ERROR ANALYSIS OF FITTED Q-ITERATION WITH RELU-ACTIVATED DEEP NEURAL NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Deep reinforcement learning (RL) has grown rapidly with the development of backbone feedforward neural networks (FNNs). However, there remains a theoretical gap when researchers conduct error analysis of the FNNs-based RL process. In this work, we provide an error analysis for deep-fitted Q-iteration applying ReLU-activated FNNs for value function approximation.

1 INTRODUCTION

Reinforcement learning (RL) has successfully trained sequential decision-making models over the last decade (Silver et al., 2016; Chen et al., 2020; Cao et al., 2022). Unlike conventional supervised training with explicit targets, RL aims to generate an agent maximizing the expected future return through implementing actions, interacting with the environment, and obtaining rewards. A well-behaved agent is to train a value function that maximizes the final reward. Such processes can be mathematically modeled as Markov Decision Processes (MDPs) defined in Appendix A.1.

Deep RL replaced prior-defined value function with FNNs (Henderson et al., 2018). As a representative value-based algorithm in deep RL, deep-fitted Q-iteration (DFQI) takes transition data as its input and approximates the target value function using FNNs (Ernst et al., 2005). The statistical properties of traditional fitted Q-iteration (FQI) with function approximation are well-studied (Murphy, 2005). For DFQI, Fan et al. (2020) tried to provide a theoretical analysis.

Our work further complements Fan et al. (2020) in the following aspects: (1) Our error bound depends on the ambient dimension d explicitly and polynomial (not implicitly and exponentially); (2) We introduce a weaker α-mixing condition (Hang & Steinwart, 2014) than β-mixing in Antos et al. (2007) to characterize the dependency of MDPs (not assumed that the batch data are independently and identically distributed ignoring the temporal dependency of MDPs); (3) We assume that the optimal action-value function Q^* is a Hölder continuous function (without a composition form of certain functions).

2 METHOD AND RESULTS

Through the tools of error propagation (Proposition C.1), statistical error analysis (Theorem C.1), and deep approximation error analysis (Theorem C.2), a non-asymptotic error bound has been established between the estimated action-value function corresponding to the estimated greedy policy and the optimal Q^* by controlling the statistical and approximation errors on MDPs assumed to be α-mixing. Then we derive the generalization bound in terms of the ReLU-activated FNNs and α-mixing data in the context of RL. Finally, we demonstrate that the error bound depends on the sample size, the ambient dimension (polynomially), the width and depth of the neural network, and the number of training iterations, and thus provide a powerful tool in hyper-parameters setting.

Since the optimal Bellman operator $\mathcal{T}^*$ defined in (5) is ζ-contraction ($\zeta \in [0, 1)$), at the population level, we can apply fixed point iteration to approximate the optimal action-value function Q^* given in (4). To be precise, suppose that $\mathcal{R}(\cdot|x, a)$ and $P(\cdot|x, a)$ are known, the following iteration

$$Q_0 \to Q_1 = \mathcal{T}^* Q_0 \to Q_2 = \mathcal{T}^* Q_1 \to \ldots \to Q_J = \mathcal{T}^* Q_{J-1}, \tag{1}$$

approximate Q^* well when J is large enough. Q_{j-1} and $Q_j, j = 1, \ldots, J$, as given in (1), obviously satisfies

$$Q_j \in \underset{measurable\ Q}{\arg\min}\ \mathcal{L}(Q) = \mathbb{E}|Q(X,A) - Y|^2$$
$$= \|Q - \mathcal{T}^* Q_{j-1}\|^2_{L^2(\mu)} + \mathbb{E}[(\mathcal{T}^* Q_{j-1}(X,A) - Y)^2], \tag{2}$$

where μ denotes the distribution of the state-action pair (X, A) and $Y = R + \zeta \max_{a' \in \mathcal{A}} Q_{j-1}(X', a')$. In practice, we only have the batch data $\{Z_i\}_{i=1}^n = \{X_i, A_i, R_i, X_i'\}_{i=1}^n$. DFQI (Ernst et al., 2005) mimics the iteration (1) via replacing $Q_j, j = 1, \ldots, J$ with $\widehat{Q}_j$, an estimator in $\mathcal{F}$ given by the following regression problem

$$\widehat{Q}_j \in \arg\min_{Q \in \mathcal{F}} \widehat{\mathcal{L}}(Q) = \frac{1}{n} \sum_{i=1}^n \left(Q(X_i, A_i) - Y_i\right)^2, \tag{3}$$

where $Y_i = R_i + \zeta \max_{a' \in \mathcal{A}} \widehat{Q}_{j-1}(X_{i+1}, a')$, $\mathcal{F}$ is the ReLU-activated FNNs, and $\widehat{Q}_0 \in \mathcal{F}$ is an initial guess. It is easy to check that the empirical loss in (3) is an unbiased estimation of the population loss in (2), i.e., $\mathbb{E}[\widehat{\mathcal{L}}(Q)] = \mathcal{L}(Q), \forall Q \in \mathcal{F}$ (See DFQI details in Algorithm 1).

Now, we give the main result in this paper, a non-asymptotic error bound of DFQI.

Theorem 2.1. *Suppose that $\{\mathcal{T}^* \widehat{Q}_{j-1}\}_{j=1}^J \in \mathcal{H}^\gamma$ in Definition C.3 with $\gamma = s + r, s \in \mathbb{N}_0$ and $r \in (0, 1]$, $\{Z_i\}_{i=1}^n$ is strictly exponentially α-mixing in Definition C.2, and the probability distribution μ of (X, A) is absolutely continuous with respect to Lebesgue measure. Then, for the ReLU-activated FNNs $\mathcal{F}$ with the width $\mathcal{W} = \mathcal{O}\left((n^{\frac{\eta}{1+\eta}})^{\frac{d}{4(d+4\gamma)}} \log n\right)$ and depth $\mathcal{D} = \mathcal{O}\left((n^{\frac{\eta}{1+\eta}})^{\frac{d}{4(d+4\gamma)}} \log n\right)$, we have*

$$\mathbb{E}\left[\|Q^* - Q^{\pi_J}\|_{L_1(\nu)}\right] \leq \frac{2\sqrt{C} C_{\nu,\mu} \zeta}{(1-\zeta)^2} \cdot \left[d^{s+(\gamma \vee 1)/2} (n^{\frac{\eta}{1+\eta}})^{\frac{-\gamma}{d+4\gamma}} (\log n)^{3/2}\right] + \frac{4\zeta^{J+1}}{(1-\zeta)^2} R_{\max},$$

where C is a constant depending on $s, B, \mathcal{B}, R_{\max}, \eta, a, \bar{\alpha}$.

The completeness condition $\{\mathcal{T}^* \widehat{Q}_{j-1}\}_{j=1}^J \in \mathcal{H}^\gamma$ used in Theorem 2.1 is mild. It holds when the underlying MDP satisfies some smoothness conditions (Fan et al., 2020). It was also pointed out that some completeness condition is indispensable (Chen & Jiang, 2019). By Theorem 2.1, when J is large enough, the non-asymptotic error bound is $\mathcal{O}\left(n^{\frac{-\gamma\eta}{(1+\eta)(d+4\gamma)}}\right)$ by omitting the logarithmic term. Therefore, we get the consistency result of DFQI when $n \to \infty$. It improves the result in Fan et al. (2020) by taking the temporal dependency of data into account and reducing the dependency on the dimension d from exponentially to polynomially.

3 Conclusion

In this paper, we studied the error bound of DFQI conducting value functions approximation via ReLU-activated FNNs in batch-based RL. The bound explicitly depends on the number of samples, the ambient dimension polynomially, the width and depth of the neural network, and the number of iterations, which provides insight into hyper-parameters setting when training DFQI or other value-based RL algorithms to achieve a desired convergence rate in RL.

URM Statement

The authors acknowledge that at least one key author of this work meets the URM criteria of ICLR 2023 Tiny Papers Track.

References

András Antos, Csaba Szepesvári, and Rémi Munos. Fitted q-iteration in continuous action-space mdps. *Advances in Neural Information Processing Systems*, 20:9–16, 2007.

András Antos, Csaba Szepesvári, and Rémi Munos. Learning near-optimal policies with bellman-residual minimization based fitted policy iteration and a single sample path. *Machine Learning*, 71(1):89–129, 2008.

Peter L Bartlett, Nick Harvey, Christopher Liaw, and Abbas Mehrabian. Nearly-tight vc-dimension and pseudodimension bounds for piecewise linear neural networks. *Journal of Machine Learning Research*, 20(1):2285–2301, 2019.

Benedikt Bauer and Michael Kohler. On deep learning as a remedy for the curse of dimensionality in nonparametric regression. *The Annals of Statistics*, 47(4):2261 – 2285, 2019.

Yang Cao, Cher Tian Ser, Marta Skreta, Kjell Jorner, Nathanael Kusanda, and Alán Aspuru-Guzik. Reinforcement learning supercharges redox flow batteries. *Nature Machine Intelligence*, pp. 1–2, 2022.

Alvin I Chen, Max L Balter, Timothy J Maguire, and Martin L Yarmush. Deep learning robotic guidance for autonomous vascular access. *Nature Machine Intelligence*, 2(2):104–115, 2020.

Jinglin Chen and Nan Jiang. Information-theoretic considerations in batch reinforcement learning. In *International Conference on Machine Learning*, pp. 1042–1051, 2019.

Damien Ernst, Pierre Geurts, and Louis Wehenkel. Tree-based batch mode reinforcement learning. *Journal of Machine Learning Research*, 6:503–556, 2005.

Jianqing Fan, Zhaoran Wang, Yuchen Xie, and Zhuoran Yang. A theoretical analysis of deep q-learning. In *Learning for Dynamics and Control*, pp. 486–489, 2020.

Amir-massoud Farahmand, Mohammad Ghavamzadeh, Csaba Szepesvari, and Shie Mannor. Regularized policy iteration with nonparametric function spaces. *Journal of Machine Learning Research*, 17(1):4809–4874, 2016.

Hanyuan Hang and Ingo Steinwart. Fast learning from α-mixing observations. *Journal of Multivariate Analysis*, 127:184–199, 2014.

Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep reinforcement learning that matters. In *AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Yuling Jiao, Guohao Shen, Yuanyuan Lin, and Jian Huang. Deep nonparametric regression on approximately low-dimensional manifolds. *Annals of Statistics*, 2021.

Jianfeng Lu, Zuowei Shen, Haizhao Yang, and Shijun Zhang. Deep network approximation for smooth functions. *SIAM Journal on Mathematical Analysis*, 53(5):5465–5506, 2021.

Rémi Munos. Error bounds for approximate policy iteration. In *International Conference on Machine Learning*, volume 3, pp. 560–567, 2003.

Susan A Murphy. A generalization error for q-learning. *Journal of Machine Learning Research*, 6: 1073–1097, 2005.

Ryumei Nakada and Masaaki Imaizumi. Adaptive approximation and generalization of deep neural network with intrinsic dimensionality. *Journal of Machine Learning Research*, 21:174–1, 2020.

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

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

Tengyang Xie and Nan Jiang. Q* approximation schemes for batch reinforcement learning: A theoretical comparison. In *Conference on Uncertainty in Artificial Intelligence*, pp. 550–559, 2020.

Tengyang Xie and Nan Jiang. Batch value-function approximation with only realizability. In *International Conference on Machine Learning*, volume 139, pp. 11404–11413, 2021.

A BACKGROUND AND NOTATIONS

A.1 MARKOV DECISION PROCESS

MDP is firstly modeled to facilitate the following derivation: A quintuple $(\mathcal{X}, \mathcal{A}, P, \mathcal{R}, \zeta)$ represents a discounted MDP, where $\mathcal{X}$ is the state space, $\mathcal{A}$ is the action space, $P : \mathcal{X} \times \mathcal{A} \subseteq \mathbb{R}^d \to \mathcal{M}(\mathcal{X})$ is the transition probability kernel, $\mathcal{R}(\cdot \mid x, a)$ refers to the distribution of immediate reward $R(x, a)$, and $\zeta \in [0, 1)$ is the discount factor. $\mathcal{M}(\mathcal{X})$ here denotes the sets of probability measure on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$, such that $P(\cdot|x, a)$ is a probability measure on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ for each pair $(x, a) \in \mathcal{X} \times \mathcal{A}$, which defines the next-state distribution upon taking action a in state x, and $P(D|\cdot, \cdot)$ is one measurable function on $\mathcal{X} \times \mathcal{A}$ for every $D \in \mathcal{B}(\mathcal{X})$. Moreover, let $\pi(\cdot|x)$ denotes the stochastic policy which is an associated distribution of the action at state x. Given one initial distribution $\nu \in \mathcal{M}(\mathcal{X})$, i.e., $X_1 \sim \nu$, the batch data $\{Z_i\}_{i=1}^n = \{X_i, A_i, R_i, X_i'\}_{i=1}^n$ with $X_i' = X_{i+1}$ is generated by

$$X_1 \sim \nu,\ A_i \sim \pi(\cdot \mid X_i),\ R_i \sim \mathcal{R}(\cdot \mid X_i, A_i),\ X_i' \sim P(\cdot \mid X_i, A_i),\ i = 1, \ldots, n.$$

In this work, we assume that the MDP $\{Z_i\}_{i=1}^n$ is strictly stationary α-mixing (see Definition C.2). Here strict stationarity means that $Z_i's$ admit the same distribution.

Denote the action-value function as

$$Q^\pi(x, a) := \mathbb{E}\left[\sum_{i=1}^{\infty} \zeta^{i-1} R_i \mid X_1 = x, A_1 = a, \pi\right].$$

For a given policy π, Q^π is the unique fixed point of the Bellman operator

$$\mathcal{T}^\pi Q(x, a) := \mathbb{E}R(x, a) + \zeta P^\pi Q(x, a),$$

with

$$P^\pi Q(x, a) := \int P(dx'|x, a)\pi(da'|x')Q(x', a').$$

Without loss of generality, we can assume that $R(x, a) \in [0, R_{\max}]$ for each pair $(x, a) \in \mathcal{X} \times \mathcal{A}$, then Q^π takes values in $\left[0, \frac{R_{\max}}{1-\zeta}\right]$. Assume that there exists a policy π^* that maximizes Q^π, such that

$$Q^* := Q^{\pi^*}. \tag{4}$$

Q^* in (4) satisfies the optimal Bellman equation $Q^* = \mathcal{T}^* Q^*$, where the optimal Bellman operator $\mathcal{T}^*$ is given by

$$\mathcal{T}^* Q(x, a) = \mathbb{E}[R(x, a)] + \zeta \mathbb{E}_{X' \sim P(\cdot|x,a)} \max_{a' \in \mathcal{A}} [Q(X', a')]. \tag{5}$$

It is easy to check that $\mathcal{T}^*$ is ζ-contraction in the sup-norm. We define the greedy policy of an action-value function Q as

$$\pi(x; Q) \in \operatorname*{argmax}_{a \in \mathcal{A}} Q(x, a),\ x \in \mathcal{X}.$$

A.2 RELU-ACTIVATED FEEDFORWARD NEURAL NETWORKS

We now introduce the FNNs with ReLU activations. We use $\mathcal{F}$ to denote the class of FNNs $f_\theta : \mathbb{R}^d \to \mathbb{R}$ with parameter θ, depth $\mathcal{D}$, width $\mathcal{W}$, where f_θ is defined as $f_\theta(x) = v_\mathcal{D} \circ \rho \circ v_{\mathcal{D}-1} \circ \rho \circ \cdots \circ \rho \circ v_1 \circ \rho \circ v_0(x),\ x \in \mathbb{R}^d$, and therefore $\|f_\theta\|_\infty \leq \mathcal{B}$ holds for some $0 < \mathcal{B} < \infty$, where $\|\cdot\|_\infty$ refers to the sup-norm, $\rho(x) = \max(0, x)$ is the ReLU activation function operates that pointwisely on x and

$$v_i(x) = \widetilde{A}_i x + b_i, \quad i = 0, 1, \ldots, \mathcal{D},$$

$\widetilde{A}_i \in \mathbb{R}^{d_{i+1} \times d_i}$ is the weight matrix, $b_i \in \mathbb{R}^{d_{i+1}}$ is the bias vector, and d_i is the width of the i-th layer. The first layer takes the input data and the last layer gives the output target. The FNNs f_θ has $\mathcal{D}$ hidden layers and in total $(\mathcal{D} + 1)$ layers. We use a $(\mathcal{D} + 1)$-vector $(d_0, d_1, \ldots, d_\mathcal{D})^\top$ to describe the width of each layer; in particular, $d_0 = d$ is the dimension of the input (X, A) and $d_\mathcal{D} = 1$ is the dimension of the output. The width $\mathcal{W}$ is defined as the maximum width of hidden layers, i.e., $\mathcal{W} = \max\{d_1, \ldots, d_\mathcal{D}\}$.

A.3 OTHER NOTATIONS

We introduce other notations used throughout this paper. For any $a, b \in \mathbb{R}$, $\lceil a \rceil$ denotes the smallest integer no less than a, $\lfloor a \rfloor$ denotes the largest integer less than a and $a \vee b := \max\{a, b\}$. Let $\mathbb{N}_0, \mathbb{N}$ refer to non-negative and strictly positive integers, respectively. $\|x\|_q = (\sum_{i=1}^d |x_i|^q)^{\frac{1}{q}}$ is the q-norm ($q \in [1, \infty]$) of a vector $x = (x_1, \ldots, x_d)^\top \in \mathbb{R}^d$. For probability measure μ and measurable function $Q : \mathbb{R}^d \to \mathbb{R}^1$, we write $\|Q\|_{L^q(\mu)}^q = \mathbb{E}_{x\sim\mu}|Q(x)|^q$.

B DFQI ALGORITHM

Algorithm 1 Deep Fitted Q-Iteration Algorithm

1: Input: Initial value $\widehat{Q}_0 \in \mathcal{F}$.
2: **for** $j = 1, \ldots, J$ **do**
3: Sampling (X_i, A_i, R_i, X_i'), $i = 1, \ldots, n$.
4: Compute $Y_i = R_i + \zeta \max_{a' \in \mathcal{A}} \widehat{Q}_{j-1}(X_i', a')$.
5: Obtain the j-step action-value function $\widehat{Q}_j$ via solving equation 3, that is,

$$\widehat{Q}_j \in \arg\min_{Q \in \mathcal{F}} \widehat{\mathcal{L}}(Q).$$

6: **end for**
7: Output: Estimator $\widehat{Q}_J$ of Q^* and the greed policy $\pi_J = \pi(\cdot; \widehat{Q}_J)$.

C ERROR ANALYSIS

In this section, we present the error analysis for DFQI by bounding $\|Q^* - Q^{\pi_J}\|_{L_1(\nu)}$ for any admissible distribution ν. We first introduce the following definition of concentration coefficients that control the distribution shift because certain concentratability is necessary for the theoretical development of the batch mode RL (Munos, 2003; Xie & Jiang, 2020; 2021; Chen & Jiang, 2019).

Definition C.1. *(Concentration Coefficients). Let $\nu_1, \nu_2 \in \mathcal{M}(\mathcal{X} \times \mathcal{A})$ be two probability measures that are absolutely continuous with respect to the Lebesgue measure on $\mathcal{X} \times \mathcal{A}$. Let $\{\pi_t\}_{t\geq 1}$ be a sequence of policies. Suppose the initial state-action pair (X_0, A_0) of the MDP has distribution ν_1, and we take action A_t according to the policy π_t. For any integer m, we denote the distribution of $\{(X_t, A_t)\}_{t=0}^m$ by $\nu_1 P^{\pi_1} P^{\pi_2} \cdots P^{\pi_m}$. The m-th concentration coefficient is defined as*

$$c_{\nu_1,\nu_2}(m) = \sup_{\pi_1,\ldots,\pi_m} \left\| \frac{d\left(\nu_1 P^{\pi_1} P^{\pi_2} \ldots P^{\pi_m}\right)}{d\nu_2} \right\|_\infty,$$

where the supremum is taken over all possible policies. Furthermore, let μ be the distribution of (X_i, A_i) in Algorithm 1 and let ν be a fixed distribution on $\mathcal{X} \times \mathcal{A}$. Denote

$$C_{\nu,\mu} := (1-\zeta)^2 \cdot \sum_{m\geq 1} m\zeta^{m-1} c_{\nu,\mu}(m), \tag{6}$$

and assume $C_{\nu,\mu} < \infty$, where $(1-\zeta)^2$ in equation 6 is a normalization term, since $\sum_{m>1} \zeta^{m-1} \cdot m = (1-\zeta)^{-2}$.

The following proposition on error propagation (Antos et al., 2008; Farahmand et al., 2016; Fan et al., 2020) connects the error bound of $\|Q^* - Q^{\pi_J}\|_{L_1(\nu)}$ with that of $\|\widehat{Q}_j - \mathcal{T}^*\widehat{Q}_{j-1}\|_{L_2(\mu)}$ which is the estimation error of the deep regression (3) in each iteration.

Proposition C.1. *(Error propagation) Let π_J be the greedy policy of $\widehat{Q}_J$ in Algorithm 1 and Q^{π_J} be the action-value function corresponding to π_J, then*

$$\|Q^* - Q^{\pi_J}\|_{L_1(\nu)} \leq \frac{2\zeta}{(1-\zeta)^2} \left(C_{\nu,\mu} \max_{1\leq j\leq J} \|\varepsilon_j\|_{L_2(\mu)} + 2\zeta^J R_{\max} \right),$$

where $\varepsilon_j = \widehat{Q}_j - \mathcal{T}^*\widehat{Q}_{j-1}$, $j = 1, \ldots, J$.

By Proposition C.1, it suffices to bound $\|\widehat{Q}_j - \mathcal{T}^*\widehat{Q}_{j-1}\|_{L_2(\mu)}$. To this end, we first decompose the excess risk $\mathcal{L}(\widehat{Q}_j) - \mathcal{L}(\mathcal{T}^*\widehat{Q}_{j-1})$ into approximation and statistical errors in Lemma C.1, and then impose the bound on each error using tools in empirical process with dependent data and deep approximation theory, respectively.

Lemma C.1. *Provided with a random sample $\{Z_i\}_{i=1}^n$, the excess risk satisfies*

$$\mathcal{L}(\widehat{Q}_j) - \mathcal{L}(\mathcal{T}^*\widehat{Q}_{j-1}) \leq 2 \sup_{Q\in\mathcal{F}} |\mathcal{L}(Q) - \mathcal{L}(Q)| + \inf_{Q\in\mathcal{F}} \|Q - \mathcal{T}^*\widehat{Q}_{j-1}\|_{L^2(\mu)}^2.$$

C.1 Statistical Error

The term $\sup_{Q\in\mathcal{F}} \left|\mathcal{L}(Q) - \widehat{\mathcal{L}}(Q)\right|$ is the statistical error of the ReLU-activated FNNs $\mathcal{F}$ with dependent data $\{Z_i\}_{i=1}^n$. We first introduce the definition of α-mixing for describing the dependence of a general stochastic process $\{U_t\}_{t\geq 1}$.

Definition C.2. ***(α-mixing)*** *Let $\{U_t\}_{t\geq 1}$ be a stochastic process. Denote by $U^{1:n}$ the collection $(U_1, \ldots, U_n)$, where we can allow $n = \infty$. Let $\sigma\left(U^{i:j}\right)$ denote the σ-algebra generated by $U^{i:j} (i \leq j)$. The m-th α-mixing coefficient of $\{U_t\}_{t\geq 1}$, α_m, is defined by*

$$\alpha_m = \sup_{t\geq 1} \sup_{A\in\sigma(U^{1:t}), B\in\sigma(U^{t+m:\infty})} |\mathbb{P}(AB) - \mathbb{P}(A)\mathbb{P}(B)|.$$

$\{U_t\}_{t\geq 1}$ is said to be α-mixing if $\alpha_m \to 0$ as $m \to \infty$. In particular, we say that a α-mixing process is exponential if there exists parameters $\bar{\alpha}, a, \eta > 0$ such that $\alpha_m \leq \bar{\alpha}\exp(-am^\eta)$ holds for all $m \geq 0$.

Following Hang & Steinwart (2014), we derive the tail probability bound of the empirical process with α-mixing data indexed by functions in $\mathcal{F}$ in term of the covering number of $\mathcal{F}$. Then, we use VC dimension to bound the covering number, which can be further bounded by the width and depth of the ReLU-activated neural networks (Bartlett et al., 2019). Finally, we obtain the bound on the statistical error $\sup_{Q\in\mathcal{F}} \left|\mathcal{L}(Q) - \widehat{\mathcal{L}}(Q)\right|$ in the following theorem.

Theorem C.1. *Suppose that $\{Z_i\}_{i=1}^n$ is strictly exponentially α-mixing, then*

$$\mathbb{E} \sup_{Q\in\mathcal{F}} \left|\mathcal{L}(Q) - \widehat{\mathcal{L}}(Q)\right| \leq C_1 \cdot \left[\left(\frac{\mathcal{D}^2\mathcal{W}^2 \log(\mathcal{W}\mathcal{D}) \log(n)}{n^{\frac{\eta}{1+\eta}}}\right)^{1/2}\right],$$

where C_1 is a constant depending on $\mathcal{B}, \eta, a, \bar{\alpha}, R_{\max}$.

Remark C.1. *To our best knowledge, Theorem C.1 is the first generalization bound on ReLU-activated FNNs with α-mixing data, which is a nontrivial extension on generalization analysis for deep networks with i.i.d data (Bauer & Kohler, 2019; Schmidt-Hieber, 2020; Nakada & Imaizumi, 2020; Jiao et al., 2021). Theorem C.1 implies that the statistical error bound is determined by n, depth $\mathcal{D}$ and width $\mathcal{W}$. It converges to 0 when n tends to ∞ for the fixed depth $\mathcal{D}$ and width $\mathcal{W}$.*

C.2 Approximation Error

The term $\inf_{Q\in\mathcal{F}} \|Q - \mathcal{T}^*\widehat{Q}_{j-1}\|_{L^2(\mu)}^2$ can be bounded by the approximation error of the ReLU-activated FNNs $\mathcal{F}$ to Hölder continuous class because $\mathcal{T}^*\widehat{Q}_{j-1}$ is contained in Hölder class by the smoothing property of $\mathcal{T}^*$. To this end, we assume that the distribution of the state-action pair (X, A) is supported on $[0, 1]^d$ without loss of generality. We also need the representation condition that the target Q^* is contained in Hölder class $\mathcal{H}^\gamma$ defined as follows.

Definition C.3. ***(Hölder class)*** *For $\gamma > 0$ with $\gamma = s + r$, where $s \in \mathbb{N}_0$ and $r \in (0, 1]$ and $d \in \mathbb{N}$, we denote Hölder class $\mathcal{H}^\gamma$ as*

$$\mathcal{H}^\gamma = \left\{ f : [0,1]^d \to \mathbb{R}, \max_{\|\widetilde{\alpha}\|_1 \leq s} \left\|\partial^{\widetilde{\alpha}} f\right\|_\infty \leq B, \max_{\|\widetilde{\alpha}\|_1 = s} \sup_{x\neq y} \frac{\left|\partial^{\widetilde{\alpha}} f(x) - \partial^{\widetilde{\alpha}} f(y)\right|}{\|x - y\|_\infty^r} \leq B \right\}.$$

We apply the approximation result of Jiao et al. (2021) giving the approximation error bound for Hölder functions using ReLU-activated FNNs, shown in the following Theorem C.2. Note that the prefactor $(d^{\lfloor\gamma\rfloor + (\gamma\vee 1)/2})$ depends on the ambient dimension d polynomially, which improves that in Lu et al. (2021) from exponentially to polynomially.

Theorem C.2. *(Theorem 3.3 of Jiao et al. (2021)) Assume that $f \in \mathcal{H}^\gamma$ with $\gamma = s + r, s \in \mathbb{N}_0$ and $r \in (0, 1]$. For any $W, L \in \mathbb{N}$, there exists a function $\tilde{f}$ belonging to the ReLU-activated FNNs $\mathcal{F}$ with width $\mathcal{W} = 38(\lfloor\gamma\rfloor + 1)^2 d^{\lfloor\gamma\rfloor+1} W \lceil\log_2(8W)\rceil$ and depth $\mathcal{D} = 21(\lfloor\gamma\rfloor + 1)^2 L \lceil\log_2(8L)\rceil$ such that*

$$|f(x) - \tilde{f}(x)| \leq 18B(\lfloor\gamma\rfloor + 1)^2 d^{\lfloor\gamma\rfloor+(\gamma\vee 1)/2}(WL)^{-2\gamma/d},$$

for all $x \in [0, 1]^d \backslash \Omega([0, 1]^d, S, \delta)$, where $\Omega([0, 1]^d, S, \delta) = \cup_{i=1}^d \{x = [x_1, x_2, \ldots, x_d]^\top : x_i \in \cup_{k=1}^{S-1}(k/S - \delta, k/S)\}$, with $S = \lceil (WL)^{2/d}\rceil$ and $\delta \in (0, 1/(3S)]$.

C.3 Bounding the excess risk $\mathcal{L}(\widehat{Q}_j) - \mathcal{L}(\mathcal{T}^*\widehat{Q}_{j-1})$

Based on the statistical and approximation error analysis in Theorems C.1-C.2, respectively, we can establish the non-asymptotic error bound for the excess risk $\mathcal{L}(\widehat{Q}_j) - \mathcal{L}(\mathcal{T}^*\widehat{Q}_{j-1})$ $\left(\|\widehat{Q}_j - \mathcal{T}^*\widehat{Q}_{j-1}\|^2_{L_2(\mu)}\right)$ by choosing appropriate width $\mathcal{W}$ and depth $\mathcal{D}$, shown in Theorem C.3.

Theorem C.3. *Suppose that $\{\mathcal{T}^*\widehat{Q}_{j-1}\}_{j=1}^J \in \mathcal{H}^\gamma$ in Definition C.3 with $\gamma = s + r, s \in \mathbb{N}_0$ and $r \in (0, 1]$, $\{Z_i\}_{i=1}^n$ is strictly exponentially α-mixing, and the probability distribution μ of (X, A) is absolutely continuous with respect to Lebesgue measure. Then, for the ReLU-activated FNNs $\mathcal{F}$ with the width $\mathcal{W} = \mathcal{O}\left((n^{\frac{\eta}{1+\eta}})^{\frac{d}{4(d+4\gamma)}} \log n\right)$ and depth $\mathcal{D} = \mathcal{O}\left((n^{\frac{\eta}{1+\eta}})^{\frac{d}{4(d+4\gamma)}} \log n\right)$, the excess risk satisfies*

$$\mathbb{E}\left[\|\widehat{Q}_j - \mathcal{T}^*\widehat{Q}_{j-1}\|^2_{L_2(\mu)}\right] \leq C\left[d^{2s+(\gamma\vee 1)}(n^{\frac{\eta}{1+\eta}})^{\frac{-2\gamma}{d+4\gamma}}(\log n)^3\right],\ j = 1, \ldots, J,$$

where C is a constant depending on $s, B, \mathcal{B}, R_{\max}, \eta, a, \bar{\alpha}$.

Proof. By Theorem C.2, for any $f^* \in \mathcal{H}^\gamma$, there exists one function $\tilde{f} \in \mathcal{F}$ with width $\mathcal{W} = 38(s + 1)^2 d^{s+1} W \lceil\log_2 8W\rceil$ and depth $\mathcal{D} = 21(s + 1)^2 L \lceil\log_2 8L\rceil$ such that

$$\left|f^*(x) - \tilde{f}(x)\right| \leq 18B(s + 1)^2 d^{s+(\gamma\vee 1)/2}(WL)^{-2\gamma/d}$$

for $x \in \cup_\theta \widetilde{Q}_\theta$, with

$$\widetilde{Q}_\theta = \left\{x : x_i \in \left[\frac{\theta_i}{S}, \frac{\theta_i + 1}{S} - \delta \cdot 1_{\{\theta_i < S-1\}}\right]\right\}$$

where $\theta = (\theta_1, \theta_2, \ldots, \theta_n) \in \{0, 1, \ldots, S - 1\}^d$, and δ being an arbitrary number in satisfying $0 < \delta \leq \frac{1}{3S}$. Then the Lebesgue measure of $[0, 1]\backslash\widetilde{Q}_\theta$ is no more than $dS\delta$ which can be arbitrarily small if δ is arbitrarily small. Since μ is absolutely continuous with respect to the Lebesgue measure and $\mathcal{T}^*\widehat{Q}_{j-1} \in \mathcal{H}^\gamma$, we have

$$\inf_{Q\in\mathcal{F}} \|Q - \mathcal{T}^*\widehat{Q}_{j-1}\|^2_{L^2(\mu)} \leq 324B^2(s + 1)^4 d^{2s+(\gamma\vee 1)}\left\lfloor(WL)^{2/d}\right\rfloor^{-2\gamma}.$$

By Lemma C.1 and Theorem C.1, it yields that

$$\begin{aligned}\mathbb{E}[\|\widehat{Q}_j - \mathcal{T}^*\widehat{Q}_{j-1}\|^2_{L_2(\mu)}] \leq C_1 \cdot &\left[\left(\frac{\mathcal{W}^2\mathcal{D}^2\log(\mathcal{W}\mathcal{D})\log(n)}{n^{\frac{\eta}{1+\eta}}}\right)^{1/2}\right] \\ &+ 324B^2(s + 1)^4 d^{2s+(\gamma\vee 1)}\left\lfloor(WL)^{2/d}\right\rfloor^{-2\gamma}.\end{aligned}$$

Moreover, set $\mathcal{W} = \mathcal{O}\left((n^{\frac{\eta}{1+\eta}})^{\frac{d}{4(d+4\gamma)}} \log n\right)$, depth $\mathcal{D} = \mathcal{O}\left((n^{\frac{\eta}{1+\eta}})^{\frac{d}{4(d+4\gamma)}} \log n\right)$, then it follows that

$$\mathbb{E}[\|\widehat{Q}_j - \mathcal{T}^*\widehat{Q}_{j-1}\|^2_{L_2(\mu)}] \leq C_2 \cdot \left[d^{2s+(\gamma\vee 1)}(n^{\frac{\eta}{1+\eta}})^{\frac{-2\gamma}{d+4\gamma}}(\log n)^3\right],$$

where C_2 is a constant depending on $s, B, \mathcal{B}, R_{\max}, \eta, a, \bar{\alpha}$.

□

C.4 PROOF OF THEOREM 2.1

Proof. By Proposition C.1, Lemma C.1 and Theorem C.3, we can conclude that

$$\mathbb{E}\left[\|Q^* - Q^{\pi_J}\|_{L_1(\nu)}\right] \leq \frac{2\sqrt{C}C_{\nu,\mu}\zeta}{(1-\zeta)^2} \cdot \left[d^{s+(\gamma\vee 1)/2}(n^{\frac{\eta}{1+\eta}})^{\frac{-\gamma}{d+4\gamma}}(\log n)^{3/2}\right] + \frac{4\zeta^{J+1}}{(1-\zeta)^2}R_{\max},$$

where C is a constant depending on $s, B, \mathcal{B}, R_{\max}, \eta, a, \bar{\alpha}$. This completes the proof. □