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.

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

We consider the offline RL setting where the agent cannot explore further the environment but has access to a fixed logged data \( D = \{(s_i, a_i, s'_i, r_i)\}_{i=1}^T \) collected a priori by certain behaviour policy \( \eta \) where \( (s_i, a_i) \sim \mu(\cdot, \cdot) \), \( r_i \sim R(s_i, a_i) \). Here \( \mu \) is the (sampling) state-action visitation distribution. The goals of OPE and OPL are to estimate \( V^\pi \) and \( V^* \), resp. from \( D \), and in this paper we measure performance by sub-optimality gaps.

For OPE. Given a fixed target policy \( \pi \), for any value estimate \( \hat{V} \) computed from the offline data \( D \), 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 \( \pi^* \) that is learned from the offline data \( D \), we define the sub-optimality of OPL as \( \text{SubOpt}(\hat{\pi}) = \mathbb{E}_\mu \left[ V^{\pi^*}(s) - Q^{\pi^*}(s, \hat{\pi}(s)) \right] \) where \( \mathbb{E}_\mu \) is the expectation w.r.t. \( s \sim \mu \).

3.1. Deep ReLU Networks as Function Approximation

A \( L \)-height, \( m \)-width ReLU network on \( \mathbb{R}^d \) takes the form
\[
f^{L,m}(x) = W^{(L)}\sigma(\ldots(\sigma(W^{(1)}\sigma(x) + b^{(1)}) + b^{(1)}) + \ldots + b^{(L)})
\]
where \( W^{(L)} \in \mathbb{R}^{1 \times m}, b^{(L)} \in \mathbb{R}, W^{(l)} \in \mathbb{R}^{m \times o}, b^{(l)} \in \mathbb{R}, \forall 1 < l < L, \theta = \{W^{(l)}, b^{(l)}\}_{1 \leq l \leq L} \), and \( \sigma \) is the element-wise ReLU function. We define \( \Phi(L, m, S, B) \) as the space of \( L \)-height, \( m \)-width ReLU functions \( f^{L,m}_\theta(x) \) with sparsity constraint \( S \), and norm constraint \( B \), i.e., \( \sum_{l=1}^L (\|W^{(l)}\|_0 + \|b^{(l)}\|_0) \leq S, \max_{1 \leq l \leq L} \|W^{(l)}\|_\infty + \|b^{(l)}\|_\infty \leq B \) where \( \|\cdot\|_0 \) is the 0-norm, i.e., the number of non-zero elements, and \( a \lor b = \max\{a, b\} \). 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 \( \mathcal{F}_{NN} := \{ f \in \Phi(L, m, S, B) : \|f\|_\infty \leq 1 \} \).

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 \( \omega_{L^p}(f) := \sup_{0 \leq h \leq t} \|D^r_h f\|_p, t > 0, r \in \mathbb{N} \) where the \( r \)-th order translation-difference operator \( D^r_h f = \Delta_h \circ \Delta_h^{r-1} \) is recursively defined as \( \Delta^r_h f(\cdot) := (f(\cdot + h) - f(\cdot))_T^r = \sum_{k=0}^\infty (-1)^r k f(\cdot + k \cdot h) \).

**Definition 2** (Besov space \( B_{p,q}^\alpha(\mathcal{X}) \)). For \( 0 \leq p, q \leq \infty \) 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
\[
|f|_{B_{p,q}^\alpha} := \left( \sum_{\nu>0} \frac{\omega_{L^p}(f)_\nu}{\nu^{\alpha}} \right)^{1/q}, \quad 1 \leq q < \infty,
\]
and \( \|f\|_{B_{p,q}^\alpha} := \|f\|_p \) if \( q = \infty \),

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 \( L^p \)-norm of the \( \alpha \)-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 Pörtsch, 2007). In particular, the Besov space \( B_{p,q}^\alpha \) reduces into the Hölder space \( C^{\alpha} \) when \( p = q = \infty \) and \( \alpha \) is a positive non-integer while it reduces into the Sobolev space \( W^{p, 0} \) when \( p = q = 2 \) and \( \alpha \) is a positive integer. We further consider the unit ball of \( B_{p,q}^\alpha(\mathcal{X}) \): \( B_{p,q}^\alpha(\mathcal{X}) := \{ g \in B_{p,q}^\alpha : \|g\|_{B_{p,q}^\alpha} \leq 1 \text{ and } \|g\|_\infty \leq 1 \} \).

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 \( \frac{d\mu}{d\nu} \leq \kappa_\mu \) for any realizable distribution \( \nu \).

Here, \( \nu \) is realizable if there exists \( t \geq 0 \) and policy \( \pi_1 \) such that \( \nu(s, a) = \mathbb{P}(s_t = s, a_t = a|s_{t-1} \sim \rho, \pi_1), \forall s, a \). Intuitively, Assumption 3.1 asserts that the sampling distribution \( \mu \) is not too far away from any realizable distribution uniformly over the state-action space. \( \kappa_\mu \) 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 \( p, q \in [1, \infty], \alpha > \frac{d}{p+2}, \forall f \in \mathcal{F}_{NN}(\mathcal{X}), \forall \pi, T^\pi f \in B_{p,q}^\alpha(\mathcal{X}) \).
The assumption signifies that for any policy \( \pi, T^\pi \) applied on any ReLU network in \( \mathcal{F}_{NN}(\mathcal{X}) \) results in a Besov function in \( \mathcal{F}^{\rho,\gamma}_{p,q}(\mathcal{X}) \). Moreover, as \( T^\pi f = T^\pi f \) where \( \pi_f \) is the greedy policy w.r.t. \( f \), Assumption 3.2 also implies that \( T^\pi f \in \mathcal{F}^{\rho,\gamma}_{p,q}(\mathcal{X}) \) 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; Duan et al., 2021; Wang et al., 2019; 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 \( \alpha \leq 1/2 \) 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 \( \alpha > \frac{d}{p+2} \) guarantees a finite bound for the compactness and the (local) Rademacher complexity of the considered Besov space.

Importantly, our Besov dynamic closure is considerably general that 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 \( \alpha \leq 1/2 \) 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 \( \alpha > \frac{d}{p+2} \) 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; 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: \( y_i \leftarrow r_i + \gamma \int_{\mathcal{A}} Q_{k-1}(s, a)\pi(da|s) \)
4.  If OPL: \( y_i \leftarrow r_i + \gamma \max_{a' \in \mathcal{A}} Q_{k-1}(s, a') \)
5.  \( Q_k \leftarrow \arg \inf_{f \in \mathcal{F}_{NN}} \sum_{i=1}^n (f(s, a_i) - y_i)^2 \)
6.  end for  
7.  If OPE, return \( V_K = \|Q_K\|_\rho^\pi \)
8.  If OPL, return the greedy policy \( \pi_K \) 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 \( x_i := (s, a) \). 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 \( [T^\pi Q_{k-1}](x_i) \) for OPL (or at \( [T^\pi Q_{k-1}](x_i) \) for OPE, resp.), i.e., \( E[[T^\pi Q_{k-1}](x_i) - y_i|x_i] \neq 0 \). This data-dependent structure hinders the use of any standard non-parametric regression analysis and concentration phenomena 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 \( D \) (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 \( f(\epsilon, n) \leq c \cdot g(\epsilon, n), \forall \epsilon > 0, n \in \mathbb{N} \). 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 \( n \gtrsim (\frac{1}{\epsilon})^{1+\frac{d}{p+2}} \log^6 n + (\frac{1}{\epsilon}) (1+\log(1/\delta) + \log \log n) \), with probability \( \geq 1 - \delta \), the sup-optimality of Algorithm 1 is

\[
\begin{align*}
\text{SubOpt}(V_K; \pi) &\lesssim \frac{\sqrt{\epsilon}}{1-\gamma} + \epsilon^K/((1-\gamma)^{1/2}) \quad \text{for OPE,} \\
\text{SubOpt}(\pi_K) &\lesssim 4\gamma\sqrt{\epsilon} + \epsilon^{1/2}/((1-\gamma)^{1/2}) \quad \text{for OPL.}
\end{align*}
\]

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 \( B \asymp N^{1/d+2/(d-1)} \), where \( \epsilon : = d(p-1 - (1+\alpha)^{-1})^{-1} \), \( \gamma \asymp m \), and \( \beta = 2(\alpha^{-1} + \gamma^{-1})^{-1} \).

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 \( \kappa_\rho \), the effective horizon \( 1/(1-\gamma) \), and the user-specified precision \( \epsilon \) for \( n \) satisfying the in-
Sample Complexity of Offline Reinforcement Learning with Deep ReLU Networks

<table>
<thead>
<tr>
<th>Function Approx.</th>
<th>Regularity</th>
<th>Tasks</th>
<th>Complexity</th>
<th>Remark</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yin and Wang (2020)</td>
<td>Tabular</td>
<td>Tabular</td>
<td>OPE</td>
<td>$O\left(\frac{1}{N}\left(</td>
</tr>
<tr>
<td>Duan and Wang (2020)</td>
<td>Linear</td>
<td>Linear</td>
<td>OPE</td>
<td>$O\left(\kappa \cdot d \cdot \epsilon^{-2}\right)$</td>
</tr>
<tr>
<td>Le et al. (2019)</td>
<td>General</td>
<td>General</td>
<td>OPE/OPL</td>
<td>N/A</td>
</tr>
<tr>
<td>Yang et al. (2019)</td>
<td>ReLU</td>
<td>Hölder</td>
<td>OPL</td>
<td>$\tilde{O}\left(K \cdot \kappa^{2+\frac{d}{\alpha}} \cdot \epsilon^{-2-\frac{d}{\alpha}}\right)$</td>
</tr>
<tr>
<td>Ours</td>
<td>ReLU</td>
<td>Besov</td>
<td>OPE/OPL</td>
<td>$\tilde{O}\left(\kappa^{1+\frac{d}{\alpha}} \cdot \epsilon^{-2-\frac{d}{\alpha}}\right)$</td>
</tr>
</tbody>
</table>

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

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 $\tilde{O}\left(\kappa^{1+d/\alpha} \cdot \epsilon^{-2-2d/\alpha}\right)$ 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 approxi-
Sample Complexity of Offline Reinforcement Learning with Deep ReLU Networks

mation, which is currently studied only in tabular and linear settings (Rashidinejad et al., 2021; Jin et al., 2020).

References


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


Sample Complexity of Offline Reinforcement Learning with Deep ReLU Networks


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

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

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

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 $T^\pi$ 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 $P^\pi : \{ \mathcal{X} \rightarrow \mathbb{R} \} \rightarrow \{ \mathcal{X} \rightarrow \mathbb{R} \}$ where

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

for any $f \in \{ \mathcal{X} \rightarrow \mathbb{R} \}$. 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 $Q^\pi$ is the (unique) fixed point of $T^\pi$, we have

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

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 $\alpha_k := \frac{(1-\gamma)^k}{1-\gamma^{K+1}}, \forall k \in [K]$ 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

\[
\|Q_K - Q^*\|_\mu^2 \leq \frac{(1 - \gamma^{K+1})^2}{(1 - \gamma)^2} \int \rho(ds) \pi(da|s) \left( \sum_{k=0}^{K} \alpha_k A_{k|\epsilon_k} \right)^2 \tag{a}
\]

\[
\leq \frac{(1 - \gamma^{K+1})^2}{(1 - \gamma)^2} \int \rho(ds) \pi(da|s) \sum_{k=0}^{K} \alpha_k^2 A_k^2 \epsilon_k^2(s, a) \tag{b}
\]

\[
\leq \frac{(1 - \gamma^{K+1})^2}{(1 - \gamma)^2} \left( \int \rho(ds) \pi(da|s) \sum_{k=0}^{K-1} \alpha_k A_k^2 \epsilon_k^2(s, a) + \alpha_K \right) \tag{c}
\]

\[
\leq \frac{(1 - \gamma^{K+1})^2}{(1 - \gamma)^2} \left( \sum_{k=0}^{K-1} \alpha_k \kappa_\mu \epsilon_k^2 + \alpha_K \right) \tag{d}
\]

\[
\leq \frac{\kappa_\mu}{(1 - \gamma)^2} \max_{0 \leq k \leq K-1} \|\epsilon_k\|_\mu^2 + \frac{\gamma^K}{(1 - \gamma)^{1/2}}.
\]

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

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

\[
\leq \mathbb{E}_{\rho,\pi}[|Q(s, a) - Q^*(s, a)|] \leq \sqrt{\mathbb{E}_{\rho,\pi}[(Q(s, a) - Q^*(s, a))^2]}
\]

\[
= \|Q_K - Q^*\|_{\rho,\pi} \leq \sqrt{\frac{\kappa_\mu}{1 - \gamma}} \max_{0 \leq k \leq K-1} \|\epsilon_k\|_\mu + \frac{\gamma^{K/2}}{(1 - \gamma)^{1/2}}.
\]

(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 $e_{k-1} = T^* Q_{k-1} - Q_k$, for all $k$ and $\pi^*(s) = \arg\max_a Q^*(s, a)$, $\forall s$, we have

\[
Q^* - Q_K = T^* Q^* - T^* Q_{K-1} + T^* Q_{K-1} - T^* Q_{K-1} + e_{K-1} \leq 0
\]

\[
\leq \gamma P^\pi (Q^* - Q_{K-1}) + e_{K-1}
\]

\[
\leq \sum_{k=0}^{K-1} \gamma^{K-k-1} (P^\pi)^{K-k-1} e_k + \gamma^K (P^\pi)^K (Q^* - Q_0) \quad \text{(by recursion).} \tag{1}
\]

Now, let $\pi_k$ be the greedy policy w.r.t. $Q_k$, we have

\[
Q^* - Q_K = \sum_{T^{\pi_k} Q^* 
\geq \gamma P^{\pi_{K-1}} (Q^* - Q_{K-1}) + e_{K-1}
\]

\[
\geq \sum_{k=0}^{K-1} \gamma^{K-k-1} (P^{\pi_{K-1}} \ldots P^{\pi_0}) e_k + \gamma^K (P^{\pi_{K-1}} \ldots P^{\pi_0}) (Q^* - Q_0). \tag{2}
\]
Now, we turn to decompose $Q^* - Q^{\pi_K}$ as
\[
Q^* - Q^{\pi_K} = (T^{\pi}Q^* - T^{\pi}Q_K) + (T^{\pi}Q_K - T^{\pi_K}Q_K) + (T^{\pi_K}Q_K - T^{\pi_K}Q^{\pi_K}) \\
\leq \gamma P^{\pi} (Q^* - Q_K) + \gamma P^{\pi_K} (Q_K - Q^* + Q^* - Q^{\pi_K}).
\]

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

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

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

Thus, we have
\[
\|Q^* - Q^{\pi_K}\|^2_{\rho^*} \leq \frac{2\gamma \sqrt{\kappa_\mu}}{(1 - \gamma)^2} \max_{0 \leq k \leq K-1} \|\epsilon_k\|_{\mu} + \frac{2\gamma^{K/2+1}}{(1 - \gamma)^{3/2}}.
\]
Finally, we have

\[
\text{SubOpt}(\pi_K) = \mathbb{E}_\rho [Q^*(s, \pi^*(s)) - Q^*(s, \pi_K(s))] \\
\leq \mathbb{E}_\rho [Q^*(s, \pi^*(s)) - Q^\pi_K(s, \pi^*(s)) + Q^\pi_K(s, \pi_K(s)) - Q^*(s, \pi_K(s))] \\
\leq \|Q^* - Q^\pi_K\|_\rho + \|Q^* - Q^\pi_K\|_\rho^K \\
\leq 4\gamma \sqrt{\frac{K}{\rho_1}} \max_{0 \leq k \leq K-1} \|\epsilon_k\|_\mu + \frac{4\gamma K^{1/2}}{(1 - \gamma)^{3/2}}.
\]

\[\square\]

**Step 2: A Bellman error decomposition**

The next step of the proof is to decompose the Bellman errors \(\|Q_{k+1} - T^* Q_k\|_\mu\) 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 \(D\). 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 \(\mathbb{E} [T^* Q_k|x_i] - y_i|x_i = 0\) 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 \(D\) 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 \(\{r_i + \gamma \max_{a'} Q_k(s'_i, a')\}_{i=1}^n\) where \(Q_k\) is a deterministic function from the covering which ensures that \(\mathbb{E} [r_i + \gamma \max_{a'} Q_k(s'_i, a') - T^* Q_k|x_i] = 0\). In particular, we denote

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

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

\[
\mathbb{E}(l_f - l_{f^Q}) = \|f - f^Q_*\|^2_\mu, \forall f,
\]

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

\[
\max_k \|Q_{k+1} - T^* Q_k\|_\mu^2 = \max_k \|f^{Q_k} - f^Q_*\|^2_\mu \leq \sup_{Q \in \mathcal{F}_{NN}} \|f^{Q} - f^Q_*\|^2_\mu \overset{(a)}{=} \sup_{Q \in \mathcal{F}_{NN}} \mathbb{E}(l_f^Q - l_{f^Q}) \overset{(b)}{=} \sup_{Q \in \mathcal{F}_{NN}} \mathbb{E}(l_f^Q - l_{f^Q}) \overset{(c)}{=} \sup_{Q \in \mathcal{F}_{NN}} \left\{ \mathbb{E}(l_f^Q - l_{f^Q}) + \mathbb{E}(l_{f^Q} - l_{f^Q}) \right\} \overset{(d)}{=} \sup_{Q \in \mathcal{F}_{NN}} \left\{ \mathbb{E}(l_{f^Q} - l_{f^Q}) + \mathbb{E}(l_{f^Q} - l_{f^Q}) \right\} \overset{(e)}{=} \sup_{Q \in \mathcal{F}_{NN}} \left\{ \mathbb{E}(l_{f^Q} - l_{f^Q}) + \mathbb{E}(l_{f^Q} - l_{f^Q}) \right\} = \sum_{Q \in \mathcal{F}_{NN}} \left\{ \mathbb{E}(l_{f^Q} - l_{f^Q}) + \mathbb{E}(l_{f^Q} - l_{f^Q}) \right\} \right.
\]

where \(a, b, c, d, e\) follows from that \(Q_k \in \mathcal{F}_{NN}\), 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 \(f^Q_\perp\) not being in the function space \(\mathcal{F}_{NN}\).
Sample Complexity of Offline Reinforcement Learning with Deep ReLU Networks

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 \( \mathcal{F} := \{ f - g : f, g \in \mathcal{G} \} \), let \( N(\epsilon, \mathcal{F}, \| \cdot \|) \) be the \( \epsilon \)-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 \( N_0(\epsilon, \mathcal{F}, \| \cdot \|) \) be the bracketing number of \( \mathcal{F} \), i.e., the minimum number of brackets of \( \| \cdot \| \)-size less than or equal to \( \epsilon \), necessary to cover \( \mathcal{F} \), let \( H_0(\epsilon, \mathcal{F}, \| \cdot \|) \) be the \( \| \cdot \| \)-bracketing metric entropy of \( \mathcal{F} \), let \( \mathcal{F} \{ x \}^n_{i=1} = \{ (f(x_1), ..., f(x_n)) \in \mathbb{R}^n : f \in \mathcal{F} \} \), and let \( T^* \mathcal{F} = \{ T^* f : f \in \mathcal{F} \} \). Finally, for sample set \( \{ x_i \}^n_{i=1} \), 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 \( d_{\mathcal{F} \mathcal{N} \mathcal{N}} := \sup_{Q \in \mathcal{F} \mathcal{N} \mathcal{N}} \inf_{f \in \mathcal{F} \mathcal{N} \mathcal{N}} \| f - T^* Q \|_\mu \). This implies that

\[
\frac{d_{\mathcal{F} \mathcal{N} \mathcal{N}}^2 := \sup_{Q \in \mathcal{F} \mathcal{N} \mathcal{N}} \inf_{f \in \mathcal{F} \mathcal{N} \mathcal{N}} \| f - T^* Q \|_\mu^2 = \sup_{Q \in \mathcal{F} \mathcal{N} \mathcal{N}} \mathbb{E}(l_{f}^Q - l_{f^Q}).
\]

We have

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

We have

\[
H(\epsilon, \{ l_{f \perp}^Q - l_{f^Q} : Q \in \mathcal{F} \mathcal{N} \mathcal{N} \})_{\{ x_i, y_i \}^n_{i=1}} \leq H\left(\frac{\epsilon}{4}, \{ f^Q - f^Q \perp : Q \in \mathcal{F} \mathcal{N} \mathcal{N} \} \right)_{\{ x_i \}^n_{i=1}} \leq H\left(\frac{\epsilon}{4}, \{ \mathcal{F} - T^* \mathcal{F} \mathcal{N} \mathcal{N} \} \right)_{\{ x_i \}^n_{i=1}} \leq H\left(\frac{\epsilon}{8}, \mathcal{F} \mathcal{N} \mathcal{N} \right)_{\{ x_i \}^n_{i=1}} + H\left(\frac{\epsilon}{8}, T^* \mathcal{F} \mathcal{N} \mathcal{N} \right)_{\{ x_i \}^n_{i=1}} \leq H\left(\frac{\epsilon}{8}, \mathcal{F} \mathcal{N} \mathcal{N} \right)_{\{ x_i \}^n_{i=1}} + H\left(\frac{\epsilon}{8}, T^* \mathcal{F} \mathcal{N} \mathcal{N} \right)_{\| \cdot \| \infty}.
\]

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

\[
\mathbb{E}(l_{f \perp}^Q - l_{f^Q}) \leq 3\mathbb{E}(l_{f \perp}^Q - l_{f^Q}^Q) + \epsilon'' \leq 3d_{\mathcal{F} \mathcal{N} \mathcal{N}}^2 + \epsilon'^2,
\]

given that

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

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 \( \epsilon' \) instead of \( \epsilon'^2 \) (i.e., \( n \) scales with \( O(1/\epsilon'^2) \) instead of \( O(1/\epsilon'^4) \)). In addition, unlike (Le et al., 2019), the uniform convergence argument hinges 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 \( \{ s_i, a_i \}^n_{i=1} \)). 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

$$|f_{f_{2}}^{Q} - f_{1}^{Q}| \leq 2|f_{2}^{Q} - f_{1}^{Q}| \leq 2,$$

$$\forall |f_{f_{2}}^{Q} - f_{1}^{Q}| \leq \mathbb{E}((f_{f_{2}}^{Q} - f_{1}^{Q})^{2}) \leq 4\mathbb{E}(f_{1}^{Q} - f_{2}^{Q})^{2}.$$

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

$$\sup\{(\mathbb{E} - \mathbb{E}_n)(f_{f_{2}} - f_{1}^{Q}) : Q \in \mathcal{F}_{NN}, \|f_{f_{2}}^{Q} - f_{1}^{Q}\|_{\mu} \leq r\}$$

$$\leq \sup\{(\mathbb{E} - \mathbb{E}_n)(f_{f_{2}} - f_{1}) : f \in \mathcal{F}_{NN}, g \in T^{*}\mathcal{F}, \|f - g\|_{\mu} \leq r\}$$

$$\leq 3\mathbb{E}R_n\{l_{f} - l_{g} : f \in \mathcal{F}_{NN}, g \in T^{*}\mathcal{F}_{NN}, \|f - g\|_{\mu} \leq r\} + 2\sqrt{\frac{2r\log(1/\delta)}{n}} + \frac{28\log(1/\delta)}{3n}.$$

**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 $\psi$ 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

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

We recall that a function 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 $\psi$ has a unique fixed point $r_*$ where $r_* = \psi(r_*)$. In addition, $\psi(r) \leq \sqrt{\mathbb{F}}r_*, \forall r \geq r_*$. In the next step, we will find a sub-root function $\psi$ that satisfies the inequality above, but for this step we just assume that we have such $\psi$ at hand. Combining (5), (7), and (8), we have: for any $r \geq r_*$ and any $\delta \in (0, 1)$, if $\|f_{Q_{k-1}}^{Q_{k-1}} - f_{*}^{Q_{k-1}}\|_{2, \mu} \leq r$, with probability at least $1 - \delta$, we have

$$\|f_{Q_{k-1}}^{Q_{k-1}} - f_{*}^{Q_{k-1}}\|_{2, \mu} \leq 2\psi(r) + 2\sqrt{\frac{2r\log(2/\delta)}{n}} + \frac{28\log(2/\delta)}{3n} + 3d_F + \epsilon^2$$

$$\leq \sqrt{r_\delta} + 2\sqrt{\frac{2r\log(2/\delta)}{n}} + \frac{28\log(2/\delta)}{3n} + (\sqrt{3}d_F + \epsilon^2),$$

where

$$n = \frac{1}{4\epsilon^2} \left(\log(8/\delta) + \log\mathbb{E}N\left(\frac{\epsilon^2}{20}, (\mathcal{F}_{NN} - T^{*}\mathcal{F}_{NN})\{x_i\}_{i=1}^{n}, n^{-1}\|\cdot\|_1\right)\right).$$

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

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

where $l = \log_2(\frac{1}{r_0}) \leq \log_2(\frac{1}{r_*})$. We have $B_0 \subseteq B_1 \subseteq \ldots \subseteq B_l$ and since $\|f - g\|_{\mu} \leq 1, \forall \|f\|_{\infty}, \|g\|_{\infty} \leq 1$, we have

$$P(B_l) = 1.$$ If $\|f_{Q_{k-1}}^{Q_{k-1}} - f_{*}^{Q_{k-1}}\|_{2, \mu} \leq 2^{k}r_0$ for some $i \leq l$, then with probability at least $1 - \delta$, we have

$$\|f_{Q_{k-1}}^{Q_{k-1}} - f_{*}^{Q_{k-1}}\|_{2, \mu} \leq 2^{k}r_0 + 2\sqrt{\frac{2^{k+1}r_0\log(2/\delta)}{n}} + \frac{28\log(2/\delta)}{3n} + (\sqrt{3}d_F + \epsilon^2)^2 \leq 2^{i-1}r_0,$$

if the following inequalities hold

$$\sqrt{2^{k}r_0} + 2\sqrt{\frac{2^{k+1}\log(2/\delta)}{n}} \leq \frac{1}{2}2^{i-1}\sqrt{r_0},$$

$$\frac{28\log(2/\delta)}{3n} + (\sqrt{3}d_F + \epsilon)^2 \leq \frac{1}{2}2^{i-1}r_0.$$
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
\[
\sqrt{r_0} = \frac{2}{2^{l-1}} \left( \sqrt{2^r r_*} + 2 \sqrt{\frac{2^{i+1} \log(2/\delta)}{n}} \right)_{i=0} + \frac{2}{2^{l-1}} \left( \frac{28 \log(2/\delta)}{3n} \right)_{i=0} + (\sqrt{\delta d_{\cal F_N}} + \epsilon')^2 \left|_{i=0} \right.
\]
\[
\lesssim d_{\cal F_N} + \epsilon' + \sqrt{\frac{\log(2/\delta)}{n}} + \sqrt{r_*}.
\]

Since \( \{B_i\} \) is a sequence of increasing events, we have
\[
P(B_0) = P(B_1) - P(B_1 \cap B_0^c) = P(B_2) - P(B_2 \cap B_1^c) - P(B_1 \cap B_0^c)
\]
\[
= P(B_l) - \sum_{i=0}^{l-1} P(B_{i+1} \cap B_i^c) \geq 1 - l \delta.
\]

Thus, with probability at least \( 1 - \delta \), we have
\[
\|f^{Q_{k-1}} - f^{Q_{k-1}}_*\|_{\mu} \lesssim d_{\cal F_N} + \epsilon' + \sqrt{\frac{\log(2/\delta)}{n}} + \sqrt{r_*} \tag{9}
\]
where
\[
n \approx \frac{1}{4\epsilon^2} \left( \log(8l/\delta) + \log \mathbb{E} N \left( \frac{\epsilon^2}{20} \| \cal F_{N_0} - T^* \|_{\cal F_{N_0}} \right) \right).
\]

**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 \( \cal F_{N_0} \) and \( T^* \cal F_{N_0} \).

For any \( \epsilon > 0 \), we have the following inequalities for entropic numbers:
\[
H(\epsilon, \cal F_{N_0} - T^* \cal F_{N_0}, \| \cdot \|_n) \leq H(\epsilon/2, \cal F_{N_0}, \| \cdot \|_n) + H(\epsilon/2, T^* \cal F_{N_0}, \| \cdot \|_n),
\]
\[
H(\epsilon, \cal F_{N_0}, \| \cdot \|_n) \leq H(\epsilon, \cal F_{N_0} \{x_i\}_{i=1}^n, \| \cdot \|_n) \lesssim N[(\log N)^2 + \log(1/\epsilon)], \tag{10}
\]
\[
H(\epsilon, T^* \cal F_{N_0}, \| \cdot \|_n) \leq H(\epsilon, T^* \cal F_{N_0}, \| \cdot \|_\infty) \leq H(2\epsilon, T^* \cal F_{N_0}, \| \cdot \|_\infty) \leq H(2\epsilon, T^* \cal F_{N_0}, \| \cdot \|_\infty) \leq (2\epsilon)^{-d/\alpha}, \tag{11}
\]
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 \( \cal H := \cal F_{N_0} - T^* \cal F_{N_0} \); it follows from Lemma 6 with \( \{\xi_k := \epsilon/2^k\}_{k \in \mathbb{N}} \) for any \( \epsilon > 0 \) that
\[
\mathbb{E}_{\xi} R_n \{ h \in \cal H : \| h \|_{\frac{1}{2k-1}} \leq \epsilon \} \leq 4 \sum_{k=1}^\infty \frac{\epsilon}{2^{k-1}} \sqrt{\frac{H(\epsilon/2^{k-1}, \cal H, \| \cdot \|_{\frac{1}{2k-1}})}{n}} \leq 4 \sum_{k=1}^\infty \frac{\epsilon}{2^{k-1}} \left( \sqrt{\frac{H(\epsilon/2^{k-1}, \cal H, \| \cdot \|_{\frac{1}{2k-1}})}{n}} + \frac{H(\epsilon/2^k, \cal F_{N_0}, \| \cdot \|_{\infty})}{n} \right) \leq \frac{4\epsilon}{\sqrt{n}} \sum_{k=1}^\infty 2^{-(k-1)} \sqrt{\frac{\epsilon}{2k-1}} + \frac{\epsilon^2}{(2k-1) \sqrt{n}} \leq \frac{1}{\sqrt{n}} \epsilon \sqrt{\frac{\epsilon}{\sqrt{n}}},
\]
where we use \( \sqrt{a + b} \leq \sqrt{a} + \sqrt{b}, \forall a, b \geq 0, \sum_{k=1}^\infty \frac{2^\epsilon}{\sqrt{2^k}} < \infty, \) and \( \sum_{k=1}^\infty \left( \frac{1}{2k-1} \right)^{k-1} < \infty.\)
It now follows from Lemma 5 that

\[
E_\sigma R_n\{f \in F, g \in T^*F : \|f - g\|_\mu^2 \leq r\}
\leq \inf_{\epsilon > 0} E_\sigma R_n\{h \in \mathcal{H} - \mathcal{H} : \|h\|_\mu \leq \epsilon\} + \sqrt{\frac{2rH(\epsilon/2, \mathcal{H}, \|\cdot\|_n)}{n}}
\]
\[
\lesssim \left[ \frac{\epsilon}{\sqrt{n}} \log N((\log N)^2 + \log(1/\epsilon)) + \frac{\epsilon^{1 - \frac{d}{2}}}{\sqrt{n}} + \sqrt{\frac{2r}{n}} \log N((\log N)^2 + \log(4/\epsilon)) + \sqrt{\frac{2r}{n}} (\epsilon/2)^{\frac{d}{2}} \right]_{\epsilon = n^{-\beta}}
\approx n^{-\beta - 1/2} \sqrt{N(\log^2 N + \log n)} + n^{-\beta(1 - \frac{d}{2})-1/2} + \frac{\sqrt{r}}{n} \sqrt{N(\log^2 N + \log n)} + \sqrt{rn}^{-\frac{1}{2}(1 - \frac{d}{2})} =: \psi_1(r),
\]

where \(\beta \in (0, \frac{d}{2})\) 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 \(f \in F_{NN}, g \in T^*F_{NN}\). 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 \(f \in F_{NN}, g \in T^*F_{NN}\) such that \(\|f - g\|_\mu^2 \leq r\),

\[
\|f - g\|_n^2 \\
\leq \|f - g\|_\mu^2 + 3E_{NN}\{f - g \in F_{NN}, g \in T^*F_{NN}, \|f - g\|_\mu^2 \leq r\} + \sqrt{\frac{2r \log n}{n}} + \frac{56 \log n}{3n}
\leq \|f - g\|_\mu^2 + 3E_{NN}\{f - g \in F_{NN}, g \in T^*F_{NN}, \|f - g\|_\mu^2 \leq r\} + \sqrt{\frac{2r \log n}{n}} + \frac{56 \log n}{3n}
\leq r + \psi(r) + r + r \leq 4r,
\]

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

\[
3E_{NN}\{f - g \in F_{NN}, g \in T^*F_{NN}, \|f - g\|_\mu^2 \leq r\}
= 3E_{NN}\{f - g \in F_{NN}, g \in T^*F_{NN}, \|f - g\|_\mu^2 \leq r\}
\leq 3E_{NN}\{f \in E_r, \|f - g\|_\mu^2 \leq r\} + (1 - 1E_r)
\leq 3E_{NN}\{f \in E_r, \|f - g\|_\mu^2 \leq r\} + (1 - 1E_r)
\leq 3(\psi_1(4r) + \frac{1}{n})
\lesssim n^{-\beta - 1/2} \sqrt{N(\log^2 N + \log n)} + n^{-\beta(1 - \frac{d}{2})-1/2} + \sqrt{\frac{r}{n}} \sqrt{N(\log^2 N + \log n)}
+ \sqrt{rn}^{-\frac{1}{2}(1 - \frac{d}{2})} + n^{-1} =: \psi(r)
\]

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

\[
\sqrt{r_*} \lesssim n^{-1/2} \sqrt{N(\log^2 N + \log n)} + n^{-\frac{d}{2}(1 - \frac{d}{2})} + n^{-\beta} - \frac{1}{2} [N(\log^2 N + \log n)]^{1/4}
+ n^{-\beta(1 - \frac{d}{2})-1/2} + n^{-1/2}
\lesssim n^{-\frac{d}{2}(1 - \frac{d}{2})} + n^{-\frac{1}{2}(1 - \frac{d}{2})} + n^{-\frac{d}{2}(1 - \frac{d}{2})-1/2} + n^{-1/2}
\]

(12)

It follows from Equation (9) (where \(l \lesssim \log(1/r_*))\), the definition of \(d_{F_{NN}}\), Lemma 10, and (12) that for any \(\epsilon' > 0\) and \(\delta \in (0, 1)\), with probability at least \(1 - \delta\), we have

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

(13)
where

\[
    n \gtrsim \frac{1}{4e^2} \left( \log(1/\delta) + \log \log n + \log \mathbb{E} N \left( \frac{\epsilon^2}{20}, (\mathcal{F}_{NN} - T^* \mathcal{F}_{NN}) | \{x_i\}_{i=1}^n, n^{-1} : \| \cdot \|_1) \right) \right) .
\]

(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 \( \beta \in (0, \frac{d}{\alpha}) \) 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 \( \beta \in (0, \frac{d}{\alpha}) \) 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

\[
    N = n^{\frac{1}{4}((2\beta \wedge 1) + 1)} \frac{d}{\alpha + d} \quad \text{and} \quad \beta = \left( 2 + \frac{d^2}{\alpha(\alpha + d)} \right)^{-1},
\]

(15)

which results in \( N \approx n^{\frac{1}{4}(2\beta + 1) + 1} \frac{d}{\alpha + d} \). At these optimal values, (13) becomes

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

(16)

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

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

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

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

\[
    n \gtrsim \left( \frac{1}{\epsilon^2} \right)^{\frac{d}{\alpha + d} + \frac{\delta}{\delta + 1}} \left( \log^2 n \right)^{\frac{d}{\alpha + d} + \frac{\delta}{\delta + 1}} \quad \text{and} \quad n \gtrsim \frac{1}{\epsilon^2} \left( \log(1/\delta) + \log \log n \right).
\]

(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

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

(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

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

**Appendix B. Technical Lemmas**

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

\[
    \mathcal{F} \subseteq \{ f : \mathcal{X} \to [a, b] : \text{Var}[f(X_1)] \leq r \}.
\]
1. For any $\lambda > 0$, we have with probability at least $1 - e^{-\lambda}$,

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

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

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

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

**Lemma 3** (Györfi et al. (2002, Theorem 11.6)). Let $B \geq 1$ and $\mathcal{F}$ be a set of functions $f : \mathbb{R}^d \rightarrow [0, B]$; 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

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

**Lemma 4** (Contraction property (Rebeschini, 2019)). Let $\phi : \mathbb{R} \rightarrow \mathbb{R}$ be a $L$-Lipschitz, then

$$
E_{\sigma} R_n (\phi \circ \mathcal{F}) \leq LE_{\sigma} R_n \mathcal{F}.
$$

**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 $X_1$), we have

$$
E_{\sigma} R_n \{ f \in \mathcal{F} : \| f \|_2^2 \leq r \} \leq \inf_{\epsilon > 0} \left[ E_{\sigma} R_n \{ f \in \mathcal{F} - \mathcal{F} : \| f \|_{2, \mu} \leq \epsilon \} + \sqrt{\frac{2r \log \mathcal{N}(\epsilon / 2, \mathcal{F}, \| \cdot \|_{2, \mu})}{n}} \right]
$$

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 $\{ \xi_k \}_{k=n}^{\infty}$ decreasing to 0, we have the following inequality for any non-negative integer $N$

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

**Lemma 7** (Pollard’s inequality). Let $\mathcal{F}$ be a set of measurable functions $f : \mathcal{X} \rightarrow [0, K]$ 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

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

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

$$H_\|\|_0(\epsilon, \mathcal{B}^\alpha_{p,q}(X), \| \cdot \|_\infty) \lesssim \epsilon^{-d/\alpha}.$$ 

Lemma 10 (Approximation power of deep ReLU networks for Besov spaces (Suzuki, 2018)). Let $1 \leq p, q \leq \infty$ and $\alpha \in \left(\frac{d}{p^2}, \infty\right)$. 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 $\delta := d(p^{-1} - (1 + |\alpha|)^{-1})^+$ such that

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