Width and Depth Guidelines for Deep Q-Learning: A Function Approximation Perspective

Anonymous Author(s)
Affiliation
Address
email

Abstract

This paper provides a theoretical study of the successful deep Q-networks (DQN) framework from the perspective of function class and neural networks architectures (e.g., width and depth). Motivated by practical deep neural networks used in DQN under an online setting with T episodes, we focus on Besov and Barron function spaces in approximating an α-smooth Q-function in a d-dimensional feature space. Based on a probabilistic sufficient exploration (PSE) assumption, which is attainable in practice, we prove that scaling the width $m = \widetilde{\mathcal{O}}(T^{\frac{d}{2\alpha+d}})$ and the depth $L = \widetilde{\mathcal{O}}(1)$ of the neural network for DQN is sufficient for learning with sublinear regret in Besov spaces. Moreover, for a two layer neural network in the Barron space, scaling the width even in sublinearly in T is sufficient. While increasing the model capacity improves the DQN efficiency in both cases, our theory suggests that the width has a more prominent effect than depth. In particular, for low smooth Q-function, the theory suggests that the depth should be chosen as the logarithm of the width. This matches the practical DQN setting. Our analysis relies on transforming the estimation of temporal difference error to a generalization problem under the conditionally independent but non-identically distributed data setting over the averaged empirical measure.

1 Introduction

Efficient reinforcement learning (RL) under the large (or even infinite) state space and action space setting is increasingly important and relevant challenge [1, 2, 3]. One of the first successful approaches towards this problem is the deep Q-network (DQN) [4, 5] framework, which deploys powerful nonlinear function approximation techniques via Deep Neural Networks (DNNs) [6] to concisely approximate state and action spaces.

Despite its impressive practical success, there is still a gap between practical uses and theoretical understanding on DQN under the online setting, e.g., the nonlinear Q-function approximation and how much the width and depth of DNNs are enough. In contrast, many works directly focus on linear function approximation [7, 8], linear mixtures [9, 10].

Existing non-linear function approximation results on RL are largely based on neural tangent kernel (NTK) [11, 12], Bellman rank [13, 14], and Eluder dimension [15, 16, 17]. Nevertheless, these approaches fail in truly capturing the highly non-linear properties of DQN. Indeed, NTK (or lazy training [18]) essentially works in a linear regime [19, 20, 21], and can not efficiently learn even a single ReLU neuron [22, 23, 24] as it requires $\Omega(\varepsilon^{-d})$ samples to achieve ε approximation error, where d is the (original) or transformed feature dimension input.

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

Interestingly, even for two-layer neural networks, the corresponding Eluder dimension is at least in an exponential order [25, 26], which appears difficult to fully demonstrate the success of practical DQN both theoretically and empirically. Regarding to ensuring efficient deep Q-learning for sublinear regret, the width of neural networks under the NTK regime [12] is required to be $m = \Omega(T^{13})$, where T is the number of episodes, which does not match the DQNs used in practice with small width/depth under large episodes [4, 27].

To bridge the large theory-practice gap, we make a first attempt to study DQN with a slight modification but broadly capture its key features under the online setting. Our analysis framework is equipped with two-layer as well as deep fully-connected neural networks in the respective Barron and Besov spaces, which can fully capture the properties of Q-functions, e.g., smoothness by neural networks. We also consider a weaker version of the independent and identically distributed (iid) assumption by experience replay in DQN, which could be of interest beyond DQN.

Our results allow for reasonably finite width and depth in practice, give precise characterization of these parameters to obtain sublinear regret, and demonstrate the relationship between the problem-dependent smoothness of Q-function and regret bounds. We believe that these results can be viewed as a guide for practitioners to choose the architecture in practical implementations of DQN.

1.1 Technical challenges and contributions

Most previous RL theory results on function approximation in the online setting work with "optimism in the face of uncertainty" principle for exploration, leading to a series of upper confidence bound (UCB)-type algorithms. However, conceptually, optimism is sometimes too aggressive and UCB-style algorithms can suffer exponential sample complexity even for nonlinear bandits [26]. Empirical model-based RL has also not found optimism to be very useful [28].

In the RL setting, linear/kernel function approximation [7, 12, 29] depends on a known feature mapping or the NTK matrix to design a UCB-type bonus function for optimism, which appears invalid for DNNs as the used confidence ellipsoid and elliptical potential lemma are not applicable for data-dependent feature mapping.

To avoid explicitly designing a bonus function, Thompson sampling [30, 31] appears promising in a Bayesian perspective by using randomized (i.e., perturbed) versions of the estimated model or value function [32]. Nevertheless, the bonus function is still implicitly included in confidence estimate of perturbations in their UCB-style algorithm.

This work analyzes an algorithm closely related to DQN in general function classes and provides a new regret analysis framework under a realistic and attainable exploratory assumption, dubbed as probabilistic sufficient exploration (PSE), to ensure how much width and depth are enough for efficient RL. We remark that PSE can be satisfied in practical cases of interest when employing DQN. Designing a provably efficient exploration mechanism is beyond the scope of our work for understanding DQN but it remains an interesting future direction.

The technical contribution in our analysis is how to control the temporal difference (TD) error without bonus function design. By a measure transform, we analyze the TD error in an L^2-integrable space wrt historical state-action pairs, and then transform to generalization bounds under the independent but non-identically distributed data setting. To this end, we need to employ proper function spaces beyond linear/kernel Markov decision processes (MDP) in our analysis: one is the Barron space [33], as the largest[1] function space for two-layer ReLU neural networks without curse of dimensionality [34]; the other is the Besov space [35, 36] for deep ReLU neural networks, which allows for continuous/non-continuous functions and thus is more general than Hölder spaces used in offline RL [37]. Under these proper hypothesis spaces, such generalization problem benefits from the (nearly) unbiased estimation by the target network in DQN, and thus can be addressed by uniform convergence via Rademacher complexity of these spaces under the averaged empirical measure.

[1]The terminology "largest" means two-layer neural networks with bounded norm and functions in the Barron space can be efficiently represented without curse of dimensionality, refer to Appendix A for details.

Under attainable assumptions, our results show that (i) the problem-dependent smoothness of Q-function affects the deep Q-learning efficiency, which can be improved by the model complexity. We use α as a parameter indicating the smoothness degree of Q-function. A larger α indicates smoother functions. (ii) for two-layer neural networks, the width $m = \Omega(\sqrt{T})$ suffices to ensure optimal minimax $\widetilde{\mathcal{O}}(\sqrt{T})$ regret. (iii) for deep neural networks, the width $m = \widetilde{\mathcal{O}}(T^{\frac{d}{2\alpha+d}})$ and the depth $L = \widetilde{\mathcal{O}}(1)$ are enough for sublinear regret, where $\widetilde{\mathcal{O}}(\cdot)$ omits the $\log$ terms. Overall, we find that the width has a more prominent effect than depth. Furthermore, our regret bounds can be (nearly) independent of the feature dimension d, supporting the premise of practical, high-dimensional RL.

1.2 Related work

Theoretical understanding of DQN is due to [38, 37, 39], which regards DQN as a neural fitted Q-iteration algorithm [40] under an offline setting. Recent work on neural network function approximation beyond NTK (or the Eluder dimension) mainly restrict on the generative setting [41, 42] by assuming a simulator in which the agent can require any state and action, and the offline setting [37, 43]. In sequel, we review RL with function approximation under the online setting that DQN falls into this regime.

RL with linear/kernel function approximation: RL with linear function approximation results in regret bounds $\widetilde{\mathcal{O}}(\sqrt{d^3H^3T})$ [7] and $\widetilde{\mathcal{O}}(dH^2\sqrt{T})$ [29], where H is the length of each episode; and linear mixtures with regret $\widetilde{\mathcal{O}}(d\sqrt{H^3T})$ [9] and improved to $\widetilde{\mathcal{O}}(dH\sqrt{T})$ [10]; kernel regime with regret $\widetilde{\mathcal{O}}(\delta_{\mathcal{F}}H^2\sqrt{T})$ [29, 12], where $\delta_{\mathcal{F}}$ is the intrinsic complexity (e.g., effective dimension) of the function class RKHS $\mathcal{F}$. The above bounds are based on confidence ellipsoid to quantify the uncertainty in an explicit bonus function by feature mapping/kernel matrix; while Thompson sampling [31, 30] utilizes an implicit bonus function in probability estimation on uncertainty quantification, which leads to $\widetilde{\mathcal{O}}(d^2H^2\sqrt{T})$ [32] regret in linear function approximation.

RL with general function approximation: One prototypical scheme uses the Eluder dimension [15], which measures the degree of dependence among action rewards, resulting in $\widetilde{\mathcal{O}}(\texttt{poly}(\delta_{\mathcal{F}}H)\sqrt{T})$ regret [16, 17], where the complexity $\delta_{\mathcal{F}}$ depends on the Eluder dimension. Bellman rank [13] is also another metric for general function approximation [14] by measuring the error of the function class under the Bellman operator. Jin et al. [44] introduce a new metric, Bellman Eluder dimension, by combining these two metrics, and achieve $\widetilde{\mathcal{O}}(H\sqrt{\delta_{\mathcal{F}}T})$, where $\delta_{\mathcal{F}}$ depends on this metric.

Overall, none of the above metrics can handle the nonlinear spaces that concern us.

2 Background and preliminaries

In this section, we provide some preliminaries with respect to online reinforcement learning based on episodic MDPs, function spaces of two (and deep)-layer ReLU neural networks.

Notation: We denote by $a(n) \lesssim b(n)$: there exists a positive constant c independent of n such that $a(n) \leqslant cb(n)$. We use the shorthand $[n] := \{1, 2, \ldots, n\}$ for some positive n. Let $\mathcal{X} = [0,1]^d$ be a domain of the functions, we denote the L^p-integrable space by $L^p(\mathcal{X})$ endowed by the norm $\|f\|_{L^p(\mathcal{X})} = \big(\int_{\mathcal{X}} |f(\boldsymbol{x})|^p \mathrm{d}\boldsymbol{x}\big)^{1/p}$, and the μ-integrable L^p space by $L^p(\mathrm{d}\mu)$ for a probability measure μ on $\mathcal{X}$ and the norm is given by $\|f\|_{L^p(\mathrm{d}\mu)} = \big(\int_{\mathcal{X}} |f(\boldsymbol{x})|^p \mathrm{d}\mu(\boldsymbol{x})\big)^{1/p}$.

2.1 Episodic Markov decision processes

A (finite-horizon) episodic MDPs is denoted as $\mathrm{MDP}(\mathcal{S}, \mathcal{A}, H, \mathbb{P}, r)$, where $\mathcal{S}$ is the state space with possibly infinite states; $\mathcal{A}$ is the action space, often with finite actions; H is the number of steps in each episode; $\mathbb{P} := \{\mathbb{P}_h\}_{h=1}^H$ is the Markov transition kernel with the transition probability $\mathbb{P}_h(\cdot|s,a)$ on action a taken at state $s \in \mathcal{S}$ in the h-th step; the reward functions $r := \{r_h\}_{h=1}^H$ are assumed to be deterministic. For notational simplicity, denote $\mathcal{X} = \mathcal{S} \times \mathcal{A}$ and $\boldsymbol{x} = (s,a)$, we assume $\mathcal{X} = [0,1]^d$ as a compact space of $\mathbb{R}^d$ and $r_h : \mathcal{S} \times \mathcal{A} \to [0,1]$ at h-th step.

A non-stationary policy π is a collection of H functions $\pi := \{\pi_h : \mathcal{S} \to \mathcal{A}\}_{h=1}^H$. Given a policy π, the (state) value function $V_h^\pi \colon \mathcal{S} \to [0, H]$ is defined as the expected cumulative reward of the MDP starting from step $h \in [H]$, i.e., $V_h^\pi(s) = \mathbb{E}_\pi\big[\sum_{h'=h}^H r_{h'}(s_{h'}, a_{h'})\big| s_h = s\big], \forall s \in \mathcal{S}, h \in [H]$ where $\mathbb{E}_\pi[\cdot]$ denotes the expectation with respect to the randomness of the trajectory $\{(s_h, a_h)\}_{h=1}^H$ obtained by the policy π. Likewise, the action-value function $Q_h^\pi : \mathcal{S} \times \mathcal{A} \to [0, H]$ is defined as $Q_h^\pi(s, a) = \mathbb{E}_\pi\big[\sum_{h'=h}^H r_{h'}(s_{h'}, a_{h'}) \,\big|\, s_h = s,\, a_h = a\big]$.

Moreover, since the action space and episode length are both finite, there always exists an optimal policy $\pi^\star$ such that $V_h^\star(s) = \sup_\pi V_h^\pi(s)$ for all $s \in \mathcal{S}$ and $h \in [H]$ [45]. To simplify the notation, denote $(\mathbb{P}_h V)(s, a) := \mathbb{E}_{s' \sim \mathbb{P}_h(\cdot|s,a)}[V(s')]$ and the Bellman operator $(\mathbb{T}_h V)(s, a) = r_h(s, a) + (\mathbb{P}_h V)(s, a)$ for any measurable function $V \colon \mathcal{S} \to [0, H]$. Using this notation, the Bellman equation associated with a policy π can be formulated as

$$Q_h^\pi(s, a) = (\mathbb{T}_h V_{h+1}^\pi)(s, a), \qquad V_h^\pi(s) = \langle Q_h^\pi(s, \cdot), \pi_h(\cdot|s)\rangle_{\mathcal{A}}, \qquad V_{H+1}^\pi(s) = 0\,. \tag{1}$$

Similarly, the Bellman optimality equation is given by

$$Q_h^\star(s, a) = (\mathbb{T}_h V_{h+1}^\star)(s, a), \qquad V_h^\star(s) = \max_{a \in \mathcal{A}} Q_h^\star(s, a), \qquad V_{H+1}^\star(s) = 0\,. \tag{2}$$

Accordingly, the optimal policy $\pi^\star$ is the greedy policy with respect to $\{Q_h^\star\}_{h=1}^H$. Due to $r_h \in [0, 1]$ yielding $Q_h \in [0, H - h + 1]$, we introduce the *projection operator* $\mathcal{K}(f)(s_h, a_h) := \min\{f(s_h, a_h), H-h+1\}_+$ in our analysis to obtain the tighter bound. Hence the Bellman optimality operator $\mathbb{T}_h^\star$ is defined as

$$(\mathbb{T}_h^\star Q)(s_h, a_h) = r_h(s_h, a_h) + \mathbb{E}_{s_{h+1} \sim \mathbb{P}_h(\cdot|s_h,a_h)} \mathcal{K}[\max_{a \in \mathcal{A}} Q(s_{h+1}, a)], \qquad \forall\, Q : \mathcal{S} \times \mathcal{A} \to [0, H]\,.$$

By definition, the Bellman equation in Eq. (2) is equivalent to $Q_h^\star = \mathbb{T}_h^\star Q_{h+1}^\star, \forall h \in [H]$.

In the **online setting**, the goal is to learn the optimal policy $\pi^\star$ by minimizing the cumulative regret under the interaction with the environment over a number of episodes. For any policy π, the difference between V_1^π and $V_1^\star$ quantifies its sub-optimality. Thus, after T (fixed but large) episodes, the total (expected) regret is defined as $\mathrm{Regret}(T) = \sum_{t=1}^T \big[V_1^\star(s_1^t) - V_1^{\pi^t}(s_1^t)\big]$, where π^t is the policy executed in the t-th episode and s_1^t is the initial state.

2.2 Barron spaces and Besov spaces

We begin with introduction on the Barron space associated with two-layer ReLU neural networks and then give an overview of Besov spaces for deep ReLU neural networks. More details refer to Appendix A. For description simplicity, we focus on the ReLU activation function in this work.

Barron spaces: A two-layer neural network with m neurons can be represented as[2] $f(\boldsymbol{x}) = \frac{1}{m}\sum_{k=1}^m b_k \sigma(\boldsymbol{w}_k^\top \boldsymbol{x})$, where $\sigma(x) = \max\{x, 0\}$ is the ReLU activation function used in this work. If the parameters $\{(b_k, \boldsymbol{w}_k)\}_{k=1}^m$ are associated with a signed Radon measure with finite total variation [22, 46], this function space is in fact a Barron space $\widetilde{\mathcal{P}}$ [33], of which the Barron norm can be roughly substituted by with the ℓ_1-path norm, i.e., $\|f\|_{\widetilde{\mathcal{P}}} \leqslant \|f\|_{\mathcal{P}} := \frac{1}{m}\sum_{k=1}^m |b_k| \|\boldsymbol{w}_k\|_1 \leqslant 2\|f\|_{\widetilde{\mathcal{P}}}$. Accordingly, it is natural to use $\|f\|_{\mathcal{P}}$ to denote the Barron norm, as the discrete version.

The Barron space can be regarded as the *largest* function space for two layer neural networks in two folds [33, 34]: 1) *direct approximation:* Any function in Barron spaces can be efficiently approximated by two-layer neural networks with bounded ℓ_1-path norm; 2) *inverse approximation:* Any continuous function that can be efficiently approximated by two-layer neural networks with bounded ℓ_1-path norm belongs to a Barron space.

Besov spaces: Different from two-layer neural networks in Barron spaces, the complete function space of deep ReLU neural networks is still open. Previous work in approximation theory focuses on the "smoothness" of the function, e.g., Hölder spaces [47, 37] and Sobolev spaces [48]. Here we consider the concept of α-smooth from modulus of smoothness [36], *cf.*, Appendix A.

[2] It allows for neural networks with bias, by taking $(\boldsymbol{x}, 1) \in \mathbb{R}^{d+1}$ as $\boldsymbol{x}$ and $(\boldsymbol{w}, b) \in \mathbb{R}^{d+1}$ as $\boldsymbol{w}$.

Based on this, we consider a more general function space beyond Hölder spaces and Sobolev spaces, i.e., Besov spaces [35, 36], which allows for spatially inhomogeneous smoothness with spikes and jumps. The Besov space is defined by $\mathcal{B}^{\alpha}_{p,q}(\mathcal{X}) = \{f \in L^p(\mathcal{X}) \mid \|f\|_{\mathcal{B}^{\alpha}_{p,q}} < \infty\}$, where the Besov norm is $\|f\|_{\mathcal{B}^{\alpha}_{p,q}} := \|f\|_{L^p(\mathcal{X})} + |f|_{\mathcal{B}^{\alpha}_{p,q}}$. The smoothness parameter α indicates which function at a certain smoothness degree can be represented. For example, if $\alpha > d/p$, then the related Besov space is continuously embedded in the set of the continuous functions; if $\alpha < d/p$, then the functions in the Besov space are no longer continuous. The formal definition and relations to Hölder spaces and Sobolev spaces are deferred to Appendix A.

Mixed smooth Besov spaces: Learning in Besov spaces [49] still faces the *curse of dimensionality* since the Besov space is too general. To tackle this issue, we use the mixed Besov space $\widetilde{\mathcal{B}}^{\alpha}_{p,q}(\mathcal{X})$ [50, 36], as defined by $\|f\|_{\widetilde{\mathcal{B}}^{\alpha}_{p,q}} := \|f\|_{L^p(\mathcal{X})} + |f|_{\widetilde{\mathcal{B}}^{\alpha}_{p,q}}$, and consider the differentiation over all coordinates, which does not heavily depend on the feature dimension d. Intuitively, the Besov space ensures a function with boundedness of derivatives, of which the "sum" of the orders are smaller than α; while the mixed smooth Besov space allows for the "max" of the orders of derivatives over coordinates smaller than α. Hence, we obtain the approximation performance and the separation on the *dimension-dependent* convergence rate for sublinear regret.

3 Vanilla DQN algorithm via value iteration

In this section, we lay out our algorithm 1 of vanilla DQN via value iteration by DNNs. After initialization, at t-th episode, at h-th time step, we have observed $t-1$ transition tuples, $\{(s_h^\tau, a_h^\tau, s_{h+1}^\tau)\}_{\tau=1}^{t-1}$ and attempt to estimate $\{Q_h^\star\}_{h=1}^H$ by the following least squares regression problem $\widehat{Q}_h^t = \operatorname{argmin}_{f\in\mathcal{F}} \widehat{\mathcal{E}}_h^t(f)$ with

$$\widehat{\mathcal{E}}_h^t(f) = \frac{1}{t-1}\sum_{\tau=1}^{t-1}\left[\mathcal{K}f(s_h^\tau, a_h^\tau) - r_h(s_h^\tau, a_h^\tau) - V_{h+1}^t(s_{h+1}^\tau)\right]^2 , \tag{3}$$

where $\mathcal{K}$ is the projection operator defined before for tight bound. The function class $\mathcal{F}$ is given by $\mathcal{F} = \mathcal{F}_1 \times \cdots \times \mathcal{F}_H$, including $\mathcal{F}_{\texttt{SNN}}$ for two-layer (Shallow) ReLU neural networks and $\mathcal{F}_{\texttt{DNN}}$ for deep ReLU neural networks as below.

$$\mathcal{F}_{\texttt{SNN}} = \left\{ f : [0,1]^d \to \mathbb{R} \middle| f(\boldsymbol{x}) = \frac{1}{m}\sum_{k=1}^m b_k \sigma(\boldsymbol{w}_k^\top \boldsymbol{x} + c_k), \|f\|_{\mathcal{P}} \leqslant B \right\} , \tag{4}$$

where $B > 0$ is the ℓ_1-path norm constraint parameter, and deep ReLU neural networks [36] as

$$\begin{aligned}\mathcal{F}_{\texttt{DNN}}(L,m,S,B) := \Bigg\{ f : [0,1]^d \to \mathbb{R} \Bigg| f(\boldsymbol{x}) = (\boldsymbol{W}^{(L)}\sigma(\cdot) + b^{(L)}) \circ \cdots \circ (\boldsymbol{W}^{(1)}\boldsymbol{x} + \boldsymbol{b}^{(1)}) , \\ \sum_{i=1}^L (\|\boldsymbol{W}^{(i)}\|_0 + \|\boldsymbol{b}^{(i)}\|_0) \leqslant S, \ \max_i \|\boldsymbol{W}^{(i)}\|_\infty \vee \|\boldsymbol{b}^{(i)}\|_\infty \leqslant B \Bigg\} ,\end{aligned} \tag{5}$$

where the weight parameters are $\boldsymbol{W}^{(1)} \in \mathbb{R}^{m\times d}$, $\boldsymbol{W}^l \in \mathbb{R}^{m\times m}$, $\forall l \in \{2,3,\ldots,L-1\}$, and $\boldsymbol{W}^L \in \mathbb{R}^m$; the bias parameter are $\boldsymbol{b}^{(l)} \in \mathbb{R}^m$, $\forall l \in [L-1]$ and $b^{(L)} \in \mathbb{R}$. The depth L, the width m, the sparsity parameter S and the norm parameter B can be further determined to achieve good approximation and estimation performance.

In our setting, we use the historical $t-1$ episodic data $\{(s_h^\tau, a_h^\tau, s_{h+1}^\tau)\}_{\tau=1}^{t-1}$ for value function approximation by neural networks. It can be found that, for each time step h, these $t-1$ examples are merely conditionally independent snf non-identically distributed. We do not employ experience replay in DQN for the iid assumption, which could expand the application scope of DQN. Formally, we assume each state-action pair is sampled from a respective (unknown) probability measure, i.e., $(s_h^\tau, a_h^\tau) \sim \mu_h^\tau, \forall \tau \in [t-1]$, where $\mu_h^\tau \in \mathscr{P}(\mathcal{S}\times\mathcal{A})$ is from the collection of all probability distribution on $\mathcal{S}\times\mathcal{A}$. Taking the averaged measure $\bar{\mu}_h^t := \frac{1}{t-1}\sum_{\tau=1}^{t-1}\mu_h^\tau$, the expected risk of Eq. (3) is defined as

$$\mathcal{E}_h(f) = \mathbb{E}_{(s_h,a_h)\sim\bar{\mu}_h^t, s_{h+1}\sim\mathbb{P}_h(\cdot|s_h,a_h)}\left[\mathcal{K}f(s_h,a_h) - r_h(s_h,a_h) - V_{h+1}^t(s_{h+1})\right]^2 , \tag{6}$$

Algorithm 1 Online Fitted Q Iteration with Deep Neural Networks

Input: Function class $\mathcal{F}$, the number of episodes T.
for episode $t = 1, \ldots, T$ **do**
  Receive the initial state s_1^t.
  Set V_{H+1}^t as the zero function.
  for step $h = H, \ldots, 1$ **do**
    Obtain $\widehat{Q}_h^t := \operatorname{argmin}_{f \in \mathcal{F}} \widehat{\mathcal{E}}_h^t(f)$ in Eq. (3).
    Obtain $Q_h^t = \min\{\widehat{Q}_h^t, H - h + 1\}^+$ and $V_h^t(\cdot) = \max_{a \in \mathcal{A}} Q_h^t(\cdot, a)$.
  end for
  for step $h = 1, \ldots, H$ **do**
    Take $a_h^t = \operatorname{argmax} Q_h^t(s_h^t, a)$.
    Observe the reward $r_h(s_h^t, a_h^t)$ and obtain the next state s_{h+1}^t.
  end for
end for

which is the exact Q-learning update [4]. Note that, Eq. (3) is a biased estimator of Eq. (6) due to an extra variance term caused by the stochastic transition [51]. However, this issue can be avoided if some schemes are employed, e.g., double sampling [52], simulator [42], and deterministic transition that is used in Atari [53] in practice. In fact, DQN [4] uses target networks to reduce the biased estimate or even eliminate it [37], and some variants [54] can also reduce the biased estimate and performs well without target networks. Accordingly, both practical uses and theoretical guarantees provide a good justification to not strictly distinguish this biased estimation in our analysis for better understanding DQN.

4 Theoretical guarantees of efficient value iteration via DNNs

In this section, we present our results for value iteration achieving sublinear regret under two-layer and deep ReLU neural networks.

To begin, we need one assumption on $\bar{\mu}_h^t$ to make it applicable to online setting. Denote $\mathcal{G}_\epsilon := \{(s,a) : |f(s,a)| \geqslant \|f\|_\infty - \epsilon\}$ for any $(s,a) \in \mathcal{S} \times \mathcal{A}$, where $\epsilon \leqslant \|f\|_\infty$. It is clear that the cardinality of $\mathcal{G}_\epsilon$ is a non-increasing function of ϵ. For any probability measure μ on $\mathcal{S} \times \mathcal{A}$, we have

$$\|f\|_{L^p(\mathrm{d}\mu)} \geqslant \left(\int_{\mathcal{G}_\epsilon} (\|f\|_\infty - \epsilon)^p \mathrm{d}\mu \right)^{1/p} = (\|f\|_\infty - \epsilon)[\mu(\mathcal{G}_\epsilon)]^{1/p}, \tag{7}$$

which is also valid to our averaged measure $\bar{\mu}_h^t$ based on the historical state-action pairs. We make a fair and attainable assumption in practice on $\bar{\mu}_h^t(\mathcal{G}_\epsilon)$ when employing DQN for some specific tasks.

Assumption 1. [Probabilistic sufficient exploration (PSE)] *We assume that* $\bar{\mu}_h^t(\mathcal{G}_\epsilon) > 0$, $\forall \epsilon \geqslant 1/\sqrt{T}$ *hold with probability at least* $1 - \delta$ *for some constant* $\delta \in (0,1)$.

Remark: We discuss the rationale of the PSE assumption in theory and in practice.
(i) In theory, our PSE assumption shares the similar spirit with [41, Theorem 4.8], i.e., at least one good state-action pair can be visited according to the historical data, which admits the minimal requirement on the data distribution with probability. This is weaker than [55, Definition 3.1] that requires a positive visitation probability of each state-action pair. In fact, general neural function approximation without extra assumptions is infeasible even with good representations [56, 57].

(ii) In practice, exploration in DQN depends on massive training data rather than designing UCB-type bonus function, which coincides with the spirit of the PSE assumption. This assumption can be satisfied in several cases of interest: (1) Assumption 1 can be satisfied using ϵ-greedy exploration as suggested in the original DQN algorithm [4]. (2) For deterministic dynamics in discrete states and actions spaces, we can use the first $|\mathcal{S}||\mathcal{A}|/H$ episodes to visit every state-action pair once. At this point we can recover a measure satisfying Assumption 1. This may be feasible in Atari [53] where the frames seen by the agent are not all the possible pixel permutations but only a small subset of them that can be easily recognized and processed by a human. (3) In navigation tasks, Assumption 1 holds when the initial distribution has full support over the state space.

4.1 Efficient value iteration under two-layer ReLU neural networks

In this setting, we consider $\widehat{Q}_h^t = \operatorname{argmin}_{f \in \mathcal{F}_{\texttt{SNN}}} \widehat{\mathcal{E}}_h^t(f)$ in Eq. (3), where $\mathcal{F}_{\texttt{SNN}}$ is the function space of two-layer ReLU neural networks defined in Eq. (4). We give an assumption for two-layer ReLU neural networks, where the Bellman optimality operator maps any bounded value function to a bounded Barron space ball. This assumption is commonly used in RL theory under various function classes [7, 12].

Assumption 2. *Let $\widetilde{R} > 0$ be a fixed constant. Define $\mathcal{P}_{\widetilde{R}} = \{f \in \mathcal{P} : \|f\|_{\mathcal{P}} \leq \widetilde{R}\}$ in the Barron space, and assume that for any $h \in [H]$ and $Q \colon \mathcal{S} \times \mathcal{A} \to [0, H]$, we have $\mathbb{T}_h^\star Q \in \mathcal{P}_{\widetilde{R}}$.*

Remark: Due to $Q_h \in [0, H-h+1]$, the radius R in fact depends on H, and thus we have $\widetilde{R} \asymp H$. Based on this assumption, we have the following regret bounds for two-layer ReLU neural networks, with the proof deferred to Appendix E.

Theorem 1. *Under Assumptions 1, 2, considering that the Q-function is learned by two-layer ReLU neural networks with width m and bounded ℓ_1 norm B in Eq. (4), then for any $\delta \in (0,1)$, the total expected regret can be upper bounded with probability at least $1-\delta$*

$$\begin{aligned}\textit{Regret}(T) &\lesssim H^2 T^{\frac{3}{4}} \left[B(\log d)^{\frac{1}{4}} + \log^{\frac{1}{4}}\left(\frac{6}{\delta}\right)\right] + \frac{H^2 T}{\sqrt{m}} + H^2\sqrt{T} + \sqrt{TH^3 \log\left(\frac{6}{\delta}\right)} \\ &\lesssim \widetilde{\mathcal{O}}(H^2 T^{\frac{3}{4}}), \quad \textit{by taking } m = \Omega(\sqrt{T}).\end{aligned}$$

Remark: To ensure sub-linear regret, the width m in two-layer ReLU neural networks is required to $m = \Omega(T^\beta)$ with $\beta > 0$. The result is complementary to scaling of $m = \Omega(T^{13})$ obtained in [12]. The results should be compared with care because the focus of the two works is different. While, [12] focuses on designing UCB like exploration in NTK, we focus on getting more realistic characterization of the width at the cost of introducing Assumption 1. More importantly, our regret bound almost gets rid of the feature dimension d, which allows for DQN for high-dimensional image data in practice.

4.2 Efficient value iteration under deep ReLU neural networks

In this setting, we consider $\widehat{Q}_h^t = \operatorname{argmin}_{f \in \mathcal{F}_{\texttt{DNN}}} \widehat{\mathcal{E}}_h^t(f)$ in Eq. (3), where $\mathcal{F}_{\texttt{DNN}}$ is the function space of deep ReLU neural networks defined in Eq. (5). Similar to Assumption 2, we have the following assumption on the Besov space $\mathcal{B}$, where the Bellman optimality operator maps any bounded value function to a bounded Besov space ball.

Assumption 3. *Let $\widetilde{R} \asymp H$ be a fixed constant. Define $\mathcal{B}_{\widetilde{R}} = \{f \in \mathcal{B}_{p,q}^\alpha(\mathcal{X}) : \|f\|_{\mathcal{B}} \leqslant \widetilde{R}\}$ in the Besov space and assume that for any $h \in [H]$ and $Q \colon \mathcal{S} \times \mathcal{A} \to [0, H]$, we have $\mathbb{T}_h^\star Q \in \mathcal{B}_{\widetilde{R}}$.*

Based on this assumption, we have the following theorem on regret bound in the Besov space for efficient deep Q-learning, with the proof deferred to Appendix F.1.

Theorem 2. *Under Assumptions 1, 3 with the smoothness parameter $\alpha > d(1/p - 1/4)_+$ in the Besov space $\mathcal{B}_{p,q}^\alpha(\mathcal{X})$, considering that the Q-function is learned by deep ReLU neural networks in Eq. (5), taking*

$$\text{the depth } L \asymp \tfrac{d}{2\alpha+d}\log T, \qquad \text{the width } m \asymp \tfrac{d}{2\alpha+d} T^{\frac{d}{2\alpha+d}} \log T,$$

then for any $\delta \in (0,1)$, the regret bound holds with probability at least $1-\delta$

$$\begin{aligned}\textit{Regret}(T) &\lesssim H^{\frac{3}{2}} T^{\frac{\alpha+d}{2\alpha+d}} \log^3 T + H^2\sqrt{T}\sqrt{\log\left(\frac{3}{\delta}\right)}\log T + \sqrt{TH^3\log\left(\frac{6}{\delta}\right)} \\ &\lesssim \widetilde{\mathcal{O}}(H^2 T^{\frac{\alpha+d}{2\alpha+d}}).\end{aligned}$$

Remark: (i) It is clear that the depth in $\widetilde{\mathcal{O}}(1)$ order and the width $\widetilde{\mathcal{O}}(T^{\frac{d}{2\alpha+d}})$ are enough to achieve sublinear regret, which coincides with DQN in practice. Under a special case with $p = q = \infty$ and

$\alpha \gg d$, which requires Q-function to be infinitely differentiable, we obtain the optimal minimax $\widetilde{\mathcal{O}}(\sqrt{T})$ regret by taking the depth and width in a constant order.

(ii) According to the width-depth ratio $m/L = T^{\frac{d}{2\alpha+d}}$, our theory indicates that less problem-dependent smoothness of Q-function requires DNNs to be wider. In practice, if we work in the limited budget of parameters N in neural networks, e.g., $N \asymp m^2 L$, our theory implies that there is a tradeoff between the depth and width on smoothness, i.e., the depth $L := N^{1/3} T^{-\frac{2d}{3(2\alpha+d)}}$ increasing with α (or T) and the width $m = N^{1/3} T^{\frac{d}{3(2\alpha+d)}}$ decreasing with α (or T).

(iii) Two-layer neural networks endowed by Barron spaces cannot be regarded as a special case of DNNs endowed by Besov spaces. This is because, our results in Barron spaces accommodate dense two-layer neural networks; while our result in Besov spaces here depends on the sparsity of deep ReLU neural networks [58] for statistical complexity estimation of $\mathcal{F}_{\texttt{DNN}}$.

It can be found that, though the sub-linear regret can be achieved by Theorem 2 by choosing $\widetilde{\mathcal{O}}(1)$ depth and $\widetilde{\mathcal{O}}(T^{\frac{d}{2\alpha+d}})$ width, the rate heavily depends on the feature dimension d, failing in the *curse of dimensionality*, which appears ineffective on high dimensional data in deep Q-learning. To tackle this issue, we consider *mixed smooth* Besov space $\widetilde{\mathcal{B}}$ to alleviate the curse of dimensionality.

Assumption 4. *Let $\widetilde{R} \asymp H$ be a fixed constant. Define $\widetilde{\mathcal{B}}_{\widetilde{R}} = \{f \in \widetilde{\mathcal{B}}^{\alpha}_{p,q} : \|f\|_{\widetilde{\mathcal{B}}} \leqslant \widetilde{R}\}$ in the mixed smooth Besov space and assume that $\forall h \in [H]$ and $Q \colon \mathcal{S} \times \mathcal{A} \to [0, H]$, then we have $\mathbb{T}^{\star}_h Q \in \widetilde{\mathcal{B}}_{\widetilde{R}}$.*

Based on this assumption, we have the following proposition on regret bound in the mixed-smooth Besov space, with the proof deferred to Appendix F.2.

Proposition 1. *Under Assumptions 1, 4 with the smoothness parameter $\alpha > 0$ and $u := (1 - 1/q)_+$ defined in the mixed smooth Besov space $\widetilde{\mathcal{B}}^{\alpha}_{p,q}$ with $p = 4$, considering that the Q-function is learned by deep ReLU neural networks in Eq.* (5)*, taking*

$$\text{the depth } L \asymp \widetilde{\mathcal{O}}(1), \qquad \text{the width } m \asymp \widetilde{\mathcal{O}}(T^{\frac{1}{2\alpha+1}}) \ ,$$

then for any $\delta \in (0, 1)$ with probability at least $1 - \delta$, we have

$$\textit{Regret}(T) \lesssim \begin{cases} \widetilde{\mathcal{O}}\left(H^2 T^{\frac{\alpha+1}{2\alpha+1}}\right), \ \textit{if } \alpha \leqslant u \log_2 e \\ \widetilde{\mathcal{O}}\left(H^2 T^{\frac{1+\alpha+(1-u)\log_2 e}{1+2\alpha+(1-2u)\log_2 e}}\right), \ \textit{if } \alpha > u \log_2 e \, . \end{cases}$$

Remark: Our regret bound gets rid of the feature dimension d and is close to the minimax square-root $\widetilde{\mathcal{O}}(\sqrt{T})$ bound if the Q-function is sufficiently smooth, which holds for DQN on high dimensional image data in practice. Further, if $\alpha > u \log_2 e$, the width can be improved to $m \asymp \widetilde{\mathcal{O}}\left(T^{\frac{1}{1+2\alpha+(1-2u)\log_2 e}}\right)$ suffice to efficient deep Q-learning with sublinear regret.

5 Proof outline

In this section, we outline the proof of value iteration via DNNs with sublinear regret bounds presented in Section 4. Our proof framework includes three parts: regret decomposition, transformation between TD error and generalization, and generalization bounds on non-iid data for TD error analysis. The last two parts are the main technical contribution in this work and provide width and depth guideline for efficient Q-learning. The complete proof is reported in the appendix.

Regret decomposition: This part is standard and commonly studied in RL theory, e.g., [59, 7, 12]. We briefly include here for self-completeness. Define the temporal-difference (TD) error as

$$\Gamma^t_h(s,a) = r_h(s,a) + (\mathbb{P}_h V^t_{h+1})(s,a) - Q^t_h(s,a) = (\mathbb{T}_h V^t_{h+1})(s,a) - Q^t_h(s,a), \quad \forall (s,a) \in \mathcal{S} \times \mathcal{A} \, , \tag{8}$$

where Γ^t_h is a function on $\mathcal{S} \times \mathcal{A}$ for all $h \in [H]$ and $t \in [T]$. Accordingly, the regret can be decomposed into two main terms related to the TD error and the statistical error, respectively (*c.f.* Lemma 1). The statistical error is based on the standard martingale difference sequences, and thus can be upper bounded by Hoeffding-Azuma inequality with $\mathcal{O}(\sqrt{H^3 T})$ regret (*c.f.* Lemma 2).

$$\begin{cases} \text{Rademacher complexity on non-i.i.d data: Lem. 6} \Leftarrow \text{Lem. 5} \\ \text{two layers: Thm. 1} \Leftarrow \mathcal{O}(H^2 B\sqrt{\log d/n}) \Leftarrow \text{Lem. 10: Rademacher complexity of Barron spaces} \\ \text{deep layers} \begin{cases} \text{Besov spaces: Thm. 2} \Leftarrow \mathcal{O}(n^{-\frac{2\alpha}{2\alpha+d}}) \Leftarrow \text{Prop. 3} \Leftarrow \text{Lem. 11 on } \texttt{LRC} \text{ for Besov spaces} \\ \text{mixed smooth Besov spaces: Prop. 1} \Leftarrow \mathcal{O}(n^{-\frac{2\alpha}{2\alpha+1}}) \text{ (or faster)} \Leftarrow \text{Prop. 4} \end{cases} \end{cases}$$

Figure 1: Proof framework of temporal difference error under generalization on non-iid data. We denote LRC by local Rademacher complexity for short. The results omitted in the main text are deferred to the appendix.

Transforming TD error to generalization: To bound the TD error, previous theory work focuses on designing bonus function to ensure TD error negative, which mismatches with practical DQN that depends on not bonus function but historical data. Accordingly, we attempt to built the connection between the regret error and $\|\Gamma_h^t\|_{L^2(\mathrm{d}\bar{\mu}_h^t)}$, see Lemma 3; and then transform estimation on $\|\Gamma_h^t\|_{L^2(\mathrm{d}\bar{\mu}_h^t)}$ to the following two terms: generalization error and approximation error, respectively (*c.f.* Lemma 4)

$$\|\Gamma_h^t\|^2_{L^2(\mathrm{d}\bar{\mu}_h^t)} \leqslant \left[\mathcal{E}_h(\widehat{Q}_h^t) - \min_{f\in\mathcal{F}} \mathcal{E}_h(f)\right] + \inf_{f\in\mathcal{F}} \|f - \mathbb{T}_h^\star Q_{h+1}^t\|^2_{L^2(\mathrm{d}\bar{\mu}_h^t)}\,.$$

The convergence rate of approximation error can be directly obtained by previous work, e.g., $\mathcal{O}(1/m)$ for two-layer ReLU networks in Barron spaces [33] and $\mathcal{O}(n^{-2\alpha/(2\alpha+d)})$ in [36] for deep ReLU networks; and in the next we elucidate the generalization error bounds.

Analysis of TD error based on generalization bounds for non-iid data: The key part left is to bound the generalization error under our non-i.i.d data for TD error, see the proof framework in Figure 1. Following a covering argument for the function approximation class used for value functions [7], we introduce two additional terms to handle conditional independence and biased estimation, which can be upper bounded by ϵ-covering of $\mathcal{F}$. By doing so, we firstly verify that the maximum error in estimating the mean of any function $f \in \mathcal{F}$ can be still bounded by the Rademacher complexity of $\mathcal{F}$ in Lemma 5, and then generalization bounds by Rademacher complexity still holds by Lemma 6 via the averaged measure, which only requires the data to be independent.

To be specific, for two-layer neural networks, by computing the Rademacher complexity of $\mathcal{F}_{\texttt{SNN}}$ in Lemma 10, we obtain the generalization error at a certain $\mathcal{O}(H^2 B\sqrt{\log d/n})$ convergence rate; combining the result of approximation error in Barron spaces, we conclude the proof of Theorem 1 with $\mathcal{O}(H^2 T^{\frac{3}{4}} B \log d)$ regret for efficient Q-learning.

For deep neural networks, by computing the local Rademacher complexity of $\mathcal{F}_{\texttt{DNN}}$ in Lemma 11, by choosing proper parameters in Eq. (5), we present the generalization bounds in Besov spaces (*c.f.* Proposition 3) with $\mathcal{O}(n^{-\frac{2\alpha}{2\alpha+d}})$ and mixed smooth Besov spaces (*c.f.* Proposition 4) with $\mathcal{O}(n^{-\frac{2\alpha}{2\alpha+1}})$ (or faster) respectively. Finally, combining the result of approximation error in these two spaces, the regret can be upper bounded by $\widetilde{\mathcal{O}}(H^2 T^{\frac{\alpha+d}{2\alpha+d}})$ in Theorem 2 and $\widetilde{\mathcal{O}}\left(H^2 T^{\frac{\alpha+1}{2\alpha+1}}\right)$ (or smaller) in Theorem 1, respectively. The above results demonstrate that, $\widetilde{\mathcal{O}}(1)$ depth and the width in sublinear T are suffice to efficient deep value iteration.

6 Conclusion and discussion

This paper provides an initial understanding on DQN with general function approximation by shallow and deep neural networks in the online setting. Our results shed light on width and depth guidelines in DQN. In particular, the theoretical analysis provides guidelines to obtain the best regret guarantees with memory constraints. For realistic low values of α, we observe that the depth L should be chosen approximately as $\log m$. Interestingly, we notice that this result is closely matching practical implementation of DQN. For example, the choices of [4] $m = 512$ and $L = 5$ can be explained with our theory, indeed $\log(512) \approx 6$.

We remark that, the analysis of this work is built on the PSE assumption for exploration, which are satisfied in practical cases when employing DQN. Nevertheless, designing a provably efficient exploration mechanism for DQN could be an interesting future direction in both practice and theory.

References

[1] Csaba Szepesvári. Algorithms for reinforcement learning. *Synthesis lectures on artificial intelligence and machine learning*, 4(1):1–103, 2010. 1

[2] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018. 1

[3] Scott Fujimoto, Herke Hoof, and David Meger. Addressing function approximation error in actor-critic methods. In *International conference on machine learning*, pages 1587–1596. PMLR, 2018. 1

[4] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015. 1, 2, 6, 9

[5] 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. 1

[6] LeCun Yann, Bengio Yoshua, and Hinton Geoffrey. Deep learning. *Nature*, 521(7553):436–444, 2015. 1

[7] Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pages 2137–2143. PMLR, 2020. 1, 2, 3, 7, 8, 9, 26, 29, 31, 33, 34

[8] Yining Wang, Ruosong Wang, Simon Shaolei Du, and Akshay Krishnamurthy. Optimism in reinforcement learning with generalized linear function approximation. In *International Conference on Learning Representations*, 2020. 1

[9] Alex Ayoub, Zeyu Jia, Csaba Szepesvari, Mengdi Wang, and Lin Yang. Model-based reinforcement learning with value-targeted regression. In *International Conference on Machine Learning*, pages 463–474. PMLR, 2020. 1, 3

[10] Dongruo Zhou, Quanquan Gu, and Csaba Szepesvari. Nearly minimax optimal reinforcement learning for linear mixture markov decision processes. In *Conference on Learning Theory*, pages 4532–4576. PMLR, 2021. 1, 3

[11] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, pages 8571–8580, 2018. 1

[12] Zhuoran Yang, Chi Jin, Zhaoran Wang, Mengdi Wang, and Michael I Jordan. On function approximation in reinforcement learning: Optimism in the face of large state spaces. In *Advances in Neural Information Processing Systems*, 2020. 1, 2, 3, 7, 8, 18

[13] Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E Schapire. Contextual decision processes with low bellman rank are pac-learnable. In *International Conference on Machine Learning*, pages 1704–1713. PMLR, 2017. 1, 3

[14] Kefan Dong, Jian Peng, Yining Wang, and Yuan Zhou. Root-n-regret for learning in markov decision processes with function approximation and low bellman rank. In *Conference on Learning Theory*, pages 1554–1557. PMLR, 2020. 1, 3

[15] Daniel Russo and Benjamin Van Roy. Eluder dimension and the sample complexity of optimistic exploration. *Advances in Neural Information Processing Systems*, 26, 2013. 1, 3

[16] Ruosong Wang, Russ R Salakhutdinov, and Lin Yang. Reinforcement learning with general value function approximation: Provably efficient approach via bounded eluder dimension. In *Advances in Neural Information Processing Systems*, volume 33, pages 6123–6135, 2020. 1, 3

[17] Haque Ishfaq, Qiwen Cui, Viet Nguyen, Alex Ayoub, Zhuoran Yang, Zhaoran Wang, Doina Precup, and Lin Yang. Randomized exploration in reinforcement learning with general value function approximation. In *International Conference on Machine Learning*, pages 4607–4616. PMLR, 2021. 1, 3

[18] Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. In *Advances in Neural Information Processing Systems*, pages 2933–2943, 2019. 1

[19] Jaehoon Lee, Lechao Xiao, Samuel Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. In *Advances in Neural Information Processing Systems*, pages 8570–8581, 2019. 1

[20] Blake Woodworth, Suriya Gunasekar, Jason D Lee, Edward Moroshko, Pedro Savarese, Itay Golan, Daniel Soudry, and Nathan Srebro. Kernel and rich regimes in overparametrized models. In *Conference on Learning Theory*, pages 3635–3673. PMLR, 2020. 1

[21] Mario Geiger, Stefano Spigler, Arthur Jacot, and Matthieu Wyart. Disentangling feature and lazy training in deep neural networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2020(11):113301, 2020. 1

[22] Francis Bach. Breaking the curse of dimensionality with convex neural networks. *Journal of Machine Learning Research*, 18(1):629–681, 2017. 1, 4, 16

[23] Gilad Yehudai and Ohad Shamir. On the power and limitations of random features for understanding neural networks. In *Advances in Neural Information Processing Systems*, pages 6594–6604, 2019. 1

[24] Michael Celentano, Theodor Misiakiewicz, and Andrea Montanari. Minimum complexity interpolation in random features models. *arXiv preprint arXiv:2103.15996*, 2021. 1

[25] Gene Li, Pritish Kamath, Dylan J Foster, and Nathan Srebro. Eluder dimension and generalized rank. *arXiv preprint arXiv:2104.06970*, 2021. 2

[26] Kefan Dong, Jiaqi Yang, and Tengyu Ma. Provable model-based nonlinear bandit and reinforcement learning: Shelve optimism, embrace virtual curvature. In *Advances in Neural Information Processing Systems*, volume 34, 2021. 2

[27] Todd Hester, Matej Vecerik, Olivier Pietquin, Marc Lanctot, Tom Schaul, Bilal Piot, Dan Horgan, John Quan, Andrew Sendonaris, Ian Osband, et al. Deep q-learning from demonstrations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. 2

[28] Yuping Luo, Huazhe Xu, Yuanzhi Li, Yuandong Tian, Trevor Darrell, and Tengyu Ma. Algorithmic framework for model-based deep reinforcement learning with theoretical guarantees. In *International Conference on Learning Representations*, 2018. 2

[29] Lin Yang and Mengdi Wang. Reinforcement learning in feature space: Matrix bandit, kernels, and regret bound. In *International Conference on Machine Learning*, pages 10746–10756. PMLR, 2020. 2, 3

[30] Daniel J Russo, Benjamin Van Roy, Abbas Kazerouni, Ian Osband, Zheng Wen, et al. A tutorial on thompson sampling. *Foundations and Trends® in Machine Learning*, 11(1):1–96, 2018. 2, 3

[31] Shipra Agrawal and Navin Goyal. Analysis of thompson sampling for the multi-armed bandit problem. In *Conference on Learning Theory*, pages 1–39. JMLR Workshop and Conference Proceedings, 2012. 2, 3

[32] Andrea Zanette, David Brandfonbrener, Emma Brunskill, Matteo Pirotta, and Alessandro Lazaric. Frequentist regret bounds for randomized least-squares value iteration. In *International Conference on Artificial Intelligence and Statistics*, pages 1954–1964. PMLR, 2020. 2, 3

[33] Weinan E, Chao Ma, and Lei Wu. The barron space and the flow-induced function spaces for neural network models. *Constructive Approximation*, pages 1–38, 2021. 2, 4, 9, 16, 25, 26

[34] Weinan E and Stephan Wojtowytsch. Representation formulas and pointwise properties for barron functions. *arXiv preprint arXiv:2006.05982*, 2020. 2, 4, 16

[35] Yoshihiro Sawano. *Theory of Besov spaces*, volume 56. Springer, 2018. 2, 5, 17

[36] Taiji Suzuki. Adaptivity of deep ReLU network for learning in Besov and mixed smooth Besov spaces: optimal rate and curse of dimensionality. In *International Conference on Learning Representations*, 2019. 2, 4, 5, 9, 16, 17, 18, 27, 29, 32, 33

[37] Jianqing Fan, Zhaoran Wang, Yuchen Xie, and Zhuoran Yang. A theoretical analysis of deep q-learning. In *Learning for Dynamics and Control*, pages 486–489, 2020. 2, 3, 4, 6

[38] Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is q-learning provably efficient? In *Advances in neural information processing systems*, 2018. 3

[39] Pan Xu and Quanquan Gu. A finite-time analysis of q-learning with neural network function approximation. In *International Conference on Machine Learning*, pages 10555–10565. PMLR, 2020. 3

[40] Martin Riedmiller. Neural fitted Q iteration–first experiences with a data efficient neural reinforcement learning method. In *European Conference on Machine Learning*, pages 317–328. Springer, 2005. 3

[41] Baihe Huang, Kaixuan Huang, Sham Kakade, Jason D Lee, Qi Lei, Runzhe Wang, and Jiaqi Yang. Going beyond linear rl: Sample efficient neural function approximation. In *Advances in Neural Information Processing Systems*, 2021. 3, 6

[42] Jihao Long, Jiequn Han, et al. An l^2 analysis of reinforcement learning in high dimensions with kernel and neural network approximation. *arXiv preprint arXiv:2104.07794*, 2021. 3, 6

[43] Thanh Nguyen-Tang, Sunil Gupta, Hung Tran-The, and Svetha Venkatesh. Sample complexity of offline reinforcement learning with deep relu networks. *arXiv preprint arXiv:2103.06671*, 2021. 3

[44] Chi Jin, Qinghua Liu, and Sobhan Miryoosefi. Bellman eluder dimension: New rich classes of rl problems, and sample-efficient algorithms. *Advances in Neural Information Processing Systems*, 34, 2021. 3

[45] Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014. 4

[46] Yoshua Bengio, Nicolas Le Roux, Pascal Vincent, Olivier Delalleau, and Patrice Marcotte. Convex neural networks. In *Advances in Neural Information Processing Systems*, pages 123–130, 2005. 4, 16

[47] Minshuo Chen, Haoming Jiang, Wenjing Liao, and Tuo Zhao. Efficient approximation of deep ReLU networks for functions on low dimensional manifolds. *Advances in neural information processing systems*, 32:8174–8184, 2019. 4

[48] Ahmed Abdeljawad and Philipp Grohs. Approximations with deep neural networks in sobolev time-space. *arXiv preprint arXiv:2101.06115*, 2020. 4

[49] Gérard Kerkyacharian and Dominique Picard. Density estimation in besov spaces. *Statistics & probability letters*, 13(1):15–24, 1992. 5, 17

[50] Winfried Sickel and Tino Ullrich. Tensor products of sobolev–besov spaces and applications to approximation from the hyperbolic cross. *Journal of Approximation Theory*, 161(2):748–786, 2009. 5, 17, 18

[51] Steven J Bradtke and Andrew G Barto. Linear least-squares algorithms for temporal difference learning. *Machine learning*, 22(1):33–57, 1996. 6

[52] Yaqi Duan, Chi Jin, and Zhiyuan Li. Risk bounds and rademacher complexity in batch reinforcement learning. In *International Conference on Machine Learning*, pages 2892–2902. PMLR, 2021. 6

[53] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013. 6

[54] Seungchan Kim, Kavosh Asadi, Michael Littman, and George Konidaris. Deepmellow: removing the need for a target network in deep q-learning. In *International Joint Conference on Artificial Intelligence*, 2019. 6

[55] Botao Hao, Tor Lattimore, Csaba Szepesvári, and Mengdi Wang. Online sparse reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pages 316–324. PMLR, 2021. 6

[56] Yuanhao Wang, Ruosong Wang, and Sham Kakade. An exponential lower bound for linearly realizable mdp with constant suboptimality gap. *Advances in Neural Information Processing Systems*, 34, 2021. 6

[57] Gellért Weisz, Philip Amortila, and Csaba Szepesvári. Exponential lower bounds for planning in mdps with linearly-realizable optimal action-value functions. In *Algorithmic Learning Theory*, pages 1237–1264. PMLR, 2021. 6

[58] Boris Hanin and David Rolnick. Deep relu networks have surprisingly few activation patterns. *Advances in Neural Information Processing Systems*, 32, 2019. 8

[59] Qi Cai, Zhuoran Yang, Chi Jin, and Zhaoran Wang. Provably efficient exploration in policy optimization. In *International Conference on Machine Learning*, pages 1283–1294, 2020. 8, 18, 19

[60] Weinan E, Chao Ma, and Lei Wu. A priori estimates of the population risk for two-layer neural networks. *Communications in Mathematical Sciences*, 17(5):1407–1425, 2019. 16

[61] Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. Norm-based capacity control in neural networks. In *Conference on Learning Theory*, pages 1376–1401. PMLR, 2015. 16

[62] Evarist Giné and Richard Nickl. *Mathematical foundations of infinite-dimensional statistical models*. Cambridge University Press, 2021. 17

[63] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018. 22

[64] Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019. 22, 26

[65] Shahar Mendelson. Improving the sample complexity using global data. *IEEE transactions on Information Theory*, 48(7):1977–1991, 2002. 22, 25

[66] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge University Press, 2014. 23, 24

[67] Peter L Bartlett, Olivier Bousquet, and Shahar Mendelson. Local rademacher complexities. *Annals of Statistics*, 33(4):1497–1537, 2005. 25

[68] Yunwen Lei, Lixin Ding, and Yingzhou Bi. Local rademacher complexity bounds based on covering numbers. *Neurocomputing*, 218:320–330, 2016. 25, 30

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] We clearly discuss the limitation of this work in the Conclusion section.
   (c) Did you discuss any potential negative societal impacts of your work? [No] Our work is theoretical and generally will have no negative societal impacts.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] The assumptions are clearly stated and well discussed.
   (b) Did you include complete proofs of all theoretical results? [Yes] All of the proofs can be found in the Appendix.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]