that $\|f_j\|_\infty \leq F$. Generate i.i.d. random variables $\{\mathbf{X}_i^{'}, i = 1,\cdots,n\}$ with the same distribution as $\mathbf{X}$ ($\mathbf{X} \overset{D}{=} \mathbf{X}_i$) and independent of $\{\mathbf{X}_i, i=1,\cdots,n\}$. Using that $\|f_j\|_\infty, \|f_0\|_\infty, \delta \leq F$,

$$\begin{aligned} &|R(\widehat{f}, f_0) - \widehat{R}_n(\widehat{f}, f_0)| \\ &= \left|\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n(\widehat{f}(\mathbf{X}_i^{'}) - f_0(\mathbf{X}_i^{'}))^2 - \frac{1}{n}\sum_{i=1}^n(\widehat{f}(\mathbf{X}_i) - f_0(\mathbf{X}_i))^2\right]\right| \\ &\leq \mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^n g_{j^*}(\mathbf{X}_i, \mathbf{X}_i^{'})\right|\right] + 9\delta F, \end{aligned}$$

with $g_{j^*}(\mathbf{X}_i, \mathbf{X}_i^{'}) := (f_{j^*}(\mathbf{X}_i^{'}) - f_0(\mathbf{X}_i^{'}))^2 - (f_{j^*}(\mathbf{X}_i) - f_0(\mathbf{X}_i))^2$. Define g_j in the same way with f_j^* replaced by f_j. Similarly, set $\gamma_j := \sqrt{n^{-1}\log N_n} \vee \mathbb{E}^{1/2}[(f_j(\mathbf{X}) - f_0(\mathbf{X}))^2]$ and define γ^* as

$$\begin{aligned} \gamma^* &= \sqrt{n^{-1}\log N_n} \vee \mathbb{E}^{1/2}[(f_{j^*}(\mathbf{X}) - f_0(\mathbf{X}))^2 | \{(\mathbf{X}_i, Y_i)\}_{i=1}^\infty] \\ &\leq \sqrt{n^{-1}\log N_n} + \mathbb{E}^{1/2}[(\widehat{f}(\mathbf{X}) - f_0(\mathbf{X}))^2 | \{(\mathbf{X}_i, Y_i)\}_{i=1}^\infty] + \delta, \end{aligned} \tag{19}$$

where the last part follows from triangle inequality and $f_{j^*} - \widehat{f} \leq \delta$.
For random variables U_1, T_1, Cauchy-Schwarz inequality gives $\mathbb{E}[U_1T_1] \leq \mathbb{E}^{1/2}[U_1^2]\mathbb{E}^{1/2}[T_1^2]$. Choose $U_1 = \mathbb{E}^{1/2}[(\widehat{f}(\mathbf{X}) - f_0(\mathbf{X}))^2 | \{(\mathbf{X}_i, Y_i)\}_{i=1}^\infty]$ and $T_1 = \max_j |\sum_{i=1}^n g_j(\mathbf{X}_i, \mathbf{X}_{i'})/\gamma_j F|$. Using that $\mathbb{E}[U_1^2] = R(\widehat{f}, f_0)$,

$$\begin{aligned} &|R(\widehat{f}, f_0) - \widehat{R}_n(\widehat{f}, f_0)| \\ &\leq \frac{F}{n}R(\widehat{f}, f_0)^{\frac{1}{2}}\mathbb{E}^{\frac{1}{2}}[T_1^2] + \frac{F}{n}(\sqrt{\frac{\log N_n}{n}} + \delta)\mathbb{E}[T_1] + 9\delta F. \end{aligned} \tag{20}$$

Next, we need to estimate the upper bound for $\mathbb{E}[T]$ and $\mathrm{E}[T^2]$. To simplify the notation, we let $v_j^2 = \mathrm{Var}(g_j(\mathbf{X}_i, \mathbf{X}_i^{'})/\gamma_j F)$ and $\tilde{v}_j^2 = \mathrm{Var}\left(\frac{g_j(\mathbf{X}_i, \mathbf{X}_i^{'})}{\gamma_j F}\right) + 2\sum_{k>i}\mathrm{Cov}\left(\frac{g_j(\mathbf{X}_k, \mathbf{X}_k^{'})}{\gamma_j F}, \frac{g_j(\mathbf{X}_i, \mathbf{X}_i^{'})}{\gamma_j F}\right)$. We know that $|g_j(\mathbf{X}_i, \mathbf{X}_i^{'})/F| \leq 4F$ and $1/\gamma_j \leq n^{1/2}$. From lemma 2, we can derive

$$\tilde{v}_j^2 \leq ([12\log(n)/c] + 3)v_j^2 + 22(4F)^2\frac{1}{n(\exp(c/3) - 1)}.$$

Since $v_j^2 = 2\mathrm{Var}((f_j(\mathbf{X}_i) - f_0(\mathbf{X}_i))^2/\gamma_j F) \leq 2\mathbb{E}[(f_j(\mathbf{X}_i) - f_0(\mathbf{X}_i))^4]/(\gamma_j^2 F^2) \leq 8$, we conclude that $\tilde{v}_j^2 \leq C_1\log(n) + C_2F^2/n$, where C_1 and C_2 are constants which only depend on c. Observe that

$\mathbb{E}[g_j(\mathbf{X}_i, \mathbf{X}_i^{'})] = 0$, $|\frac{g_j(X_i, X_i^{'})}{\gamma_j F}| \leq 4F/\gamma_j$. Also, we know that $\{g_j(\mathbf{X}_i, \mathbf{X}_{i'})/\gamma_j F,\ i = 1, \cdots, n\}$ is an exponentially α-mixing process. By Bernstein inequality (Theorem 2, [2]) with a union bound over j, we have

$$\begin{aligned}
&\mathbb{P}(T_1 \geq t) \leq 1 \wedge \left(2N_n \max_j \exp\left(-\frac{\tilde{c}t^2}{n\tilde{v}_j^2 + \frac{16F^2}{\gamma_j^2} + t\frac{4F}{\gamma_j}(\log n)^2}\right)\right) \\
&\leq 1 \wedge \left(2N_n \max_j \exp\left(-\frac{\tilde{c}t^2}{C_1 n\log(n) + C_2F^2 + 16F^2\frac{n}{\log(N_n)} + 4tF(\log n)^2\sqrt{\frac{n}{\log(N_n)}}}\right)\right) \\
&\leq 1 \wedge \left(2N_n \max_j \exp\left(-\frac{\tilde{c}t^2}{n(C_1\log(n) + (16 + C_2)F^2) + 4tF(\log n)^2\sqrt{\frac{n}{\log(N_n)}}}\right)\right).
\end{aligned}$$

Therefore we can estimate the upper of $\mathbb{E}[T_1]$ by $\mathbb{P}(T_1 > t)$

$$\begin{aligned}
\mathbb{E}[T_1] &= \int_0^\infty \mathbb{P}(T_1 > t)dt \\
&\leq \theta\sqrt{n\log(N_n)} + 2N_n \int_{\theta\sqrt{n\log(N_n)}}^\infty \exp\left(-\frac{\tilde{c}t}{\frac{((16+C_2)F^2+C_1\log(n))n}{\theta\sqrt{n\log(N_n)}} + 4\sqrt{\frac{n}{\log(N_n)}}(\log n)^2F}\right) \\
&= \theta\sqrt{n\log(N_n)} + 2N_n\frac{1}{\tilde{c}}\left(\frac{((16 + C_2)F^2 + C_1\log(n))n}{\theta\sqrt{n\log(N_n)}} + 4\sqrt{\frac{n}{\log(N_n)}}(\log n)^2F\right) \\
&\qquad \exp\left(-\frac{\tilde{c}\theta\sqrt{n\log(N_n)}}{\frac{((16+C_2)F^2+C_1\log(n))n}{\theta\sqrt{n\log(N_n)}} + 4\sqrt{\frac{n}{\log(N_n)}}(\log n)^2F}\right).
\end{aligned}$$

Let $\theta = \sqrt{\frac{(32+2C_2)F^2+2C_1\log(n)}{\tilde{c}}} \vee \frac{8F(\log n)^2}{\tilde{c}}$, we have

$$\begin{aligned}
\mathbb{E}[T] &\leq \theta\sqrt{n\log(N_n)} + \left(\frac{(32 + 2C_2)F^2 + 2C_1\log(n)}{\theta\tilde{c}} + \frac{8F(\log n)^2}{\tilde{c}}\right)\sqrt{\frac{n}{\log(N_n)}} \\
&= \theta\sqrt{n\log(N_n)} + A_1\sqrt{\frac{n}{\log(N_n)}}.
\end{aligned}$$

Next we estimate the upper bound of $\mathbb{E}[T_1^2]$ in a similar way.

$$\begin{aligned}
&\mathbb{E}[T_1^2] = \int_0^\infty \mathbb{P}(T_1 > \sqrt{t})dt \\
&\leq \theta^2 n\log(N_n) + 2N_n\int_{\theta^2 n\log(N_n)}^\infty \exp\left(-\frac{\tilde{c}\sqrt{t}}{\frac{((16+C_2)F^2+C_1\log(n))n}{\theta\sqrt{n\log(N_n)}} + 4\sqrt{\frac{n}{\log(N_n)}}(\log n)^2F}\right)dt.
\end{aligned}$$

Using the fact that $\int_{b^2}^\infty \exp(-\sqrt{u}a)du = (2ab+1)\exp(-ab)/a^2$ and with the same choice of θ as above, we estimate the upper bound of $\mathbb{E}[T_1]$. Also, we can prove that

$$\begin{aligned}
\mathbb{E}[T_1^2] &\leq \theta^2 n\log(N_n) + 6n\left(\frac{(16 + C_2)F^2 + C_1\log(n)}{\theta\tilde{c}^2} + \frac{4\log^2(n)F}{\tilde{c}^2}\right) \\
&= \theta^2 n\log(N_n) + 3nA_1/\tilde{c} \\
&\leq (\theta^2 + \frac{3A_1}{\tilde{c}})n\log N_n.
\end{aligned}$$

With eq(20) and the upper bound for $\mathbb{E}T_1$ and $\mathbb{E}T_1^2$, we have

$$\begin{aligned}|R(\widehat{f},f_0)-\widehat{R}_n(\widehat{f},f_0)| \leq{}& \frac{F}{n}R(\widehat{f},f_0)^{\frac{1}{2}}\sqrt{\theta^2 n\log(N_n)+3nA_1/\tilde{c}}+\\ & \frac{F}{n}(\sqrt{\frac{\log N_n}{n}}+\delta)\left(\theta\sqrt{n\log N_n}+A_1\sqrt{\frac{n}{\log N_n}}\right)+9\delta F\\ ={}& \frac{F}{n}R(\widehat{f},f_0)^{\frac{1}{2}}\sqrt{\theta^2 n\log(N_n)+3nA_1/\tilde{c}}+\frac{\theta F\log N_n}{n}+\frac{A_1F}{n}+\\ & F\delta(\theta\sqrt{\frac{\log N_n}{n}}+A_1\sqrt{\frac{1}{n\log N_n}}+9)\\ \leq{}& \frac{F}{n}R(\widehat{f},f_0)^{\frac{1}{2}}\sqrt{(\theta^2+\frac{3A_1}{\tilde{c}})n\log N_n}+\frac{(\theta+A_1)F\log N_n}{n}+\\ & (\theta+A_1+9)F\delta.\end{aligned} \tag{21}$$

Since there exists some constant such that $\frac{(\theta+A_1+9)F}{\log^2 n}\vee\frac{\theta^2+3A_1/\tilde{c}}{\log^4 n}\leq C_F$, where C_F depends on F and c, eq(21) can be simplified as

$$\begin{aligned}|R(\widehat{f},f_0)-\widehat{R}_n(\widehat{f},f_0)| \leq{}& \frac{F}{n}R(\widehat{f},f_0)^{\frac{1}{2}}\sqrt{C_F n\log^4 n\log N_n}+C_F\frac{\log^2 n\log N_n}{n}+\\ & C_F\delta\log^2 n.\end{aligned}$$

From eq(43) in [3], we know that for positive real numbers a,b,c,d being such that $|a-b|\leq 2\sqrt{a}c+d$, then for any $0<\epsilon<1$,

$$(1-\epsilon)b-d-\frac{c^2}{\epsilon}\leq a\leq(1+\epsilon)(b+d)+\frac{(1+\epsilon)^2}{\epsilon}c^2. \tag{22}$$

Using (22) with $a=R(\widehat{f},f_0)$ and $b=\widehat{R}(\widehat{f},f_0)$, we can derive the following bounds for $R(\widehat{f},f_0)$ from (21):

$$\begin{aligned}&(1-\epsilon)\widehat{R}_n(\widehat{f},f_0)-C_F\frac{\log^2 n\log N_n}{n}-C_F\delta\log^2 n-\frac{1}{\epsilon}\frac{C_F F^2\log^4 n\log N_n}{n}\\ &\leq R(\widehat{f},f_0)\leq\\ &(1+\epsilon)\widehat{R}_n(\widehat{f},f_0)+(1+\epsilon)\left(C_F\frac{\log^2 n\log N_n}{n}+C_F\delta\log^2 n\right)\\ &+\frac{(1+\epsilon)^2}{\epsilon}\frac{C_F F^2\log^4 n\log N_n}{n}.\end{aligned}$$

(II): Similar to the proof of (I), there exists a random j^* such that $\|f_{j^*}-\tilde{f}\|_\infty\leq\delta$. We have $|\mathbb{E}[\sum_{i=1}^n\epsilon_i(\tilde{f}(\mathbf{X}_i)-f_{j^*}(\mathbf{X}_i))]|\leq\delta\mathbb{E}[\sum_{i=1}^n|\epsilon_i|]\leq n\delta$. Since $\mathbb{E}[\epsilon_i f_0(\mathbf{X}_i)]=\mathbb{E}[\mathbb{E}[\epsilon_i f_0(X_i)|X_i]]=0$, we also find

$$\begin{aligned}\left|\mathbb{E}[\frac{2}{n}\sum_{i=1}^n\epsilon_i\tilde{f}(\mathbf{X}_i)]\right| &= \left|\mathbb{E}[\frac{2}{n}\sum_{i=1}^n\epsilon_i(\tilde{f}(\mathbf{X}_i)-f_0(\mathbf{X}_i))]\right|\\ &\leq 2\delta+\frac{2}{n}\mathbb{E}\left|\sum_{i=1}^n\epsilon_i(f_{j^*}(\mathbf{X}_i)-f_0(\mathbf{X}_i))\right|.\end{aligned} \tag{23}$$

Recall that $\gamma_j:=\sqrt{n^{-1}\log N_n}\vee\mathbb{E}^{1/2}[(f_j(\mathbf{X})-f(\mathbf{X}))^2]$ and the definition of γ^* is

$$\begin{aligned}\gamma^* &= \sqrt{n^{-1}\log N_n} \vee \mathbb{E}^{1/2}[(f_{j^*}(\mathbf{X}) - f_0(\mathbf{X}))^2 | \{(\mathbf{X}_i, Y_i)\}_{i=1}^{\infty}] \\ &\le \sqrt{n^{-1}\log N_n} + \mathbb{E}^{1/2}[(\tilde{f}(\mathbf{X}) - f_0(\mathbf{X}))^2 | \{(\mathbf{X}_i, Y_i)\}_{i=1}^{\infty}] + \delta.\end{aligned}$$

Using Cauchy-Schwarz inequality $\mathbb{E}[U_2 T_2] \le \mathbb{E}^{1/2}[U_2^2]\mathbb{E}^{1/2}[T_2^2]$ with $T_2 = \max_j |\frac{\epsilon_j(f_j(\mathbf{X}_i) - f_0(\mathbf{X}_i))}{\gamma_j}|$ and $U_2 = \mathbb{E}^{1/2}[(\tilde{f}(\mathbf{X}) - f_0(\mathbf{X}))^2 | \{(\mathbf{X}_i, Y_i)\}_{i=1}^{\infty}]$, we have that

$$\begin{aligned}\mathbb{E}\left|\sum_{i=1}^{n} \epsilon_i (f_{j^*}(\mathbf{X}_i) - f_0(\mathbf{X}_i))\right| &= \mathbb{E}\left|\frac{\sum_{i=1}^{n} \epsilon_i (f_{j^*}(\mathbf{X}_i) - f_0(\mathbf{X}_i))}{\gamma^*} \cdot \gamma^*\right| \\ &\le \mathbb{E}[|T_2(\sqrt{n^{-1}\log N_n} + \mathbb{E}^{1/2}[(\tilde{f}(\mathbf{X}) - f_0(\mathbf{X}))^2 | \{(\mathbf{X}_i, Y_i)\}_{i=1}^{\infty}] + \delta)|] \\ &\le \mathbb{E}[T_2](\delta + \sqrt{n^{-1}\log N_n}) + \mathbb{E}^{1/2}[T_2^2]\mathbb{E}^{1/2}[(\tilde{f}(\mathbf{X}) - f_0(\mathbf{X}))^2].\end{aligned} \tag{24}$$

Notice that $\mathbb{E}^{1/2}[(\tilde{f}(\mathbf{X}) - f_0(\mathbf{X}))^2] = R^{1/2}(\widehat{f}, f_0)$. Now, using eq(23), eq(24),

$$\mathbb{E}\left|\frac{2}{n}\sum_{i=1}^{n} \epsilon_i \tilde{f}(\mathbf{X_i})\right| \le 2\delta + \frac{2}{n}\mathbb{E}[T_2](\delta + \sqrt{n^{-1}\log N_n}) + \frac{2}{n}\mathbb{E}^{1/2}[T_2^2]R^{1/2}(\tilde{f}, f_0). \tag{25}$$

Let $Z_{ij} := \frac{f_j(\mathbf{X}_i) - f_0(\mathbf{X}_i)}{\gamma_j}$. From $\gamma_j \ge \sqrt{n^{-1}\log N_n}$ and $\gamma_j \ge \mathbb{E}^{1/2}[(f_j(\mathbf{X}) - f(\mathbf{X}))^2]$, we know that (1) $|Z_{ij}| \le 2F/\sqrt{n^{-1}\log N_n}$; (2) $\mathbb{E}Z_{ij}^2 \le 1$. Observe that $\{\epsilon_i Z_{ij}\}_{i \ge 1}$ is a martingale difference sequence with respect to the filtration $\mathcal{F}_t = \sigma((\mathbf{X}_i), i \le t)$. Then we check the moment conditions in theorem 1.2B in [1]. Since ϵ_i and $\mathcal{F}_i$ are independent, $\mathbb{E}(\epsilon_i^2 Z_{ij}^2 | \mathcal{F}_i) = Z_{ij}^2\sigma^2$. For the m-th moment of $|\epsilon_i Z_{ij}|$ given $\mathcal{F}_i$, we have that

$$\mathbb{E}[|\epsilon_i^m Z_{ij}^m||\mathcal{F}_i] \le \sigma^2 m! c_0^{m-2}|Z_{ij}|^m \le \sigma^2 Z_{ij}^2 m! \left(\frac{2c_0 F}{\sqrt{n^{-1}\log N_n}}\right)^{m-2}.$$

We split $\mathbb{P}(|\sum_{i=1}^{n} \epsilon_i Z_{ij}| \ge x)$ into two parts:

$$\begin{aligned}\mathbb{P}(|\sum_{i=1}^{n} \epsilon_i Z_{ij}| \ge x) \le &\underbrace{\mathbb{P}(|\sum_{i=1}^{n} \epsilon_i Z_{ij}| \ge x, \sigma^2 \sum_{i=1}^{n} Z_{ij}^2 \le \sqrt{\frac{n}{\log N_n}} x)}_{a(x)} + \\ &\underbrace{\mathbb{P}(\sum_{i=1}^{n} Z_{ij}^2 \ge \frac{1}{\sigma^2}\sqrt{\frac{n}{\log N_n}} x).}_{b(x)}\end{aligned} \tag{26}$$

Using theorem 1.2B in [1] with $c = 2c_0 F/\sqrt{(n^{-1})\log N_n}$, $V_n = \sigma^2 \sum_{i=1}^{n} Z_{ij}^2$, $y = \sqrt{n^{-1}\log N_n} x$, we obtain

$$a(x) \le \exp\left(-\frac{x}{(2 + 2c_0 F)\sqrt{\frac{n}{\log N_n}}}\right). \tag{27}$$

To simplify the notation, we let $v_j^2 = \mathrm{Var}(Z_{ij}^2)$ and $\tilde{v}_j^2 = \mathrm{Var}\left(Z_{ij}^2\right) + 2\sum_{k>i} \mathrm{Cov}\left(Z_{ij}^2, Z_{kj}^2\right)$. We know that $(f_j(\mathbf{X}_i) - f_0(\mathbf{X}_i))^2 \le 4F^2$ and $1/\gamma_j^2 \le n$. From lemma 2, we have

$$\tilde{v}_j^2 \le ([24\log(n)/c]+3)v_j^2 + 22(16F^4)\frac{1}{n^2(\exp(c/3)-1)}.$$

Since $v_j^2 \le \mathbb{E}Z_{ij}^4 \le \mathbb{E}(f_j(\mathbf{X}_i)-f_0(\mathbf{X}_i))^4/(n^{-1}\log N_n \mathbb{E}(f_j(\mathbf{X}_i)-f_0(\mathbf{X}_i))^2) \le 4F^2 n/\log N_n$, we derive that $\tilde{v}_j^2 \le C_1 F^2 n\log(n)/\log N_n + C_2 F^4/n^2$, where C_1 and C_2 only depend on c. Since $\mathbb{E}Z_{ij}^2 \le 1$, we have $b(x) \le \mathbb{P}(\sum_{i=1}^n (Z_{ij}^2 - \mathbb{E}Z_{ij}^2) \ge \frac{1}{\sigma^2}\sqrt{\frac{n}{\log N_n}}x - n)$. Observe that $|Z_{ij}^2 - \mathbb{E}Z_{ij}^2| \le 4F^2 n/\log N_n$ and $\{Z_{ij}^2\}_i$ is an exponentially α-mixing process, using Bernstein inequality (Theorem 2, [2]), again we have that

$$\begin{aligned} b(x) &\le \exp\left(-\frac{C_3(\frac{1}{\sigma^2}\sqrt{\frac{n}{\log N_n}}x-n)^2}{n\tilde{v}_j^2 + \frac{16n^2F^4}{\log N_n} + (\frac{1}{\sigma^2}\sqrt{\frac{n}{\log N_n}}x-n)\frac{n\log^2 n}{\log N_n}}\right) \\ &\le \exp\left(-\frac{C_3(\frac{1}{\sigma^2}\sqrt{\frac{n}{\log N_n}}x-n)^2}{\frac{C_1F^2n^2\log n}{\log N_n} + \frac{C_2F^4}{n} + \frac{16n^2F^4}{\log N_n} + (\frac{1}{\sigma^2}\sqrt{\frac{n}{\log N_n}}x-n)\frac{n\log^2 n}{\log N_n}}\right) \qquad (28) \\ &\le \exp\left(-\frac{C_3(\frac{1}{\sigma^2}\sqrt{\frac{n}{\log N_n}}x-n)^2}{\frac{(C_1F^2+C_2F^4+16F^4)n^2\log n}{\log N_n} + (\frac{1}{\sigma^2}\sqrt{\frac{n}{\log N_n}}x-n)\frac{n\log^2 n}{\log N_n}}\right). \end{aligned}$$

The last inequality uses the assumption that $n \ge \log N_n$.

Next, we use eq(26), eq(27), eq(28) to estimate the upper bound for $\mathbb{E}T_2$ and $\mathbb{E}T_2^2$ in eq(25). Taking the union bound over all $j = 1, \cdots, N_n$, we have

$$\mathbb{P}(T_2 \ge x) \le 1 \wedge 2N_n(a(x)+b(x)).$$

Therefore we find that for all θ

$$\mathbb{E}T_2 \le \theta\sqrt{n\log N_n} + 2N_n\int_{\theta\sqrt{n\log N_n}}^{\infty} a(x)dx + 2N_n\int_{\theta\sqrt{n\log N_n}}^{\infty} b(x)dx. \qquad (29)$$

From eq(27), we have

$$2N_n\int_{\theta\sqrt{n\log N_n}}^{\infty} a(x)dx \le 2N_n\left((2+2c_0F)\sqrt{\frac{n}{\log N_n}}\right)\exp\left(-\frac{\theta\log N_n}{2+2c_0F}\right). \qquad (30)$$

When $\theta \ge 2+2c_0F$, it follows that

$$2N_n\int_{\theta\sqrt{n\log N_n}}^{\infty} a(x)dx \le 2\left((2+2c_0F)\sqrt{\frac{n}{\log N_n}}\right).$$

From eq(28), we have that for $\theta > \sigma^2$

$$
\begin{aligned}
&\int_{\theta\sqrt{n\log N_n}}^{\infty} b(x)dx \\
&\leq \int_{\theta\sqrt{n\log N_n}}^{\infty} \exp\left(-\frac{C_3(\frac{1}{\sigma^2}\sqrt{\frac{n}{\log N_n}}x-n)^2}{\frac{(C_1F^2+C_2F^4+16F^4)n^2\log n}{\log N_n}+(\frac{1}{\sigma^2}\sqrt{\frac{n}{\log N_n}}x-n)\frac{n\log^2 n}{\log N_n}}\right)dx \\
&\leq \int_{\theta\sqrt{n\log N_n}}^{\infty} \exp\left(-\frac{C_3(\frac{1}{\sigma^2}\sqrt{\frac{n}{\log N_n}}x-n)}{\frac{(C_1F^2+C_2F^4+16F^4)n\log n}{(\theta/\sigma^2-1)\log N_n}+\frac{n\log^2 n}{\log N_n}}\right)dx \\
&\leq \left(\frac{(C_1F^2+C_2F^4+16F^4)}{(\theta/\sigma^2-1)}+1\right)\frac{\sigma^2\sqrt{n}\log^2 n}{C_3\sqrt{\log N_n}}\exp\left(-\frac{C_3(\theta/\sigma^2-1)\log N_n}{\frac{C_1F^2+C_2F^4+16F^4}{\theta/\sigma^2-1}\log n+\log^2 n}\right).
\end{aligned}
$$

When $\theta \geq \sigma^2(1+2\log^2 n/C_3 \vee \sqrt{(2C_1F^2+2C_2F^4+32F^4)\log n/C_3})$,

$$
\exp\left(-\frac{C_3(\theta/\sigma^2-1)\log N_n}{\frac{C_1F^2+C_2F^4+16F^4}{\theta/\sigma^2-1}\log n+\log^2 n}\right). \leq \frac{1}{N_n}
$$

Therefore,

$$
2N_n\int_{\theta\sqrt{n\log N_n}}^{\infty} b(x)dx \leq (C_1F^2+C_2F^4+16F^4)\sigma^2\sqrt{\frac{n}{\log N_n}}+\frac{2\sigma^2\log^2 n}{C_3}\sqrt{\frac{n}{\log N_n}}. \quad (31)
$$

We choose $\theta = (2+2c_0F) \vee \sigma^2(1+2\log^2 n/C_3 \vee \sqrt{(2C_1F^2+2C_2F^4+32F^4)\log n/C_3})$. Combining eq(29), eq(30) and eq(31) gives

$$
\mathbb{E}T \leq C(F,\sigma^2,c)\sqrt{n\log N_n}\log^2 n, \quad (32)
$$

where $C(F,\sigma^2,c)$ is a constant depending on F,σ^2,c.

Similar to eq(29), we can prove that

$$
\mathbb{E}T_2^2 \leq \theta^2 n\log N_n + 2N_n\int_{\theta^2 n\log N_n}^{\infty} a(\sqrt{x})dx + 2N_n\int_{\theta^2 n\log N_n}^{\infty} b(\sqrt{x})dx. \quad (33)
$$

We still choose $\theta = (2+2c_0F) \vee \sigma^2(1+2\log^2 n/C_3 \vee \sqrt{(2C_1F^2+2C_2F^4+32F^4)\log n/C_3})$. Using the fact that $\int_{b^2}^{\infty}\exp(-(\sqrt{u}-c)a)du = 2(b/a+1/a^2)\exp(-a(b-c))$, we can estimate the upper for both $\int_{\theta^2 n\log N_n}^{\infty} a(\sqrt{x})dx$ and $\int_{\theta^2 n\log N_n}^{\infty} b(\sqrt{x})dx$. From eq(27), we have

$$
\begin{aligned}
\int_{\theta^2 n\log N_n}^{\infty} a(\sqrt{x})dx &\leq \int_{\theta^2 n\log N_n}^{\infty} \exp\left(-\frac{\sqrt{x}}{(2+2c_0F)\sqrt{\frac{n}{\log N_n}}}\right)dx \\
&= 2\left(\theta n(2+2c_0F)+(2+2c_0F)^2\frac{n}{\log N_n}\right)\exp\left(-\frac{\theta\sqrt{n\log N_n}}{(2+2c_0F)\sqrt{n/\log N_n}}\right) \\
&\leq C(F,c)\frac{n\log^2 n}{N_n}.
\end{aligned}
$$

From eq(28), we have

$$
\begin{aligned}
\int_{\theta^2 n\log N_n}^{\infty} b(\sqrt{x})dx &\le \int_{\theta^2 n\log N_n}^{\infty} \exp\left(-\frac{C_3(\frac{1}{\sigma^2}\sqrt{\frac{n}{\log N_n}}\sqrt{x}-n)^2}{\frac{(C_1F^2+C_2F^4+16F^4)n^2\log n}{\log N_n}+(\frac{1}{\sigma^2}\sqrt{\frac{n}{\log N_n}}\sqrt{x}-n)\frac{n\log^2 n}{\log N_n}}\right)dx \\
&\le \int_{\theta^2 n\log N_n}^{\infty} \exp\left(-\frac{C_3(\frac{1}{\sigma^2}\sqrt{\frac{n}{\log N_n}}\sqrt{x}-n)}{\frac{(C_1F^2+C_2F^4+16F^4)n\log n}{(\theta/\sigma^2-1)\log N_n}+\frac{n\log^2 n}{\log N_n}}\right)dx \\
&\le 2\left[\theta(\frac{C_1F^2+C_2F^4+16F^4}{\theta/\sigma^2-1}+1)n\log^2 n+(\frac{C_1F^2+C_2F^4+16F^4}{\theta/\sigma^2-1}+1)^2\frac{n\log^4 n}{\log N_n}\right] \\
&\qquad \cdot\exp\left(-\frac{C_3(\theta/\sigma^2-1)\log N_n}{\frac{C_1F^2+C_2F^4+16F^4}{\theta/\sigma^2-1}\log n+\log^2 n}\right) \\
&\le C(F,\sigma^2,c)n\log^4 n/N_n.
\end{aligned}
$$

Therefore, eq(33) gives

$$
\mathbb{E}T_2^2 \le C(F,\sigma^2,c)n\log^4 n\log N_n.
$$

With eq(25) and upper bound for $\mathbb{E}T_2, \mathbb{E}T_2^2$,

$$
\begin{aligned}
\mathbb{E}\left|\frac{2}{n}\sum_{i=1}^{n}\epsilon_i\tilde{f}(\mathbf{X})\right| &\le 2\delta+\frac{2}{n}C(F,\sigma^2,c)\sqrt{n\log N_n}\log^2 n(\delta+\sqrt{n^{-1}\log N_n}) \\
&\quad +\frac{2}{n}\sqrt{C(F,\sigma^2,c)n\log^4 n\log N_n}R^{1/2}(\tilde{f},f_0) \\
&\le \delta C(F,\sigma^2,c)\log^2 n+C(F,\sigma^2,c)\ \frac{\log N_n}{n}\log^2 n \\
&\quad +\frac{2}{n}\sqrt{C(F,\sigma^2,c)n\log^4 n\log N_n}R^{1/2}(\tilde{f},f_0).
\end{aligned} \tag{34}
$$

From the upper bound of inequality in (I), setting $\epsilon = 1$ gives

$$
\begin{aligned}
\sqrt{R(\tilde{f},f_0)} &\le \sqrt{2\widehat{R}_n(\tilde{f},f_0)}+\sqrt{2C_F\frac{\log^2 n\log N_n}{n}}+\sqrt{2C_F\delta\log^2 n} \\
&\qquad +\sqrt{\frac{4C_FF^2\log^4 n\log N_n}{n}} \\
&\le \sqrt{2\widehat{R}_n(\tilde{f},f_0)}+\sqrt{C_F\frac{\log^4 n\log N_n}{n}}+\sqrt{C_F\delta\log^2 n}.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
\frac{2}{n}\sqrt{C(F,\sigma^2,c)n\log^4 n\log N_n}R^{1/2}(\tilde{f},f_0) &\le \sqrt{\frac{C(F,\sigma^2,c)\log^4 n\log N_n}{n}}\widehat{R}_n^{1/2}(\tilde{f},f_0) \\
&\quad +C(F,\sigma^2,c)\frac{\log^4 n\log N_n}{n}+2\sqrt{\frac{\delta C_F\log^6 n\log N_n}{n}}.
\end{aligned} \tag{35}
$$

Using the fact that $2\sqrt{\delta C_F\log^6 n\log N_n/n} \le \delta\log^2 n + C_F\log^4 n\log N_n/n$, then inequality (II) follows from eq(34), eq(35).

(III): For any fixed $f \in \mathcal{F}$, $\mathbb{E}[\frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{f}(\mathbf{X}_i))^2] \le \mathbb{E}[\frac{1}{n}\sum_{i=1}^{n}(Y_i - f(\mathbf{X}_i))^2] + \Delta_n$. Because of $\mathbf{X}_i \overset{\mathcal{D}}{=} \mathbf{X}$ and f being deterministic, we have $\mathbb{E}[\|f - f_0\|_n^2] = \mathbb{E}(f(\mathbf{X}) - f_0(\mathbf{X}))^2$. Since also $\mathbb{E}[\epsilon_i f(\mathbf{X}_i)] = 0$,

$$\begin{aligned}
\widehat{R}_n(\widehat{f}, f_0) &\le \mathbb{E}[\|f - f_0\|_n^2] + \mathbb{E}[\frac{2}{n}\sum_{i=1}^{n}\epsilon_i \widehat{f}(\mathbf{X}_i)] + \Delta_n \\
&\le \mathbb{E}[\|f - f_0\|_n^2] + \Delta_n + \delta C(F, \sigma^2, c)\log^2 n + C(F, \sigma^2, c)\log^4 n \frac{\log N_n}{n} \\
&\quad + 2C(F, \sigma^2, c)\sqrt{\frac{\log^4 n \log N_n}{n}}\widehat{R}^{1/2}(\widehat{f}, f_0).
\end{aligned} \tag{36}$$

Observe that

$$\begin{aligned}
2C(F, \sigma^2, c)\sqrt{\frac{\log^4 n \log N_n}{n}}\widehat{R}^{1/2}(\widehat{f}, f_0) \le \frac{1+\epsilon}{\epsilon}C^2(F, \sigma^2, c)\log^4 n \log N_n / n \\
+ \frac{\epsilon}{1+\epsilon}\widehat{R}(\widehat{f}, f_0).
\end{aligned} \tag{37}$$

Combining eq(36) with eq(37) and rearranging $\widehat{R}(\widehat{f}, f_0)$ to one side give inequality (III).

(IV): Let $\tilde{f} \in \mathrm{argmin}_{f \in \mathcal{F}} \sum_{i=1}^{n}(Y_i - f(\mathbf{X}_i))^2$ be an empirical risk minimizer. We have

$$\begin{aligned}
&\widehat{R}_n(\widehat{f}, f_0) - \widehat{R}_n(\tilde{f}, f_0) \\
&= \Delta_n + 2\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\epsilon_i \widehat{f}(\mathbf{X}_i)\right] - 2\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\epsilon_i \tilde{f}(\mathbf{X}_i)\right] \\
&\ge \Delta_n - 2\delta C(F, \sigma^2, c)\log^2 n - 2C(F, \sigma^2, c)\log^4 n \frac{\log N_n}{n} \\
&- 2C(F, \sigma^2, c)\sqrt{\frac{\log^4 n \log N_n}{n}}\widehat{R}_n^{1/2}(\widehat{f}, f_0) - 2C(F, \sigma^2, c)\sqrt{\frac{\log^4 n \log N_n}{n}}\widehat{R}_n^{1/2}(\tilde{f}, f_0).
\end{aligned} \tag{38}$$

Observe that

$$\begin{aligned}
2C(F, \sigma^2, c)\sqrt{\frac{\log^4 n \log N_n}{n}}\widehat{R}^{1/2}(\widehat{f}, f_0) &\le C^2(F, \sigma^2, c)\frac{\log^4 n \log N_n}{n} + \widehat{R}(\widehat{f}, f_0) \\
2C(F, \sigma^2, c)\sqrt{\frac{\log^4 n \log N_n}{n}}\widehat{R}_n^{1/2}(\tilde{f}, f_0) &\le \frac{1-\epsilon}{\epsilon}C^2(F, \sigma^2, c)\frac{\log^4 n \log N_n}{n} + \frac{\epsilon}{1-\epsilon}\widehat{R}_n(\tilde{f}, f_0).
\end{aligned} \tag{39}$$

Inequality (IV) follows from eq(38), eq(39). ☐

lemma 4. *Consider the d-variate nonparametric regression model* $Y_i = f_0(\mathbf{X}_i) + \epsilon_i$ *with unknown regression function* f_0*, satisfying* $\|f_0\|_\infty \le F$ *for some* $F \ge 1$*. Let* $\widehat{f}_n$ *be any estimator taking values in the class* $\mathcal{F}(L, p, s, F)$ *and* Δ_n *is defined in lemma 3. Under Assumptions 1-3, for any* $\epsilon \in (0, 1]$*, there exists a constant* C_ϵ*, only depending on* ϵ*, such that with*

$$\gamma_{\epsilon,n} := C_\epsilon F^2 \frac{(s+1)\log(n(s+1)^L)p_0 p_{L+1}\log^4 n}{n},$$

$$
\begin{aligned}
&(1-\epsilon)^2\Delta_n(\widehat{f}_n, f_0) - \gamma_{\epsilon,n} \\
&\leq R(\widehat{f}_n, f_0) \leq (1+\epsilon)^2 \left(\inf_{f\in\mathcal{F}(L,p,s,F)} \|f - f_0\|_\infty^2 + \Delta_n(\widehat{f}_n, f_0) \right) + \gamma_{\epsilon,n}
\end{aligned}
$$

Proof. Lemma 4 follows from Lemma 3 with the choice of $\delta = 1/n$, $\mathcal{F} = \mathcal{F}(L,p,s,\infty)$ and Remark 1 in [3]

$$
\log\mathcal{N}\left(\delta, \mathcal{F}(L,p,s,\infty), \|\cdot\|_\infty\right) \leq (s+1)\log\left(2^{2L+5}\delta^{-1}(L+1)p_0^2 p_{L+1}^2 s^{2L}\right)
$$

□

A.2 Proof of Theorem 1

Proof of theorem 1. Combining lemma 4 with the assumed bounds on L and s, it follows that

$$
\begin{aligned}
&\frac{1}{4}\Delta_n(\widehat{f}_n, f_0) - C^{'}\phi_n L\log^6 n \leq R(\widehat{f}, f_0) \\
&4 \inf_{f^*\in\mathcal{F}(L,p,s,F)} \|f^* - f_0\|_\infty^2 + 4\Delta_n(\widehat{f}_n, f_0) + C^{'}\phi_n L\log^6 n,
\end{aligned}
\tag{40}
$$

where we used $\epsilon = 1/2$ for the lower bound and $\epsilon = 1$ for the upper bound. Let $C = 8C^{'}$, then $C^{'}\phi_n L\log^6 n \leq \widehat{R}(\widehat{f}, f_0)$ whenever $\Delta_n(\widehat{f}, f_0) \geq C\phi_n L\log^6 n$. The lower bound on (7) is proved.

To get the upper bound, we need to control $\inf_{f^*\in\mathcal{F}(L,p,s,F)} \|f^* - f_0\|_\infty^2$. From eq(26) in [3], we know that

$$
\inf_{f^*\in\mathcal{F}(L,p,s)} \|f^* - f_0\|_\infty^2 \leq C^{'} \max_{i=0,\cdots,q} N^{-\frac{2\beta_i^*}{t_i}} \leq C^{'} \max_{i=0,\cdots,q} C_0^{-\frac{2\beta_i^*}{t_i}} n^{-\frac{\beta_i^*}{2\beta_i^*+t_i}}.
$$

Hence, $\inf_{f^*\in\mathcal{F}(L,p,s,F)} \|f^* - f_0\|_\infty^2 \leq C_1\phi_n$. Therefore, there exists $\tilde{f} \in \mathcal{F}(L,p,s,F)$ such that $\|f^* - f_0\|_\infty^2 \leq C_1\phi_n$. Define $f^* = \tilde{f}(\|f_0\|_\infty/\|\tilde{f}\|_\infty)$. Then $\|f^*\|_\infty \leq \|f_0\|_\infty = \|g_q\|_\infty \leq K \leq F$, which implies that $\tilde{f} \in \mathcal{F}(L,p,s,F)$. Writing $f^* - f_0 = (f^* - \tilde{f}) + (\tilde{f}_n - f_0)$, we obtain $\|f^* - f_0\|_\infty \leq 2\|\tilde{f} - f_0\|_\infty \leq 2C_1\phi_0$. From eq(40), we obtain that

$$
R(\widehat{f}, f_0) \leq 8C_1\phi_0 + 4\Delta_n(\widehat{f}_n, f_0) + C^{'}\phi_n L\log^6 n. \tag{41}
$$

The upper bounds on eq(6) and eq(7) follow from eq(41). □

A.3 Proof of Theorem 2

Proof of Theorem 2. From lemma 3, we know that

$$
\begin{aligned}
&(1-\epsilon)^2\Delta_n - C(F,\sigma^2,c)\frac{\log^4 n\log N_n}{n\epsilon} - \delta C(F,\sigma^2,c)\log^2 n \\
&\leq R(\widehat{f}, f_0) \leq \\
&(1+\epsilon)^2\left(\inf_{f^*\in\mathcal{F}} \|f^* - f_0\|_\infty^2 + \Delta_n(\widehat{f}, f) + C(F,\sigma^2,c)\delta\log^2 n\right) \\
&+ \frac{(1+\epsilon)^3}{\epsilon}C(F,\sigma^2,c)\frac{\log^4 n\log N_n}{n}.
\end{aligned}
$$

Let $\mathcal{F} := \mathcal{F}(L,p,s,\infty)$. From remark 1 in [3],

$$
\log\mathcal{N}\left(\delta, \mathcal{F}(L,p,s,\infty), \|\cdot\|_\infty\right) \leq (s+1)\log\left(2^{2L+5}\delta^{-1}(L+1)p_0^2 p_{L+1}^2 s^{2L}\right).
$$

According to the assumption $s \asymp dL$ and $d \lesssim p$ with the choice of $\delta = \frac{1}{n}$, we have that $\log N_n \asymp dL\log n$. Therefore, using the result of lemma 3 with $\epsilon = \frac{1}{2}$, we have that

$$\begin{aligned}&\frac{1}{4}\Delta_n - C' \frac{dL\log^5 n}{n} \le R(\widehat{f}, f_0) \le \\ &4 \inf_{f^* \in \mathcal{F}} \|f^* - f_0\|_\infty^2 + 4\Delta_n(\widehat{f}, f) + C' \frac{dL\log^5 n}{n}.\end{aligned} \tag{42}$$

The lower bound of eq(11) can be derived from the left side of eq(42) with the assumption $d \asymp n^{\frac{1}{\alpha+1}}$. To derive the upper bound of eq(10) and eq(11), we need the upper bound of $\inf_{f^* \in \mathcal{F}(L,\mathbf{p},s,F)} \|f^* - f_0\|_\infty$. Let $\tilde{f} = \sum_{i=1}^d \phi_i Y_{t-i} \in \mathcal{F}(3, (d, 2Kd, 1), s), s = 4Kd$ (For each Y_{t-i} in input layer, it maps to $2\lceil \phi_i \rceil$ units in hidden layer. $\lceil \phi_i \rceil$ units are $\phi_i(Y_{t-i})_+/\lceil \phi_i \rceil$ and other $\lceil \phi_i \rceil$ units are $-\phi_i(-Y_{t-i})_+/\lceil \phi_i \rceil$. Assumption $\phi_i < K$ implies that there are at most $2Kd$ units in hidden layer. The output $\tilde{f}$ equals to the summation of all hidden units). We have that

$$\|\tilde{f} - f_0\|_\infty = \| \sum_{i=d+1}^{\infty} \phi_i Y_{t-i} \|_\infty \le K \sum_{d+1}^{\infty} |\phi_i| \lesssim \frac{1}{(d+2)^\alpha}. \tag{43}$$

Define $f^* = \tilde{f}(\|f_0\|_\infty/\|\tilde{f}\|_\infty)$. Then $\|f^*\|_\infty \le \|f_0\|_\infty \le K \sum_{i=1}^{\infty} |\phi_i| \le KM/2^\alpha \le F$. Therefore, $f^* \in \mathcal{F}(4, \mathbf{p}^*, 4Kd+1, F)$. To extend the layer of neural network from 4 to L, we can add $(L-4)$ additional identical layers before $\mathcal{F}(4, \mathbf{p}, (6K+2)d, F)$. Note that the dimension of input vector is d, the deepened network belongs to

$$\mathcal{F}(L, (d, \cdots, d, \mathbf{p}^*), (4K+L)d+1, F).$$

Therefore, the neural network $f^* \subset \mathcal{F}(L, \mathbf{p}, s, F)$. Writing $f^* - f_0 = (f^* - \tilde{f}) + (\tilde{f}_n - f_0)$, we obtain $\|f^* - f_0\|_\infty \le 2\|\tilde{f} - f_0\|_\infty$. With eq(42) and eq(43), we have that

$$R(\widehat{f}, f_0) \le C\frac{1}{d^\alpha} + 4\Delta_n(\widehat{f}, f) + C' \frac{dL\log^5 n}{n}. \tag{44}$$

Using the assumption $d \asymp n^{\frac{1}{\alpha+1}}$ again, we can derive the upper bound of eq(10) and eq(11). $\square$

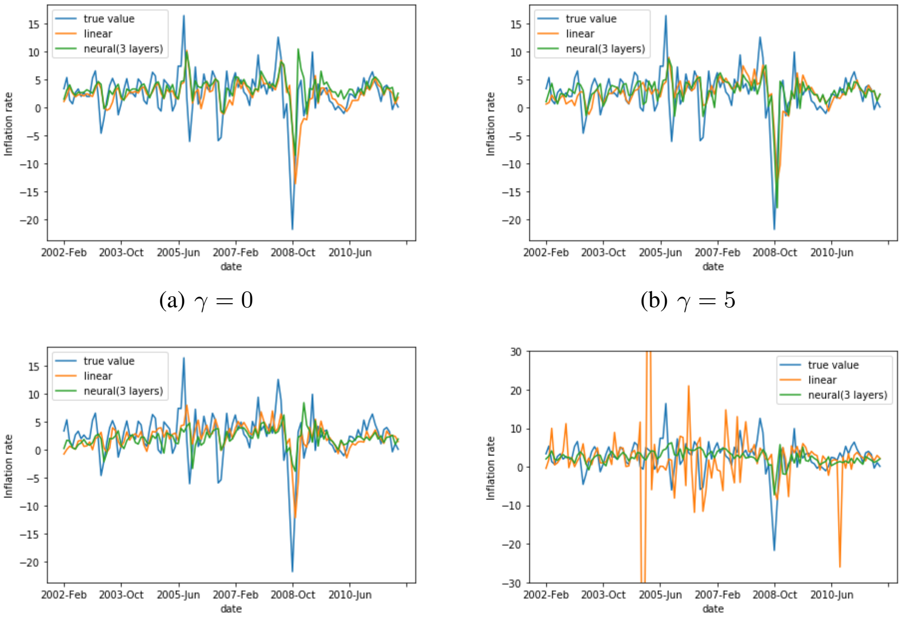
B Additional details on numerical experiments

Computer information:
Processor: Intel(R) Core(TM) i7-9750H CPU @ 2.60GHz 2.59 GHz
Installed RAM: 16.0 GB (15.9 GB usable)
System type: Windows 10 Home 64-bit operating system, x64-based processor
Disk: Samsung mzvlb512hbjq-000h1
GPU: NVIDIA GeForce GTX 1660 Ti

Additional figure in Section 5: Figure 3.

C Numerical comparison between DNN and LSTM

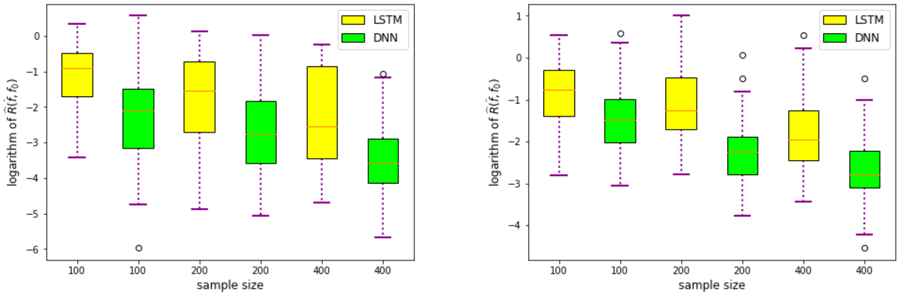
In this section, we compare the predicting performance of DNN with LSTM. We consider the same set of models as in section 4.2. The input dimension of both DNN and LSTM are determined by AIC and the number of hidden layer units are set to be 10. The results are shown in figure 4. As we can see from this figure, DNN has a slightly faster convergence rate than LSTM in linear AR models while for non-linear AR models, LSTM performs slightly better. In fact, the difference between the results of DNN and LSTM in these simulation settings are not statistically significant.



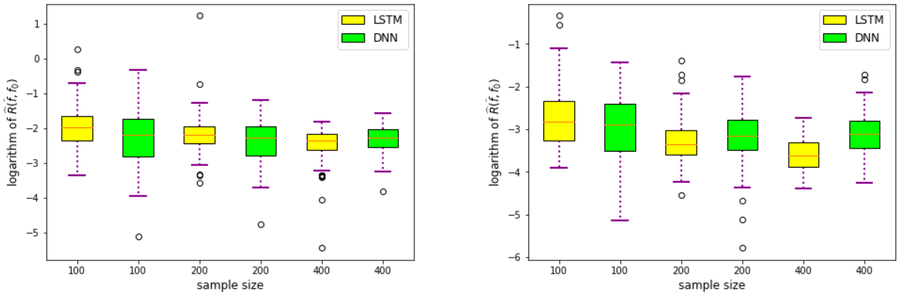
(a) $\gamma = 0$ (b) $\gamma = 5$

(c) $\gamma = 7.5$ (d) $\gamma = 10$

Figure 3: Estimation for monthly inflation rate. The blue line reflects the true change of inflation rate from 2002 Feb to 2011 Nov. Other two lines correspond to two estimators.



(a) model (1) (b) model (2)



(c) model (3) (d) model (4)

Figure 4: : Box plots of logarithm of the mean square error on the testing set as a function of sample size.

References

[1] VH De La Pena. A general class of exponential inequalities for martingales and ratios. 1999. *Ann. Probab*, 36:1902–1938.

[2] Florence Merlevède, Magda Peligrad, and Emmanuel Rio. Bernstein inequality and moderate deviations under strong mixing conditions. In *High dimensional probability V: the Luminy volume*, pages 273–292. Institute of Mathematical Statistics, 2009.

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