Sample Complexity of Offline Reinforcement Learning with Deep ReLU Networks

Thanh Nguyen-Tang [1] **Sunil Gupta** [1] **Hung Tran-The** [1] **Svetha Venkatesh** [1]

Abstract

We study the statistical theory of offline reinforcement learning (RL) with deep ReLU network function approximation. We analyze a variant of fitted-Q iteration (FQI) algorithm under a new dynamic condition that we call Besov dynamic closure, which encompasses the conditions from prior analyses for deep neural network function approximation. Under Besov dynamic closure, we prove that the FQI-type algorithm enjoys an improved sample complexity than the prior results. Importantly, our sample complexity is obtained under the new general dynamic condition and a data-dependent structure where the latter is either ignored in prior algorithms or improperly handled by prior analyses. This is the first comprehensive analysis for offline RL with deep ReLU network function approximation under a general setting.

1. Introduction

Offline reinforcement learning (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 are 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 use 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, but relatively few work provide a comprehensive and adequate analysis of the statistical efficiency for the problems.

[1]The Applied Artificial Intelligence Institute (A^2I^2), Deakin University, Australia. Correspondence to: Thanh Nguyen-Tang <nguyent2792@gmail.com>.

The ICML Workshop on Reinforcement Learning Theory, 2021.

On the theoretical side, predominant sample efficiency results in offline RL focus on tabular environments with small finite state spaces (Yin and Wang, 2020; Yin et al., 2021; Yin and Wang, 2021), but as these methods scale with the number of states, they are infeasible for infinitely large state space settings. While this tabular setting has been extended to large state spaces via linear environments (Duan and Wang, 2020), the linearity assumption does not hold for many RL problems in practice. More relevant theoretical progress has been achieved for more complex environments with general and deep neural network function approximations, but these results are either inadequate or relatively disconnected from practical settings (Munos and Szepesvári, 2008; Yang et al., 2019; Le et al., 2019). In particular, their finite-sample results either (i) depend on a so-called inherent Bellman error (Munos and Szepesvári, 2008; Le et al., 2019), which could be arbitrarily large or uncontrollable in practice, (ii) avoid the data-dependent structure in their algorithms at the cost of losing sample efficiency (Yang et al., 2019) or improperly ignore it in their analysis (Le et al., 2019), or (iii) rely on relatively strong dynamics assumption (Yang et al., 2019).

In this paper, we study a variation of fitted-Q iteration (FQI) (Bertsekas et al., 1995; Sutton and Barto, 2018) for OPE and OPL where we approximate the target Q-function from an offline data using a deep ReLU network. The algorithm is appealingly simple: it iteratively estimates the target Q-function via regression on the offline data and the previous estimate. This procedure forms the core of many current offline RL methods. With linear function approximation, (Duan and Wang, 2020) show that this procedure yields a minimax-optimal sample efficient algorithm, provided the environment dynamics satisfy certain linear properties. While their assumptions generalize the tabular settings, they are restrictive for more complex environment dynamics where non-linear function approximation is required. Moreover, as they highly exploit the linearity structure, it is unclear how their analysis can accommodate non-linear function approximation such as deep ReLU networks.

In this paper, we provide the statistical theory of a FQI-type algorithm for both OPE and OPL problems with deep ReLU networks. In particular, we provide the first comprehensive analysis for offline RL under deep ReLU network function approximation. We achieve this generality in our result via two novel considerations. First, we introduce Besov dynamic closure which is, to our knowledge, the most general assumption that encompasses the previous dynamic assumptions in offline RL. Unlike the previous dynamic conditions that permit only integer smoothness of the underlying MDP, our Besov dynamic closure generalizes the previous dynamic conditions by allowing fractional smoothness that describes the regularity of the MDP more precisely. Moreover, the MDP under our Besov dynamic closure needs not be continuous, differentiable or spatially homogeneous in smoothness. Second, as each estimate in a regression-based offline RL algorithm depends on the previous estimates and the entire offline dataset, a complicated data-dependent structure is induced. This data-dependent structure plays a central role in the statistical efficiency of the algorithm. While the prior results ignore such a structure, either in their algorithm or their analysis, resulting a loss of sample efficiency or improper analysis, resp., we consider it in a FQI-type algorithm and effectively handle it in our analysis. Under these considerations, we establish the sample complexity of offline RL with deep ReLU network function approximation that is both more general and more sample-efficient than the prior results.

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 and Li, 2015; Thomas and Brunskill, 2016; Farajtabar et al., 2018; Kallus and 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 efficiency guarantees for offline RL are obtained in tabular settings in (Xie et al., 2019; Yin and Wang, 2020; Yin et al., 2021; Yin and Wang, 2021). A lower bound for tabular offline RL is obtained in (Jiang and Li, 2016) which in particular show a 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 and Tsitsiklis, 1995; Jong and Stone, 2007; Lagoudakis and Parr, 2003; Grünewälder et al., 2012; Munos, 2003; Munos and 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 data-dependent 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 controllable 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 fresh batch of data for each iteration. Such approach is considerably sample-inefficient in offline RL as it undesirably does not leverage the past data. As a result, their sample complexity 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. We summarize the key differences between our work and the prior results in Table 1 which will be elaborated further in Subsection 4.3.

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 focus on a practical setting of deep neural network function approximation with 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 data-dependent 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 applies to more complex models such as a deep ReLU network. Importantly, we propose a new Besov dynamic closure and a uniform-convergence argument which appear absent in Duan et al. (2021).

3. Preliminaries

We consider MDP$(\mathcal{S}, \mathcal{A}, P, R, \gamma, \rho)$ where $R : \mathcal{S} \times \mathcal{A} \to \mathcal{P}([0, 1])$ is the reward distribution supported on $[0, 1]$. We consider continuous state space and action space, and for notational simplicity, we assume that $\mathcal{X} := \mathcal{S} \times \mathcal{A} \subseteq [0, 1]^d$. Denote T^* and T^π be the optimality Bellman operator and the Bellman operator, resp. for any policy π. Let Q^π be the Q-function of policy π and $Q^* = \arg\max_\pi Q^\pi$, $V^\pi(s) = \langle Q^\pi(s,\cdot), \pi(\cdot|s)\rangle_{\mathcal{A}}$, and $V^*(\cdot) = \max_a Q^*(\cdot, a)$.

We consider the offline RL setting where the agent cannot explore further 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 where $(s_i, a_i) \overset{i.i.d.}{\sim} \mu(\cdot,\cdot) := \frac{1}{1-\gamma}\sum_{t=0}^{\infty}\gamma^t P(s_t = \cdot, a_t = \cdot|\rho, \eta), s_i' \sim P(\cdot|s_i, a_i)$ and $r_i \sim R(s_i, a_i)$. Here μ is the (sampling) state-action visitation distribution. The goals of OPE and OPL are to estimate V^π and V^*, resp. from $\mathcal{D}$, and in this paper we measure performance by 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 $\mathrm{SubOpt}(\hat{V}; \pi) = |V^\pi - \hat{V}|$.

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 $\mathrm{SubOpt}(\hat{\pi}) = \mathbb{E}_\rho\left[V^*(s) - Q^*(s, \hat{\pi}(s))\right]$ where $\mathbb{E}_\rho$ is the expectation w.r.t. $s \sim \rho$.

3.1. Deep ReLU Networks as Function Approximation

A L-height, m-width ReLU network on $\mathbb{R}^d$ takes the form $f_\theta^{L,m}(x) = W^{(L)}\sigma(\ldots\sigma(W^{(1)}\sigma(x) + b^{(1)})\ldots) + b^{(L)}$, where $W^{(L)} \in \mathbb{R}^{1\times m}$, $b^{(L)} \in \mathbb{R}$, $W^{(1)} \in \mathbb{R}^{m\times d}$, $b^{(1)} \in \mathbb{R}^m$, $W^{(l)} \in \mathbb{R}^{m\times m}$, $b^{(l)} \in \mathbb{R}^m$, $\forall 1 < l < L$, $\theta = \{W^{(l)}, b^{(l)}\}_{1\le l\le L}$, and σ is the element-wise ReLU function. 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., $\sum_{l=1}^{L}(\|W^{(l)}\|_0 + \|b^{(l)}\|_0) \le S, \max_{1\le l\le L}\|W^{(l)}\|_\infty \vee \|b^{(l)}\|_\infty \le B$ where $\|\cdot\|_0$ is the 0-norm, i.e., the number of non-zero elements, and $a \vee b = \max\{a, b\}$. 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 $\mathcal{F}_{NN} := \{f \in \Phi(L, m, S, B) : \|f\|_\infty \le 1\}$.

3.2. Regularity

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 encompasses Lipschitz, Hölder and Sobolev spaces. In particular in this work, we consider Besov spaces. Unlike the previous spaces considered in offline RL such as Hölder and Sobolev spaces that permit only integer smoothness, Besov spaces allow *fractional* smoothness that describes the regularity of a function more precisely and generalizes the previous smoothness notions. We characterize the smoothness in Besov spaces via moduli of smoothness, following (Giné and Nickl, 2016).

Definition 1 (*Moduli of smoothness*)**.** For $f \in L^p(\mathcal{X})$ for some $p \in [1, \infty]$, we define its r-th modulus of smoothness as $\omega_r^{t,p}(f) := \sup_{0\le h\le t}\|\Delta_h^r(f)\|_p, t > 0, r \in \mathbb{N}$ where the r-th order translation-difference operator $\Delta_h^r = \Delta_h \circ \Delta_h^{r-1}$ is recursively defined as $\Delta_h^r(f)(\cdot) := (f(\cdot + h) - f(\cdot))^r = \sum_{k=0}^{r}\binom{r}{k}(-1)^{r-k}f(\cdot + k\cdot h)$.

Definition 2 (*Besov space* $B_{p,q}^\alpha(\mathcal{X})$)**.** For $1 \le p, q \le \infty$ and $\alpha > 0$, we define the norm $\|\cdot\|_{B_{p,q}^\alpha}$ of the Besov space $B_{p,q}^\alpha(\mathcal{X})$ as $\|f\|_{B_{p,q}^\alpha} := \|f\|_p + |f|_{B_{p,q}^\alpha}$ where

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

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

Intuitively, the Besov seminorm $|f|_{B_{p,q}^\alpha}$ roughly describes the L^q-norm of the l^p-norm of the α-order smoothness of f. 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 and Pötscher, 2007). In particular, the Besov space $B_{p,q}^\alpha$ reduces into the Hölder space C^α when $p = q = \infty$ and α is a positive non-integer while it reduces into the Sobolev space W_2^α when $p = q = 2$ and α is a positive integer. We further consider the unit ball of $B_{p,q}^\alpha(\mathcal{X})$: $\bar{B}_{p,q}^\alpha(\mathcal{X}) := \{g \in B_{p,q}^\alpha : \|g\|_{B_{p,q}^\alpha} \le 1 \text{ and } \|g\|_\infty \le 1\}$.

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

Assumption 3.1 (*Concentration coefficient*)**.** $\exists\kappa_\mu < \infty$ such that $\|\frac{d\nu}{d\mu}\|_\infty \le \kappa_\mu$ for any *realizable* distribution ν.

Here, ν is realizable if there exists $t \ge 0$ and policy π_1 such that $\nu(s, a) = \mathbb{P}(s_t = s, a_t = a|s_1 \sim \rho, \pi_1), \forall s, a$. Intuitively, 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 and Szepesvári, 2008). Chen and Jiang (2019) further provide 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.

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

The assumption signifies that for any policy π, T^π applied on any ReLU network 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 $f \in \mathcal{F}_{NN}(\mathcal{X})$. Such an assumption is relatively standard and common in the offline RL literature (Chen and Jiang, 2019). A natural example for this assumption to hold is when both the expected reward function $r(s,a)$ and the transition density $P(s'|s,a)$ for each fixed s' are Besov functions.

Importantly, our Besov dynamic closure is considerably general that encompasses the conditions considered in prior results (Yang et al., 2019). In particular, the Besov dynamic closure only requires the boundedness of a very general notion of local oscillations of the underlying MDP which are allowed to be discontinuous or non-differentiable (e.g., when $\alpha \le 1/2$ and $p = 2$), or even have spatially inhomogeneous smoothness (e.g., when $p < 2$). This flexibility is novel w.r.t. the prior results. 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.

4. Algorithm and Main Result

4.1. Algorithm

Now we turn to the main algorithm and the main result. We study least-squares value iteration (LSVI) for both OPE and offline learning, with the pseudo-code presented in Algorithm 1 where we denote $\rho^\pi(s,a) = \rho(s)\pi(a|s)$.

On the computational side, solving the non-convex optimization at line 1 of Algorithm 1 can be highly involved and (stochastic) gradient descent (GD) is a dominant optimization method for such a task in deep learning. In particular, GD is guaranteed to converge to a global minimum under certain structural assumptions (Nguyen, 2021). Here, as we focus on the statistical properties of LSVI, not on the optimization problem, we assume that a global minimizer at line 1 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 and Jiang, 2019; Duan et al., 2021; Wang et al., 2019; 2020; Jin et al., 2021).

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**: $y_i \leftarrow r_i + \gamma \int_{\mathcal{A}} Q_{k-1}(s'_i, a)\pi(da|s'_i)$
4: If **OPL**: $y_i \leftarrow r_i + \gamma \max_{a' \in \mathcal{A}} Q_{k-1}(s'_i, a')$
5: $Q_k \leftarrow \arg\inf_{f \in \mathcal{F}_{NN}} \sum_{i=1}^n (f(s_i, a_i) - y_i)^2$
6: **end for**
7: If **OPE**, return $V_K = \|Q_K\|_{\rho^\pi}$
8: If **OPL**, return the greedy policy π_K w.r.t. Q_K.

4.2. Data-dependent structure

The target variable y_i computed in 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 data-dependent structure across all iterations where the current estimate depends on all the previous estimates and the past data. Specifically, 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, resp.), i.e., $\mathbb{E}\left[[T^* Q_{k-1}](x_i) - y_i | x_i\right] \neq 0$. This data-dependent structure hinders the use of any standard non-parametric regression analysis and concentration phenomenon typically used in supervised learning. Prior results either improperly ignore the data-dependent structure in their analysis (Le et al., 2019) or directly avoid it by estimating each Q_k on a separate subset of $\mathcal{D}$ (Yang et al., 2019). While the latter removes the data-dependent structure, it pays the cost of scaling the sample complexity with the number of iterations K as it requires splitting the original data into K disjoint subsets. In our work, we consider the data-dependent structure in LSVI and effectively handle it via a uniform-convergence argument and local Rademacher complexities.

4.3. Main Result

Our main result is a sup-optimality bound for LSVI in both OPE and OPL settings under Assumption 3.1 and Assumption 3.2. Before stating the main result, we introduce the necessary notations of asymptotic relations: we write $f(\epsilon, n) \lesssim g(\epsilon, n)$ if there is an absolute constant c such that $f(\epsilon, n) \le c \cdot g(\epsilon, n), \forall \epsilon > 0, n \in \mathbb{N}$. We write $f(\epsilon, n) \asymp g(\epsilon, n)$ if $f(\epsilon, n) \lesssim g(\epsilon, n)$ and $g(\epsilon, n) \lesssim f(\epsilon, n)$.

Theorem 1. *Under Assumptions 3.1-3.2, for any $\epsilon, \delta, K > 0$, for $n \gtrsim \left(\frac{1}{\epsilon^2}\right)^{1+\frac{d}{\alpha}} \log^6 n + \frac{1}{\epsilon^2}(\log(1/\delta) + \log\log n)$, with probability $\ge 1 - \delta$, the sup-optimality of Algorithm 1 is*

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

In addition, the optimal deep ReLU network $\Phi(L, m, S, B)$ that obtains such sample complexity (for both OPE and OPL) satisfies: $L \asymp \log N, m \asymp N, S \asymp N$, and $B \asymp N^{1/d + (2\iota)/(\alpha - \iota)}$, where $\iota := d(p^{-1} - (1 + \lfloor \alpha \rfloor)^{-1})_+$, $N \asymp n^{\frac{(\beta+1/2)d}{2\alpha+d}}$, and $\beta = (2 + \frac{d^2}{\alpha(\alpha+d)})^{-1}$.

The result 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 in-

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

Table 1. Recent advances in the sample complexity of offline RL with various function approximations where κ is a distributional shift measure, ϵ is the user-specified precision, d is the dimension of the input space, α is the smoothness parameter of the MDP, and K is the algorithmic iteration number.

equality given in Theorem 1. In particular, the sample complexity does not depend on the number of states as in tabular MDPs (Yin and Wang, 2020; Yin et al., 2021; Yin and Wang, 2021) or the inherent Bellman error as in the general function approximation (Munos and Szepesvári, 2008; Le et al., 2019). Instead, it explicitly scales with the smoothness α of the MDP and the dimension d of the input space. Importantly, this guarantee is established under the general Besov dynamic closure and the data-dependent structure of the algorithm. Thus, Theorem 1 is the most comprehensive result we are aware of for offline RL with deep neural networks.

Moreover, to develop further intuition on our sample complexity, we compare it with 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 and Picard, 1992; Giné and Nickl, 2016) even though we deal with a more complicated data-dependent structure in a value iteration problem instead of a standard non-parametric regression problem. This gap is expected due to the data-dependent 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 disjoint subsets and estimate each Q_k in Algorithm 1 using a separate disjoint subsets. This however comes at the cost that the sample complexity scales with K which could be arbitrarily large in practice.

To show the significance of our sample complexity, we summarize our result and compare it with the prior results in Table 1. The "data reuse" in Table 1 means the data is reused across all iterations instead of being splitted into disjoint subsets for each iteration. 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 while they are too strong as a condition to hold in practice. Second, as remarked earlier, our Besov smoothness is more general than Hölder smoothness considered in (Yang et al., 2019), thus our setting is more practical. Moreover, the LSVI algorithm 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 an unifying analysis that accounts for the data-dependent 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 data-dependent structure in their analysis. In addition, (Le et al., 2019) does not provide an explicit sample complexity as it depends on an unknown inherent Bellman error. Thus, our sample complexity is one of the most general result in practical and comprehensive settings with an improved performance. We provide a detailed proof for Theorem 1 in the **supplementary**.

5. Conclusion and Discussion

This paper presents the sample complexity of offline RL with deep ReLU network function approximation. We prove that the FQI-type algorithm with the data-dependent structure obtains an improved sample complexity of $\tilde{\mathcal{O}}\left(\kappa^{1+d/\alpha} \cdot \epsilon^{-2-2d/\alpha}\right)$ under a standard condition of distributional shift and a new dynamic condition namely Besov dynamic closure which encompasses the dynamic conditions considered in the prior results. Established under the data-dependent structure and the Besov dynamic closure, our sample complexity is the most general result for offline RL with deep ReLU network function approximation.

We close with some open problems. First, although the finite concentration coefficient is a uniform data coverage assumption that 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 difficult to generalize this condition to function approximation. Another important direction is to investigate the sample complexity of *pessimism* principle (Buckman et al., 2020) in offline RL with non-linear function approximation, which is currently studied only in tabular and linear settings (Rashidinejad et al., 2021; Jin et al., 2020).

References

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.

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.

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, pages 560–564. IEEE, 1995.

Dimitri P Bertsekas, Dimitri P Bertsekas, Dimitri P Bertsekas, and Dimitri P Bertsekas. *Dynamic programming and optimal control*, volume 1. Athena scientific Belmont, MA, 1995.

Jacob Buckman, Carles Gelada, and Marc G Bellemare. The importance of pessimism in fixed-dataset policy optimization. *arXiv preprint arXiv:2009.06799*, 2020.

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

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

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.

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*, pages 652–661. JMLR.org, 2016.

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.

Ying Jin, Zhuoran Yang, and Zhaoran Wang. Is pessimism provably efficient for offline rl? *arXiv preprint arXiv:2012.15085*, 2020.

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

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

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

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

Hoang Minh Le, Cameron Voloshin, and Yisong Yue. Batch policy learning under constraints. In *ICML*, volume 97 of *Proceedings of Machine Learning Research*, pages 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*, pages 5361–5371, 2018.

Rémi Munos. Error bounds for approximate policy iteration. In *ICML*, pages 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, page 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.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. A Bradford Book, Cambridge, MA, USA, 2018. ISBN 0262039249.

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*, pages 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*, pages 3434–3443. PMLR, 2017.

H. Triebel. Theory of function spaces. 1983.

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.

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*, pages 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, editors, *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*, pages 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.

Appendix A. Proof

We now provide a complete proof of Theorem 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 1, that applies generally to any least-squares Q-iteration methods.

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

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

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 and Szepesvári, 2008; Le et al., 2019).

Proof of Lemma 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

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

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

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

(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

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

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

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

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

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

Thus, we have

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

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

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

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

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

Using the triangle inequality, the above inequality becomes

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

where

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

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

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

Thus, we have

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

Finally, we have

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

□

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

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

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 $f_\perp^Q := \operatorname{arg\,inf}_{f \in \mathcal{F}_{NN}} \|f - f_*^Q\|_{2,\mu}$ 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 and 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

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

We have

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

We have

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

For any $\epsilon' > 0$ and $\delta' \in (0, 1)$, it follows from Lemma 3 with $\epsilon = 1/2$ and $\alpha = \epsilon'^2$, with probability at least $1 - \delta'$, for any $Q \in \mathcal{F}_{NN}$, we have

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

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 and Szepesvári, 2008) in the place of Lemma 3, 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 (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

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

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

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

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

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

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:

$$H(\epsilon, \mathcal{F}_{NN} - T^*\mathcal{F}_{NN}, \|\cdot\|_n) \leq H(\epsilon/2, \mathcal{F}_{NN}, \|\cdot\|_n) + H(\epsilon/2, T^*\mathcal{F}_{NN}, \|\cdot\|_n),$$

$$H(\epsilon, \mathcal{F}_{NN}, \|\cdot\|_n) \leq H(\epsilon, \mathcal{F}_{NN}|\{x_i\}_{i=1}^n, \|\cdot\|_\infty) \overset{(a)}{\lesssim} N[(\log N)^2 + \log(1/\epsilon)], \tag{10}$$

$$\begin{aligned} H(\epsilon, T^*\mathcal{F}_{NN}, \|\cdot\|_n) &\leq H(\epsilon, T^*\mathcal{F}_{NN}, \|\cdot\|_\infty) \leq H_{[]}(2\epsilon, T^*\mathcal{F}_{NN}, \|\cdot\|_\infty) \\ &\overset{(b)}{\leq} H_{[]}(2\epsilon, \bar{B}_{p,q}^\alpha(\mathcal{X}), \|\cdot\|_\infty) \overset{(c)}{\lesssim} (2\epsilon)^{-d/\alpha}, \end{aligned} \tag{11}$$

where N is a hyperparameter of the deep ReLU network described in Lemma 10, (a) follows from Lemma 10, and (b) follows from Assumption 3.2, and (c) follows from Lemma 9. Let $\mathcal{H} := \mathcal{F}_{NN} - T^*\mathcal{F}_{NN}$, it follows from Lemma 6 with $\{\xi_k := \epsilon/2^k\}_{k\in\mathbb{N}}$ for any $\epsilon > 0$ that

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

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 5 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 2 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 10, and (12) that for any $\epsilon' > 0$ and $\delta \in (0,1)$, with probability at least $1-\delta$, we have

$$\begin{aligned}
\max_k \|Q_{k+1} - T^*Q_k\|_\mu &\lesssim N^{-\alpha/d} + \epsilon' + n^{-\frac{1}{4}((2\beta)\wedge 1)+1)}\sqrt{N(\log^2 N + \log n)} + n^{-\frac{1}{2}(1-\frac{\beta d}{\alpha})} \\
&+ n^{-\frac{\beta}{2}(1-\frac{d}{2\alpha})-\frac{1}{2}} + n^{-1/2}\sqrt{\log(1/\delta) + \log\log n} \qquad (13)
\end{aligned}$$

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 10. 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, (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

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

Appendix B. Technical Lemmas

Lemma 2 (Bartlett et al. (2005)). *Let $r > 0$ and let*

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

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 3 (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 4 (*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 5 (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*

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

where $\mathcal{F} - \mathcal{F} := \{f - g : f, g \in \mathcal{F}\}$.

Lemma 6 (*Refined entropy integral* (modified from (Lei et al., 2016)))**.** *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*

$$\mathbb{E}_\sigma R_n\{f \in \mathcal{F} : \|f\|_n \leq \xi_0\} \leq 4\sum_{k=1}^N \xi_{k-1}\sqrt{\frac{\log\mathcal{N}(\xi_k, \mathcal{F}, \|\cdot\|_{2,n})}{n}} + \xi_N.$$

Lemma 7 (*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 8 (*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 9 (*Entropic number of bounded Besov spaces* (Nickl and Pötscher, 2007, Corollary 2.2))**.** *For $1 \leq p, q \leq \infty$ and $\alpha > d/p$, we have*

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

Lemma 10 (*Approximation power of deep ReLU networks for Besov spaces* (Suzuki, 2018))**.** *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

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