SAMPLE COMPLEXITY OF OFFLINE REINFORCEMENT LEARNING WITH DEEP RELU NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Offline reinforcement learning (RL) leverages previously collected data for policy optimization without any further active exploration. Despite the recent interest in this problem, its theoretical foundations in neural network function approximation setting remain limited. In this paper, we study the statistical theory of offline RL with deep ReLU network function approximation. In particular, we establish the sample complexity of $\tilde{\mathcal{O}}\left(\kappa^{1+d/\alpha} \cdot \epsilon^{-2-2d/\alpha}\right)$ for offline RL with deep ReLU networks, where κ is a measure of distributional shift, d is the dimension of the state-action space, α is a (possibly fractional) smoothness parameter of the underlying Markov decision process (MDP), and ϵ is a user-specified error. Notably, our sample complexity holds under two novel considerations, namely the Besov dynamic closure and the correlated structure that arises from value regression for offline RL. While the Besov dynamic closure generalizes the dynamic conditions for offline RL in the prior works, the correlated structure renders the existing analyses improper or inefficient. To our knowledge, our work is the first to provide such a comprehensive analysis for offline RL with deep ReLU network function approximation.

1 INTRODUCTION

Offline reinforcement learning (Lange et al., 2012; Levine et al., 2020) is a practical paradigm of reinforcement learning (RL) where logged experiences are abundant but a new interaction with the environment is limited or even prohibited. The fundamental offline RL problems concern with how well previous experiences could be used to evaluate a new target policy, known as off-policy evaluation (OPE) problem, or to learn the optimal policy, known as off-policy learning (OPL) problem. We study these offline RL problems with infinitely large state spaces, where the agent must rely on function approximation such as deep neural networks to generalize across states from an offline dataset without any further exploration. Such problems form the core of modern RL in practical settings (Levine et al., 2020; Kumar et al., 2020; Singh et al., 2020), but no work has provided a comprehensive and adequate analysis of the statistical efficiency for offline RL with neural network function approximation.

On the theoretical side, predominant sample-efficient results in offline RL focus on tabular environments with small finite state spaces (Yin & Wang, 2020; Yin et al., 2021; Yin & Wang, 2021), but as these methods scale with the number of states, they are infeasible for the settings with infinitely large state spaces. While this tabular setting has been extended to large state spaces via linear environments (Duan & Wang, 2020; Tran-The et al., 2021), the linearity assumption often does not hold for many RL problems in practice. Theoretical guarantees for offline RL with general and deep neural network function approximations have also been derived, but these results are either inadequate or relatively disconnected from practical settings. In particular, while the finite-sample results for offline RL with general function approximation (Munos & Szepesvári, 2008; Le et al., 2019) depend on an inherent Bellman error which could be large or uncontrollable in practice, other analyses (Yang et al., 2019) rely on an inefficient data splitting technique to deal with the highly correlated structures arisen in value regression for offline RL and use a relatively strong dynamic assumption. It therefore remains unclear whether offline RL can provably work in a more general dynamic condition and the highly correlated structure of value regression.

In this paper, we provide a statistical theory of both OPE and OPL with neural network function approximation in a broad generality. In particular, our contributions are:

- *First*, we achieve a generality for the guarantees of offline RL with neural network function approximation via two novel considerations: (i) we introduce a new structural condition namely *Besov dynamic closure* which generalizes the existing dynamic conditions for offline RL with neural network function approximation and even includes MDPs that need not be continuous, differentiable or spatially homogeneous in smoothness; (ii) we take into account the highly correlated structure of the value estimate produced by a regression-based algorithm from the offline data. This correlated structure plays a central role in the statistical efficiency of an offline algorithm but the prior results (Munos & Szepesvári, 2008; Le et al., 2019; Yang et al., 2019) improperly ignore this structure or avoid it using an inefficient data splitting approach.
- *Second*, we prove that an offline RL algorithm based on fitted-Q iteration (FQI) can achieve the sample complexity of $\tilde{\mathcal{O}}\left(\kappa^{1+d/\alpha} \cdot \epsilon^{-2-2d/\alpha}\right)$ where κ measures the distributional shift in the offline data, d is the input dimension, α is a smoothness parameter of the underlying MDP, and ϵ is a user-specified error. Notably, our guarantee holds under a general condition encompassing the dynamic conditions in the existing works while it does not require any data splitting as in (Yang et al., 2019). The data splitting approach splits the offline data into K disjoint folds where K is the number of iterations in their algorithm. As the sample complexity of such data splitting scales linearly with K where K can be arbitrarily large in practice, the guarantee in (Yang et al., 2019) is highly inefficient for offline RL. Moreover, our analysis also improves upon the analysis in (Le et al., 2019) that incorrectly ignores the correlated structure of offline value estimate.

Notation. Let $L^p(\mathcal{X}, \mu) = \{f : \mathcal{X} \to \mathbb{R} \mid \|f\|_{p,\mu} := (\int_{\mathcal{X}} |f|^p d\mu)^{1/p} < \infty\}$ be the space of measurable functions for which the p-th power of the absolute value is μ-measurable, $C^0(\mathcal{X}) = \{f : \mathcal{X} \to \mathbb{R} \mid f \text{ is continuous and } \|f\|_\infty < \infty\}$ be the space of bounded continuous functions, $C^\alpha(\mathcal{X})$ be the Hölder space with smoothness parameter $\alpha \in (0, \infty) \backslash \mathbb{N}$, $W_p^m(\mathcal{X})$ be the Sobolev space with regularity $m \in \mathbb{N}$ and parameter $p \in [1, \infty]$, and $X \hookrightarrow Y$ be *continuous embedding* from a metric space X to a metric space Y. Denote by $\mathcal{P}(\Omega)$ the set of probability measures supported in domain Ω. For simplicity, we use $\|\cdot\|_\mu$ for $\|\cdot\|_{p,\mu}$ when $p = 2$. Denote by $\|\cdot\|_0$ the 0-norm, i.e., the number of non-zero elements, and $a \vee b = \max\{a, b\}$. For any two real-valued functions f and g, we write $f(\cdot) \lesssim g(\cdot)$ if there is an absolute constant c independent of the function parameters $(\cdot)$ such that $f(\cdot) \leq c \cdot g(\cdot)$. We write $f(\cdot) \asymp g(\cdot)$ if $f(\cdot) \lesssim g(\cdot)$ and $g(\cdot) \lesssim f(\cdot)$. We write $f(\cdot) \simeq g(\cdot)$ if there exists an absolute constant c such that $f(\cdot) = c \cdot g(\cdot)$.

2 Related Work

The majority of the theoretical results for offline RL focus on tabular settings and mostly on OPE task where the state space is finite and an importance sampling -related approach is possible (Precup et al., 2000; Dudík et al., 2011; Jiang & Li, 2015; Thomas & Brunskill, 2016; Farajtabar et al., 2018; Kallus & Uehara, 2019). The main drawback of the importance sampling-based approach is that it suffers high variance in long horizon problems. The high variance problem is later mitigated by the idea of formulating the OPE problem as a density ratio estimation problem (Liu et al., 2018; Nachum et al., 2019a; Zhang et al., 2020a;b; Nachum et al., 2019b) but these results do not provide sample complexity guarantees. The sample-efficient guarantees for offline RL are obtained in tabular settings in (Xie et al., 2019; Yin & Wang, 2020; Yin et al., 2021; Yin & Wang, 2021). Jiang & Li (2016) derive Cramer-Rao lower bound for discrete-tree MDPs.

For the function approximation setting, as the state space of MDPs is often infinite or continuous, some form of function approximation is deployed in approximate dynamic programming such as fitted Q-iteration, least squared policy iteration (Bertsekas & Tsitsiklis, 1995; Jong & Stone, 2007; Lagoudakis & Parr, 2003; Grünewälder et al., 2012; Munos, 2003; Munos & Szepesvári, 2008; Antos et al., 2008; Tosatto et al., 2017), and fitted Q-evaluation (FQE) (Le et al., 2019). A recent line of work studies offline RL in non-linear function approximation (e.g, general function approximation and deep neural network function approximation) (Le et al., 2019; Yang et al., 2019). In particular, Le et al. (2019) provide an error bound of OPE and OPL with general function approximation but they ignore the correlated structure in the FQI-type algorithm, resulting in an improper analysis. Moreover, their error bounds depend on the inherent Bellman error that can be large and uncontrollable in practical settings. More closely related to our work is (Yang et al., 2019) which also considers deep neural network approximation. In particular, Yang et al. (2019) focused on analyzing deep Q-learning using a disjoint fold of offline data for each iteration. Such approach is considerably sample-inefficient for offline RL as their sample complexity linearly scales with the number of iterations K which is very large in practice. In addition, they rely on a relatively restricted smoothness assumption of the underlying MDPs that hinders their results from being widely applicable in more practical settings.

Since the initial version of this paper appeared, a concurrent work studies offline RL with general function approximation via local Rademacher complexities (Duan et al., 2021). While both papers independently have the same idea of using local Rademacher complexities as a tool to study sample complexities in offline RL, our work differs from (Duan et al., 2021) in three main aspects. First, we focus on infinite-horizon MDPs while (Duan et al., 2021) work in finite-horizon MDPs. Second, we derive an explicit sample complexity while the sample complexity in (Duan et al., 2021) depends on the critical radius of local Rademacher complexity. Bounding the critical radius for a complex model under the correlated structure is highly non-trivial. Duan et al. (2021) provided the specialized sample complexity for finite classes, linear classes, kernel spaces and sparse linear spaces but it is unclear how their result can apply to more complex models such as a deep ReLU network. Moreover, we propose a new Besov dynamic closure and establish the sample compelxity using a uniform convergence argument which appear absent in Duan et al. (2021).

3 PRELIMINARIES

We consider reinforcement learning in an infinite-horizon discounted Markov decision process (MDP) with possibly infinitely large state space $\mathcal{S}$, continuous action space $\mathcal{A}$, initial state distribution $\rho \in \mathcal{P}(\mathcal{S})$, transition operator $P : \mathcal{S} \times \mathcal{A} \to \mathcal{P}(\mathcal{S})$, reward distribution $R : \mathcal{S} \times \mathcal{A} \to \mathcal{P}([0, 1])$, and a discount factor $\gamma \in [0, 1)$. For notational simplicity, we assume that $\mathcal{X} := \mathcal{S} \times \mathcal{A} \subseteq [0, 1]^d$ but our results readily generalizes to the case when $\mathcal{A}$ is finite.

A policy $\pi : \mathcal{S} \to \mathcal{P}(\mathcal{A})$ induces a distribution over the action space conditioned on states. The Q-value function for policy π at state-action pair (s, a), denoted by $Q^\pi(s, a) \in [0, 1]$, is the expected discounted total reward the policy collects if it initially starts in the state-action pair,

$$Q^\pi(s, a) := \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t r_t | s_0 = s, a_0 = a \right],$$

where $r_t \sim R(s_t, a_t), a_t \sim \pi(\cdot|s_t)$, and $s_t \sim P(\cdot|s_{t-1}, a_{t-1})$. The value for a policy π is $V^\pi = \mathbb{E}_{s\sim\rho, a\sim\pi(\cdot|s)} [Q^\pi(s, a)]$, and the optimal value is $V^* = \max_\pi V^\pi$ where the maximization is taken over all stationary policies. Alternatively, the optimal value V^* can be obtained via the optimal Q-function $Q^* = \max_\pi Q^\pi$ as $V^* = \mathbb{E}_{s\sim\rho} [\max_a Q^*(s, a)]$. Denote by T^π and T^* the Bellman operator and the optimality Bellman operator, respectively, i.e., for any $f : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$

$$[T^\pi f](s, a) = \mathbb{E}_{r\sim R(s,a)}[r] + \gamma \mathbb{E}_{s'\sim P(\cdot|s,a), a'\sim\pi(\cdot|s')} [f(s', a')]$$
$$[T^* f](s, a) = \mathbb{E}_{r\sim R(s,a)}[r] + \gamma \mathbb{E}_{s'\sim P(\cdot|s,a)} \left[\max_{a'} f(s', a') \right],$$

we have $T^\pi Q^\pi = Q^\pi$ and $T^* Q^* = Q^*$.

We consider the offline RL setting where a learner cannot explore the environment but has access to a fixed logged data $\mathcal{D} = \{(s_i, a_i, s'_i, r_i)\}_{i=1}^n$ collected a priori by certain behaviour policy η. For simplicity, we assume that $\{s_i\}_{i=1}^n$ are independent and η is stationary. Equivalently, $\{(s_i, a_i)\}_{i=1}^n$ are i.i.d. samples from the normalized discounted stationary distribution over state-actions with respect to η, i.e., $(s_i, a_i) \overset{i.i.d.}{\sim} \mu(\cdot, \cdot) := (1 - \gamma) \sum_{t=0}^{\infty} \gamma^t \mathbb{P}(s_t = \cdot, a_t = \cdot|\rho, \eta)$ where $s'_i \sim P(\cdot|s_i, a_i)$ and $a_i \sim \eta(\cdot|s_i)$. This assumption is relatively standard in the offline RL setting (Munos & Szepesvári, 2008; Chen & Jiang, 2019; Yang et al., 2019) and is used merely for the sake of theoretical analysis. The goals of OPE and OPL are to estimate V^π and V^*, respectively from $\mathcal{D}$. The performance of OPE and OPL estimates are measured via sub-optimality gaps.

For OPE. Given a fixed target policy π, for any value estimate $\hat{V}$ computed from the offline data $\mathcal{D}$, the sub-optimality of OPE is defined as

$$\text{SubOpt}(\hat{V}; \pi) = |V^{\pi} - \hat{V}|.$$

For OPL. For any estimate $\hat{\pi}$ of the optimal policy π^* that is learned from the offline data $\mathcal{D}$, we define the sup-optimality of OPL as

$$\text{SubOpt}(\hat{\pi}) = \mathbb{E}_{\rho}\left[V^*(s) - Q^*(s, \hat{\pi}(s))\right],$$

where $\mathbb{E}_{\rho}$ is the expectation with respect to (w.r.t.) $s \sim \rho$.

3.1 Deep ReLU Networks as Function Approximation

In practice, the state space is often very large and complex, and thus function approximation is required to ensure generalization across different states. Deep networks with the ReLU activation offer a rich class of parameterized functions with differentiable parameters. Deep ReLU networks are state-of-the-art in many applications, e.g., (Krizhevsky et al., 2012; Mnih et al., 2015), including offline RL with deep ReLU networks that can yield superior empirical performance (Voloshin et al., 2019). In this section, we describe the architecture of deep ReLU networks and the associated function space which we use throughout this paper. Specifically, a L-height, m-width ReLU network on $\mathbb{R}^d$ takes the form of

$$f_{\theta}^{L,m}(x) = W^{(L)}\sigma\left(W^{(L-1)}\sigma\left(\dots\sigma\left(W^{(1)}\sigma(x) + b^{(1)}\right)\dots\right) + b^{(L-1)}\right) + b^{(L)},$$

where $W^{(L)} \in \mathbb{R}^{1\times m}, b^{(L)} \in \mathbb{R}, W^{(1)} \in \mathbb{R}^{m \times d}, b^{(1)} \in \mathbb{R}^m$, $W^{(l)} \in \mathbb{R}^{m\times m}, b^{(l)} \in \mathbb{R}^m, \forall 1 < l < L, \theta = \{W^{(l)}, b^{(l)}\}_{1\leq l \leq L}$, and $\sigma(x) = \max\{x, 0\}$ is the (element-wise) ReLU activation. We define $\Phi(L, m, S, B)$ as the space of L-height, m-width ReLU functions $f_{\theta}^{L,m}(x)$ with sparsity constraint S, and norm constraint B, i.e., $\sum_{l=1}^{L}(\|W^{(l)}\|_0 + \|b^{(l)}\|_0) \leq S, \max_{1\leq l\leq L}\|W^{(l)}\|_{\infty} \vee \|b^{(l)}\|_{\infty} \leq B$. Finally, for some $L, m \in \mathbb{N}$ and $S, B \in (0, \infty)$, we define the unit ball of ReLU network function space $\mathcal{F}_{NN}$ as

$$\mathcal{F}_{NN} := \left\{f \in \Phi(L, m, S, B) : \|f\|_{\infty} \leq 1\right\}.$$

We further write $\mathcal{F}_{NN}(\mathcal{X})$ to emphasize the domain $\mathcal{X}$ of deep ReLU functions in $\mathcal{F}_{NN}$ but often use $\mathcal{F}_{NN}$ when the domain context is clear. The main benefit of deep ReLU networks is that in standard non-parametric regression, they outperform any non-adaptive linear estimator due to their higher adaptivity to spatial inhomogeneity (Suzuki, 2018).

3.2 Regularity

In this section, we define a function space for the target functions for which we study offline RL. Note that a regularity assumption on the target function is necessary to obtain a nontrivial rate of convergence (Györfi et al., 2002). A common way to measure regularity of a function is through the L^p-norm of its local oscillations (e.g., of its derivatives if they exist). This regularity notion encompasses the classical Lipschitz, Hölder and Sobolev spaces. In particular in this work, we consider Besov spaces. Besov spaces allow *fractional* smoothness that describes the regularity of a function more precisely and generalizes the previous smoothness notions. There are several ways to characterize the smoothness in Besov spaces. Here, we pursue a characterization via moduli of smoothness as it is more intuitive, following (Giné & Nickl, 2016).

Definition 3.1 (*Moduli of smoothness*). For a function $f \in L^p(\mathcal{X})$ for some $p \in [1, \infty]$, we define its r-th modulus of smoothness as

$$\omega_r^{t,p}(f) := \sup_{0\leq h\leq t}\|\Delta_h^r(f)\|_p, t > 0, r \in \mathbb{N},$$

where the r-th order translation-difference operator $\Delta_h^r = \Delta_h \circ \Delta_h^{r-1}$ is recursively defined as

$$\Delta_h^r(f)(\cdot) := (f(\cdot + h) - f(\cdot))^r = \sum_{k=0}^{r}\binom{r}{k}(-1)^{r-k}f(\cdot + k\cdot h).$$

Remark 3.1. The quantity $\Delta_h^r(f)$ captures the local oscillation of function f which is not necessarily differentiable. In the case the r-th order weak derivative $D^r f$ exists and is locally integrable, we have

$$\lim_{h\to 0}\frac{\Delta_h^r(f)(x)}{h^r} = D^r f(x), \frac{\omega_r^{t,p}(f)}{t^r} \le \|D^r f\|_p \text{ and } \frac{\omega_{r+r'}^{t,p}(f)}{t^r} \le \omega_{r'}^{t,p}(D^r f).$$

Definition 3.2 (*Besov space* $B^\alpha_{p,q}(\mathcal{X})$). For $1 \le p, q \le \infty$ and $\alpha > 0$, we define the norm $\|\cdot\|_{B^\alpha_{p,q}}$ of the Besov space $B^\alpha_{p,q}(\mathcal{X})$ as $\|f\|_{B^\alpha_{p,q}} := \|f\|_p + |f|_{B^\alpha_{p,q}}$ where

$$|f|_{B^\alpha_{p,q}} := \begin{cases} \left(\int_0^\infty (\frac{\omega^{t,p}_{\lfloor\alpha\rfloor+1}(f)}{t^\alpha})^q \frac{dt}{t}\right)^{1/q}, & 1 \le q < \infty, \\ \sup_{t>0} \frac{\omega^{t,p}_{\lfloor\alpha\rfloor+1}(f)}{t^\alpha}, & q = \infty, \end{cases}$$

is the Besov seminorm. Then, $B^\alpha_{p,q} := \{f \in L^p(\mathcal{X}) : \|f\|_{B^\alpha_{p,q}} < \infty\}$.

Intuitively, the Besov seminorm $|f|_{B^\alpha_{p,q}}$ roughly describes the L^q-norm of the l^p-norm of the α-order smoothness of f. Having defined Besov spaces, a natural question is what properties Besov spaces have and how these spaces are related to other function spaces considered in the current literature of offline RL? It turns out that Besov spaces are considerably general that encompass Hölder spaces and Sobolev spaces as well as functions with spatially inhomogeneous smoothness (Triebel, 1983; Sawano, 2018; Suzuki, 2018; Cohen, 2009; Nickl & Pötscher, 2007). We summarize the key intriguing characteristics of Besov spaces and their relation with other spaces:

- (Monotonicity in q) For $1 \le p \le \infty, 1 \le q_1 \le q_2 \le \infty$ and $\alpha \in \mathbb{R}$, $B^\alpha_{p,q_1}(\mathcal{X}) \hookrightarrow B^\alpha_{p,q_2}(\mathcal{X})$;
- (With L^p spaces) $L^2(\mathcal{X}) \hookrightarrow B^0_{2,2}(\mathcal{X})$, $B^0_{p,1}(\mathcal{X}) \hookrightarrow L^p(\mathcal{X}) \hookrightarrow B^0_{p,\infty}(\mathcal{X})$ for $1 \le p \le \infty$, and $B^\alpha_{p,q}(\mathcal{X}) \hookrightarrow L^r(\mathcal{X})$ for $\alpha > d(1/p - 1/r)_+$ where $r = \lfloor\alpha\rfloor + 1$;
- (With $C^0(\mathcal{X})$) $B^\alpha_{p,q}(\mathcal{X}) \hookrightarrow C^0(\mathcal{X})$ for $1 \le p, q \le \infty, \alpha > d/p$;
- (With Sobolev spaces) $B^m_{2,2}(\mathcal{X}) = W^m_2(\mathcal{X})$ for $m \in \mathbb{N}$;
- (With Hölder spaces) $B^\alpha_{\infty,\infty}(\mathcal{X}) = C^\alpha(\mathcal{X})$ for $\alpha = (0,\infty)\backslash\mathbb{N}$.

In particular, the Besov space $B^\alpha_{p,q}$ reduces into the Hölder space C^α when $p = q = \infty$ and α is positive and non-integer while it reduces into the Sobolev space W^α_2 when $p = q = 2$ and α is a positive integer. We further consider the unit ball of $B^\alpha_{p,q}(\mathcal{X})$:

$$\bar{B}^\alpha_{p,q}(\mathcal{X}) := \{g \in B^\alpha_{p,q} : \|g\|_{B^\alpha_{p,q}} \le 1 \text{ and } \|g\|_\infty \le 1\}.$$

To obtain a non-trivial guarantee, certain assumptions on the distribution shift and the MDP regularity are necessary. Here, we introduce such assumptions. The first assumption is a common restriction that quantifies the distribution shift in offline RL.

Assumption 3.1 (*Concentration coefficient*). There exists $\kappa_\mu < \infty$ such that $\|\frac{d\nu}{d\mu}\|_\infty \le \kappa_\mu$ for any *realizable* distribution ν, where a distribution ν is said to be realizable if there exist $t \ge 0$ and policy $\bar{\pi}$ such that $\nu(s,a) = \mathbb{P}(s_t = s, a_t = a | s_1 \sim \rho, \bar{\pi}), \forall s, a$.

Intuitively, the finite κ_μ in Assumption 3.1 asserts that the sampling distribution μ is not too far away from any realizable distribution uniformly over the state-action space. κ_μ is finite for a reasonably large class of MDPs, e.g., for any finite MDP, any MDP with bounded transition kernel density, and equivalently any MDP whose top-Lyapunov exponent is negative (Munos & Szepesvári, 2008). Chen & Jiang (2019) further provided natural problems with rich observations generated from hidden states that has low concentration coefficients. These suggest that low concentration coefficients can be found in fairly many interesting problems in practice. We present a simple (though stronger than necessary) example for which Assumption 3.1 holds.

Example 3.1. If the transition density $P(s'|s,a)$ is sufficiently stochastic and the behaviour policy ν has a sufficient uniform coverage over the action space, i.e., there exist absolute constants $c_1, c_2 > 0$ such that for any $s, s' \in \mathcal{S}$, there exists an action $a \in \mathcal{A}$ such that $P(s'|s,a) \ge 1/c_1$ and $\eta(a|s) \ge 1/c_2, \forall s, a$, then we can choose $\kappa_\mu = c_1 c_2$.

Next, we introduce a completeness assumption.

Assumption 3.2 (*Besov dynamic closure*)**.** $\forall f \in \mathcal{F}_{NN}(\mathcal{X}), \forall \pi, T^{\pi} f \in \bar{B}^{\alpha}_{p,q}(\mathcal{X})$ for some $p, q \in [1, \infty]$ and $\alpha > \frac{d}{p \wedge 2}$.

Assumption 3.2 signifies that for any policy π, the Bellman operator T^{π} applied on any ReLU network function in $\mathcal{F}_{NN}(\mathcal{X})$ results in a Besov function in $\bar{B}^{\alpha}_{p,q}(\mathcal{X})$. Moreover, as $T^{\pi_f} f = T^* f$ where π_f is the greedy policy w.r.t. f, Assumption 3.2 also implies that $T^* f \in \bar{B}^{\alpha}_{p,q}(\mathcal{X})$ if $f \in \mathcal{F}_{NN}(\mathcal{X})$. This kind of completeness assumption is relatively standard and common in the offline RL literature (Chen & Jiang, 2019); yet our Besov dynamic closure is sufficiently general that encompasses almost all the previous completeness assumptions in the literature. For example, a simple (yet considerably stronger than necessary) sufficient condition for Assumption 3.2 is that the expected reward function $r(s, a)$ and the transition density $P(s'|s, a)$ for each fixed s' are the functions in the Besov space $B^{\alpha}_{p,q}(\mathcal{X})$, regardless of any function approximator f and any policy π. Such a condition on the transition dynamic is common in the RL literature; for example, linear MDPs Jin et al. (2020) posit a linear structure on the expected reward and the transition density as $r(s, a) = \langle \phi(s, a), \theta \rangle$ and $P(s'|s, a) = \langle \phi(s, a), \lambda(s') \rangle$ for some feature map $\phi : \mathcal{X} \to \mathbb{R}^{d_0}$ and signed measures $\lambda(s') = (\lambda(s')_1, \ldots, \lambda(s')_{d_0})$. To make it even more concrete, we present a simple example for the sufficient condition above.

Example 3.2 (*Reproducing kernel Hilbert space (RKHS) with Matérn kernels*)**.** Define $k_{h,l}$ the Matérn kernel with smoothness parameter $h > 0$ and length scale $l > 0$. If both the expected reward function $r(\cdot)$ and the transition density $g_{s'}(\cdot) := P(s'|\cdot)$ at any $s' \in \mathcal{S}$ are functions in the RKHS of Matérn kernel $k_{h,l}$ where $h = \alpha - d/2 > 0$ and $l > 0$, then Assumption 3.2 holds for $p = q = 2$. This is due to the norm-equivalence between the above RKHS and the Sobolev space $W^{\alpha}_2(\mathcal{X})$ (Kanagawa et al., 2018) and the degeneration from Besov spaces to Sobolev spaces as $B^{\alpha}_{2,2}(\mathcal{X}) = W^{\alpha}_2(\mathcal{X})$.

More generally, our Besov dynamic closure assumption also encompasses the dynamic condition considered in the prior result (Yang et al., 2019). In particular, as remarked earlier, the Besov space $B^{\alpha}_{p,q}$ reduces into the Hölder space C^{α} and Sobolev space W^{α}_2 at $p = q = \infty, \alpha \in (0, \infty) \backslash \mathbb{N}$, and at $p = q = 2, \alpha \in \mathbb{N}$, respectively. Moreover, our dynamic assumption only requires the boundedness of a very general notion of local oscillations of the underlying MDP; that is, the underlying MDP can be discontinuous or non-differentiable (e.g., when $\alpha \leq 1/2$ and $p = 2$), or even have spatially inhomogeneous smoothness (e.g., when $p < 2$).

The condition $\alpha > \frac{d}{p \wedge 2}$ guarantees a finite bound for the compactness and the (local) Rademacher complexity of the considered Besov space. When $p < 2$ (thus the condition above becomes $\alpha > d/p$), a function in the corresponding Besov space contains both spiky parts and smooth parts, i.e., the Besov space has inhomogeneous smoothness (Suzuki, 2018). In particular, when $\alpha > d/p$, each equivalence class $[f]_{\lambda}, f \in B^{\alpha}_{p,q}(\mathbb{R}^d)$, i.e., modulo equality λ-almost everywhere, contains a unique continuous representative. In addition, this representative has partial derivatives of order at least $\alpha - d/p$; thus $\alpha - d/p$ is called the *differential dimension* of the Besov space. Finally, we remark that linear MDPs (Jin et al., 2020) corresponds to Assumption 3.2 with $\alpha = 1$ and $p = q$ on a p-norm bounded domain. However, the additional condition $\alpha > \frac{d}{p \wedge 2}$ is not necessary for the particular case of linear MDPs. This is due to the fact that there is a closed-form solution to the value regression problem in linear MDPs and the size of the linear models for MDP is controllably small without any additional smoothness assumption (rather than the completeness assumption). Of course, our analysis addresses significantly more complex and general settings than linear MDPs which we believe is more important than recovering the optimal condition in linear MDPs.

4 Algorithm and Theory

4.1 Algorithm

Now we turn to the main algorithm and the main result. We study a FQI-type algorithm, namely least-squares value iteration (LSVI) for both OPE and OPL with the pseudo-code presented in Algorithm 1 where we denote $\rho^{\pi}(s, a) = \rho(s)\pi(a|s)$. The algorithm is nearly identical to (Duan & Wang, 2020) but with deep neural network function approximation instead of linear models. As such, it can be considered as a generalization.

Algorithm 1 Least-squares value iteration (LSVI)

1: Initialize $Q_0 \in \mathcal{F}_{NN}$.
2: **for** $k = 1$ **to** K **do**
3: If **OPE** (for a fixed policy π): $y_i \leftarrow r_i + \gamma \int_{\mathcal{A}} Q_{k-1}(s'_i, a)\pi(da|s'_i), \forall i$
4: If **OPL**: $y_i \leftarrow r_i + \gamma \max_{a' \in \mathcal{A}} Q_{k-1}(s'_i, a'), \forall i$
5: $Q_k \leftarrow \arg\min_{f \in \mathcal{F}_{NN}} \frac{1}{n} \sum_{i=1}^{n} (f(s_i, a_i) - y_i)^2$
6: **end for**
7: If **OPE**, return $V_K = \|Q_K\|_{\rho^\pi} = \sqrt{\mathbb{E}_{\rho(s)\pi(a|s)}\left[Q_K(s,a)^2\right]}$
8: If **OPL**, return the greedy policy π_K w.r.t. Q_K.

The idea of LSVI is appealingly simple: it does the best it could with all the offline data using least-squares regression over a function space. The algorithm arbitrarily initializes $Q_0 \in \mathcal{F}_{NN}$ and iteratively computes Q_k as follows: at each iteration k, the algorithm constructs a new regression data $\{(x_i, y_i)\}_{i=1}^n$ where the covariates x_i are (s_i, a_i) and the Bellman targets y_i are computed following dynamic programming style. In particular, depending on whether this is an OPE or OPL problem, y_i are computed according to line 3 and line 4 of Algorithm 1, respectively. It then fits the function class $\mathcal{F}_{NN}$ to the constructed regression data by minimizing the mean squared error at line 5. This type of algorithm belongs to the fitted Q-iteration family (Munos & Szepesvári, 2008; Le et al., 2019) that iteratively uses least-squares (value) regression to estimate the value functions. The main difference in the algorithm is here we use deep neural networks as function approximation for generalization to unseen states and actions in a complex MDP.

On the computational side, solving the non-convex optimization at line 5 of Algorithm 1 can be highly involved and stochastic gradient descent is a dominant optimization method for such a task in deep learning. In particular, (stochastic) gradient descent is guaranteed to converge to a global minimum under certain structural assumptions (Du et al., 2019a;b; Allen-Zhu et al., 2019; Nguyen, 2021). Here, as we only focus on the statistical properties of LSVI, we assume that the minimizer at line 5 is attainable. Such a oracle assumption is common when analyzing the statistical properties of an RL algorithm with non-linear function approximation (Yang et al., 2019; Chen & Jiang, 2019; Duan et al., 2021; Wang et al., 2019; 2020; Jin et al., 2021).

4.2 Correlated Structure

We remark the correlated structure in Algorithm 1. The target variable y_i computed at line 3 and line 4 of the algorithm depends on the previous estimate Q_{k-1} which in turn depends on the covariate $x_i := (s_i, a_i)$. This induces a complex correlated structure across all iterations where the current estimate depends on all the previous estimates and the past data. In particular, one of the main difficulties caused by such correlated structure is that conditioned on each x_i, the target variable y_i is no longer centered at $[T^* Q_{k-1}](x_i)$ for OPL (or at $[T^\pi Q_{k-1}](x_i)$ for OPE, respectively), i.e., $\mathbb{E}\left[[T^* Q_{k-1}](x_i) - y_i | x_i\right] \neq 0$. This correlated structure hinders a direct use of the standard concentration inequalities (e.g. Hoeffding's inequality, Bernstein inequality). Prior results either improperly ignore the correlated structure in their analysis (Le et al., 2019) or directly avoid it by estimating each Q_k on a separate fold of the original data (Yang et al., 2019). The data splitting approach in (Yang et al., 2019), which splits the original data into K disjoint folds, helps remove the correlated structure but scales the sample complexity linearly with K where K can be arbitrarily large. In contrast, we overcome the correlated structure via a uniform convergence argument by considering deterministic coverings of the target function space $T^* \mathcal{F}_{NN}$ without the need for the inefficient data splitting.

4.3 Theoretical Analysis

Our main result is a sup-optimality bound for LSVI in both OPE and OPL settings.

Theorem 4.1. *Under Assumption 3.1 and Assumption 3.2, for any $\epsilon > 0, \delta \in (0, 1], K > 0$, if n satisfies that $n \gtrsim \left(\frac{1}{\epsilon^2}\right)^{1+\frac{d}{\alpha}} \log^6 n + \frac{1}{\epsilon^2}(\log(1/\delta) + \log\log n)$, then with probability at least $1 - \delta$,* *the sup-optimality of Algorithm 1 is*

$$\begin{cases} SubOpt(V_K;\pi) \leq \frac{\sqrt{\kappa_\mu}}{1-\gamma}\epsilon + \frac{\gamma^{K/2}}{(1-\gamma)^{1/2}} & \textit{for OPE}, \\ SubOpt(\pi_K) \leq \frac{4\gamma\sqrt{\kappa_\mu}}{(1-\gamma)^2}\epsilon + \frac{4\gamma^{1+K/2}}{(1-\gamma)^{3/2}} & \textit{for OPL}. \end{cases}$$

In addition, the optimal deep ReLU network $\Phi(L, m, S, B)$ that obtains such sample complexity (for both OPE and OPL) satisfies

$$L \asymp \log N, m \asymp N \log N, S \asymp N, \textit{ and } B \asymp N^{1/d+(2\iota)/(\alpha-\iota)},$$

where $\iota := d(p^{-1} - (1 + \lfloor\alpha\rfloor)^{-1})_+$ and $N \asymp n^{\frac{\frac{1}{2}+\left(2+\frac{d^2}{\alpha(\alpha+d)}\right)^{-1}}{1+\frac{2\alpha}{d}}}$ is the number of parameters to approximate a function in the Besov space.

Remark 4.1. The role of deep ReLU networks in offline RL is to guarantee a maximal adaptivity to the (spatial) regularity of the functions in Besov space and obtain an optimal approximation error rate that otherwise were not possible with other function approximation such as kernel methods (Suzuki, 2018). Moreover, by the equivalence in the functions that a neural architecture can compute (Yarotsky, 2017), Theorem 4.1 also readily holds for any other continuous piece-wise linear activation functions with finitely many line segments M where the optimal network architecture only increases the number of units and weights by constant factors depending only on M.

Remark 4.2. The optimal ReLU network that realizes our sample complexity can be further simplified as $L = \mathcal{O}(\log n)$ and $m = \mathcal{O}(\sqrt{n} \log n)$. That is, the optimal ReLU network is relatively "thinner" than overparameterized neural networks that have been recently studied in the literature (Arora et al., 2019; Allen-Zhu et al., 2019; Hanin & Nica, 2019; Cao & Gu, 2019; Belkin, 2021) where the width m is a high-order polynomial of n. As overparameterization is a key feature for such overparameterized neural networks to obtain a good generalization, it is natural to ask why a thinner neural network in Theorem 4.1 also guarantees a strong generalization for offline RL even when the network is not in the overparameterization regime? Intuitively, it is due to that the optimal ReLU network in Theorem 4.1 is regularized by a strong sparsity which resonates with our practical wisdom that we can use a sparsity-based regularization to prevent over-fitting and achieve a better generalization. In particular, as the total number of parameters in the considered neural network is $p = md + m + m^2(L-2) = \mathcal{O}(N^2 \log^3 N)$ while the number of non-zeros parameters S only scales with N, the optimal ReLU network in Theorem 4.1 is relatively sparse.

Theorem 4.1 states that LSVI incurs a sub-optimality which consists of the statistical error (the first term) and the algorithmic error (the second term). While the algorithmic error enjoys the fast linear convergence to 0, the statistical error reflects the fundamental difficulty of the problems. The statistical errors for both OPE and OPL cases are bounded by the distributional shift κ_μ, the effective horizon $1/(1-\gamma)$, and the user-specified precision ϵ for n satisfying the inequality given in Theorem 4.1. In particular, the sample complexity does not depend on the number of states as in tabular MDPs (Yin & Wang, 2020; Yin et al., 2021; Yin & Wang, 2021) or the inherent Bellman error as in the general function approximation (Munos & Szepesvári, 2008; Le et al., 2019). Instead, it explicitly scales with the (possible fractional) smoothness α of the underlying MDP and the dimension d of the input space. Importantly, this guarantee is established under the correlated structure of the value estimate in the algorithm and the Besov dynamic closure encompassing the dynamic conditions of the prior results. Thus, Theorem 4.1 is the most comprehensive result we are aware of for offline RL with deep neural network function approximation.

Moreover, to further develop an intuition on our sample complexity, we compare it with the prior results. Regarding the tightness of our result, our sample complexity $\epsilon^{-2-2d/\alpha}$ (ignoring the log factor and the factor pertaining to κ_μ and effective horizon) nearly matches the nonparametric regression's minimax-optimal sample complexity $\epsilon^{-2-d/\alpha}$ (Kerkyacharian & Picard, 1992; Giné & Nickl, 2016) even though in our case we deal with a more complicated correlated structure in a value iteration problem instead of a standard non-parametric regression problem. This gap is necessary and expected due to the correlated structure in the algorithm. We remark that it is possible to retain the rate $\epsilon^{-2-d/\alpha}$ if we split the offline data $\mathcal{D}$ into K (given in Algorithm 1) disjoint subsets and estimate each Q_k in Algorithm 1 using a separate disjoint subsets. This however scales the sample complexity linearly with K which could be arbitrarily large in practice.

Table 1: The state-of-the-art (SOTA) statistical theory of offline RL with function approximation. Here, the distributional shift measure κ can be defined differently in different works.

Work	**Function**	**Regularity**	**Tasks**	**Sample complexity**	**Remark**
Yin & Wang (2020)	Tabular	Tabular	OPE	$\tilde{\mathcal{O}}\left(\frac{\kappa}{\epsilon^2} \cdot \lvert\mathcal{S}\rvert^2 \cdot \lvert\mathcal{A}\rvert^2\right)$	minimax-optimal
Duan & Wang (2020)	Linear	Linear	OPE	$\tilde{\mathcal{O}}\left(\frac{\kappa}{\epsilon^2} \cdot d\right)$	minimax-optimal
Le et al. (2019)	General	General	OPE/OPL	N/A	improper analysis
Yang et al. (2019)	ReLU nets	Hölder	OPL	$\tilde{\mathcal{O}}\left(K \cdot \kappa^{2+\frac{d}{\alpha}} \cdot \epsilon^{-2-\frac{d}{\alpha}}\right)$	no data reuse
This work	ReLU nets	Besov	OPE/OPL	$\tilde{\mathcal{O}}\left(\kappa^{1+\frac{d}{\alpha}} \cdot \epsilon^{-2-2\frac{d}{\alpha}}\right)$	data reuse

To show the significance of our sample complexity, we summarize our result and compare it with the prior results in Table 1. From the leftmost column to the rightmost one, the table describes the related works, the function approximations being employed, the regularity conditions considered to establish theoretical guarantees, the offline RL tasks considered, the sample complexity obtained, and the important remarks or features of each work. Here, $|\mathcal{S}|$ and $|\mathcal{A}|$ are the cardinalities of the state and action space when they are finite. Specifically, the "data reuse" in Table 1 means that an algorithm reuses the data across all iterations instead of splitting the original offline data into disjoint subsets for each iteration and the regularity column specifies the regularity assumption on the underlying MDP. Based on this comparison, we make the following observations. First, with simpler models such as tabular and linear MDPs, it requires less samples to achieve the same sub-optimality precision ϵ than more complex environments such as Hölder and Besov MDPs. This should not come as a surprise as the simpler regularities are much easier to learn but they are too strong as a condition to hold in practice. Second, as remarked earlier that Besov smoothness is more general than Hölder smoothness considered in (Yang et al., 2019), our setting is more practical and comprehensive as it covers more scenarios of the regularity of the underlying MDPs than the prior results. Third, our result obtains an improved sample complexity as compared to that in (Yang et al., 2019) where we are able to get rid of the dependence on the algorithmic iteration number K which can be arbitrarily large in practice. On the technical side, we provide a unifying analysis that allows us to account for the complicated correlated structure in the algorithm and handle the complex deep ReLU network function approximation. This can also be considered as a substantial technical improvement over (Le et al., 2019) as Le et al. (2019) improperly ignores the correlated structure in their analysis. In addition, the result in (Le et al., 2019) does not provide an explicit sample complexity as it depends on an unknown inherent Bellman error. Thus, our sample complexity improves over the result of the data splitting method and holds with in a broader context by our Besov dynamic closure.

Finally, we provide a detailed proof for Theorem 4.1 in Section A. The proof has four main components: a sub-optimality decomposition for error propagation across iterations, a Bellman error decomposition using a uniform convergence argument, a deviation analysis for least-squares value regression with deep ReLU networks using local Rademacher complexities via a localization argument, and an upper bound minimization step to obtain an optimal deep ReLU architecture.

5 Conclusion

We presented the sample complexity of offline RL with deep ReLU network function approximation. We proved that the FQI-type algorithm can achieve the sample complexity of $\tilde{\mathcal{O}}\left(\kappa^{1+d/\alpha} \cdot \epsilon^{-2-2d/\alpha}\right)$ under highly correlated structures and a general dynamic condition namely the Besov dynamic closure. We also provided various insights into the benefits and the effects of deep neural networks in offline RL.

We close with a future direction. Although the finite concentration coefficient assumption is relatively standard in offline RL, can we develop a weaker, non-uniform assumption that can still accommodate offline RL with non-linear function approximation? While such a weaker data coverage assumptions do exist for offline RL in tabular settings (Rashidinejad et al., 2021), it seems non-trivial to generalize this condition to the function approximation setting.

REFERENCES

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *International Conference on Machine Learning*, pp. 242–252. PMLR, 2019.

András Antos, Csaba Szepesvári, and Rémi Munos. Learning near-optimal policies with bellman-residual minimization based fitted policy iteration and a single sample path. *Mach. Learn.*, 71 (1):89–129, April 2008. ISSN 0885-6125. doi: 10.1007/s10994-007-5038-2. URL https://doi.org/10.1007/s10994-007-5038-2.

Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, Ruslan Salakhutdinov, and Ruosong Wang. On exact computation with an infinitely wide neural net. *arXiv preprint arXiv:1904.11955*, 2019.

Peter L. Bartlett, Olivier Bousquet, and Shahar Mendelson. Local rademacher complexities. *Ann. Statist.*, 33(4):1497–1537, 08 2005. doi: 10.1214/009053605000000282. URL https://doi.org/10.1214/009053605000000282.

Mikhail Belkin. Fit without fear: remarkable mathematical phenomena of deep learning through the prism of interpolation. *arXiv preprint arXiv:2105.14368*, 2021.

Dimitri P Bertsekas and John N Tsitsiklis. Neuro-dynamic programming: an overview. In *Proceedings of 1995 34th IEEE Conference on Decision and Control*, volume 1, pp. 560–564. IEEE, 1995.

Yuan Cao and Quanquan Gu. Generalization bounds of stochastic gradient descent for wide and deep neural networks. *Advances in Neural Information Processing Systems*, 32:10836–10846, 2019.

Jinglin Chen and Nan Jiang. Information-theoretic considerations in batch reinforcement learning. In *ICML*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1042–1051. PMLR, 2019.

Albert Cohen. A primer on besov spaces, 2009. URL http://cnx.org/content/col10679/1.2/>.

Simon S. Du, Jason D. Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In *ICML*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1675–1685. PMLR, 2019a.

Simon S. Du, Xiyu Zhai, Barnabás Póczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. In *ICLR (Poster)*. OpenReview.net, 2019b.

Yaqi Duan and Mengdi Wang. Minimax-optimal off-policy evaluation with linear function approximation. *CoRR*, abs/2002.09516, 2020.

Yaqi Duan, Chi Jin, and Zhiyuan Li. Risk bounds and rademacher complexity in batch reinforcement learning. *arXiv preprint arXiv:2103.13883*, 2021.

Miroslav Dudík, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. *arXiv preprint arXiv:1103.4601*, 2011.

Mehrdad Farajtabar, Yinlam Chow, and Mohammad Ghavamzadeh. More robust doubly robust off-policy evaluation. *arXiv preprint arXiv:1802.03493*, 2018.

Max H Farrell, Tengyuan Liang, and Sanjog Misra. Deep neural networks for estimation and inference: Application to causal effects and other semiparametric estimands. *arXiv preprint arXiv:1809.09953*, 2018.

Evarist Giné and Richard Nickl. *Mathematical foundations of infinite-dimensional statistical models*, volume 40. Cambridge University Press, 2016.

Steffen Grünewälder, Guy Lever, Luca Baldassarre, Massimiliano Pontil, and Arthur Gretton. Modelling transition dynamics in mdps with RKHS embeddings. In *ICML*. icml.cc / Omnipress, 2012.

László Györfi, Michael Kohler, Adam Krzyzak, and Harro Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer series in statistics. Springer, 2002.

Boris Hanin and Mihai Nica. Finite depth and width corrections to the neural tangent kernel. *arXiv preprint arXiv:1909.05989*, 2019.

Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning, 2015.

Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. In *ICML*, volume 48 of *JMLR Workshop and Conference Proceedings*, pp. 652–661. JMLR.org, 2016.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pp. 2137–2143. PMLR, 2020.

Chi Jin, Qinghua Liu, and Sobhan Miryoosefi. Bellman eluder dimension: New rich classes of RL problems, and sample-efficient algorithms. *CoRR*, abs/2102.00815, 2021. URL https://arxiv.org/abs/2102.00815.

Nicholas K. Jong and Peter Stone. Model-based function approximation in reinforcement learning. In *AAMAS*, pp. 95. IFAAMAS, 2007.

Nathan Kallus and Masatoshi Uehara. Double reinforcement learning for efficient off-policy evaluation in markov decision processes, 2019.

Motonobu Kanagawa, Philipp Hennig, Dino Sejdinovic, and Bharath K Sriperumbudur. Gaussian processes and kernel methods: A review on connections and equivalences. *arXiv preprint arXiv:1807.02582*, 2018.

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

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pp. 1097–1105, 2012.

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning. *arXiv preprint arXiv:2006.04779*, 2020.

Michail G. Lagoudakis and Ronald Parr. Least-squares policy iteration. *J. Mach. Learn. Res.*, 4: 1107–1149, 2003.

Sascha Lange, Thomas Gabel, and Martin Riedmiller. Batch reinforcement learning. In *Reinforcement learning*, pp. 45–73. Springer, 2012.

Hoang Minh Le, Cameron Voloshin, and Yisong Yue. Batch policy learning under constraints. In *ICML*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3703–3712. PMLR, 2019.

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

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

Qiang Liu, Lihong Li, Ziyang Tang, and Dengyong Zhou. Breaking the curse of horizon: Infinite-horizon off-policy estimation. In *NeurIPS*, pp. 5361–5371, 2018.

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.

Rémi Munos. Error bounds for approximate policy iteration. In *ICML*, pp. 560–567. AAAI Press, 2003.

Rémi Munos and Csaba Szepesvári. Finite-time bounds for fitted value iteration. *J. Mach. Learn. Res.*, 9:815–857, 2008.

Ofir Nachum, Yinlam Chow, Bo Dai, and Lihong Li. Dualdice: Behavior-agnostic estimation of discounted stationary distribution corrections, 2019a.

Ofir Nachum, Bo Dai, Ilya Kostrikov, Yinlam Chow, Lihong Li, and Dale Schuurmans. Algaedice: Policy gradient from arbitrary experience. *ArXiv*, abs/1912.02074, 2019b.

Quynh Nguyen. On the proof of global convergence of gradient descent for deep relu networks with linear widths. *arXiv preprint arXiv:2101.09612*, 2021.

R. Nickl and B. M. Pötscher. Bracketing metric entropy rates and empirical central limit theorems for function classes of besov- and sobolev-type. *Journal of Theoretical Probability*, 20:177–199, 2007.

Doina Precup, Richard S. Sutton, and Satinder P. Singh. Eligibility traces for off-policy policy evaluation. In *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML '00, pp. 759–766, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc. ISBN 1558607072.

Paria Rashidinejad, Banghua Zhu, Cong Ma, Jiantao Jiao, and Stuart Russell. Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *arXiv preprint arXiv:2103.12021*, 2021.

Patrick Rebeschini. Oxford Algorithmic Foundations of Learning, Lecture Notes: Maximal Inequalities and Rademacher Complexity, 2019. URL: http://www.stats.ox.ac.uk/~rebeschi/teaching/AFoL/20/material/lecture02.pdf. Last visited on Sep. 14, 2020.

Yoshihiro Sawano. *Theory of Besov Spaces*, volume 56. Springer, 2018. ISBN 978-981-13-0835-2.

Avi Singh, Albert Yu, Jonathan Yang, Jesse Zhang, Aviral Kumar, and Sergey Levine. Cog: Connecting new skills to past experience with offline reinforcement learning. *arXiv preprint arXiv:2010.14500*, 2020.

Taiji Suzuki. Adaptivity of deep relu network for learning in besov and mixed smooth besov spaces: optimal rate and curse of dimensionality. *arXiv preprint arXiv:1810.08033*, 2018.

Philip Thomas and Emma Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. In *International Conference on Machine Learning*, pp. 2139–2148, 2016.

Samuele Tosatto, Matteo Pirotta, Carlo D'Eramo, and Marcello Restelli. Boosted fitted q-iteration. In *ICML*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3434–3443. PMLR, 2017.

Hung Tran-The, Sunil Gupta, Thanh Nguyen-Tang, Santu Rana, and Svetha Venkatesh. Combining online learning and offline learning for contextual bandits with deficient support. *arXiv preprint arXiv:2107.11533*, 2021.

H. Triebel. Theory of function spaces. 1983.

Cameron Voloshin, Hoang M Le, Nan Jiang, and Yisong Yue. Empirical study of off-policy policy evaluation for reinforcement learning. *arXiv preprint arXiv:1911.06854*, 2019.

Ruosong Wang, Ruslan Salakhutdinov, and Lin F. Yang. Provably efficient reinforcement learning with general value function approximation. *CoRR*, abs/2005.10804, 2020. URL https://arxiv.org/abs/2005.10804.

Yining Wang, Ruosong Wang, Simon S. Du, and Akshay Krishnamurthy. Optimism in reinforcement learning with generalized linear function approximation. *CoRR*, abs/1912.04136, 2019. URL http://arxiv.org/abs/1912.04136.

Tengyang Xie, Yifei Ma, and Yu-Xiang Wang. Towards optimal off-policy evaluation for reinforcement learning with marginalized importance sampling, 2019.

Zhuoran Yang, Yuchen Xie, and Zhaoran Wang. A theoretical analysis of deep q-learning. *CoRR*, abs/1901.00137, 2019.

Dmitry Yarotsky. Error bounds for approximations with deep relu networks. *Neural Networks*, 94: 103–114, 2017.

Ming Yin and Yu-Xiang Wang. Asymptotically efficient off-policy evaluation for tabular reinforcement learning. In *AISTATS*, volume 108 of *Proceedings of Machine Learning Research*, pp. 3948–3958. PMLR, 2020.

Ming Yin and Yu-Xiang Wang. Characterizing uniform convergence in offline policy evaluation via model-based approach: Offline learning, task-agnostic and reward-free, 2021.

Ming Yin, Yu Bai, and Yu-Xiang Wang. Near-optimal provable uniform convergence in offline policy evaluation for reinforcement learning. In Arindam Banerjee and Kenji Fukumizu (eds.), *The 24th International Conference on Artificial Intelligence and Statistics, AISTATS 2021, April 13-15, 2021, Virtual Event*, volume 130 of *Proceedings of Machine Learning Research*, pp. 1567–1575. PMLR, 2021. URL http://proceedings.mlr.press/v130/yin21a.html.

Ruiyi Zhang, Bo Dai, Lihong Li, and Dale Schuurmans. Gendice: Generalized offline estimation of stationary values. *ArXiv*, abs/2002.09072, 2020a.

Shangtong Zhang, Bo Liu, and Shimon Whiteson. Gradientdice: Rethinking generalized offline estimation of stationary values. *ArXiv*, abs/2001.11113, 2020b.

A Proof of Theorem 4.1

We now provide a complete proof of Theorem 4.1. The proof has four main components: a sub-optimality decomposition for error propagation across iterations, a Bellman error decomposition using a uniform convergence argument, a deviation analysis for least squares with deep ReLU networks using local Rademacher complexities and a localization argument, and a upper bound minimization step to obtain an optimal deep ReLU architecture.

Step 1: A sub-optimality decomposition

The first step of the proof is a sub-optimality decomposition, stated in Lemma A.1, that applies generally to any least-squares Q-iteration methods.

Lemma A.1 (*A sub-optimality decomposition*). *Under Assumption 3.1, the sub-optimality of V_K returned by Algorithm 1 is bounded as*

$$\textit{SubOpt}(V_K) \leq \begin{cases} \frac{\sqrt{\kappa_\mu}}{1-\gamma} \max\limits_{0 \leq k \leq K-1} \|Q_{k+1} - T^\pi Q_k\|_\mu + \dfrac{\gamma^{K/2}}{(1-\gamma)^{1/2}} & \text{for OPE,} \\ \frac{4\gamma\sqrt{\kappa_\mu}}{(1-\gamma)^2} \max\limits_{0 \leq k \leq K-1} \|Q_{k+1} - T^* Q_k\|_\mu + \dfrac{4\gamma^{1+K/2}}{(1-\gamma)^{3/2}} & \text{for OPL.} \end{cases}$$

where we denote $\|f\|_\mu := \sqrt{\int \mu(dsda) f(s,a)^2}, \forall f : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$.

The lemma states that the sub-optimality decomposes into a statistical error (the first term) and an algorithmic error (the second term). While the algorithmic error enjoys the fast linear convergence rate, the statistical error arises from the distributional shift in the offline data and the estimation error of the target Q-value functions due to finite data. Crucially, the contraction of the (optimality) Bellman operators T^π and T^* allows the sup-optimality error at the final iteration K to propagate across all iterations $k \in [0, K-1]$. Note that this result is agnostic to any function approximation form and does not require Assumption 3.2. The result uses a relatively standard argument that appears in a number of works on offline RL (Munos & Szepesvári, 2008; Le et al., 2019).

Proof of Lemma A.1. We will prove the sup-optimality decomposition for both settings: OPE and OPL.

(i) For OPE. We denote the right-linear operator by $P^\pi \cdot : \{\mathcal{X} \to \mathbb{R}\} \to \{\mathcal{X} \to \mathbb{R}\}$ where

$$(P^\pi f)(s,a) := \int_{\mathcal{X}} f(s', a') \pi(da'|s') P(ds'|s,a),$$

for any $f \in \{\mathcal{X} \to \mathbb{R}\}$. Denote Denote $\rho^\pi(dsda) = \rho(ds)\pi(da|s)$. Let $\epsilon_k := Q_{k+1} - T^\pi Q_k, \forall k \in [0, K-1]$ and $\epsilon_K = Q_0 - Q^\pi$. Since Q^π is the (unique) fixed point of T^π, we have

$$Q_k - Q^\pi = T^\pi Q_{k-1} - T^\pi Q^\pi + \epsilon_{k-1} = \gamma P^\pi (Q_{k-1} - Q^\pi) + \epsilon_{k-1}.$$

By recursion, we have

$$Q_K - Q^\pi = \sum_{k=0}^{K} (\gamma P^\pi)^k \epsilon_k = \frac{1-\gamma^{K+1}}{1-\gamma} \sum_{k=0}^{K} \alpha_k A_k \epsilon_k$$

where $\alpha_k := \frac{(1-\gamma)\gamma^k}{1-\gamma^{K+1}}, \forall k \in [K]$ and $A_k := (P^\pi)^k, \forall k \in [K]$. Note that $\sum_{k=0}^{K} \alpha_k = 1$ and A_k's are probability kernels. Denoting by $|f|$ the point-wise absolute value $|f(s,a)|$, we have that the following inequality holds point-wise:

$$|Q_K - Q^\pi| \leq \frac{1-\gamma^{K+1}}{1-\gamma} \sum_{k=0}^{K} \alpha_k A_k |\epsilon_k|.$$

We have

$$\|Q_K - Q^\pi\|^2_{\rho^\pi} \leq \frac{(1-\gamma^{K+1})^2}{(1-\gamma)^2} \int \rho(ds)\pi(da|s) \left(\sum_{k=0}^{K} \alpha_k A_k |\epsilon_k|(s,a) \right)^2$$

$$
\begin{aligned}
&\overset{(a)}{\leq} \frac{(1-\gamma^{K+1})^2}{(1-\gamma)^2} \int \rho(ds)\pi(da|s) \sum_{k=0}^{K} \alpha_k A_k^2 \epsilon_k^2(s,a) \\
&\overset{(b)}{\leq} \frac{(1-\gamma^{K+1})^2}{(1-\gamma)^2} \int \rho(ds)\pi(da|s) \sum_{k=0}^{K} \alpha_k A_k \epsilon_k^2(s,a) \\
&\overset{(c)}{\leq} \frac{(1-\gamma^{K+1})^2}{(1-\gamma)^2} \left(\int \rho(ds)\pi(da|s) \sum_{k=0}^{K-1} \alpha_k A_k \epsilon_k^2(s,a) + \alpha_K \right) \\
&\overset{(d)}{\leq} \frac{(1-\gamma^{K+1})^2}{(1-\gamma)^2} \left(\int \mu(ds,da) \sum_{k=0}^{K-1} \alpha_k \kappa_\mu \epsilon_k^2(s,a) + \alpha_K \right) \\
&= \frac{(1-\gamma^{K+1})^2}{(1-\gamma)^2} \left(\sum_{k=0}^{K-1} \alpha_k \kappa_\mu \|\epsilon_k\|_\mu^2 + \alpha_K \right) \\
&\leq \frac{\kappa_\mu}{(1-\gamma)^2} \max_{0 \leq k \leq K-1} \|\epsilon_k\|_\mu^2 + \frac{\gamma^K}{(1-\gamma)}.
\end{aligned}
$$

The inequalities (a) and (b) follow from Jensen's inequality, (c) follows from $\|Q_0\|_\infty, \|Q^\pi\|_\infty \leq 1$, and (d) follows from Assumption 3.1 that $\rho^\pi A_k = \rho^\pi (P^\pi)^k \leq \kappa_\mu \mu$. Thus we have

$$
\begin{aligned}
\text{SubOpt}(V_K; \pi) &= |V_K - V^\pi| \\
&= \left| \mathbb{E}_{\rho,\pi}[Q_K(s,a)] - \mathbb{E}_\rho[Q^\pi(s,a)] \right| \\
&\leq \mathbb{E}_{\rho,\pi}\left[|Q_K(s,a) - Q^\pi(s,a)| \right] \\
&\leq \sqrt{\mathbb{E}_{\rho,\pi}\left[(Q_K(s,a) - Q^\pi(s,a))^2 \right]} \\
&= \|Q_K - Q^\pi\|_{\rho^\pi} \\
&\leq \frac{\sqrt{\kappa_\mu}}{1-\gamma} \max_{0 \leq k \leq K-1} \|\epsilon_k\|_\mu + \frac{\gamma^{K/2}}{(1-\gamma)^{1/2}}.
\end{aligned}
$$

(ii) For OPL. The sup-optimality for the OPL setting is more complex than the OPE setting but the technical steps are relatively similar. In particular, let $\epsilon_{k-1} = T^* Q_{k-1} - Q_k, \forall k$ and $\pi^*(s) = \arg\max_a Q^*(s,a), \forall s$, we have

$$
\begin{aligned}
Q^* - Q_K &= T^{\pi^*} Q^* - T^{\pi^*} Q_{K-1} + \underbrace{T^{\pi^*} Q_{K-1} - T^* Q_{K-1}}_{\leq 0} + \epsilon_{K-1} \\
&\leq \gamma P^{\pi^*} (Q^* - Q_{K-1}) + \epsilon_{K-1} \\
&\leq \sum_{k=0}^{K-1} \gamma^{K-k-1} (P^{\pi^*})^{K-k-1} \epsilon_k + \gamma^K (P^{\pi^*})^K (Q^* - Q_0) (\text{by recursion}). \qquad (1)
\end{aligned}
$$

Now, let π_k be the greedy policy w.r.t. Q_k, we have

$$
\begin{aligned}
Q^* - Q_K &= \underbrace{T^{\pi^*} Q^*}_{\geq T^{\pi_{K-1}} Q^*} - T^{\pi_{K-1}} Q_{K-1} + \underbrace{T^{\pi_{K-1}} Q_{K-1} - T^* Q_{K-1}}_{\geq 0} + \epsilon_{K-1} \\
&\geq \gamma P^{\pi_{K-1}} (Q^* - Q_{K-1}) + \epsilon_{K-1} \\
&\geq \sum_{k=0}^{K-1} \gamma^{K-k-1} (P^{\pi_{K-1}} \dots P^{\pi_{k+1}}) \epsilon_k + \gamma^K (P^{\pi_{K-1}} \dots P^{\pi_0}) (Q^* - Q_0). \qquad (2)
\end{aligned}
$$

Now, we turn to decompose $Q^* - Q^{\pi_K}$ as

$$
\begin{aligned}
Q^* - Q^{\pi_K} &= (T^{\pi^*} Q^* - T^{\pi^*} Q_K) + \underbrace{(T^{\pi^*} Q_K - T^{\pi_K} Q_K)}_{\leq 0} + (T^{\pi_K} Q_K - T^{\pi_K} Q^{\pi_K}) \\
&\leq \gamma P^{\pi^*} (Q^* - Q_K) + \gamma P^{\pi_K} (Q_K - Q^* + Q^* - Q^{\pi_K}).
\end{aligned}
$$

Thus, we have

$$(I - \gamma P^{\pi_K})(Q^* - Q^{\pi_K}) \le \gamma(P^{\pi^*} - P^{\pi_K})(Q^* - Q_K).$$

Note that the operator $(I - \gamma P^{\pi_K})^{-1} = \sum_{i=0}^{\infty} (\gamma P^{\pi_K})^i$ is monotone, thus

$$Q^* - Q^{\pi_K} \le \gamma (I - \gamma P^{\pi_K})^{-1} P^{\pi^*}(Q^* - Q_K) - \gamma (I - \gamma P^{\pi_K})^{-1} P^{\pi_K}(Q^* - Q_K). \tag{3}$$

Combining Equation (3) with Equations (1) and (2), we have

$$\begin{aligned} Q^* - Q^{\pi_K} \le (I - \gamma P^{\pi_K})^{-1} \left(\sum_{k=0}^{K-1} \gamma^{K-k} (P^{\pi^*})^{K-k} \epsilon_k + \gamma^{K+1} (P^{\pi^*})^{K+1} (Q^* - Q_0) \right) - \\ (I - \gamma P^{\pi_K})^{-1} \left(\sum_{k=0}^{K-1} \gamma^{K-k} (P^{\pi_K} \dots P^{\pi_{k+1}}) \epsilon_k + \gamma^{K+1} (P^{\pi_K} \dots P^{\pi_0}) (Q^* - Q_0) \right). \end{aligned}$$

Using the triangle inequality, the above inequality becomes

$$Q^* - Q^{\pi_K} \le \frac{2\gamma(1-\gamma^{K+1})}{(1-\gamma)^2} \left(\sum_{k=0}^{K-1} \alpha_k A_k |\epsilon_k| + \alpha_K A_K |Q^* - Q_0| \right),$$

where

$$\begin{aligned} A_k &= \frac{1-\gamma}{2} (I - \gamma P^{\pi_K})^{-1} \left((P^{\pi^*})^{K-k} + P^{\pi_K} \dots P^{\pi_{k+1}} \right), \forall k < K, \\ A_K &= \frac{1-\gamma}{2} (I - \gamma P^{\pi_K})^{-1} \left((P^{\pi^*})^{K+1} + P^{\pi_K} \dots P^{\pi_0} \right), \\ \alpha_k &= \gamma^{K-k-1}(1-\gamma)/(1-\gamma^{K+1}), \forall k < K, \\ \alpha_K &= \gamma^K (1-\gamma)/(1-\gamma^{K+1}). \end{aligned}$$

Note that A_k is a probability kernel for all k and $\sum_k \alpha_k = 1$. Thus, similar to the steps in the OPE setting, for any policy π, we have

$$\begin{aligned} \|Q^* - Q^{\pi_K}\|_{\rho^\pi}^2 &\le \left[\frac{2\gamma(1-\gamma^{K+1})}{(1-\gamma)^2} \right]^2 \left(\int \rho(ds)\pi(da|s) \sum_{k=0}^{K-1} \alpha_k A_k \epsilon_k^2(s,a) + \alpha_K \right) \\ &\le \left[\frac{2\gamma(1-\gamma^{K+1})}{(1-\gamma)^2} \right]^2 \left(\int \mu(ds,da) \sum_{k=0}^{K-1} \alpha_k \kappa_\mu \epsilon_k^2(s,a) + \alpha_K \right) \\ &= \left[\frac{2\gamma(1-\gamma^{K+1})}{(1-\gamma)^2} \right]^2 \left(\sum_{k=0}^{K-1} \alpha_k \kappa_\mu \|\epsilon_k\|_\mu^2 + \alpha_K \right) \\ &\le \frac{4\gamma^2 \kappa_\mu}{(1-\gamma)^4} \max_{0 \le k \le K-1} \|\epsilon_k\|_\mu^2 + \frac{4\gamma^{K+2}}{(1-\gamma)^3}. \end{aligned}$$

Thus, we have

$$\|Q^* - Q^{\pi_K}\|_{\rho^\pi} \le \frac{2\gamma\sqrt{\kappa_\mu}}{(1-\gamma)^2} \max_{0 \le k \le K-1} \|\epsilon_k\|_\mu + \frac{2\gamma^{K/2+1}}{(1-\gamma)^{3/2}}.$$

Finally, we have

$$\begin{aligned} \text{SubOpt}(\pi_K) &= \mathbb{E}_\rho \left[Q^*(s, \pi^*(s)) - Q^*(s, \pi_K(s)) \right] \\ &\le \mathbb{E}_\rho \left[Q^*(s, \pi^*(s)) - Q^{\pi_K}(s, \pi^*(s)) + Q^{\pi_K}(s, \pi_K(s)) - Q^*(s, \pi_K(s)) \right] \\ &\le \|Q^* - Q^{\pi_K}\|_{\rho^{\pi^*}} + \|Q^* - Q^{\pi_K}\|_{\rho^{\pi_K}} \\ &\le \frac{4\gamma\sqrt{\kappa_\mu}}{(1-\gamma)^2} \max_{0 \le k \le K-1} \|\epsilon_k\|_\mu + \frac{4\gamma^{K/2+1}}{(1-\gamma)^{3/2}}. \end{aligned}$$

□

STEP 2: A BELLMAN ERROR DECOMPOSITION

The next step of the proof is to decompose the Bellman errors $\|Q_{k+1} - T^{\pi} Q_k\|_{\mu}$ for OPE and $\|Q_{k+1} - T^* Q_k\|_{\mu}$ for OPL. Since these errors can be decomposed and bounded similarly, we only focus on OPL here.

The difficulty in controlling the estimation error $\|Q_{k+1} - T^* Q_k\|_{2,\mu}$ is that Q_k itself is a random variable that depends on the offline data $\mathcal{D}$. In particular, at any fixed k with Bellman targets $\{y_i\}_{i=1}^n$ where $y_i = r_i + \gamma \max_{a'} Q_k(s_i', a')$, it is not immediate that $\mathbb{E}\left[[T^* Q_k](x_i) - y_i | x_i\right] = 0$ for each covariate $x_i := (s_i, a_i)$ as Q_k itself depends on x_i (thus the tower law cannot apply here). A naive and simple approach to break such data dependency of Q_k is to split the original data $\mathcal{D}$ into K disjoint subsets and estimate each Q_k using a separate subset. This naive approach is equivalent to the setting in (Yang et al., 2019) where a fresh batch of data is generated for different iterations. This approach is however not efficient as it uses only n/K samples to estimate each Q_k. This is problematic in high-dimensional offline RL when the number of iterations K can be very large as it is often the case in practical settings. We instead prefer to use all n samples to estimate each Q_k. This requires a different approach to handle the complicated data dependency of each Q_k. To circumvent this issue, we leverage a uniform convergence argument by introducing a deterministic covering of $T^* \mathcal{F}_{NN}$. Each element of the deterministic covering induces a different regression target $\{r_i + \gamma \max_{a'} \tilde{Q}(s_i', a')\}_{i=1}^n$ where $\tilde{Q}$ is a deterministic function from the covering which ensures that $\mathbb{E}\left[r_i + \gamma \max_{a'} \tilde{Q}(s_i', a') - [T^* \tilde{Q}](x_i) | x_i\right] = 0$. In particular, we denote

$$y_i^{Q_k} = r_i + \gamma \max_{a'} Q_k(s_i', a'), \forall i \text{ and } \hat{f}^{Q_k} := Q_{k+1} = \arg\inf_{f \in \mathcal{F}_{NN}} \sum_{i=1}^n l(f(x_i), y_i^{Q_k}), \text{ and } f_*^{Q_k} = T^* Q_k,$$

where $l(x, y) = (x - y)^2$ is the squared loss function. Note that for any deterministic $Q \in \mathcal{F}_{NN}$, we have $f_*^Q(x_1) = \mathbb{E}[y_1^Q | x_1], \forall x_1$, thus

$$\mathbb{E}(l_f - l_{f_*^Q}) = \|f - f_*^Q\|_{\mu}^2, \forall f, \tag{4}$$

where l_f denotes the random variable $(f(x_1) - y_1^Q)^2$. Now letting $f_\perp^Q := \arg\inf_{f \in \mathcal{F}_{NN}} \|f - f_*^Q\|_{2,\mu}$ be the projection of f_*^Q onto the function class $\mathcal{F}_{NN}$, we have

$$\begin{aligned}
\max_k \|Q_{k+1} - T^* Q_k\|_{\mu}^2 = \max_k \|\hat{f}^{Q_k} - f_*^{Q_k}\|_{\mu}^2 &\overset{(a)}{\leq} \sup_{Q \in \mathcal{F}_{NN}} \|\hat{f}^Q - f_*^Q\|_{\mu}^2 \overset{(b)}{=} \sup_{Q \in \mathcal{F}_{NN}} \mathbb{E}(l_{\hat{f}^Q} - l_{f_*^Q}) \\
&\overset{(c)}{\leq} \sup_{Q \in \mathcal{F}_{NN}} \left\{ \mathbb{E}(l_{\hat{f}^Q} - l_{f_*^Q}) + \mathbb{E}_n(l_{f_\perp^Q} - l_{\hat{f}^Q}) \right\} \\
&= \sup_{Q \in \mathcal{F}_{NN}} \left\{ (\mathbb{E} - \mathbb{E}_n)(l_{\hat{f}^Q} - l_{f_*^Q}) + \mathbb{E}_n(l_{f_\perp^Q} - l_{f_*^Q}) \right\} \\
&\leq \underbrace{\sup_{Q \in \mathcal{F}_{NN}} (\mathbb{E} - \mathbb{E}_n)(l_{\hat{f}^Q} - l_{f_*^Q})}_{I_1, \text{empirical process term}} + \underbrace{\sup_{Q \in \mathcal{F}_{NN}} \mathbb{E}_n(l_{f_\perp^Q} - l_{f_*^Q})}_{I_2, \text{bias term}},
\end{aligned} \tag{5}$$

where (a) follows from that $Q_k \in \mathcal{F}_{NN}$, (b) follows from Equation (4), and (c) follows from that $\mathbb{E}_n[l_{\hat{f}^Q}] \leq \mathbb{E}_n[l_{f^Q}], \forall f, Q \in \mathcal{F}_{NN}$. That is, the error is decomposed into two terms: the first term I_1 resembles the empirical process in statistical learning theory and the second term I_2 specifies the bias caused by the regression target f_*^Q not being in the function space $\mathcal{F}_{NN}$.

STEP 3: A DEVIATION ANALYSIS

The next step is to bound the empirical process term and the bias term via an intricate concentration, local Rademacher complexities and a localization argument. First, the bias term in Equation (5) is taken uniformly over the function space, thus standard concentration arguments such as Bernstein's inequality and Pollard's inequality used in (Munos & Szepesvári, 2008; Le et al., 2019) do not apply here. Second, local Rademacher complexities (Bartlett et al., 2005) are data-dependent complexity measures that exploit the fact that only a small subset of the function class will be used. Leveraging a localization argument for local Rademacher complexities (Farrell et al., 2018), we localize an empirical Rademacher ball into smaller balls by which we can handle their complexities more effectively. Moreover, we explicitly use the sub-root function argument to derive our bound and extend the technique to the uniform convergence case. That is, reasoning over the sub-root function argument makes our proof more modular and easier to incorporate the uniform convergence argument.

Localization is particularly useful to handle the complicated approximation errors induced by deep ReLU network function approximation.

STEP 3.A: BOUNDING THE BIAS TERM VIA A UNIFORM CONVERGENCE CONCENTRATION INEQUALITY

Before delving into our proof, we introduce relevant notations. Let $\mathcal{F}-\mathcal{G} := \{f-g : f \in \mathcal{F}, g \in \mathcal{G}\}$, let $N(\epsilon, \mathcal{F}, \|\cdot\|)$ be the ϵ-covering number of $\mathcal{F}$ w.r.t. $\|\cdot\|$ norm, $H(\epsilon, \mathcal{F}, \|\cdot\|) := \log N(\epsilon, \mathcal{F}, \|\cdot\|)$ be the entropic number, let $N_{[]}(\epsilon, \mathcal{F}, \|\cdot\|)$ be the bracketing number of $\mathcal{F}$, i.e., the minimum number of brackets of $\|\cdot\|$-size less than or equal to ϵ, necessary to cover $\mathcal{F}$, let $H_{[]}(\epsilon, \mathcal{F}, \|\cdot\|) = \log N_{[]}(\epsilon, \mathcal{F}, \|\cdot\|)$ be the $\|\cdot\|$-bracketing metric entropy of $\mathcal{F}$,let $\mathcal{F}|\{x_i\}_{i=1}^n = \{(f(x_1), ..., f(x_n)) \in \mathbb{R}^n | f \in \mathcal{F}\}$, and let $T^*\mathcal{F} = \{T^* f : f \in \mathcal{F}\}$. Finally, for sample set $\{x_i\}_{i=1}^n$, we define the empirical norm $\|f\|_n := \sqrt{\frac{1}{n}\sum_{i=1}^n f(x_i)^2}$.

We define the inherent Bellman error as $d_{\mathcal{F}_{NN}} := \sup_{Q\in\mathcal{F}_{NN}} \inf_{f\in\mathcal{F}_{NN}} \|f - T^*Q\|_\mu$. This implies that

$$d^2_{\mathcal{F}_{NN}} := \sup_{Q\in\mathcal{F}_{NN}} \inf_{f\in\mathcal{F}_{NN}} \|f - T^*Q\|^2_\mu = \sup_{Q\in\mathcal{F}_{NN}} \mathbb{E}(l_{f^Q_\perp} - l_{f^Q_*}). \tag{6}$$

We have

$$|l_f - l_g| \leq 4|f-g| \text{ and } |l_f - l_g| \leq 8.$$

We have

$$\begin{aligned}
&H(\epsilon, \{l_{f^Q_\perp} - l_{f^Q_*} : Q \in \mathcal{F}_{NN}\}|\{x_i, y_i\}_{i=1}^n, n^{-1}\|\cdot\|_1) \\
&\leq H(\frac{\epsilon}{4}, \{f^Q_\perp - f^Q_* : Q \in \mathcal{F}_{NN}\}|\{x_i\}_{i=1}^n, n^{-1}\|\cdot\|_1) \\
&\leq H(\frac{\epsilon}{4}, (\mathcal{F} - T^*\mathcal{F}_{NN})|\{x_i\}_{i=1}^n, n^{-1}\|\cdot\|_1) \\
&\leq H(\frac{\epsilon}{8}, \mathcal{F}_{NN}|\{x_i\}_{i=1}^n, n^{-1}\|\cdot\|_1) + H(\frac{\epsilon}{8}, T^*\mathcal{F}_{NN}|\{x_i\}_{i=1}^n, n^{-1}\|\cdot\|_1) \\
&\leq H(\frac{\epsilon}{8}, \mathcal{F}_{NN}|\{x_i\}_{i=1}^n, \|\cdot\|_\infty) + H(\frac{\epsilon}{8}, T^*\mathcal{F}_{NN}, \|\cdot\|_\infty)
\end{aligned}$$

For any $\epsilon' > 0$ and $\delta' \in (0,1)$, it follows from Lemma B.2 with $\epsilon = 1/2$ and $\alpha = \epsilon'^2$, with probability at least $1-\delta'$, for any $Q \in \mathcal{F}_{NN}$, we have

$$\mathbb{E}_n(l_{f^Q_\perp} - l_{f^Q_*}) \leq 3\mathbb{E}(l_{f^Q_\perp} - l_{f^Q_*}) + \epsilon'^2 \leq 3d^2_{\mathcal{F}_{NN}} + \epsilon'^2, \tag{7}$$

given that

$$n \approx \frac{1}{\epsilon'^2}\left(\log(4/\delta') + \log \mathbb{E}N(\frac{\epsilon'^2}{40}, (\mathcal{F}_{NN} - T^*\mathcal{F}_{NN})|\{x_i\}_{i=1}^n, n^{-1}\|\cdot\|_1)\right).$$

Note that if we use Pollard's inequality (Munos & Szepesvári, 2008) in the place of Lemma B.2, the RHS of Equation (7) is bounded by ϵ' instead of ϵ'^2(i.e., n scales with $O(1/\epsilon'^4)$ instead of $O(1/\epsilon'^2)$). In addition, unlike (Le et al., 2019), the uniform convergence argument hinders the application of Bernstein's inequality. We remark that Le et al. 2019 makes a mistake in their proof by ignoring the data-dependent structure in the algorithm (i.e., they wrongly assume that Q^k in Algorithm 1 is fixed and independent of $\{s_i, a_i\}_{i=1}^n$). Thus, the uniform convergence argument in our proof is necessary.

Step 3.b: Bounding the empirical process term via local Rademacher complexities

For any $Q \in \mathcal{F}_{NN}$, we have

$$|l_{f_\perp^Q} - l_{f_*^Q}| \leq 2|f_\perp^Q - f_*^Q| \leq 2,$$
$$\mathbb{V}[l_{f_\perp^Q} - l_{f_*^Q}] \leq \mathbb{E}[(l_{f_\perp^Q} - l_{f_*^Q})^2] \leq 4\mathbb{E}(f_\perp^Q - f_*^Q)^2.$$

Thus, it follows from Lemma 1 (with $\alpha = 1/2$) that with any $r > 0, \delta \in (0,1)$, with probability at least $1-\delta$, we have

$$\begin{aligned}
&\sup\{(\mathbb{E} - \mathbb{E}_n)(l_{\hat{f}^Q} - l_{f_*^Q}) : Q \in \mathcal{F}_{NN}, \|\hat{f}^Q - f_*^Q\|_\mu^2 \leq r\} \\
&\leq \sup\{(\mathbb{E} - \mathbb{E}_n)(l_f - l_g) : f \in \mathcal{F}_{NN}, g \in T^*\mathcal{F}, \|f-g\|_\mu^2 \leq r\} \\
&\leq 3\mathbb{E}R_n\left\{l_f - l_g : f \in \mathcal{F}_{NN}, g \in T^*\mathcal{F}_{NN}, \|f-g\|_\mu^2 \leq r\right\} + 2\sqrt{\frac{2r\log(1/\delta)}{n}} + \frac{28\log(1/\delta)}{3n} \\
&\leq 6\mathbb{E}R_n\left\{f - g : f \in \mathcal{F}_{NN}, g \in T^*\mathcal{F}_{NN}, \|f-g\|_\mu^2 \leq r\right\} + 2\sqrt{\frac{2r\log(1/\delta)}{n}} + \frac{28\log(1/\delta)}{3n}.
\end{aligned}$$

Step 3.c: Bounding $\|Q_{k+1} - T^*Q_k\|_\mu$ using localization argument via sub-root functions

We bound $\|Q_{k+1} - T^*Q_k\|_\mu$ using the localization argument, breaking down the Rademacher complexities into local balls and then build up the original function space from the local balls. Let ψ be a sub-root function (Bartlett et al., 2005, Definition 3.1) with the fixed point r_* and assume that for any $r \geq r_*$, we have

$$\psi(r) \geq 3\mathbb{E}R_n\left\{f - g : f \in \mathcal{F}_{NN}, g \in T^*\mathcal{F}_{NN}, \|f-g\|_\mu^2 \leq r\right\}. \tag{8}$$

We recall that a function $\psi : [0,\infty) \to [0,\infty)$ is *sub-root* if it is non-negative, non-decreasing and $r \mapsto \psi(r)/\sqrt{r}$ is non-increasing for $r > 0$. Consequently, a sub-root function ψ has a unique fixed point r_* where $r_* = \psi(r_*)$. In addition, $\psi(r) \leq \sqrt{rr_*}, \forall r \geq r_*$. In the next step, we will find a sub-root function ψ that satisfies the inequality above, but for this step we just assume that we have such ψ at hand. Combining Equations (5), (7), and (8), we have: for any $r \geq r_*$ and any $\delta \in (0,1)$, if $\|\hat{f}^{Q_{k-1}} - f_*^{Q_{k-1}}\|_{2,\mu}^2 \leq r$, with probability at least $1-\delta$,

$$\begin{aligned}
\|\hat{f}^{Q_{k-1}} - f_*^{Q_{k-1}}\|_{2,\mu}^2 &\leq 2\psi(r) + 2\sqrt{\frac{2r\log(2/\delta)}{n}} + \frac{28\log(2/\delta)}{3n} + 3d_{\mathcal{F}}^2 + \epsilon'^2 \\
&\leq \sqrt{rr_*} + 2\sqrt{\frac{2r\log(2/\delta)}{n}} + \frac{28\log(2/\delta)}{3n} + (\sqrt{3}d_{\mathcal{F}} + \epsilon')^2,
\end{aligned}$$

where

$$n \approx \frac{1}{4\epsilon'^2}\left(\log(8/\delta) + \log \mathbb{E}N(\frac{\epsilon'^2}{20}, (\mathcal{F}_{NN} - T^*\mathcal{F}_{NN})|\{x_i\}_{i=1}^n, n^{-1}\|\cdot\|_1)\right).$$

Consider $r_0 \geq r_*$ (to be chosen later) and denote the events

$$B_k := \{\|\hat{f}^{Q_{k-1}} - f_*^{Q_{k-1}}\|_{2,\mu}^2 \leq 2^k r_0\}, \forall k \in \{0, 1, ..., l\},$$

where $l = \log_2(\frac{1}{r_0}) \leq \log_2(\frac{1}{r_*})$. We have $B_0 \subseteq B_1 \subseteq ... \subseteq B_l$ and since $\|f-g\|_\mu^2 \leq 1, \forall |f|_\infty, |g|_\infty \leq 1$, we have $P(B_l) = 1$. If $\|\hat{f}^{Q_{k-1}} - f_*^{Q_{k-1}}\|_\mu^2 \leq 2^i r_0$ for some $i \leq l$, then with probability at least $1-\delta$, we have

$$\begin{aligned}
\|\hat{f}^{Q_{k-1}} - f_*^{Q_{k-1}}\|_{2,\mu}^2 &\leq \sqrt{2^i r_0 r_*} + 2\sqrt{\frac{2^{i+1} r_0 \log(2/\delta)}{n}} + \frac{28\log(2/\delta)}{3n} + (\sqrt{3}d_{\mathcal{F}_{NN}} + \epsilon')^2 \\
&\leq 2^{i-1} r_0,
\end{aligned}$$

if the following inequalities hold

$$\sqrt{2^i r_*} + 2\sqrt{\frac{2^{i+1}\log(2/\delta)}{n}} \le \frac{1}{2}2^{i-1}\sqrt{r_0},$$
$$\frac{28\log(2/\delta)}{3n} + (\sqrt{3}d_{\mathcal{F}_{NN}} + \epsilon')^2 \le \frac{1}{2}2^{i-1}r_0.$$

We choose $r_0 \ge r_*$ such that the inequalities above hold for all $0 \le i \le l$. This can be done by simply setting

$$\sqrt{r_0} = \frac{2}{2^{i-1}}\left(\sqrt{2^i r_*} + 2\sqrt{\frac{2^{i+1}\log(2/\delta)}{n}}\right)\Bigg|_{i=0} + \sqrt{\frac{2}{2^{i-1}}\left(\frac{28\log(2/\delta)}{3n} + (\sqrt{3}d_{\mathcal{F}_{NN}} + \epsilon')^2\right)}\Bigg|_{i=0}$$
$$\lesssim d_{\mathcal{F}_{NN}} + \epsilon' + \sqrt{\frac{\log(2/\delta)}{n}} + \sqrt{r_*}.$$

Since $\{B_i\}$ is a sequence of increasing events, we have

$$P(B_0) = P(B_1) - P(B_1 \cap B_0^c) = P(B_2) - P(B_2 \cap B_1^c) - P(B_1 \cap B_0^c)$$
$$= P(B_l) - \sum_{i=0}^{l-1} P(B_{i+1} \cap B_i^c) \ge 1 - l\delta.$$

Thus, with probability at least $1-\delta$, we have

$$\|\hat{f}^{Q_{k-1}} - f_*^{Q_{k-1}}\|_\mu \lesssim d_{\mathcal{F}_{NN}} + \epsilon' + \sqrt{\frac{\log(2l/\delta)}{n}} + \sqrt{r_*} \tag{9}$$

where

$$n \approx \frac{1}{4\epsilon'^2}\left(\log(8l/\delta) + \log \mathbb{E}N(\frac{\epsilon'^2}{20}, (\mathcal{F}_{NN} - T^*\mathcal{F}_{NN})|\{x_i\}_{i=1}^n, n^{-1}\|\cdot\|_1))\right).$$

STEP 3.D: FINDING A SUB-ROOT FUNCTION AND ITS FIXED POINT

It remains to find a sub-root function $\psi(r)$ that satisfies Equation (8) and thus its fixed point. The main idea is to bound the RHS, the local Rademacher complexity, of Equation (8) by its empirical counterpart as the latter can then be further bounded by a sub-root function represented by a measure of compactness of the function spaces $\mathcal{F}_{NN}$ and $T^*\mathcal{F}_{NN}$.

For any $\epsilon > 0$, we have the following inequalities for entropic numbers:

$$H(\epsilon, \mathcal{F}_{NN} - T^*\mathcal{F}_{NN}, \|\cdot\|_n) \le H(\epsilon/2, \mathcal{F}_{NN}, \|\cdot\|_n) + H(\epsilon/2, T^*\mathcal{F}_{NN}, \|\cdot\|_n),$$
$$H(\epsilon, \mathcal{F}_{NN}, \|\cdot\|_n) \le H(\epsilon, \mathcal{F}_{NN}|\{x_i\}_{i=1}^n, \|\cdot\|_\infty) \overset{(a)}{\lesssim} N[(\log N)^2 + \log(1/\epsilon)], \tag{10}$$
$$H(\epsilon, T^*\mathcal{F}_{NN}, \|\cdot\|_n) \le H(\epsilon, T^*\mathcal{F}_{NN}, \|\cdot\|_\infty) \le H_{[]}(2\epsilon, T^*\mathcal{F}_{NN}, \|\cdot\|_\infty)$$
$$\overset{(b)}{\le} H_{[]}(2\epsilon, \bar{B}^\alpha_{p,q}(\mathcal{X}), \|\cdot\|_\infty) \overset{(c)}{\lesssim} (2\epsilon)^{-d/\alpha}, \tag{11}$$

where N is a hyperparameter of the deep ReLU network described in Lemma B.9, (a) follows from Lemma B.9, and (b) follows from Assumption 3.2, and (c) follows from Lemma B.8. Let $\mathcal{H} := \mathcal{F}_{NN} - T^*\mathcal{F}_{NN}$, it follows from Lemma B.5 with $\{\xi_k := \epsilon/2^k\}_{k\in\mathbb{N}}$ for any $\epsilon > 0$ that

$$\mathbb{E}_\sigma R_n\{h \in \mathcal{H} - \mathcal{H} : \|h\|_n \le \epsilon\} \le 4\sum_{k=1}^\infty \frac{\epsilon}{2^{k-1}}\sqrt{\frac{H(\epsilon/2^{k-1}, \mathcal{H}, \|\cdot\|_n)}{n}}$$
$$\le 4\sum_{k=1}^\infty \frac{\epsilon}{2^{k-1}}\sqrt{\frac{H(\epsilon/2^k, \mathcal{F}_{NN}, \|\cdot\|_\infty)}{n}} + 4\sum_{k=1}^\infty \frac{\epsilon}{2^{k-1}}\sqrt{\frac{H(\epsilon/2^k, T^\pi\mathcal{F}_{NN}, \|\cdot\|_\infty)}{n}}$$
$$\le \frac{4\epsilon}{\sqrt{n}}\sum_{k=1}^\infty 2^{-(k-1)}\sqrt{N\left((\log N)^2 + \log(2^k/\epsilon)\right)} + \frac{4\epsilon}{\sqrt{n}}\sum_{k=1}^\infty 2^{-(k-1)}\sqrt{\left(\frac{\epsilon}{2^{k-1}}\right)^{-d/\alpha}}$$

$$\lesssim \frac{\epsilon}{\sqrt{n}}\sqrt{N((\log N)^2+\log(1/\epsilon))}+\frac{\epsilon^{1-\frac{d}{2\alpha}}}{\sqrt{n}},$$

where we use $\sqrt{a+b}\leq\sqrt{a}+\sqrt{b},\forall a,b\geq 0$, $\sum_{k=1}^{\infty}\frac{\sqrt{k}}{2^{k-1}}<\infty$, and $\sum_{k=1}^{\infty}\left(\frac{1}{2^{1-\frac{d}{2\alpha}}}\right)^{k-1}<\infty$.

It now follows from Lemma B.4 that

$$\begin{aligned}
&\mathbb{E}_\sigma R_n\{f\in\mathcal{F},g\in T^*\mathcal{F}:\|f-g\|_n^2\leq r\}\\
&\leq \inf_{\epsilon>0}\left[\mathbb{E}_\sigma R_n\{h\in\mathcal{H}-\mathcal{H}:\|h\|_\mu\leq\epsilon\}+\sqrt{\frac{2rH(\epsilon/2,\mathcal{H},\|\cdot\|_n)}{n}}\right]\\
&\lesssim \left[\frac{\epsilon}{\sqrt{n}}\sqrt{N((\log N)^2+\log(1/\epsilon))}+\frac{\epsilon^{1-\frac{d}{2\alpha}}}{\sqrt{n}}+\sqrt{\frac{2r}{n}}\sqrt{N((\log N)^2+\log(4/\epsilon))}+\sqrt{\frac{2r}{n}}(\epsilon/2)^{\frac{-d}{2\alpha}}\right]\Bigg|_{\epsilon=n^{-\beta}}\\
&\asymp n^{-\beta-1/2}\sqrt{N(\log^2 N+\log n)}+n^{-\beta(1-\frac{d}{2\alpha})-1/2}+\sqrt{\frac{r}{n}}\sqrt{N(\log^2 N+\log n)}+\sqrt{r}n^{-\frac{1}{2}(1-\frac{\beta d}{\alpha})}=:\psi_1(r),
\end{aligned}$$

where $\beta\in(0,\frac{\alpha}{d})$ is an absolute constant to be chosen later.

Note that $\mathbb{V}[(f-g)^2]\leq\mathbb{E}[(f-g)^4]\leq\mathbb{E}[(f-g)^2]$ for any $f\in\mathcal{F}_{NN},g\in T^*\mathcal{F}_{NN}$. Thus, for any $r\geq r_*$, it follows from Lemma B.1 that with probability at least $1-\frac{1}{n}$, we have the following inequality for any $f\in\mathcal{F}_{NN},g\in T^*\mathcal{F}_{NN}$ such that $\|f-g\|_\mu^2\leq r$,

$$\begin{aligned}
&\|f-g\|_n^2\\
&\leq\|f-g\|_\mu^2+3\mathbb{E}R_n\{(f-g)^2:f\in\mathcal{F}_{NN},g\in T^*\mathcal{F}_{NN},\|f-g\|_\mu^2\leq r\}+\sqrt{\frac{2r\log n}{n}}+\frac{56}{3}\frac{\log n}{n}\\
&\leq\|f-g\|_\mu^2+3\mathbb{E}R_n\{f-g:f\in\mathcal{F}_{NN},g\in T^*\mathcal{F}_{NN},\|f-g\|_\mu^2\leq r\}+\sqrt{\frac{2r\log n}{n}}+\frac{56}{3}\frac{\log n}{n}\\
&\leq r+\psi(r)+r+r\leq 4r,
\end{aligned}$$

if $r\geq r_*\vee\frac{2logn}{n}\vee\frac{56logn}{3n}$. For such r, denote $E_r=\{\|f-g\|_n^2\leq 4r\}\cap\{\|f-f_*\|_\mu^2\leq r\}$, we have $P(E_r)\geq 1-1/n$ and

$$\begin{aligned}
&3\mathbb{E}R_n\{f-g:f\in\mathcal{F}_{NN},g\in T^*\mathcal{F}_{NN},\|f-g\|_\mu^2\leq r\}\\
&=3\mathbb{E}\mathbb{E}_\sigma R_n\{f-g:f\in\mathcal{F}_{NN},g\in T^*\mathcal{F}_{NN},\|f-g\|_\mu^2\leq r\}\\
&\leq 3\mathbb{E}\left[1_{E_r}\mathbb{E}_\sigma R_n\{f-g:f\in\mathcal{F}_{NN},g\in T^*\mathcal{F}_{NN},\|f-g\|_\mu^2\leq r\}+(1-1_{E_r})\right]\\
&\leq 3\mathbb{E}\left[\mathbb{E}_\sigma R_n\{f-g:f\in\mathcal{F}_{NN},g\in T^*\mathcal{F}_{NN},\|f-g\|_n^2\leq 4r\}+(1-1_{E_r})\right]\\
&\leq 3(\psi_1(4r)+\frac{1}{n})\\
&\lesssim n^{-\beta-1/2}\sqrt{N(\log^2 N+\log n)}+n^{-\beta(1-\frac{d}{2\alpha})-1/2}+\sqrt{\frac{r}{n}}\sqrt{N(\log^2 N+\log n)}\\
&+\sqrt{r}n^{-\frac{1}{2}(1-\frac{\beta d}{\alpha})}+n^{-1}=:\psi(r)
\end{aligned}$$

It is easy to verify that $\psi(r)$ defined above is a sub-root function. The fixed point r_* of $\psi(r)$ can be solved analytically via the simple quadratic equation $r_*=\psi(r_*)$. In particular, we have

$$\begin{aligned}
\sqrt{r_*}&\lesssim n^{-1/2}\sqrt{N(\log^2 N+\log n)}+n^{-\frac{1}{2}(1-\frac{\beta d}{\alpha})}+n^{-\frac{\beta}{2}-\frac{1}{4}}[N(\log^2 N+\log n)]^{1/4}\\
&+n^{-\frac{\beta}{2}(1-\frac{d}{2\alpha})-\frac{1}{2}}+n^{-1/2}\\
&\lesssim n^{-\frac{1}{4}((2\beta)\wedge 1)+1)}\sqrt{N(\log^2 N+\log n)}+n^{-\frac{1}{2}(1-\frac{\beta d}{\alpha})}+n^{-\frac{\beta}{2}(1-\frac{d}{2\alpha})-\frac{1}{2}}+n^{-1/2} \qquad (12)
\end{aligned}$$

It follows from Equation (9) (where $l\lesssim\log(1/r_*)$), the definition of $d_{\mathcal{F}_{NN}}$, Lemma B.9, and Equation (12) that for any $\epsilon'>0$ and $\delta\in(0,1)$, with probability at least $1-\delta$, we have

$$\max_k\|Q_{k+1}-T^*Q_k\|_\mu\lesssim N^{-\alpha/d}+\epsilon'+n^{-\frac{1}{4}((2\beta)\wedge 1)+1)}\sqrt{N(\log^2 N+\log n)}+n^{-\frac{1}{2}(1-\frac{\beta d}{\alpha})}$$

$$+ n^{-\frac{\beta}{2}(1-\frac{d}{2\alpha})-\frac{1}{2}} + n^{-1/2}\sqrt{\log(1/\delta) + \log\log n} \tag{13}$$

where

$$n \gtrsim \frac{1}{4\epsilon'^2}\left(\log(1/\delta) + \log\log n + \log \mathbb{E}N(\frac{\epsilon'^2}{20}, (\mathcal{F}_{NN} - T^*\mathcal{F}_{NN})|\{x_i\}_{i=1}^n, n^{-1}\cdot\|\cdot\|_1))\right). \tag{14}$$

STEP 4: MINIMIZING THE UPPER BOUND

The final step for the proof is to minimize the upper error bound obtained in the previous steps w.r.t. two free parameters $\beta \in (0, \frac{\alpha}{d})$ and $N \in \mathbb{N}$. Note that N parameterizes the deep ReLU architecture $\Phi(L, m, S, B)$ given Lemma B.9. In particular, we optimize over $\beta \in (0, \frac{\alpha}{d})$ and $N \in \mathbb{N}$ to minimize the upper bound in the RHS of Equation (13). The RHS of Equation (13) is minimized (up to $\log n$-factor) by choosing

$$N \asymp n^{\frac{1}{2}((2\beta\wedge 1)+1)\frac{d}{2\alpha+d}} \text{ and } \beta = \left(2 + \frac{d^2}{\alpha(\alpha+d)}\right)^{-1}, \tag{15}$$

which results in $N \asymp n^{\frac{1}{2}(2\beta+1)\frac{d}{2\alpha+d}}$. At these optimal values, Equation (13) becomes

$$\max_k \|Q_{k+1} - T^*Q_k\|_\mu \lesssim \epsilon' + n^{-\frac{1}{2}\left(\frac{2\alpha}{2\alpha+d}+\frac{d}{\alpha}\right)^{-1}}\log n + n^{-1/2}\sqrt{\log(1/\delta) + \log\log n}, \tag{16}$$

where we use inequalities $n^{-\frac{\beta}{2}(1-\frac{d}{2\alpha})-\frac{1}{2}} \leq n^{-\frac{1}{2}(1-\frac{\beta d}{\alpha})} \asymp N^{-\alpha/d} = n^{-\frac{1}{2}\left(\frac{2\alpha}{2\alpha+d}+\frac{d}{\alpha}\right)^{-1}}$.

Now, for any $\epsilon > 0$, we set $\epsilon' = \epsilon/3$ and let

$$n^{-\frac{1}{2}\left(\frac{2\alpha}{2\alpha+d}+\frac{d}{\alpha}\right)^{-1}}\log n \lesssim \epsilon/3 \text{ and } n^{-1/2}\sqrt{\log(1/\delta) + \log\log n} \lesssim \epsilon/3.$$

It then follows from Equation (16) that with probability at least $1-\delta$, we have $\max_k \|Q_{k+1} - T^*Q_k\|_\mu \leq \epsilon$ if n simultaneously satisfies Equation (14) with $\epsilon' = \epsilon/3$ and

$$n \gtrsim \left(\frac{1}{\epsilon^2}\right)^{\frac{2\alpha}{2\alpha+d}+\frac{d}{\alpha}} (\log^2 n)^{\frac{2\alpha}{2\alpha+d}+\frac{d}{\alpha}} \text{ and } n \gtrsim \frac{1}{\epsilon^2}\left(\log(1/\delta) + \log\log n\right). \tag{17}$$

Next, we derive an explicit formula of the sample complexity satisfying Equation (14). Using Equations (13), (17), and (15), we have that n satisfies Equation (14) if

$$\begin{cases} n & \gtrsim \frac{1}{\epsilon^2}\left[n^{\frac{2\beta+1}{2}\frac{d}{2\alpha+d}}(\log^2 n + \log(1/\epsilon))\right], \\ n & \gtrsim \left(\frac{1}{\epsilon^2}\right)^{1+\frac{d}{\alpha}}, \\ n & \gtrsim \frac{1}{\epsilon^2}\left(\log(1/\delta) + \log\log n\right). \end{cases} \tag{18}$$

Note that $\beta \leq 1/2$ and $\frac{d}{\alpha} \leq 2$; thus, we have

$$\left(1 - \frac{2\beta+1}{2}\frac{d}{2\alpha+d}\right)^{-1} \leq 1 + \frac{d}{\alpha} \leq 3.$$

Hence, n satisfies Equations (17) and (18) if

$$n \gtrsim \left(\frac{1}{\epsilon^2}\right)^{1+\frac{d}{\alpha}}\log^6 n + \frac{1}{\epsilon^2}(\log(1/\delta) + \log\log n).$$

B TECHNICAL LEMMAS

Lemma B.1 (Bartlett et al. (2005)). *Let $r > 0$ and let*

$$\mathcal{F} \subseteq \{f : \mathcal{X} \to [a, b] : \mathbb{V}[f(X_1)] \leq r\}.$$

1. *For any $\lambda > 0$, we have with probability at least $1 - e^{-\lambda}$,*

$$\sup_{f \in \mathcal{F}} (\mathbb{E}f - \mathbb{E}_n f) \leq \inf_{\alpha > 0} \left(2(1+\alpha)\mathbb{E}\left[R_n \mathcal{F}\right] + \sqrt{\frac{2r\lambda}{n}} + (b-a)\left(\frac{1}{3} + \frac{1}{\alpha}\right)\frac{\lambda}{n} \right).$$

2. *With probability at least $1 - 2e^{-\lambda}$,*

$$\sup_{f \in \mathcal{F}} (\mathbb{E}f - \mathbb{E}_n f) \leq \inf_{\alpha \in (0,1)} \left(\frac{2(1+\alpha)}{(1-\alpha)} \mathbb{E}_\sigma \left[R_n \mathcal{F}\right] + \sqrt{\frac{2r\lambda}{n}} + (b-a)\left(\frac{1}{3} + \frac{1}{\alpha} + \frac{1+\alpha}{2\alpha(1-\alpha)}\right)\frac{\lambda}{n} \right).$$

Moreover, the same results hold for $\sup_{f \in \mathcal{F}} (\mathbb{E}_n f - \mathbb{E}f)$.

Lemma B.2 (Györfi et al. (2002, Theorem 11.6))**.** *Let $B \geq 1$ and $\mathcal{F}$ be a set of functions $f : \mathbb{R}^d \to [0, B]$. Let $Z_1, ..., Z_n$ be i.i.d. $\mathbb{R}^d$-valued random variables. For any $\alpha > 0$, $0 < \epsilon < 1$, and $n \geq 1$, we have*

$$P\left\{ \sup_{f \in \mathcal{F}} \frac{\frac{1}{n}\sum_{i=1}^n f(Z_i) - \mathbb{E}[f(Z)]}{\alpha + \frac{1}{n}\sum_{i=1}^n f(Z_i) + \mathbb{E}[f(Z)]} > \epsilon \right\} \leq 4\mathbb{E}N(\frac{\alpha\epsilon}{5}, \mathcal{F}|Z_1^n, n^{-1}\|\cdot\|_1) \exp\left(\frac{-3\epsilon^2 \alpha n}{40B}\right).$$

Lemma B.3 (*Contraction property* (Rebeschini, 2019))**.** *Let $\phi : \mathbb{R} \to \mathbb{R}$ be a L-Lipschitz, then*

$$\mathbb{E}_\sigma R_n (\phi \circ \mathcal{F}) \leq L\mathbb{E}_\sigma R_n \mathcal{F}.$$

Lemma B.4 (Lei et al. (2016, Lemma 1))**.** *Let $\mathcal{F}$ be a function class and P_n be the empirical measure supported on $X_1, ..., X_n \sim \mu$, then for any $r > 0$ (which can be stochastic w.r.t X_i), we have*

$$\mathbb{E}_\sigma R_n\{f \in \mathcal{F} : \|f\|_n^2 \leq r\} \leq \inf_{\epsilon > 0} \left[\mathbb{E}_\sigma R_n\{f \in \mathcal{F} - \mathcal{F} : \|f\|_\mu \leq \epsilon\} + \sqrt{\frac{2r \log N(\epsilon/2, \mathcal{F}, \|\cdot\|_n)}{n}} \right].$$

Lemma B.5 (Lei et al. (2016, modification))**.** *Let $X_1, ..., X_n$ be a sequence of samples and P_n be the associated empirical measure. For any function class $\mathcal{F}$ and any monotone sequence $\{\xi_k\}_{k=0}^\infty$ decreasing to 0, we have the following inequality for any non-negative integer N*

$$\mathbb{E}_\sigma R_n\{f \in \mathcal{F} : \|f\|_n \leq \xi_0\} \leq 4\sum_{k=1}^N \xi_{k-1}\sqrt{\frac{\log \mathcal{N}(\xi_k, \mathcal{F}, \|\cdot\|_n)}{n}} + \xi_N.$$

Lemma B.6 (*Pollard's inequality*)**.** *Let $\mathcal{F}$ be a set of measurable functions $f : \mathcal{X} \to [0, K]$ and let $\epsilon > 0, N$ arbitrary. If $\{X_i\}_{i=1}^N$ is an i.i.d. sequence of random variables taking values in $\mathcal{X}$, then*

$$P\left(\sup_{f \in \mathcal{F}} \left| \frac{1}{N}\sum_{i=1}^N f(X_i) - \mathbb{E}[f(X_1)] \right| > \epsilon \right) \leq 8\mathbb{E}\left[N(\epsilon/8, \mathcal{F}|_{X_{1:N}})\right] e^{\frac{-N\epsilon^2}{128K^2}}.$$

Lemma B.7 (*Properties of (bracketing) entropic numbers*)**.** *Let $\epsilon \in (0, \infty)$. We have*

1. *$H(\epsilon, \mathcal{F}, \|\cdot\|) \leq H_{[]}(2\epsilon, \mathcal{F}, \|\cdot\|)$;*

2. *$H(\epsilon, \mathcal{F}|\{x_i\}_{i=1}^n, n^{-1/p} \cdot \|\cdot\|_p) = H(\epsilon, \mathcal{F}, \|\cdot\|_{p,n}) \leq H(\epsilon, \mathcal{F}|\{x_i\}_{i=1}^n, \|\cdot\|_\infty) \leq H(\epsilon, \mathcal{F}, \|\cdot\|_\infty)$ for all $\{x_i\}_{i=1}^n \subset dom(\mathcal{F})$.*

3. *$H(\epsilon, \mathcal{F} - \mathcal{F}, \|\cdot\|) \leq 2H(\epsilon/2, \mathcal{F}, \|\cdot\|)$), where $\mathcal{F} - \mathcal{F} := \{f - g : f, g \in \mathcal{F}\}$.*

Lemma B.8 (*Entropic number of bounded Besov spaces* (Nickl & Pötscher, 2007, Corollary 2.2))**.** *For $1 \leq p, q \leq \infty$ and $\alpha > d/p$, we have*

$$H_{[]}(\epsilon, \bar{B}_{p,q}^\alpha(\mathcal{X}), \|\cdot\|_\infty) \lesssim \epsilon^{-d/\alpha}.$$

Lemma B.9 (*Approximation power of deep ReLU networks for Besov spaces* (Suzuki, 2018, a modified version))**.** *Let* $1 \leq p, q \leq \infty$ *and* $\alpha \in (\frac{d}{p \wedge 2}, \infty)$*. For sufficiently large* $N \in \mathbb{N}$*, there exists a neural network architecture* $\Phi(L, m, S, B)$ *with*

$$L \asymp \log N, m \asymp N \log N, S \asymp N, \text{ and } B \asymp N^{d^{-1}+\nu^{-1}},$$

where $\nu := \frac{\alpha-\delta}{2\delta}$ *and* $\delta := d(p^{-1} - (1 + \lfloor \alpha \rfloor)^{-1})_+$ *such that*

$$\sup_{f_* \in \bar{B}^{\alpha}_{p,q}(\mathcal{X})} \inf_{f \in \Phi(L,W,S,B)} \|f - f_*\|_\infty \lesssim N^{-\alpha/d}.$$