Approximate Policy Iteration With Deep Minimax Average Bellman Error Minimization

Lican Kang, Yuhui Liu, Yuan Luo, Jerry Zhijian Yang, Han Yuan, and Chang Zhu

***Abstract*— In this work, we investigate the utilization of deep approximate policy iteration (DAPI) in estimating the optimal action-value function Q^* within the context of reinforcement learning, employing rectified linear unit (ReLU) ResNet as the underlying framework. The iterative process of DAPI incorporates the minimax average Bellman error minimization principle. It employs ReLU ResNet to estimate the fixed point of the Bellman equation, which is aligned with the estimated greedy policy. Through error propagation, we derive nonasymptotic error bounds between Q^* and the estimated Q function induced by the output greedy policy in DAPI. To effectively control the Bellman residual error, we address both the statistical and approximation errors associated with the α-mixing dependent data derived from Markov decision processes, using the techniques of empirical process and deep approximation theory, respectively. Furthermore, we present a novel generalization bound for ReLU ResNet in the presence of dependent data, as well as an approximation bound for ReLU ResNet within the Hölder class. Notably, this approximation bound contributes to a significant improvement in the dependence on the ambient dimension, transitioning from an exponential relationship to a polynomial one. The derived nonasymptotic error bounds explicitly depend on factors such as the sample size, the ambient dimension (in polynomial terms), and the width and depth of the neural networks. Consequently, these bounds serve as valuable theoretical guidelines for appropriately setting the hyperparameters, thereby enabling the achievement of the desired convergence rate during the training process of DAPI.**

***Index Terms*— α-mixing, deep approximate policy iteration (DAPI), deep neural networks, minimax loss, nonasymptotic error bound, reinforcement learning (RL).**

NOMENCLATURE

$\lceil a \rceil$	Smallest integer no less than a, $a \in \mathbb{R}$.
$\lfloor a \rfloor$	Largest integer less than a, $a \in \mathbb{R}$.
$a \vee b$	$\max\{a, b\}$, $a, b \in \mathbb{R}$.
$a \lesssim b$	$a \leq Cb$ for some constant $C > 0$, $a, b \in \mathbb{R}$.
$a \asymp b$	$a \lesssim b \lesssim a$, $a, b \in \mathbb{R}$.
$\mathbb{N}_0$	Nonnegative integers.
$\mathbb{N}$	Strictly positive integers.
$\|x\|_q$	ℓ_q-norm of the vector $x = (x_1, \ldots, x_p)^\top \in \mathbb{R}^p$: $\|x\|_q = (\sum_{i=1}^p \|x_i\|^q)^{\frac{1}{q}}$, $q \in [1, \infty]$.
$\|f\|^q_{L^q(\mu)}$	ℓ_q-norm of the measurable function $f : \mathbb{R}^p \to \mathbb{R}$: $\|f\|^q_{L^q(\mu)} = \mathbb{E}_{x\sim\mu}\|f(x)\|^q$ with the probability measure μ on $\mathbb{R}^p$.

Manuscript received 26 April 2022; revised 11 December 2022, 23 May 2023, and 22 October 2023; accepted 21 December 2023. This work was supported by the National Science Foundation of China under Grant 12125103 and Grant 12071362. *(Corresponding author: Chang Zhu.)*

Lican Kang is with the Cardiovascular and Metabolic Disorders Program, Duke-NUS Medical School, Singapore 169857 (e-mail: kanglican@duke-nus.edu.sg).

Yuhui Liu and Yuan Luo are with the School of Mathematics and Statistics, Wuhan University, Wuhan 430072, China (e-mail: liu_yuhui@whu.edu.cn; yuanluo@whu.edu.cn).

Jerry Zhijian Yang is with the School of Mathematics and Statistics and the Hubei Key Laboratory of Computational Science, Wuhan University, Wuhan 430072, China (e-mail: zjyang.math@whu.edu.cn).

Han Yuan is with the Centre for Quantitative Medicine, Duke-NUS Medical School, Singapore 169857 (e-mail: yuan.han@u.duke.nus.edu).

Chang Zhu is with the Department of Anesthesiology, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan 430030, China (e-mail: changzhu@hust.edu.cn).

Digital Object Identifier 10.1109/TNNLS.2023.3346992

I. INTRODUCTION

REINFORCEMENT learning (RL) [1], [2] is a prominent field in machine learning that addresses the challenge of sequential decision-making. RL can be formalized as an agent interacting with an environment to maximize its expected cumulative rewards by selecting appropriate actions based on its current state. The mathematical framework for RL is often based on the concept of a Markov decision process (MDP). In an MDP, the agent interacts with the environment by taking actions, transitioning between different states, and receiving rewards. The dynamics of the environment are modeled as a Markov process, where the future state and reward depend only on the current state and action, disregarding the past history. This Markovian assumption allows for the application of well-established mathematical tools and algorithms in RL. In recent years, deep reinforcement learning has made significant progress by utilizing deep neural networks to approximate value functions [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17]. This approach has demonstrated remarkable achievements across various domains, including video games [18], [19], AlphaGo [20], natural language processing [21], [22], [23], [24], and robotics [25], [26], [27], [28].

Approximate policy iteration (API) is a widely studied method for estimating value functions in RL [29], [30], [31], [32]. Among the existing API techniques, the least-square policy iteration (LSPI) [29] and the policy iteration with Bellman residual minimization (BRM) [30] have received considerable attention. LSPI employs least-squares regression to estimate value functions, while BRM introduces an unbiased minimax loss function to minimize the Bellman residual. Several extensions and variations of LSPI and BRM have also been proposed, such as regularized policy iteration [33], [34] and kernel-based LSPI [35], [36], [37]. For

more in-depth information, readers can refer to [31] and the references therein. Notably, certain modified versions of API have been introduced to enhance the estimation of value functions. Examples include conservative API [38], which incorporates conservatism into the policy iteration process, and gradient-based API with stochastic policies [39], [40], [41], [42]. However, it is worth noting that the existing API approaches primarily focus on linear functions or functions residing in reproducing kernel Hilbert space for value function approximation. Consequently, there is a lack of theoretical studies exploring policy iteration with deep neural networks, which offer more expressive function approximators. In the realm of error analysis, the existing studies have predominantly focused on statistical errors while assuming that approximation errors are negligible. Additionally, many analyses assume that the MDP sequence is independently and identically distributed (i.i.d.) [29]. However, this assumption can be restrictive and impractical in real-world scenarios. To address this, Antos et al. [30] introduced the concept of β-mixing [43] as a relaxation of the i.i.d. assumption. It is important to note that the α-mixing assumption [44], [45] employed in this article is less restrictive than β-mixing, allowing for more realistic modeling of dependencies in the MDP sequence.

In this work, we explore deep approximate policy iteration (DAPI), which employs deep neural networks to estimate the optimal action-value function Q^* by minimizing the minimax average Bellman error. We conduct a theoretical analysis considering temporal dependencies in MDPs with the α-mixing condition [44], [45], which is a weaker assumption compared to the commonly used β-mixing condition. We provide nonasymptotic error bounds that encompass both the statistical and approximation errors, enabling the determination of suitable hyperparameter settings such as depth, width, and the number of iterations for achieving desired convergence rates based on the training sample size and the ambient dimension. Additionally, we derive a novel approximation error bound for the Hölder class using the rectified linear unit (ReLU) activation [46] in ResNet [47], as well as an error bound for nonparametric fittings using ReLU ResNet on dependent data. These findings hold potential for independent exploration and investigation.

Technically, our error analysis of DAPI proceeds as follows: we initially bound the Bellman residual error of the minimax estimator at each iteration. Subsequently, we derive nonasymptotic error bounds between Q^* and the Q function induced by the greedy policy through the process of error propagation [30], [31], [48], [49], [50]. The main assumptions employed in the theoretical analysis are the mild distribution shift condition and the realizability-type condition, the necessity of which has been discussed recently [51], [52], [53]. To bound the fitting error at each iteration, we need to determine the approximation error of ReLU ResNet on the Hölder class and the statistical error (generalization error) of ReLU ResNet with dependent data.

Recent studies have extensively investigated the approximation power of deep neural networks [54], [55], [56], [57], [58], [59], [60], [61], [62], [63], along with numerous references therein. In alignment with [60], we offer a novel approximation error bound for the Hölder class $\mathcal{H}^\zeta$ using ResNet with ReLU activation, presenting independent interest and novelty. Regarding generalization error, concentration inequalities and learning theory for dependent data have been extensively studied [30], [43], [44], [45], [64], [65], [66], [67], [68], [69] along with numerous references. A substantial body of the literature also exists on generalization analysis for deep neural networks with i.i.d. data [60], [70], [71], [72], [73], [74], [75]. However, to the best of authors' knowledge, we present the first generalization error bound for ReLU ResNet with α-mixing dependent data [44], [45] based on empirical process with dependent data [76] and the pseudodimension of ReLU network [77], which constitutes an original and independent contribution.

The rest of this article is organized as follows. In Section II, we give a brief introduction to the background of RL and deep neural networks. In Section III, we present DAPI in detail. In Section IV, we provide a detailed description of the error analysis. In Section V, we establish the nonasymptotic error bounds for DAPI. We summarize in Section VI. Proofs for all the lemmas and theorems are provided in the Appendix.

II. Background

A. Markov Decision Process

A discounted MDP is defined by a quintuple $(\mathcal{X}, \mathcal{A}, P, \mathcal{R}, \gamma)$, 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 the immediate reward $R(x, a)$, and $\gamma \in [0, 1)$ is the discount factor. Here, d denotes the dimension of state-action pairs (X, A), and $\mathcal{M}(\mathcal{X})$ denotes the set of probability measures on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$, where $\mathcal{B}(\mathcal{X})$ is the Borel σ-algebra on $\mathcal{X}$. Specifically, for each pair $(x, a) \in \mathcal{X} \times \mathcal{A}$, $P(\cdot|x, a)$ is a probability measure on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ that defines the next-state distribution upon taking action a in state x, and $P(D|\cdot, \cdot)$ is a measurable function on $\mathcal{X} \times \mathcal{A}$ for every $D \in \mathcal{B}(\mathcal{X})$. Moreover, let $\pi(\cdot|x)$ denote the stochastic policy which is an associated distribution of the action at state x. Given an 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

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

We assume that the MDP $\{Z_i\}_{i=1}^n$ is strictly stationary α-mixing (see Definition 2), which indicates that Z_is share the same distribution. Let μ be the distribution of (X_i, A_i) for each $i \in \{1, \ldots, n\}$. Then, $\mu = \nu \circ \pi$ is the stationary distribution of this Markov chain $\{X_i, A_i\}_{i=1}^n$, where $\mu = \nu \circ \pi$ is defined by

$$\mu(E) = \int_E \pi(da|x)\nu(dx), \quad E \in \mathcal{B}(\mathcal{X}) \times \mathcal{B}(\mathcal{A}).$$

Denote the action-value function as

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

For a given policy π, Q^π is the unique fixed point of the Bellman operator $\mathcal{T}^\pi$, that is,

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

with

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

Without loss of generality, suppose that $R(x,a) \in [0, R_{\max}]$ for each pair $(x,a) \in \mathcal{X}\times\mathcal{A}$, and thus, Q^π takes values in $[0,(R_{\max}/1-\gamma)]$. Assume that there exists a policy π^* that maximizes Q^π, such that $Q^* := Q^{\pi^*}$. Q^* satisfies the optimal Bellman equation $Q^* = \mathcal{T}^*Q^*$, where the optimal Bellman operator $\mathcal{T}^*$ is given by

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

It can be observed that $\mathcal{T}^*$ is γ-contraction in the sup-norm. We can define the greedy policy for an action-value function Q as

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

These definitions and properties provide the foundation for the study of MDPs and their associated optimal policies and action-value functions.

B. ResNet With ReLU Activation

We now introduce the feedforward neural networks and ResNet [47] with ReLU activation, respectively.

We use $\mathcal{F}$ to denote the class of feedforward neural networks $f_\theta : \mathbb{R}^d \to \mathbb{R}$ with parameter θ, depth $\mathcal{D}$, and width $\mathcal{W}$. Each network f_θ is defined as

$$f_\theta(x) = v_{\mathcal{D}}\circ\rho\circ v_{\mathcal{D}-1}\circ\rho\circ\cdots\circ\rho\circ v_1\circ\rho\circ v_0(x),\ \ x\in\mathbb{R}^d$$

where $\|f_\theta\|_\infty \le \mathcal{B}$ holds for some $0 < \mathcal{B} < \infty$. Here, $\|\cdot\|_\infty$ refers to the sup-norm, $\rho(x) = \max(0,x)$ is the ReLU activation function operating pointwisely on x, and

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

with the weight matrix $\widetilde{A}_i \in \mathbb{R}^{d_{i+1}\times d_i}$ and the bias vector $b_i \in \mathbb{R}^{d_{i+1}}$, where d_i is the width of the ith layer. The input data correspond to the first layer, and the output corresponds to the last layer. The feedforward neural network f_θ consists of $\mathcal{D}$ hidden layers, resulting in a total of $(\mathcal{D}+1)$ layers. We represent the width of each layer using a $(\mathcal{D}+1)$-vector $(d_0,d_1,\ldots,d_{\mathcal{D}})^\top$, where $d_0 = d$ is the dimension of the input (X,A), and $d_{\mathcal{D}+1} = 1$ is the dimension of the output. The width $\mathcal{W}$ is defined as the maximum width among the hidden layers, specifically, $\mathcal{W} = \max\{d_1,\ldots,d_{\mathcal{D}}\}$.

Let $m\in\mathbb{N}$ be a positive integer. A residual block $\mathbf{R}:\mathbb{R}^m\to\mathbb{R}^m$ with depth N and width $M = \max\{\widetilde{N}_\kappa : \kappa = 0,1,\ldots,N\}$ is defined as

$$\begin{aligned}\mathbf{R}^{(0)}(x) &= x\\ \mathbf{R}^{(\kappa)}(x) &= \rho\big(\widetilde{A}_\kappa\mathbf{R}^{(\kappa-1)}(x) + b_\kappa\big), \kappa = 1,\ldots,N\\ \mathbf{R}(x) &= \mathbf{R}^{(N)}(x) + x\end{aligned} \tag{1}$$

where $\widetilde{A}_\kappa \in \mathbb{R}^{\widetilde{N}_\kappa\times\widetilde{N}_{\kappa-1}}$, $b_\kappa\in\mathbb{R}^{\widetilde{N}_\kappa}$, $\widetilde{N}_0 = \widetilde{N}_N = m$, $\widetilde{N}_1,\ldots,\widetilde{N}_{N-1}\in\mathbb{N}$, and ρ is the ReLU activation function. A function $f_\phi:\mathbb{R}^d\to\mathbb{R}$ implemented by a ReLU ResNet with $K-1$ residual block is defined by

$$\begin{aligned}f_{(0)}(x) &= x\\ f_{(\kappa)}(x) &= \rho\big(\widetilde{A}_\kappa\mathbf{R}_\kappa(f_{\kappa-1}(x)) + b_\kappa\big), \kappa = 1,\ldots,K-1\\ f_\phi(x) &:= f_{(K)}(x) = \widetilde{A}_K f_{(K-1)}(x) + b_K\end{aligned} \tag{2}$$

where $\mathbf{R}_\kappa:\mathbb{R}^{d_\kappa}\to\mathbb{R}^{d_\kappa}, \kappa = 1,\ldots,K-1$, is the residual block with depth $\widetilde{N}_\kappa$ and width M_κ, with $d_1 = d$, $M_1 = d$, $\widetilde{A}_\kappa\in\mathbb{R}^{d_{\kappa+1}\times d_\kappa}$, and $\widetilde{A}_{K+1}\in\mathbb{R}^{1\times d_{K+1}}$. Without loss of generality, we assume that the depth of each residual block is small and equal such that $\widetilde{N}_\kappa = N = \mathcal{O}(1)$ for $\kappa = 1,\ldots,K-1$. We use $\mathcal{F}^K$ to denote the class of ReLU ResNet $f_\phi:\mathbb{R}^d\to\mathbb{R}$ with parameter ϕ, depth $\mathcal{D} = (N+1)(K-1)+1$, and width $\mathcal{W} = \max\{M_\kappa : \kappa = 1,\ldots,K-1\}$, where $\|f_\phi\|_\infty\le\mathcal{B}$.

III. Deep Approximate Policy Iteration

In DAPI, it follows that:

$$Q_0 \xrightarrow{\text{I}} \pi_1 \xrightarrow{\text{E}} Q_1 \xrightarrow{\text{I}} \pi_2 \xrightarrow{\text{E}} \cdots \xrightarrow{\text{E}} Q_{J-1} \xrightarrow{\text{I}} \pi_J. \tag{3}$$

Here, π_j is the greedy policy with respect to Q_{j-1}, Q_j is the approximation action-value function for the policy π_j, and J is a positive integer large enough. The symbol $\xrightarrow{\text{E}}$ denotes a policy evaluation, and $\xrightarrow{\text{I}}$ denotes a policy improvement. Given the policy π_j in (3), to obtain the estimation Q_j of the action-value function Q^{π_j}, we introduce the minmax loss [51] related to the average Bellman errors. It is formulated as follows:

$$\min_{Q\in\mathcal{U}_1}\max_{O\in\mathcal{U}_2}|\mathcal{L}(Q,O)|.$$

Here, $\mathcal{U}_1$ and $\mathcal{U}_2$ are some measurable function classes, and

$$\mathcal{L}(Q,O) := \mathbb{E}_{(X,A)\sim\mu}\big[O(X,A)(Q(X,A) - \mathcal{T}^{\pi_j}Q(X,A))\big]$$

where $Q\in\mathcal{U}_1, O\in\mathcal{U}_2$. If $Q^{\pi_j}\in\mathcal{U}_1$ and $\mathcal{U}_1 = \mathcal{U}_2$ are supposed, it yields that

$$Q^{\pi_j}\in\arg\min_{Q\in\mathcal{U}_1}\max_{O\in\mathcal{U}_2}|\mathcal{L}(Q,O)|.$$

In this article, we assume that both $\mathcal{U}_1$ and $\mathcal{U}_2$ are Hölder classes as defined in Definition 4, and that Q^{π_j} is contained within a Hölder class. Meanwhile, at sample level, we denote the empirical minimax estimator as

$$Q_j\in\arg\min_{Q\in\mathcal{G}_1}\max_{O\in\mathcal{G}_2}\big|\widehat{\mathcal{L}}(Q,O)\big| \tag{4}$$

where

$$\widehat{\mathcal{L}}(Q,O) := \frac{1}{n}\sum_{i=1}^{n}O(X_i,A_i)(Q(X_i,A_i) - Y_i)$$

is the empirical minimax loss function with

$$Y_i := R_i + \gamma\int Q(X_{i+1},a)\pi_j(da|X_{i+1})$$

and $Q \in \mathcal{G}_1$, $O \in \mathcal{G}_2$, and $\mathcal{G}_1, \mathcal{G}_2$ are the ReLU ResNet class $\mathcal{F}^K$ for approximating the Hölder class. Moreover, this empirical loss $\widehat{\mathcal{L}}$ is unbiased, that is, $\mathbb{E}_{\{Z_i\}_{i=1}^n}\widehat{\mathcal{L}}(Q, O) = \mathcal{L}(Q, O)$. This property is demonstrated in Lemma 1. For notational simplicity, we denote by

$$\mathcal{L}_{\mathcal{U}_2}(Q) := \sup_{O\in\mathcal{U}_2} |\mathcal{L}(Q, O)|$$
$$\mathcal{L}_{\mathcal{G}_2}(Q) := \sup_{O\in\mathcal{G}_2} |\mathcal{L}(Q, O)|$$
$$\widehat{\mathcal{L}}_{\mathcal{G}_2}(Q) := \sup_{O\in\mathcal{G}_2} |\widehat{\mathcal{L}}(Q, O)|.$$

Then, it can be deduced that

$$Q_j \in \operatorname*{argmin}_{Q\in\mathcal{G}_1} \widehat{\mathcal{L}}_{\mathcal{G}_2}(Q) = \operatorname*{argmin}_{Q\in\mathcal{G}_1} \sup_{O\in\mathcal{G}_2} \left|\widehat{\mathcal{L}}(Q, O)\right|.$$

Lemma 1 (Unbiased Loss): $\widehat{\mathcal{L}}(Q, O)$ is an unbiased loss, that is, $\mathbb{E}_{\{Z_i\}_{i=1}^n}\widehat{\mathcal{L}}(Q, O) = \mathcal{L}(Q, O)$ for each pair $(Q, O) \in \mathcal{U}_1 \times \mathcal{U}_2$.

The detailed procedure of DAPI is summarized in Algorithm 1.

Algorithm 1 Deep Approximate Policy Iteration

1: Input: MDP $(\mathcal{X}, \mathcal{A}, R, P, \gamma)$, function class $\mathcal{F}^K$, initial value Q_0.
2: **for** $j = 1, \ldots, J-1$ **do**
3: Obtain greedy policy π_j.
4: Sample $(X_i, A_i, R_i, X_i'), i = 1, \ldots, n$.
5: Compute $Y_i := R_i + \gamma \int Q(X_{i+1}, a)\pi_j(da|X_{i+1})$, and obtain the j-step action-value function Q_j via (4), that is,

$$Q_j \in \arg\min_{Q\in\mathcal{F}^K} \max_{O\in\mathcal{F}^K} |\widehat{\mathcal{L}}(Q, O)|.$$

6: **end for**
7: Output: The greed policy $\pi_J := \pi(\cdot; Q_{J-1})$ with respect to Q_{J-1}.

In jth iteration of DAPI (Algorithm 1), $j = 1, \ldots, J-1$, it mainly solves the minimax optimization problem (4), which can be regarded as obtaining a minimax estimator of a nonparametric regression problem with a minimax loss. Therefore, to investigate the computational complexity of DAPI is equivalent to derive the total cost of solving $J-1$ times nonparametric regression problems with a minimax loss.

IV. Error Analysis

In this section, we present the theoretical analysis of DAPI. That is to bound $\|Q^* - Q^{\pi_J}\|_{L_2(\nu)}$, where ν is an admissible distribution and is allowed to differ from the sample distribution μ in Algorithm 1. To establish the nonasymptotic error bound for $\|Q^* - Q^{\pi_J}\|_{L_2(\nu)}$, we initially employ error propagation techniques to transform it into the control of the Bellman residual error, represented as $\|Q_j - \mathcal{T}^{\pi_j}Q_j\|_{L_2(\mu)}$, as demonstrated in Proposition 1. Subsequently, this Bellman residual error can be effectively controlled through the concept of excess risk, denoted as $\mathcal{L}_{\mathcal{U}_2}(Q_j) - \mathcal{L}_{\mathcal{U}_2}(Q^{\pi_j})$, leading to a further decomposition that enables us to control the excess risk in terms of bounding statistical and approximation errors, as expounded in Lemma 2. Finally, we bound the statistical and approximation errors by employing techniques from empirical process theory with dependent data [30], [43], [44], [45], [64], [65], [66], [67], [68], [69] and deep approximation theory [54], [55], [56], [57], [58], [59], [60], [61], [62], [63], respectively. These results are presented in Theorems 1 and 2. In this intricate endeavor, the pivotal role of concentration coefficients takes center stage within the error propagation. To this end, we first introduce the following definition of concentration coefficients, designed to control distributional shifts. This control is imperative, as a certain level of concentratability is deemed necessary for the theoretical development of batch mode RL, as acknowledged in [51], [52], [78], and [53].

Definition 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 mth concentration coefficient is defined as

$$c_{\nu_1,\nu_2}(m) = \sup_{\pi_1,\ldots,\pi_m} \left\| \frac{\mathbf{d}(\nu_1 P^{\pi_1} P^{\pi_2} \ldots P^{\pi_m})}{\mathbf{d}\nu_2} \right\|_\infty \tag{5}$$

where the supremum is taken over all possible policies. In (5), the notation $(\mathbf{d}\tilde{\mu}/\mathbf{d}\tilde{\nu})$ refers to the Radon–Nikodym derivative of $\tilde{\mu}$ with respect to $\tilde{\nu}$, where $\tilde{\mu}$ and $\tilde{\nu}$ are the two probability measures, see [79], [80], and [81] for more details. 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

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

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

In order to bound $\|Q^* - Q^{\pi_J}\|_{L_2(\nu)}$ for any admissible probability distribution ν, we can propagate this error, as analyzed in [30] and [31]. As a result, it can be controlled by the Bellman residual error $\|Q_j - \mathcal{T}^{\pi_j}Q_j\|_{L_2(\mu)}$, as shown in Proposition 1.

Proposition 1 (Error Propagation): Let J be a positive integer and $Q_{\max} \leq R_{\max}/(1-\gamma)$. Then, for any sequence of functions $\{Q_j\}_{j=0}^{J-1}$ with $|Q_j| \leq Q_{\max}$, we have

$$\|Q^* - Q^{\pi_J}\|_{L_2(\nu)} \leq \frac{2\gamma}{(1-\gamma)^2}\left(C_{\nu,\mu}^{1/2} \max_{0<j<J} \|\varepsilon_j\|_{L_2(\mu)} + \gamma^{J/2} R_{\max}\right)$$

where $\varepsilon_j := Q_j - \mathcal{T}^{\pi_j}Q_j$ denotes the Bellman residual of Q_j for $j = 1, \ldots, J-1$, and $Q_{\max}$ and $R_{\max}$ are two positive constants bounding the action-value function and the immediate reward, respectively.

Proposition 1 indicates that it suffices to bound $\|Q_j - \mathcal{T}^{\pi_j}Q_j\|_{L_2(\mu)}$. To accomplish this objective, we first decompose the excess risk $\mathcal{L}_{\mathcal{U}_2}(Q_j) - \mathcal{L}_{\mathcal{U}_2}(Q^{\pi_j})$ into three terms, including one statistical error and two

approximation errors, as shown in Lemma 2. It is worth noting that we can straightforwardly demonstrate that

$$\mathbb{E}\|Q_j - \mathcal{T}^{\pi_j} Q_j\|^2_{L^2(\mu)} \leq \mathbb{E}\big(\mathcal{L}_{\mathcal{U}_2}(Q_j) - \mathcal{L}_{\mathcal{U}_2}\big).$$

Please refer to Appendix J for comprehensive proofs and detailed elaboration of this assertion. We can then establish bounds for each of these errors using tools from empirical process theory with dependent data and deep approximation theory.

Lemma 2 : Provided with a random sample $\{Z_i\}_{i=1}^n$, the excess risk satisfies

$$\begin{aligned}\mathcal{L}_{\mathcal{U}_2}(Q_j) - \mathcal{L}_{\mathcal{U}_2}(Q^{\pi_j}) &\leq \underbrace{2\sup_{\phi\in\mathcal{G}_1}|\widehat{\mathcal{L}}_{\mathcal{G}_2}(\phi) - \mathcal{L}_{\mathcal{G}_2}(\phi)|}_{\text{statistical error: } \varepsilon_{\text{sta}}} \\ &+ \underbrace{\inf_{\phi\in\mathcal{G}_1}\mathcal{L}_{\mathcal{U}_2}(\phi) - \mathcal{L}_{\mathcal{U}_2}(Q^{\pi_j})}_{\text{app error of } \mathcal{G}_1: \varepsilon_{\mathcal{G}_1}} \\ &+ \underbrace{2\sup_{\phi\in\mathcal{G}_1}|\mathcal{L}_{\mathcal{U}_2}(\phi) - \mathcal{L}_{\mathcal{G}_2}(\phi)|}_{\text{app error of } \mathcal{G}_2\text{: } \varepsilon_{\mathcal{G}_2}}.\end{aligned}$$

A. Statistical Error

The term $\mathcal{E}_{\text{sta}}$ represents the statistical error with dependent data $\{Z_i\}_{i=1}^n$. To analyze this error, we first recall the definition α-mixing [44], [45] for measuring the dependence of a general stochastic process $\{W_t\}_{t\geq 1}$.

Definition 2 (α-Mixing): Let $\{W_t\}_{t\geq 1}$ be a stochastic process. Denote by $W^{1:n}$ the collection $(W_1, \ldots, W_n)$, where we can allow $n = \infty$. Let $\sigma(W^{i:j})$ denote the σ-algebra generated by $W^{i:j} (i \leq j)$. The mth α-mixing coefficient of $\{W_t\}_{t\geq 1}$, denoted as α_m, is defined as follows:

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

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

Next, we introduce the concept of covering number [70]. The covering number holds significant importance in our statistical error analysis, serving as a measure of the complexity of a function class. As investigated in [30], [43], [44], [45], [64], [65], [66], [67], [68], and [69], through the utilization of empirical process techniques with dependent data, we can express the upper bound of the statistical error $\mathcal{E}_{\text{sta}}$ in terms of the complexity of the function class $\mathcal{G}_1$, exemplified by the covering number.

Definition 3 (Covering Number): The covering number $\mathcal{N}(\varepsilon, \mathbf{F}, d)$ related to a semimetric d on the set $\mathbf{F}$ is defined as

$$\mathcal{N}(\varepsilon, \mathbf{F}, d) = \min_{\kappa}\Big\{\text{there are } g_1, \ldots, g_\kappa \text{ such that } \min_{1\leq j\leq\kappa} d(f, g_j) \leq \varepsilon \text{ for any } f \text{ in } \mathbf{F}\Big\}.$$

Following [44] and [45], we derive the tail probability bound of the empirical process with α-mixing data indexed by functions in $\mathcal{F}^K$ in terms of the covering number of $\mathcal{F}^K$. Then, we use Vapnik–Chervonenkis (VC) dimension [70] to bound the covering number, which can be further bounded by the width and depth of the ReLU ResNet [77]; details are shown in the Appendix. Finally, we obtain the upper bound for the statistical error $\mathcal{E}_{\text{sta}}$ in the following theorem.

Theorem 1 : Suppose that $\mathcal{G}_1$ and $\mathcal{G}_2$ are ReLU ResNet classes $\mathcal{F}^K$, and $\{Z_i\}_{i=1}^n$ is strictly and exponentially α-mixing with parameters $\bar{\alpha}, a, \eta > 0$ as defined in Definition 2. Then,

$$\mathbb{E}\sup_{\phi\in\mathcal{G}_1}|\widehat{\mathcal{L}}_{\mathcal{G}_2}(\phi) - \mathcal{L}_{\mathcal{G}_2}(\phi)| \leq C_{\mathcal{B},\eta,a,\bar{\alpha},R_{\max}}\cdot\Big(\mathcal{W}\mathcal{D}\sqrt{\log(\mathcal{W}\mathcal{D})}\Big)\sqrt{\frac{\log n}{n^{\eta/(1+\eta)}}}$$

for a constant $C_{\mathcal{B},\eta,a,\bar{\alpha},R_{\max}}$ depending on $\mathcal{B}, \eta, a, \bar{\alpha}$, and $R_{\max}$.

Remark 1 : To the best of authors' knowledge, Theorem 1 represents the first generalization bound for ReLU ResNets with dependent data. This marks a nontrivial extension of generalization analysis for deep neural networks with i.i.d. data [60], [70], [71], [72], [73], [74], [75]. Theorem 1 implies that the statistical error bound depends on n, the width $\mathcal{W}$, and the depth $\mathcal{D}$, and it converges to 0 as n approaches infinity for fixed values of $\mathcal{W}$ and $\mathcal{D}$.

B. Approximation Error

To derive the respective upper bounds for the approximation errors $\mathcal{E}_{\mathcal{G}_1}$ and $\mathcal{E}_{\mathcal{G}_2}$ based on the existing approximation theory, we need to bridge the gap between $\mathcal{E}_{\mathcal{G}_1}$, $\mathcal{E}_{\mathcal{G}_2}$, and the following quantities:

$$\inf_{\phi\in\mathcal{G}_1}\|Q^{\pi_j} - \phi\|_{L^1(\mu)}$$
$$\sup_{v\in\mathcal{U}_2}\inf_{O\in\mathcal{G}_2}\|O - v\|_{L^1(\mu)}.$$

This bridging can indeed be achieved using Lemma 3.

Lemma 3 : Assume that function classes $\mathcal{U}_1, \mathcal{U}_2, \mathcal{G}_1$, and $\mathcal{G}_2$ are uniformly bounded by a constant $\mathcal{B}$. Then, we have

$$\mathcal{E}_{\mathcal{G}_1} \leq \mathcal{B}\big(1 + c_{\mu,\mu}(1)\big)\inf_{\phi\in\mathcal{G}_1}\|Q^{\pi_j} - \phi\|_{L^1(\mu)}$$
$$\mathcal{E}_{\mathcal{G}_2} \leq (2\mathcal{B} + R_{\max})\cdot\sup_{v\in\mathcal{U}_2}\inf_{O\in\mathcal{G}_2}\|O - v\|_{L^1(\mu)}$$

where $c_{\mu,\mu}(1)$ is a constant defined in (5).

As a consequence of Lemma 3, we only need to provide upper bounds for the approximation errors $\inf_{\phi\in\mathcal{G}_1}\|Q^{\pi_j} - \phi\|_{L^1(\mu)}$ and $\sup_{v\in\mathcal{U}_2}\inf_{O\in\mathcal{G}_2}\|O - v\|_{L^1(\mu)}$. Both of these errors can be bounded by the approximation error of the ReLU ResNet class $\mathcal{F}^K$ to the Hölder class under the assumption that the function classes $\mathcal{U}_1$ and $\mathcal{U}_2$ are Hölder classes, $\mathcal{G}_1$ and $\mathcal{G}_2$ are ReLU ResNet classes $\mathcal{F}^K$, and Q^{π_j} is contained in the Hölder class, as defined in Definition 4. To that end, we assume that the distribution of the state action (X, A) is supported on $[0, 1]^d$ without loss of generality.

Definition 4 (Hölder Class): For $\zeta > 0$ with $\zeta = s + r$, where $s \in \mathbb{N}_0$ and $r \in (0, 1]$ and $d \in \mathbb{N}$, we denote Hölder class $\mathcal{H}^\zeta$ as

$$\mathcal{H}^\zeta = \Big\{f : [0, 1]^d \to \mathbb{R}, \max_{\|\tilde{\alpha}\|_1\leq s}\big\|\partial^{\tilde{\alpha}} f\big\|_\infty \leq B$$

$$\max_{\|\tilde{\alpha}\|_1=s} \sup_{x\neq y} \frac{\left|\partial^{\tilde{\alpha}} f(x) - \partial^{\tilde{\alpha}} f(y)\right|}{\|x-y\|_\infty^r} \leq B \Bigg\}.$$

We give a novel approximation error bound for Hölder class using ResNet with the ReLU activation, which is of independent interest. Furthermore, we improve the result of dependence on the ambient dimension d from $\mathcal{O}(8^\zeta(\zeta+1)^d)$ [59] to $\mathcal{O}(d^\zeta)$ by following [60].

Theorem 2 : Assume $f \in \mathcal{H}^\zeta$ with $\zeta = s + r, s \in \mathbb{N}_0$ and $r \in (0, 1]$. For any $W, L \in \mathbb{N}$, there exists a function $\tilde{f}$ belonging to the ReLU ResNet class $\mathcal{F}^K$ with width $\mathcal{W} \asymp (s+1)^2 d^{s+1} W \lceil \log_2 8W \rceil$ and depth $\mathcal{D} \asymp (s+1)^2 L \lceil \log_2 8L \rceil$ such that

$$|f(x) - \tilde{f}(x)| \lesssim B(s+1)^2 d^{s+(\zeta\vee 1)/2} \lfloor (WL)^{2/d} \rfloor^{-\zeta}$$

for all $x \in [0,1]^d \backslash \Omega([0,1]^d, S, \delta)$, where

$$\Omega([0,1]^d, S, \delta) = \cup_{i=1}^d \Big\{ x = [x_1, x_2, \ldots, x_d]^\top : x_i \in \cup_{k=1}^{S-1} (k/S - \delta, k/S) \Big\}$$

with $S = \lceil (WL)^{2/d} \rceil$ and $\delta \in (0, 1/(3S)]$.

As a consequence of Theorem 2, we can derive the approximation results in L^p-norm with $1 \leq p < \infty$, where the width and depth depend on the ambient dimension d polynomially. To establish the validity of Theorem 2, our proof proceeds in two key steps. First, we demonstrate that ReLU ResNet can be expressed as ReLU FNN. Subsequently, we establish the approximation capabilities of ReLU FNN for the Hölder class. The latter result directly follows from [60]. For a comprehensive presentation of the proof, we refer readers to the Appendix.

V. Nonasymptotic Error Bounds

Building upon the comprehensive analyses of the statistical error provided in Theorem 1 and the approximation error as detailed in Theorem 2, we proceed to establish a nonasymptotic error bound for the excess risk, denoted as $\mathcal{L}_{\mathcal{U}_2}(Q_j) - \mathcal{L}_{\mathcal{U}_2}(Q^{\pi_j})$. This can be achieved by judiciously selecting appropriate width $\mathcal{W}$ and depth $\mathcal{D}$ within the ReLU ResNet class $\mathcal{F}^K$. Our findings are encapsulated in the ensuing Theorem 3, which offers a precise characterization of the error bound. Moreover, we extend our analysis to bound $\|Q^* - Q^{\pi_J}\|_{L_2(\nu)}$ by incorporating the completeness assumption (Assumption 1), as delineated in Theorem 4.

Theorem 3 : Suppose that $\mathcal{G}_1$ and $\mathcal{G}_2$ are the ReLU ResNet class $\mathcal{F}^K$, $\mathcal{U}_1$ and $\mathcal{U}_2$ are the Hölder class $\mathcal{H}^\zeta$, $Q^{\pi_j} \in \mathcal{H}^\zeta$ with $\zeta = s + r, s \in \mathbb{N}_0$ and $r \in (0, 1]$, $\{Z_i\}_{i=1}^n$ is strictly and exponentially α-mixing with parameters $\bar{\alpha}, a, \eta > 0$ defined in Definition 2, and the probability distribution μ of (X, A) is absolutely continuous with respect to the Lebesgue measure. Then, for the ReLU ResNet class $\mathcal{F}^K$ with width $\mathcal{W} = \mathcal{O}((n^{(\eta/1+\eta)})^{(d/4(d+2\zeta))} \log n)$ and depth $\mathcal{D} = \mathcal{O}((n^{(\eta/1+\eta)})^{(d/4(d+2\zeta))} \log n)$, the excess risk satisfies

$$\mathbb{E}\big(\mathcal{L}_{\mathcal{U}_2}(Q_j) - \mathcal{L}_{\mathcal{U}_2}(Q^{\pi_j})\big) \leq C_{B,s,R_{\max},\mathcal{B},\eta,a,\bar{\alpha},c_{\mu,\mu}(1)} \cdot \left[d^{s+(\zeta\vee 1)/2} \left(n^{\frac{\eta}{1+\eta}} \right)^{\frac{-\zeta}{d+2\zeta}} (\log n)^3 \right]$$

where $C_{B,s,R_{\max},\mathcal{B},\eta,a,\bar{\alpha},c_{\mu,\mu}(1)}$ is a constant depending on $B, s, R_{\max}, \mathcal{B}, \eta, a, \bar{\alpha}, c_{\mu,\mu}(1)$.

Next, we construct the nonasymptotic error bound of $\|Q^* - Q^{\pi_J}\|_{L_2(\nu)}$. To that end, we introduce the following completeness assumption.

Assumption 1 (Completeness): $\forall Q \in \mathcal{F}^K$, $\mathcal{T}^{\pi_j} Q \in \mathcal{H}^\zeta$.

Theorem 4 : Assume that, in addition to the conditions of Theorem 3, Assumption 1 holds. Then,

$$\begin{aligned}\mathbb{E}\|Q^* - Q^{\pi_J}\|_{L_2(\nu)} &\leq \frac{C_{B,s,R_{\max},\mathcal{B},\eta,a,\bar{\alpha},c_{\mu,\mu}(1)} \gamma C_{\nu,\mu}^{1/2}}{(1-\gamma)^2} \\ &\quad \times d^{s/2+(\zeta\vee 1)/4} \left(n^{\frac{\eta}{1+\eta}} \right)^{\frac{-\zeta}{2d+4\zeta}} (\log n)^{3/2} \\ &\quad + \frac{2\gamma^{(2+J)/2}}{(1-\gamma)^2} \cdot R_{\max}.\end{aligned}$$

Remark 2 : The completeness assumption outlined in Assumption 1, as employed in Theorem 4, is considered to be a relatively lenient requirement. It is satisfied when the underlying MDP adheres to certain smoothness conditions, as extensively discussed in [82]. Notably, Chen and Jiang [53] emphasized the indispensable nature of such completeness conditions. Furthermore, Theorem 4 provides an insight that the nonasymptotic error bound, neglecting other terms, is on the order of $\mathcal{O}(n^{(-\zeta/2d+4\zeta)} + \gamma^{(2+J)/2})$. Consequently, we can establish the consistency of DAPI as n and J tend to infinity. However, it is essential to acknowledge that this convergence rate is affected by the curse of dimensionality, especially when dealing with high-dimensional problems. As a result, we consider this aspect as a subject for future research, warranting further exploration.

VI. Conclusion

In this work, we delve into DAPI, employing the ReLU ResNet to estimate the optimal action-value function Q^*. Our primary objective revolves around establishing a nonasymptotic error bound for $\|Q^* - Q^{\pi_J}\|_{L_2(\nu)}$. This bound, crucial to our analysis, is intricately tied to the Bellman residual error $\|Q_j - \mathcal{T}^{\pi_j} Q_j\|_{L_2(\mu)}$ via error propagation mechanisms. To dissect and ultimately bound the quantity $\|Q_j - \mathcal{T}^{\pi_j} Q_j\|_{L_2(\mu)}$, we disentangle it into two distinct sources of error: statistical and approximation errors. To tackle these, we enlist the aid of tools from empirical process theory, particularly suited for handling dependent data, and draw upon the rich domain of deep approximation theory. One of our significant contributions is the derivation of a generalization bound tailored for ReLU ResNet operating in conjunction with α-mixing dependent data. This departures from the traditional reliance on β-mixing assumptions, as commonly seen in prior studies focusing on MDPs, and expands the scope of applicable scenarios. Moreover, we derive a novel approximation error bound for the Hölder class utilizing ReLU ResNet. This bound exhibits improved scalability with respect to the ambient dimension, transitioning from an exponential dependence to a polynomial one. However, it is essential to acknowledge the persistent challenge posed by the curse of dimensionality, which still affects our nonasymptotic error bounds. To address this,

we posit that leveraging the inherent low-dimensional structure of the data might hold the key to mitigating this issue. Exploring these avenues represents an interesting and challenging direction for future research. Additionally, we recognize the need for empirical validation through numerical experiments and practical applications. Consequently, we intend to embark on such endeavors in future work to further assess and refine the performance of DAPI in real-world settings.

APPENDIX

In this appendix, we prove Lemmas 1–3, Proposition 1, and Theorems 1–4.

A. Proof of Lemma 1

Proof: Recall that $Y_i := R_i + \gamma \int Q(X_{i+1}, a)\pi_j(da|X_{i+1})$ and $Z_i := \{X_i, A_i, R_i, X_{i+1}\}$. Some elementary computation yields that for each pair $(Q, O) \in \mathcal{U}_1 \times \mathcal{U}_2$

$$\begin{aligned}
&\mathbb{E}_{Z_i}[O(X_i, A_i)(Q(X_i, A_i) - Y_i)] \\
&= \mathbb{E}_{(X_i,A_i)}\big[\mathbb{E}_{(Z_i|X_i,A_i)}(O(X_i, A_i)(Q(X_i, A_i) - Y_i)|X_i, A_i)\big] \\
&= \mathbb{E}_{(X_i,A_i)}\big[O(X_i, A_i)(Q(X_i, A_i) - \mathcal{T}^{\pi_j}Q(X_i, A_i))\big] \\
&= \mathbb{E}_{(X,A)\sim\mu}\big[O(X, A)(Q(X, A) - \mathcal{T}^{\pi_j}Q(X, A))\big].
\end{aligned}$$

Hence, we have

$$\mathbb{E}_{\{Z_i\}_{i=1}^n}\widehat{\mathcal{L}}(Q, O) = \mathcal{L}(Q, O).$$

□

B. Proof of Proposition 1

Proof: This proposition is directly followed from [30] and [31]. □

C. Proof of Lemma 2

Proof: For any $\phi \in \mathcal{G}_1$, we can deduce that

$$\begin{aligned}
\mathcal{L}_{\mathcal{U}_2}(Q_j) - \mathcal{L}_{\mathcal{U}_2}(Q^{\pi_j}) = &\big[\mathcal{L}_{\mathcal{U}_2}(Q_j) - \mathcal{L}_{\mathcal{G}_2}(Q_j)\big] \\
&+ \big[\mathcal{L}_{\mathcal{G}_2}(Q_j) - \widehat{\mathcal{L}}_{\mathcal{G}_2}(Q_j)\big] \\
&+ \big[\widehat{\mathcal{L}}_{\mathcal{G}_2}(Q_j) - \widehat{\mathcal{L}}_{\mathcal{G}_2}(\phi)\big] \\
&+ \big[\widehat{\mathcal{L}}_{\mathcal{G}_2}(\phi) - \mathcal{L}_{\mathcal{G}_2}(\phi)\big] \\
&+ \big[\mathcal{L}_{\mathcal{G}_2}(\phi) - \mathcal{L}_{\mathcal{U}_2}(\phi)\big] \\
&+ \big[\mathcal{L}_{\mathcal{U}_2}(\phi) - \mathcal{L}_{\mathcal{U}_2}(Q^{\pi_j})\big]
\end{aligned}$$

where the first and fifth terms can be bounded by the approximation error of $\mathcal{U}_2$, that is, $\sup_{\phi\in\mathcal{G}_1}|\mathcal{L}_{\mathcal{U}_2}(\phi) - \mathcal{L}_{\mathcal{G}_2}(\phi)|$, the second and fourth terms can be bounded by the statistical error $\sup_{\phi\in\mathcal{G}_1}|\widehat{\mathcal{L}}_{\mathcal{G}_2}(\phi) - \mathcal{L}_{\mathcal{G}_2}(\phi)|$, and the third term satisfies $\widehat{\mathcal{L}}_{\mathcal{G}_2}(Q_j) - \widehat{\mathcal{L}}_{\mathcal{G}_2}(\phi) \le 0$ by the optimality of Q_j. Taking infimum over all $\phi \in \mathcal{G}_1$ yields the desired result, i.e.,

$$\begin{aligned}
\mathcal{L}_{\mathcal{U}_2}(Q_j) - \mathcal{L}_{\mathcal{U}_2}(Q^{\pi_j}) \le &\, 2\sup_{\phi\in\mathcal{G}_1}|\widehat{\mathcal{L}}_{\mathcal{G}_2}(\phi) - \mathcal{L}_{\mathcal{G}_2}(\phi)| \\
&+ 2\sup_{\phi\in\mathcal{G}_1}|\mathcal{L}_{\mathcal{U}_2}(\phi) - \mathcal{L}_{\mathcal{G}_2}(\phi)| \\
&+ \inf_{\phi\in\mathcal{G}_1}\mathcal{L}_{\mathcal{U}_2}(\phi) - \mathcal{L}_{\mathcal{U}_2}(Q^{\pi_j}).
\end{aligned}$$

□

D. Proof of Lemma 3

Proof: For notation simplicity, let $\mathbb{E}_\mu Q$ denote $\mathbb{E}_{(X,A)\sim\mu}Q(X, A)$ for any measurable function Q. Then, some elementary algebraic computations show that

$$\begin{aligned}
\mathcal{E}_{\mathcal{G}_1} &= \inf_{\phi\in\mathcal{G}_1}\sup_{v\in\mathcal{U}_2}\big\{|\mathbb{E}_\mu(v(\phi - \mathcal{T}^{\pi_j}\phi))|\big\} \\
&\le \mathcal{B}\inf_{\phi\in\mathcal{G}_1}\|Q^{\pi_j} - \phi\|_{L^1(\mu)} \\
&\quad + \mathcal{B}\inf_{\phi\in\mathcal{G}_1}\|\mathcal{T}^{\pi_j}Q^{\pi_j} - \mathcal{T}^{\pi_j}\phi\|_{L^1(\mu)} \\
&\le \mathcal{B}\big(1 + c_{\mu,\mu}(1)\big)\inf_{\phi\in\mathcal{G}_1}\|Q^{\pi_j} - \phi\|_{L^1(\mu)}
\end{aligned}$$

where the first inequality holds since function classes $\mathcal{U}_1, \mathcal{U}_2, \mathcal{G}_1$, and $\mathcal{G}_2$ are uniformly bounded by constant $\mathcal{B}$, the last inequality follows from:

$$\|\mathcal{T}^{\pi_j}Q^{\pi_j} - \mathcal{T}^{\pi_j}\phi\|_{L^1(\mu)} \le \gamma c_{\mu,\mu}(1)\|Q^{\pi_j} - \phi\|_{L^1(\mu)}.$$

Similarly, we also have

$$\begin{aligned}
\mathcal{E}_{\mathcal{G}_2} &= \sup_{\phi\in\mathcal{G}_1}\Big|\sup_{v\in\mathcal{U}_2}\big\{|\mathbb{E}_\mu(v(\phi - \mathcal{T}^{\pi_j}\phi))|\big\} \\
&\qquad - \sup_{O\in\mathcal{G}_2}\big\{|\mathbb{E}_\mu(O(\phi - \mathcal{T}^{\pi_j}\phi))|\big\}\Big| \\
&\le \sup_{\phi\in\mathcal{G}_1}\sup_{v\in\mathcal{U}_2}\inf_{O\in\mathcal{G}_2}\|(v - O)(\phi - \mathcal{T}^{\pi_j}\phi)\|_{L^1(\mu)} \\
&\le (2\mathcal{B} + R_{\max})\cdot\sup_{v\in\mathcal{U}_2}\inf_{O\in\mathcal{G}_2}\|O - v\|_{L^1(\mu)}.
\end{aligned}$$

□

E. Preliminary Lemmas for Theorem 1

First, we introduce the definition of uniform covering number [70]. Let $\mathcal{F}$ be a class of measurable functions mapping $\mathbb{R}^d$ to $\mathbb{R}$. For a given sequence $z = (z_1, \ldots, z_n)$ with $z_i \in \mathbb{R}^d$, $i = 1, \ldots, n$, let $\mathcal{F}|_z := \{(f(z_1), \ldots, f(z_n)) : f \in \mathcal{F}\}$ be the subset of $\mathbb{R}^n$. For a positive number ε, let $\mathcal{N}(\varepsilon, \mathcal{F}|_z, \|\cdot\|_\infty)$ be the covering number of $\mathcal{F}|_z$ under the uniform norm $\|\cdot\|_\infty$ with radius ε. Denote the uniform covering number $\mathcal{N}_n(\varepsilon, \mathcal{F}, \|\cdot\|_\infty)$ as the maximum over all $z \in \mathbb{R}^{dn}$ of the covering number $\mathcal{N}(\varepsilon, \mathcal{F}|_z, \|\cdot\|_\infty)$, that is,

$$\begin{aligned}
&\mathcal{N}_n(\varepsilon, \mathcal{F}, \|\cdot\|_\infty) \\
&:= \max\big\{\mathcal{N}(\varepsilon, \mathcal{F}|_z, \|\cdot\|_\infty) : z = (z_1, \ldots, z_n) \in \mathbb{R}^{dn}\big\}. \quad \text{(E1)}
\end{aligned}$$

Lemma A.1 [44, Theorem 4.3]: Let $(Z_n)_{n\ge1}$ be an $\mathbb{R}^d$-valued and exponentially decayed α-mixing process with parameters $\bar{\alpha}, a, \eta > 0$. Furthermore, we assume that $h : \mathbb{R}^d \to \mathbb{R}$ is a bounded measurable function for which there exists constants $\widetilde{M} > 0$ and $\sigma \ge 0$ such that $\mathbb{E}[h(Z_n)] = 0$ and $\mathbb{E}[h^2(Z_n)] \le \sigma^2$ for all $n \ge 1$ and $\|h\|_\infty \le \widetilde{M}$. For $n \ge 1$, we define

$$n^{(\eta)} := \left\lfloor n\left\lceil\left(\frac{8n}{a}\right)^{\frac{1}{\eta+1}}\right\rceil^{-1}\right\rfloor. \quad \text{(E2)}$$

Then, for any $n \ge 1$ and all $\varepsilon > 0$, we have

$$\mathbb{P}\left(\{\omega \in \Omega : \frac{1}{n}\sum_{i=1}^n h(Z_i(\omega)) \ge \varepsilon\}\right.$$

$$\leq \left(1+4e^{-2}\bar{\alpha}\right)\exp\left(-\frac{3\varepsilon^2 n^{(\eta)}}{6\sigma^2+2\varepsilon\widetilde{M}}\right).$$

Lemma A.2 : Suppose that $\{Z_1,\ldots,Z_n\}$ is a strictly stationary and exponentially decayed α-mixing process with parameters $\bar{\alpha}, a, \eta > 0$. Then, for all fixed $\varepsilon > 0$ and the measurable function class $\mathbf{F}_{\widetilde{M}}$ bounded by $\widetilde{M}$, we have

$$\mathbb{P}\left(\sup_{f\in\mathbf{F}_{\widetilde{M}}}|\frac{1}{n}\sum_{i=1}^{n}f(Z_i)-\mathbb{E}\big[f(Z_1)\big]|>\varepsilon\right) \leq 2\big(1+4e^{-2}\bar{\alpha}\big)\mathcal{N}_{2n}\Big(\frac{\varepsilon}{4},\mathbf{F}_{\widetilde{M}},\|\cdot\|_\infty\Big) \times\exp\left(-\frac{3\varepsilon^2 n^{(\eta)}}{96\widetilde{M}^2+32\varepsilon\widetilde{M}}\right).$$

Proof: To begin, we introduce a separate set of random variables $\{\widetilde{Z}_1,\ldots,\widetilde{Z}_n\}$, which serve as an independent copy of $\{Z_1,\ldots,Z_n\}$. Next, we consider another set of independent random variables $\{\sigma_1,\ldots,\sigma_n\}$, which are uniformly distributed over the set $\{-1,1\}$. These variables are independent of both $\{Z_1,\ldots,Z_n\}$ and $\{\widetilde{Z}_1,\ldots,\widetilde{Z}_n\}$. As a result, the redefined sequence $\{(\sigma_1,Z_1),\ldots,(\sigma_n,Z_n)\}$ retains the property of being α-mixing, just like the original sequence $\{Z_1,\ldots,Z_n\}$, as long as the latter possesses this property. Now, let us define $\mathcal{G}(\varepsilon/4,\mathbf{F}_{\widetilde{M}},\|\cdot\|_\infty)$ as an $\varepsilon/4$-uniform covering of the function class $\mathbf{F}_{\widetilde{M}}$. The uniform covering number of this set is denoted as $\mathcal{N}_{2n}(\varepsilon/4,\mathbf{F}_{\widetilde{M}},\|\cdot\|_\infty)$ and is defined in (E1). For any fixed function f belonging to the function class $\mathbf{F}_{\widetilde{M}}$, there exists a function g in $\mathcal{G}(\varepsilon/4,\mathbf{F}_{\widetilde{M}},\|\cdot\|_\infty)$ that satisfies the following conditions for $i\in\{1,\ldots,n\}$:

$$|f(Z_i)-g(Z_i)|<\varepsilon/4 \text{ and } |f\big(\widetilde{Z}_i\big)-g\big(\widetilde{Z}_i\big)|<\varepsilon/4.$$

Furthermore, for any function g belonging to $\mathcal{G}(\varepsilon/4,\mathbf{F}_{\widetilde{M}},\|\cdot\|_\infty)$, the following properties hold: $\|g\|_\infty\leq\widetilde{M}$, $\mathbb{E}[\sigma_i g(Z_i)]=0$, $\mathbb{E}[\sigma_i g(\widetilde{Z}_i)]=0$, $\mathbb{E}[g^2(Z_i)]\leq\widetilde{M}^2$, and $\mathbb{E}[g^2(\widetilde{Z}_i)]\leq\widetilde{M}^2$. Consequently, utilizing the Bernstein-type inequality presented in Lemma A.2, we can deduce that

$$\begin{aligned}
&\mathbb{P}\left(\sup_{f\in\mathbf{F}_{\widetilde{M}}}\left|\frac{1}{n}\sum_{i=1}^{n}f(Z_i)-\mathbb{E}\big[f(Z_1)\big]\right|>\varepsilon\right)\\
&\leq\mathbb{P}\left(\sup_{f\in\mathbf{F}_{\widetilde{M}}}\left|\frac{1}{n}\sum_{i=1}^{n}f(Z_i)-\frac{1}{n}\sum_{i=1}^{n}f(\widetilde{Z}_i)\right|>\varepsilon\right)\\
&\leq\mathbb{P}\left(\sup_{f\in\mathbf{F}_{\widetilde{M}}}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i\big(f(Z_i)-f(\widetilde{Z}_i)\big)\right|>\varepsilon\right)\\
&\leq\mathbb{P}\left(\sup_{g\in\mathcal{G}(\varepsilon/4,\mathbf{F}_{\widetilde{M}},\|\cdot\|_\infty)}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i(g(Z_i)-g(\widetilde{Z}_i))\right|>\varepsilon/2\right)\\
&\leq\mathbb{P}\left(\sup_{g\in\mathcal{G}(\varepsilon/4,\mathbf{F}_{\widetilde{M}},\|\cdot\|_\infty)}\left(\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i g(Z_i)\right|\right)>\varepsilon/4\right)\\
&\leq 2\big(1+4e^{-2}\bar{\alpha}\big)\mathcal{N}_{2n}\Big(\frac{\varepsilon}{4},\mathbf{F}_{\widetilde{M}},\|\cdot\|_\infty\Big)\\
&\quad\times\exp\left(-\frac{3\varepsilon^2 n^{(\eta)}}{96\widetilde{M}^2+32\varepsilon\widetilde{M}}\right)
\end{aligned}$$

where $n^{(\eta)}$ is given in (E2) and $\bar{\alpha}, a$, and η are the given parameters in exponentially decayed α-mixing process. Here, the first inequality is established through the utilization of Jensen's inequality, the second inequality is justified based on the fact that the difference between $f(Z_i)$ and $f(\widetilde{Z}_i)$, as well as the product of this difference and σ_i follow the same probability distribution, the third inequality is a direct consequence of the definition of the uniform covering number, the fourth inequality is derived through elementary algebraic calculations, and the last inequality is supported by invoking Lemma A.2 on the sequence $\{(\sigma_1,Z_1),\ldots,(\sigma_n,Z_n)\}$. □

F. Proof of Theorem 1

Proof: As shown in Lemma A.3, the ReLU FNN class $\mathcal{F}$ and the ResNet ReLU class $\mathcal{F}^K$ are equivalent. In the following proof, we use $\mathcal{F}$ to denote the neural network class for simplicity. Denote the composite function class

$$\ell\circ(\mathcal{G}_1,\mathcal{G}_2):=\Big\{\ell_{Q,O}:\ell_{Q,O}(x,a,r,x')=O(x,a)\Big[Q(x,a)-r-\gamma\int Q(x',a)\pi_j(da|x')\Big] \quad Q\in\mathcal{G}_1, O\in\mathcal{G}_2\Big\}.$$

Then, it follows that:

$$\begin{aligned}
\mathcal{E}_{\text{sta}}&=\sup_{Q\in\mathcal{G}_1}\Big|\sup_{O\in\mathcal{G}_2}\left|\frac{1}{n}\sum_{i=1}^{n}O(X_i,A_i)(Q(X_i,A_i)-Y_i)\right|-\sup_{O\in\mathcal{G}_2}|\mathbb{E}_\mu(O(Q-\mathcal{T}^{\pi_j}Q))|\Big|\\
&\leq\sup_{Q\in\mathcal{G}_1,O\in\mathcal{G}_2}\left|\frac{1}{n}\sum_{i=1}^{n}O(X_i,A_i)(Q(X_i,A_i)-Y_i)-\mathbb{E}_\mu(O(Q-\mathcal{T}^{\pi_j}Q))\right|\\
&=\sup_{Q\in\mathcal{G}_1,O\in\mathcal{G}_2}\left|\frac{1}{n}\sum_{i=1}^{n}\ell_{Q,O}(X_i,A_i,R_i,X_{i+1})-\mathbb{E}\ell_{Q,O}(X_i,A_i,R_i,X_{i+1})\right|.
\end{aligned}$$

Denote $\text{VC}_{\mathcal{G}_1}$, and $\text{VC}_{\mathcal{G}_2}$ as the VC-dimension of $\mathcal{G}_1$, and $\mathcal{G}_2$, respectively. Thus, for any $\delta\geq 0$, we have

$$\begin{aligned}
&\mathbb{E}\sup_{Q\in\mathcal{G}_1,O\in\mathcal{G}_2}\left|\frac{1}{n}\sum_{i=1}^{n}\ell_{Q,O}(X_i,A_i,R_i,X_{i+1})-\mathbb{E}\ell_{Q,O}(X_i,A_i,R_i,X_{i+1})\right|\\
&\leq\delta+\int_\delta^{2\widetilde{M}}P\Big(\sup_{Q\in\mathcal{G}_1,O\in\mathcal{G}_2}\left|\frac{1}{n}\sum_{i=1}^{n}\ell_{Q,O}(X_i,A_i,R_i,X_{i+1})-\mathbb{E}\ell_{Q,O}(X_1,A_1,R_1,X_2)\right|>\varepsilon\Big)d\varepsilon\\
&\leq\delta+\int_\delta^{2\widetilde{M}}2C\mathcal{N}_{2n}(\varepsilon/4,\ell\circ(\mathcal{G}_1,\mathcal{G}_2),\|\cdot\|_\infty)\times\exp\left(-\frac{3n^{(\eta)}\varepsilon^2}{96\widetilde{M}^2+32\widetilde{M}\varepsilon}\right)d\varepsilon\\
&\leq\delta+\int_\delta^{2\widetilde{M}}2C\mathcal{N}_{2n}\Big(\frac{\varepsilon}{4\lambda},\mathcal{G}_1,\|\cdot\|_\infty\Big)\times\mathcal{N}_{2n}\Big(\frac{\varepsilon}{4\lambda},\mathcal{G}_2,\|\cdot\|_\infty\Big)\times\exp\left(-\frac{3n^{(\eta)}\varepsilon^2}{96\widetilde{M}^2+32\widetilde{M}\varepsilon}\right)d\varepsilon
\end{aligned}$$

$$
\begin{aligned}
&\le \delta + \int_{\delta}^{2\widetilde{M}} 2C\left(\frac{2e\mathcal{B}n}{\frac{\varepsilon}{4\lambda}\cdot \mathrm{VC}_{\mathcal{G}_1}}\right)^{\mathrm{VC}_{\mathcal{G}_1}} \\
&\qquad \times \left(\frac{2e\mathcal{B}n}{\frac{\varepsilon}{4\lambda}\cdot \mathrm{VC}_{\mathcal{G}_2}}\right)^{\mathrm{VC}_{\mathcal{G}_2}} \\
&\qquad \times \exp\left(-\frac{3n^{(\eta)}\varepsilon^2}{96\widetilde{M}^2+32\widetilde{M}\varepsilon}\right)d\varepsilon \\
&\le \delta + \int_{\delta}^{2\widetilde{M}} 2C\left(\frac{2e\mathcal{B}n}{\frac{\delta}{4\lambda}\cdot \mathrm{VC}_{\mathcal{F}}}\right)^{2\mathrm{VC}_{\mathcal{F}}} \\
&\qquad \times \exp\left(-\frac{3n^{(\eta)}\varepsilon^2}{96\widetilde{M}^2+32\widetilde{M}\varepsilon}\right)d\varepsilon \\
&\le \delta + 4C\widetilde{M}\left(\frac{8\lambda e\mathcal{B}n}{\delta \mathrm{VC}_{\mathcal{F}}}\right)^{2\mathrm{VC}_{\mathcal{F}}} \exp\left(-\frac{3n^{(\eta)}\delta^2}{160\widetilde{M}^2}\right) \\
&\le C_{\mathcal{B},\eta,a,\bar{\alpha},R_{\max}} n^{-\frac{\eta}{2(1+\eta)}}\cdot\sqrt{\log n\cdot \mathrm{VC}_{\mathcal{F}}}.
\end{aligned}
$$

Here, the first inequality holds since $\ell\circ(\mathcal{G}_1,\mathcal{G}_2)$ is bounded by $\widetilde{M} := \mathcal{B}(2\mathcal{B}+R_{\max})$, the second inequality holds by Lemma A.2 with $C := 1+4e^{-2}\bar{\alpha}$, the third inequality is established by considering the Lipschitz continuity of the composite function $\ell_{Q,O}$ which possesses a Lipschitz constant denoted as $\lambda := 2\mathcal{B}+R_{\max}$, the fourth inequality is justified through a relation connecting the covering number and the VC-dimension of the ReLU neural networks $\mathcal{F}$, as expounded in [70], i.e.,

$$
\begin{aligned}
&\mathcal{N}_{2n}\left(\frac{\varepsilon}{4\lambda},\mathcal{F},\|\cdot\|_{\infty}\right) \\
&\qquad \le \left(\frac{2e\mathcal{B}n}{\frac{\varepsilon}{4\lambda}\cdot \mathrm{VC}_{\mathcal{F}}}\right)^{\mathrm{VC}_{\mathcal{F}}}
\end{aligned}
$$

the sixth inequality holds by some algebraic calculations, and the last inequality holds with constant $C_{\mathcal{B},\eta,a,\bar{\alpha},R_{\max}}$ depending on $\mathcal{B},\eta,a,\bar{\alpha}$, and $R_{\max}$ due to the fact that $n^{(\eta)} \ge 2^{-(2\eta+5/1+\eta)}a^{(1/1+\eta)}n^{(\eta/1+\eta)}$ when $\lceil t\rceil \le 2t$ for all $t\ge 1$ and $\lfloor t\rfloor \ge t/2$ for all $t\ge 2$ and setting

$$
\delta^2 = \frac{160\widetilde{M}^2}{n^{\frac{\eta}{1+\eta}}}\mathrm{VC}_{\mathcal{F}}\log\left(\frac{8\lambda e\mathcal{B}n}{\mathrm{VC}_{\mathcal{F}}}\right).
$$

Then, we have

$$
\begin{aligned}
&\mathbb{E}\sup_{Q\in\mathcal{G}_1,O\in\mathcal{G}_2}\Big|\frac{1}{n}\sum_{i=1}^{n}\ell_{Q,O}(X_i,A_i,R_i,X_{i+1}) \\
&\qquad -\mathbb{E}\ell_{Q,O}(X_i,A_i,R_i,X_{i+1})\Big| \\
&\le C_{\mathcal{B},\eta,a,\bar{\alpha},R_{\max}}\left(\mathcal{W}\mathcal{D}\sqrt{\log(\mathcal{W}\mathcal{D})}\right)n^{-\frac{\eta}{2(1+\eta)}}\sqrt{\log n}
\end{aligned}
$$

where the inequality holds since the VC-dimension of the ReLU neural networks $\mathcal{F}$ satisfies

$$
c_1\cdot\mathcal{W}^2\mathcal{D}^2\log(\mathcal{W})\le \mathrm{VC}_{\mathcal{F}}\le c_2\cdot\mathcal{W}^2\mathcal{D}^2\log(\mathcal{W}\mathcal{D})
$$

where $c_1,c_2>0$ are two universal constants, as detailed in [77].

□

G. Preliminary Lemma for Theorem 2

We first give a lemma on representing ReLU FNN with ReLU ResNet.

Lemma A.3 : Given $m\in\mathbb{N}$, let $\mathbf{f}:\mathbb{R}^m\to\mathbb{R}^m$ with width W and depth L, then there exists a residual block $\mathbf{g}:\mathbb{R}^m\to\mathbb{R}^m$ with width $3W$ and depth $L+1$ such that $\mathbf{f}(x)=\mathbf{g}(x)$. Moreover, there exist absolute constants $c_1,c_2>0$ such that, for any $f\in\mathcal{F}_{\mathcal{W},\mathcal{D},\mathcal{B}}$, there exists a function $g\in\mathcal{F}^K_{c_1\mathcal{W},c_2\mathcal{D},\mathcal{B}}$ such that $g(x)=f(x)$.

Proof: Note the fact that $\rho\circ\rho(x)=\rho(x)$ and $\rho(x)-\rho(-x)=x$, we know that $\mathbf{f}(x)-x$ can be represented by a ReLU FNN with width $3W$ and depth $L+1$. Then, by the definition in (1), $\mathbf{f}(x)=x+(\mathbf{f}(x)-x)$ can be represented by a residual block $\mathbf{g}$ with depth $L+1$ and width $3W$.

Let $f\in\mathcal{F}_{\mathcal{D},\mathcal{W},\mathcal{B}}$ be defined as

$$
\begin{aligned}
&f^{(0)}(x)=x \\
&f^{(\ell)}(x)=\rho\big(A_\ell f^{(\ell-1)}(x)+b_\ell\big) \quad \text{for } \ell=1,\dots,\mathcal{D}-1 \\
&f(x):=f^{(\mathcal{D})}(x)=A_{\mathcal{D}}f^{(\mathcal{D}-1)}(x)+b_{\mathcal{D}}.
\end{aligned}
$$

Recall that $g:\mathbb{R}^d\to\mathbb{R}$ implemented by a ReLU ResNet with $K-1$ residual blocks is defined by

$$
\begin{aligned}
&g_{(0)}(x)=x \\
&g_{(k)}(x)=\rho(A_k\mathbf{R}_k(g_{k-1}(x))+b_k),k=1,\dots,K-1 \\
&g(x):=g_{(K)}(x)=A_Kg_{(K-1)}(x)+b_K.
\end{aligned}
$$

We can reformulate f into g by setting $K=\mathcal{D}$ and choosing $\mathbf{R}_k(x)=x,k=1,\dots,\mathcal{D}$. □

H. Proof of Theorem 2

Proof: By Lemma A.3 and [60, Theorem 3.3], for any $W,L\in\mathbb{N}$, there exists a function $\tilde{f}$ belonging to the ReLU ResNet class $\mathcal{F}^K$ with width $\mathcal{W}\asymp(s+1)^2d^{s+1}W\lceil\log_2(8W)\rceil$ and depth $\mathcal{D}\asymp(s+1)^2L\lceil\log_2(8L)\rceil$ such that

$$
\begin{aligned}
&|f(x)-\tilde{f}(x)| \\
&\lesssim B(s+1)^2d^{s+(\zeta\vee1)/2}\big\lfloor(WL)^{2/d}\big\rfloor^{-\zeta}
\end{aligned}
$$

for all $x\in[0,1]^d\backslash\Omega([0,1]^d,S,\delta)$. Here,

$$
\begin{aligned}
\Omega([0,1]^d,S,\delta)=\cup_{i=1}^{d}\Big\{&x=[x_1,x_2,\dots,x_d]^\top: \\
&x_i\in\cup_{k=1}^{S-1}(k/S-\delta,k/S)\Big\}
\end{aligned}
$$

where $S=\lceil(WL)^{2/d}\rceil$ and $\delta\in(0,1/(3S)]$.

□

I. Proof of Theorem 3

Proof: By Theorem 2, for any $f^*\in\mathcal{H}^\zeta([0,1]^d)$, there exists a function $\phi_0\in\mathcal{F}^K$ with width $\mathcal{W}\asymp(s+1)^2d^{s+1}W\lceil\log_2(8W)\rceil$ and depth $\mathcal{D}\asymp(s+1)^2L\lceil\log_2(8L)\rceil$ such that

$$
\begin{aligned}
&\big|f^*(x)-\phi_0(x)\big| \\
&\le 18B(s+1)^2d^{s+(\zeta\vee1)/2}\big\lfloor(WL)^{2/d}\big\rfloor^{-\zeta}
\end{aligned}
$$

for $x\in\cup_\theta\widetilde{Q}_\theta$. Recall that

$$
\begin{aligned}
\widetilde{Q}_\theta:=\Big\{&x=(x_1,x_2,\dots,x_d): \\
&x_i\in\Big[\frac{\theta_i}{S},\frac{\theta_i+1}{S}-\delta\cdot1_{\{\theta_i<S-1\}}\Big],i=1,2,\dots,d\Big\}
\end{aligned}
$$

with $\theta = (\theta_1, \theta_2, \ldots, \theta_d) \in \{0, 1, \ldots, S-1\}^d$, and δ being an arbitrary number satisfying $0 < \delta \leq 1/3S$. Then, we can conclude that the Lebesgue measure of $[0,1]\backslash\widetilde{\mathcal{Q}}_\theta$ is no more than $dS\delta$. This can also be arbitrarily small if δ is arbitrarily small. Since μ is absolutely continuous with respect to the Lebesgue measure, we have

$$\begin{aligned}
&\sup_{v\in\mathcal{U}_2}\inf_{O\in\mathcal{G}_2}\|O-v\|_{L^1(\mu)}\\
&\quad\leq 18B(s+1)^2 d^{s+(\zeta\vee 1)/2}\lfloor (WL)^{2/d}\rfloor^{-\zeta}\\
&\inf_{\phi\in\mathcal{G}_1}\|Q^{\pi_j}-\phi\|_{L^1(\mu)}\\
&\quad\leq 18B(s+1)^2 d^{s+(\zeta\vee 1)/2}\lfloor (WL)^{2/d}\rfloor^{-\zeta}.
\end{aligned}$$

By Lemma 2 and Theorem 1, it yields that

$$\begin{aligned}
&\mathbb{E}\big(\mathcal{L}_{\mathcal{U}_2}(Q_j)-\mathcal{L}_{\mathcal{U}_2}(Q^{\pi_j})\big)\\
&\leq C_{\mathcal{B},\eta,a,\bar{\alpha},R_{\max}}\Big(\mathcal{W}\mathcal{D}\sqrt{\log(\mathcal{W}\mathcal{D})}\Big)n^{-\frac{\eta}{2(1+\eta)}}\sqrt{\log n}\\
&\quad+(4\mathcal{B}+2R_{\max})\cdot 18B(s+1)^2 d^{s+(\zeta\vee 1)/2}\lfloor (WL)^{2/d}\rfloor^{-\zeta}\\
&\quad+\mathcal{B}(1+c_{\mu,\mu}(1))\cdot 18B(s+1)^2 d^{s+(\zeta\vee 1)/2}\lfloor (WL)^{2/d}\rfloor^{-\zeta}.
\end{aligned}$$

Moreover, setting width $\mathcal{W}=\mathcal{O}\Big((n^{\frac{\eta}{1+\eta}})^{\frac{d}{4(d+2\zeta)}}\log n\Big)$ and depth $\mathcal{D}=\mathcal{O}\Big((n^{\frac{\eta}{1+\eta}})^{\frac{d}{4(d+2\zeta)}}\log n\Big)$, then it follows that:

$$\begin{aligned}
&\mathbb{E}\big(\mathcal{L}_{\mathcal{U}_2}(Q_j)-\mathcal{L}_{\mathcal{U}_2}(Q^{\pi_j})\big)\\
&\leq C_{B,s,R_{\max},\mathcal{B},\eta,a,\bar{\alpha},c_{\mu,\mu}(1)}\cdot\Big[d^{s+(\zeta\vee 1)/2}(n^{\frac{\eta}{1+\eta}})^{\frac{-\zeta}{d+2\zeta}}(\log n)^3\Big]
\end{aligned}$$

where $C_{B,s,R_{\max},\mathcal{B},\eta,a,\bar{\alpha},c_{\mu,\mu}(1)}$ is a constant depending on $B, s, R_{\max}, \mathcal{B}, \eta, a, \bar{\alpha}, c_{\mu,\mu}(1)$. □

J. Proof of Theorem 4

Proof: Recall that

$$\mathcal{L}_{\mathcal{U}_2}(Q)-\mathcal{L}_{\mathcal{U}_2}(Q^{\pi_j})=\sup_{O\in\mathcal{U}_2}\big|\mathbb{E}_\mu(O(Q-\mathcal{T}^{\pi_j}Q))\big|.$$

Moreover, the deep neural networks $\mathcal{F}^K$ have the capacity to approximate the Hölder class $\mathcal{H}^\zeta$. Consequently, by leveraging Theorem 2, Theorem 3, and Assumption 1, we have

$$\begin{aligned}
&\mathbb{E}\|Q_j-\mathcal{T}^{\pi_j}Q_j\|^2_{L^2(\mu)}\\
&\leq\mathbb{E}\big(\mathcal{L}_{\mathcal{U}_2}(Q_j)-\mathcal{L}_{\mathcal{U}_2}(Q^{\pi_j})\big)\\
&\leq C_{B,s,R_{\max},\mathcal{B},\eta,a,\bar{\alpha},c_{\mu,\mu}(1)}\cdot\Big[d^{s+(\zeta\vee 1)/2}(n^{\frac{\eta}{1+\eta}})^{\frac{-\zeta}{d+2\zeta}}(\log n)^3\Big].
\end{aligned}$$

By Proposition 1, we have

$$\begin{aligned}
&\mathbb{E}\big\|Q^*-Q^{\pi_J}\big\|_{L_2(\nu)}\\
&\quad\leq\frac{\gamma C^{1/2}_{\nu,\mu}\cdot C_{B,s,R_{\max},\mathcal{B},\eta,a,\bar{\alpha},c_{\mu,\mu}(1)}}{(1-\gamma)^2}\\
&\qquad\times\Big[d^{s/2+(\zeta\vee 1)/4}(n^{\frac{\eta}{1+\eta}})^{\frac{-\zeta}{2d+4\zeta}}(\log n)^{3/2}\Big]\\
&\qquad+\frac{2\gamma}{(1-\gamma)^2}\cdot\gamma^{J/2}R_{\max}.
\end{aligned}$$

□

ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to the editor, the associate editor, and the four anonymous reviewers for their valuable and constructive comments. Their insights and suggestions have significantly contributed to the improvement of this article.

REFERENCES

[1] L. P. Kaelbling, M. L. Littman, and A. W. Moore, "Reinforcement learning: A survey," *J. Artif. Intell. Res.*, vol. 4, no. 1, pp. 237–285, Jan. 1996.
[2] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: MIT Press, 2018.
[3] Y. Li, "Deep reinforcement learning: An overview," 2017, *arXiv:1701.07274*.
[4] P. Henderson, R. Islam, P. Bachman, J. Pineau, D. Precup, and D. Meger, "Deep reinforcement learning that matters," in *Proc. AAAI Conf. Artif. Intell.*, vol. 32, no. 1, 2018, pp. 3207–3214.
[5] V. François-Lavet, P. Henderson, R. Islam, M. G. Bellemare, and J. Pineau, "An introduction to deep reinforcement learning," *Found. Trends Mach. Learn.*, vol. 11, nos. 3–4, pp. 219–354, Nov. 2018.
[6] M. Mahmud, M. S. Kaiser, A. Hussain, and S. Vassanelli, "Applications of deep learning and reinforcement learning to biological data," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 6, pp. 2063–2079, Jun. 2018.
[7] Y. Keneshloo, T. Shi, N. Ramakrishnan, and C. K. Reddy, "Deep reinforcement learning for sequence-to-sequence models," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 7, pp. 2469–2489, Jul. 2020.
[8] X. Wang, Y. Gu, Y. Cheng, A. Liu, and C. L. P. Chen, "Approximate policy-based accelerated deep reinforcement learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 6, pp. 1820–1830, Jun. 2020.
[9] Z. Cao, K. Wong, and C.-T. Lin, "Weak human preference supervision for deep reinforcement learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 12, pp. 5369–5378, Dec. 2021.
[10] T. T. Nguyen and V. J. Reddi, "Deep reinforcement learning for cyber security," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 8, pp. 3779–3795, 2021.
[11] P. Ladosz et al., "Deep reinforcement learning with modulated Hebbian plus Q-network architecture," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 5, pp. 2045–2056, May 2022.
[12] Z. Rao et al., "Visual navigation with multiple goals based on deep reinforcement learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 12, pp. 5445–5455, Dec. 2021.
[13] D. L. Elliott and C. Anderson, "The wisdom of the crowd: Reliable deep reinforcement learning through ensembles of Q-functions," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 1, pp. 43–51, Jan. 2023.
[14] S. Yun, J. Choi, Y. Yoo, K. Yun, and J. Y. Choi, "Action-driven visual object tracking with deep reinforcement learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 6, pp. 2239–2252, Jun. 2018.
[15] Z. Huang, J. Wu, and C. Lv, "Efficient deep reinforcement learning with imitative expert priors for autonomous driving," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 10, pp. 7391–7403, 2023.
[16] X. Wang et al., "Deep reinforcement learning: A survey," *IEEE Trans. Neural Netw. Learn. Syst.*, 2022.
[17] A. Tsantekidis, N. Passalis, A.-S. Toufa, K. Saitas-Zarkias, S. Chairistanidis, and A. Tefas, "Price trailing for financial trading using deep reinforcement learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 7, pp. 2837–2846, Jul. 2021.
[18] V. Mnih et al., "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.
[19] P. Xu, Q. Yin, J. Zhang, and K. Huang, "Deep reinforcement learning with part-aware exploration bonus in video games," *IEEE Trans. Games*, vol. 14, no. 4, pp. 644–653, Dec. 2022.
[20] D. Silver et al., "Mastering the game of go with deep neural networks and tree search," *Nature*, vol. 529, no. 7587, pp. 484–489, Jan. 2016.
[21] M. Ranzato, S. Chopra, M. Auli, and W. Zaremba, "Sequence level training with recurrent neural networks," 2015, *arXiv:1511.06732*.
[22] J. He et al., "Deep reinforcement learning with a natural language action space," 2015, *arXiv:1511.04636*.
[23] D. B. P. Brakel, K. X. A. Goyal, R. L. J. Pineau, A. Courville, and Y. Bengio, "An actor-critic algorithm for sequence prediction," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2017, pp. 1–17.

[24] W. Y. Wang, J. Li, and X. He, "Deep reinforcement learning for NLP," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics, Tutorial Abstr.*, 2018, pp. 19–21.

[25] S. Levine, C. Finn, T. Darrell, and P. Abbeel, "End-to-end training of deep visuomotor policies," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 1334–1373, 2015.

[26] S. Levine, P. Pastor, A. Krizhevsky, J. Ibarz, and D. Quillen, "Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection," *Int. J. Robot. Res.*, vol. 37, nos. 4–5, pp. 421–436, Apr. 2018.

[27] Z. Yang, K. Merrick, L. Jin, and H. A. Abbass, "Hierarchical deep reinforcement learning for continuous action control," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 11, pp. 5174–5184, Nov. 2018.

[28] H. Li, Q. Zhang, and D. Zhao, "Deep reinforcement learning-based automatic exploration for navigation in unknown environment," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 6, pp. 2064–2076, Jun. 2020.

[29] M. G. Lagoudakis and R. Parr, "Least-squares policy iteration," *J. Mach. Learn. Res.*, vol. 4, pp. 1107–1149, Dec. 2003.

[30] A. Antos, C. Szepesvári, and R. Munos, "Learning near-optimal policies with bellman-residual minimization based fitted policy iteration and a single sample path," *Mach. Learn.*, vol. 71, no. 1, pp. 89–129, Apr. 2008.

[31] A. M. Farahmand, R. Munos, and C. Szepesvári, "Error propagation for approximate policy and value iteration," in *Proc. Adv. Neural Inf. Process. Syst.*, 2010.

[32] D. P. Bertsekas, "Approximate policy iteration: A survey and some new methods," *J. Control Theory Appl.*, vol. 9, no. 3, pp. 310–335, Aug. 2011.

[33] A. Farahmand, M. Ghavamzadeh, S. Mannor, and C. Szepesvári, "Regularized policy iteration," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 21, 2008.

[34] J. Z. Kolter and A. Y. Ng, "Regularization and feature selection in least-squares temporal difference learning," in *Proc. 26th Annu. Int. Conf. Mach. Learn.*, Jun. 2009, pp. 521–528.

[35] T. Jung and D. Polani, "Least squares SVM for least squares TD learning," in *Proc. ECAI*. Princeton, NJ, USA: CiteSeerX, 2006, pp. 499–503.

[36] X. Xu, D. Hu, and X. Lu, "Kernel-based least squares policy iteration for reinforcement learning," *IEEE Trans. Neural Netw.*, vol. 18, no. 4, pp. 973–992, Jul. 2007.

[37] G. Taylor and R. Parr, "Kernelized value function approximation for reinforcement learning," in *Proc. 26th Annu. Int. Conf. Mach. Learn.*, Jun. 2009, pp. 1017–1024.

[38] S. Kakade and J. Langford, "Approximately optimal approximate reinforcement learning," in *Proc. 19th Int. Conf. Mach. Learn.*, 2002, pp. 267–274.

[39] J. Baxter and P. L. Bartlett, "Infinite-horizon policy-gradient estimation," *J. Artif. Intell. Res.*, vol. 15, pp. 319–350, Nov. 2001.

[40] S. M. Kakade, "A natural policy gradient," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 14, 2001.

[41] J. A. Bagnell and J. Schneider, "Covariant policy search," Tech. Rep., 2003.

[42] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz, "Trust region policy optimization," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 1889–1897.

[43] B. Yu, "Rates of convergence for empirical processes of stationary mixing sequences," *Ann. Probab.*, vol. 22, no. 1, pp. 94–116, Jan. 1994.

[44] D. S. Modha and E. Masry, "Minimum complexity regression estimation with weakly dependent observations," *IEEE Trans. Inf. Theory*, vol. 42, no. 6, pp. 2133–2145, 1996.

[45] H. Hang and I. Steinwart, "Fast learning from α-mixing observations," *J. Multivariate Anal.*, vol. 127, pp. 184–199, May 2014.

[46] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. 27th Int. Conf. Mach. Learn. (ICML)*, 2010, pp. 807–814.

[47] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[48] B. Scherrer, M. Ghavamzadeh, V. Gabillon, B. Lesner, and M. Geist, "Approximate modified policy iteration and its application to the game of Tetris," *J. Mach. Learn. Res.*, vol. 16, no. 49, pp. 1629–1676, 2015.

[49] A. Lazaric, M. Ghavamzadeh, and R. Munos, "Analysis of classification-based policy iteration algorithms," *J. Mach. Learn. Res.*, vol. 17, no. 19, pp. 1–30, 2016.

[50] A.-M. Farahmand, M. Ghavamzadeh, C. Szepesvári, and S. Mannor, "Regularized policy iteration with nonparametric function spaces," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 4809–4874, 2016.

[51] T. Xie and N. Jiang, "q^* approximation schemes for batch reinforcement learning: A theoretical comparison," in *Proc. 36th Conf. Uncertainty Artif. Intell. (UAI)*, vol. 124, 2020, pp. 550–559.

[52] T. Xie and N. Jiang, "Batch value-function approximation with only realizability," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 11404–11413.

[53] J. Chen and N. Jiang, "Information-theoretic considerations in batch reinforcement learning," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 1042–1051.

[54] D. Yarotsky, "Error bounds for approximations with deep ReLU networks," *Neural Netw.*, vol. 94, pp. 103–114, Oct. 2017.

[55] D. Yarotsky, "Optimal approximation of continuous functions by very deep ReLU networks," in *Proc. Conf. Learn. Theory*, 2018, pp. 639–649.

[56] D.-X. Zhou, "Universality of deep convolutional neural networks," *Appl. Comput. Harmon. Anal.*, vol. 48, no. 2, pp. 787–794, Mar. 2020.

[57] Z. Shen, H. Yang, and S. Zhang, "Nonlinear approximation via compositions," *Neural Netw.*, vol. 119, pp. 74–84, Nov. 2019.

[58] Z. Shen, "Deep network approximation characterized by number of neurons," *Commun. Comput. Phys.*, vol. 28, no. 5, pp. 1768–1811, Jun. 2020.

[59] J. Lu, Z. Shen, H. Yang, and S. Zhang, "Deep network approximation for smooth functions," *SIAM J. Math. Anal.*, vol. 53, no. 5, pp. 5465–5506, Jan. 2021.

[60] Y. Jiao, G. Shen, Y. Lin, and J. Huang, "Deep nonparametric regression on approximately low-dimensional manifolds," 2021, *arXiv:2104.06708*.

[61] H. Lin and S. Jegelka, "ResNet with one-neuron hidden layers is a universal approximator," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018.

[62] F. Fan, J. Xiong, and G. Wang, "Universal approximation with quadratic deep networks," *Neural Netw.*, vol. 124, pp. 383–392, Apr. 2020.

[63] P. Kidger and T. Lyons, "Universal approximation with deep narrow networks," in *Proc. Conf. Learn. Theory*, 2020, pp. 2306–2327.

[64] M. Mohri and A. Rostamizadeh, "Rademacher complexity bounds for non-I.I.D. processes," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 21, 2008, pp. 1097–1104.

[65] I. Steinwart, D. Hush, and C. Scovel, "Learning from dependent observations," *J. Multivariate Anal.*, vol. 100, no. 1, pp. 175–194, Jan. 2009.

[66] M. Mohri and A. Rostamizadeh, "Stability bounds for stationary φ-mixing and β-mixing processes," *J. Mach. Learn. Res.*, vol. 11, no. 2, pp. 789–814, 2010.

[67] L. Ralaivola and M.-R. Amini, "Entropy-based concentration inequalities for dependent variables," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2436–2444.

[68] H. Hang and I. Steinwart, "A Bernstein-type inequality for some mixing processes and dynamical systems with an application to learning," *Ann. Statist.*, vol. 45, no. 2, pp. 708–743, Apr. 2017.

[69] A. Roy, K. Balasubramanian, and M. A. Erdogdu, "On empirical risk minimization with dependent and heavy-tailed data," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021.

[70] M. Anthony, P. L. Bartlett, and P. L. Bartlett, *Neural Network Learning: Theoretical Foundations*, vol. 9. Cambridge, U.K.: Cambridge Univ. Press, 1999.

[71] K. Oono and T. Suzuki, "Approximation and non-parametric estimation of ResNet-type convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 4922–4931.

[72] B. Bauer and M. Kohler, "On deep learning as a remedy for the curse of dimensionality in nonparametric regression," *Ann. Statist.*, vol. 47, no. 4, pp. 2261–2285, Aug. 2019.

[73] J. Schmidt-Hieber, "Nonparametric regression using deep neural networks with ReLU activation function," *Ann. Statist.*, vol. 48, no. 4, pp. 1875–1897, Aug. 2020.

[74] R. Nakada and M. Imaizumi, "Adaptive approximation and generalization of deep neural network with intrinsic dimensionality," *J. Mach. Learn. Res.*, vol. 21, no. 1, pp. 7018–7055, 2020.

[75] M. H. Farrell, T. Liang, and S. Misra, "Deep neural networks for estimation and inference," *Econometrica*, vol. 89, no. 1, pp. 181–213, 2021.

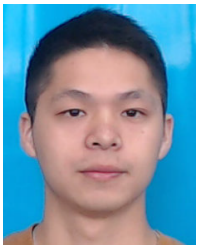
[76] H. Dehling and W. Philipp, "Empirical process techniques for dependent data," in *Empirical Process Techniques for Dependent Data*. Berlin, Germany: Springer, 2002, pp. 3–113.

[77] P. L. Bartlett, N. Harvey, C. Liaw, and A. Mehrabian, "Nearly-tight VC-dimension and pseudodimension bounds for piecewise linear neural networks," *J. Mach. Learn. Res.*, vol. 20, no. 1, pp. 2285–2301, 2019.

[78] R. Munos, "Error bounds for approximate policy iteration," in *Proc. ICML*, vol. 3, 2003, pp. 560–567.

[79] O. Kallenberg and O. Kallenberg, *Foundations of Modern Probability*, vol. 2, Springer, 1997.

[80] J. L. Doob, *Measure Theory*, vol. 143, Springer, 2012.
[81] P. R. Halmos, *Measure Theory*, vol. 18, Springer, 2013.
[82] J. Fan, Z. Wang, Y. Xie, and Z. Yang, "A theoretical analysis of deep Q-learning," in *Proc. 2nd Annu. Conf. Learn. Dyn. Control*, 2020, pp. 486–489.

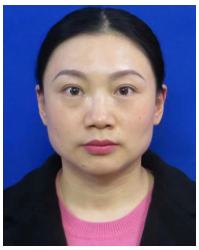


Lican Kang received the B.Sc. degree in applied mathematics from Jiangxi Normal University, Nanchang, China, in 2015, and the Ph.D. degree in statistics from Wuhan University, Wuhan, China, in 2021.

He is currently a Research Fellow with the Duke-NUS Medical School, Singapore. He has authored more than ten research articles including IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, *SIAM Journal on Control and Optimization*, *Information Sciences*, *Computational Statistics*, *Journal of Statistical Computation and Simulation*, *Computational Statistics & Data Analysis*, *Statistics & Probability Letters*, *East Asian Journal on Applied Mathematics*, and *Communications in Computational Physics*. His current research interests include machine learning, high-dimensional data analysis, and statistical computing.

Yuhui Liu received the B.S. and M.S. degrees in statistics from Wuhan University, Wuhan, China, in 2013 and 2016, respectively, where he is currently pursuing the Ph.D. degree with the School of Mathematics and Statistics.

His research interests include machine learning and regression analysis.



Yuan Luo received the B.Sc. degree in applied mathematics from Shangqiu Normal University, Shangqiu, China, in 2008, and the M.Sc. degree in applied mathematics from the Wuhan University of Science and Technology, Wuhan, China, in 2011. She is currently pursuing the Ph.D. degree in statistics with Wuhan University, Wuhan.

She has authored several research articles including IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, *Information Sciences*, *Journal of Statistical Computation and Simulation*, and *East Asian Journal on Applied Mathematics*. Her current research interests include machine learning and statistical computing.

Jerry Zhijian Yang received the B.S. and M.S. degrees from Peking University, Beijing, China, in 1999 and 2001, respectively, and the Ph.D. degree in applied and computational mathematics from Princeton University, Princeton, NJ, USA, in 2006.

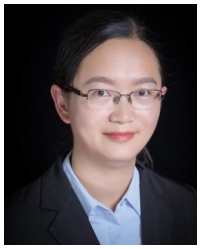
He completed his post-doctoral research at the California Institute of Technology, Pasadena, CA, USA, in 2008. He is currently a Full Professor with the School of Mathematics and Statistics as well as the Artificial Intelligence Institute, Wuhan University, Wuhan, China. He has authored or coauthored research articles, including *Numerische Mathematik*, *Physical Review B*, *SIAM Multiscale Modeling and Simulation*, *SIAM Journal on Control and Optimization*, *Journal of Computational Physics*, *The Journal of Chemical Physics*, *International Journal for Numerical Methods in Engineering*, and *Communications in Computational Physics*. His research interests include multiscale modeling and simulation, machine learning, and scientific computing.

Dr. Yang is currently the President of East Asia Section of SIAM.

Han Yuan received the double B.S. degrees in biology and applied mathematics from Nankai University, Tianjin, China. He is currently pursuing the Ph.D. degree in health data science with the Duke-NUS Medical School, Singapore, with a focus on developing innovative methodologies that inform evidence-based decision-making.

He worked as a Consultant with Harvard University, Cambridge, MA, USA, and a Visiting Researcher with the University of Zurich, Zürich, Switzerland, and Duke University, Durham, NC, USA. He has authored several articles in *Artificial Intelligence in Medicine*, *Journal of Biomedical Informatics*, and *Journal of Experimental & Clinical Cancer Research*.

Mr. Yuan serves as a Referee for *Expert Systems With Applications*, *Data Science Journal*, and Machine Learning for Health Symposium.



Chang Zhu received the M.D. degree from the Huazhong University of Science and Technology, Wuhan, China, in 2008.

From 2013 to 2015, she went to the University of Texas Health Science Center at San Antonio, San Antonio, TX, USA, to engage in post-doctoral research. She is currently an Associate Professor with the Department of Anesthesiology, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology. She has authored or coauthored more than 20 articles in academic journals including *Stem Cell Reports*, *Journal of Thrombosis and Thrombolysis*, and *Neurochemistry International*. Her research interests include machine learning in anesthesia, anesthesia, and brain protection.