On Non-Asymptotic Bounds for Off-Policy Evaluation with Deep ReLU Networks

Anonymous Author
Anonymous Institution

Abstract

This paper studies the statistical theory of batch reinforcement learning with deep ReLU networks. We consider the off-policy evaluation (OPE) problem in which the goal is to estimate the expected discounted reward of a target policy given the logged data generated by unknown behaviour policies. We study a regression-based fitted Q evaluation (FQE) method using deep ReLU networks and characterize a finite-sample bound on the estimation error of this method under mild assumptions. In particular, by leveraging local Rademacher complexity techniques, contraction property of Markov processes, and approximation theory of neural networks, we provide a finite-sample bound that characterizes how distributional shift, the regularity of the regression function, the state-action space dimension and the sample size contribute to the estimation error. This result offers a new understanding of the role of deep learning in OPE as a promising alternative to the existing methods.

1 Introduction

Batch 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 often prohibited. A fundamental question in this setting is how well we could use the previous experiences to evaluate and improve the performance of new policies. This problem is known as off-policy evaluation (OPE) where the goal is to evaluate the value of a new policy given only fixed logged datasets from previous interactions of different (possibly unknown) behaviour policies.

In this paper, we study off-policy evaluation using deep ReLU network function approximation. In particular, we analyze a regression-based approach known as fitted Q-evaluation (FQE), a variant of the basic fitted Q-iteration (Bertsekas et al., 1995; Sutton and Barto, 2018). This approach works by iteratively estimating Q-functions via regression using the batch data. Though FQE with linear function approximations and general function approximation have been studied in (Duan and Wang, 2020) and (Le et al., 2019), respectively, an analysis of FQE with deep ReLU networks has not been formally investigated.

In this work, we provide a finite-sample error bound for this FQE policy value estimator with deep ReLU network function approximations. Our bound takes the form

$$|v_K - v^{\pi}| \asymp \frac{\sqrt{\kappa(\rho\|\mu)}}{1-\gamma} n^{-\frac{\beta}{d+2\beta}} + O(1/\sqrt{n}),$$

where n is the number of observed state transitions, μ is the behaviour (sampling) state-action distribution for the offline dataset, ρ is the initial state-action distribution under the evaluation policy π, d is the dimension of the state-action space, β measures the smoothness induced by the Bellman operator into deep ReLU networks, and the concentration coefficient $\kappa(\rho\|\mu)$ measures the distributional shift. In addition, the deep ReLU network that realizes this bound has at most $L \leq (2+\lceil\log_2(\beta)\rceil)(11+\beta/d)$ layers, independent of the sample size n, and $W_n \asymp n^{\frac{d}{d+2\beta}}$ nonzero, quantized weights. Our result suggests that OPE could be data-efficient when one can leverage the expressive power of deep ReLU networks on the batch data.

Outline. In Section 2, we present the necessary background and position our work within the related literature of OPE. Section 3 presents FQE algorithm with deep ReLU networks followed by our finite-time error bound analysis. In particular, in Section 3.2, we present all the assumptions for our analysis and all the

Preliminary work. Under review by OFFLINERL2020 Workshop. Do not distribute.

relevant results which culminate into our main finite-sample error bound in Theorem 1. Finally, Section 4 discusses our contributions and concludes our work.

2 Background and Related Literature

2.1 Off-Policy Evaluation Problem

In this paper, we study off-policy evaluation (OPE) problem of an Markov Decision Process (MDP) where we only have a fixed logged dataset of empirical transitions by unknown behaviour policies. An MDP$(\mathcal{S}, \mathcal{A}, P, R, \gamma, \rho)$ consists of a state space $\mathcal{S}$, an action space $\mathcal{A}$, a transition kernel measure $P(s'|s,a)$ which specifies the probability of transitioning to state s' given the state-action (s,a), the reward distribution $R(s,a) \in \mathcal{P}([-R_{max}, R_{\max}])$, a discount factor $\gamma \in [0,1]$ and an initial state distribution ρ. A (stationary) policy induces a distribution over $\mathcal{A}$ at each given state $s \in \mathcal{S}$. Let $V_{max} = R_{\max}/(1-\gamma)$.

The goal of OPE is to evaluate the performance of an evaluation policy π at a fixed initial state distribution ρ where the transition kernel is unknown and only a fixed logged dataset $D_n = \{s_i, a_i, s_i', r_i\}_{i=1}^n$ where $(s_i, a_i) \overset{i.i.d.}{\sim} \mu$ and $s_i' \sim P(\cdot|s_i, a_i)$ and $r_i \sim R(s_i, a_i)$ is available. Here μ is the sampling state-action distribution which is also unknown. The value to be estimated is the expected discounted reward given by

$$v^\pi = \mathbb{E}_{\rho,\pi}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)\right],$$

where $s_0 \sim \rho, a_t \sim \pi(\cdot|s_t)$, and $s_{t+1} \sim P(\cdot|s_t, a_t)$. Since we evaluate the expected discounted reward of a fixed evaluation policy π, we slightly abuse the notation by using ρ as the initial state-action distribution induced by π. It is convenient to define the state-action value $Q^\pi(s,a)$ and the Bellman operator T^π as

$$Q^\pi(s,a) := \mathbb{E}_{\rho,\pi}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) | s_0 = s, a_0 = a\right],$$
$$T^\pi Q(s,a) := r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a), a' \sim \pi(\cdot|s')}[Q(s', a')],$$

where $r(s,a) = \mathbb{E}[R(s,a)]$ and $Q \in \mathbb{R}^{\mathcal{S} \times \mathcal{A}}$.

Additional Notations. We write $f(n) \lesssim g(n)$ if there exists $C > 0$ and n_0 such that $f(n) \leq Cg(n), \forall n > n_0$. We write $f(n) \asymp g(n)$ if $f(n) \lesssim g(n)$ and $g(n) \lesssim f(n)$. Denote by $\frac{d\nu}{d\mu} : \Omega \to [0, \infty)$ the Radon-Nikodym derivative for any two probability measures ν and μ on the same measurable space (Ω, Σ) such that $\nu \ll \mu$ (i.e., $\forall A \in \Sigma, \mu(A) = 0 \implies \nu(A) = 0$). We slightly abuse notation C to represent a universal constant whose specific value can vary every time the notation C is overloaded.

Denote by $\mathcal{X} = \mathcal{S} \times \mathcal{A}$ the Cartesian product space, $\mathcal{P}(\mathcal{X})$ the space of Borel probability measures on $\mathcal{X}$, and $\mathcal{B}(\mathcal{X}; V_{\max})$ the space of functions $f : \mathcal{X} \to \mathbb{R}$ such that $\|f\|_\infty \leq V_{\max}$. For any $\eta \in \mathcal{P}(\mathcal{X})$, denote $\|f\|_\eta := \sqrt{\int f(x)^2 \eta(dx)}$ for any measurable $f : \mathcal{X} \to \mathbb{R}$. For any stationary policy π, the left-linear operator $\cdot P^\pi : \mathcal{P}(\mathcal{X}) \to \mathcal{P}(\mathcal{X})$ is defined by

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

for any $\rho \in \mathcal{P}(\mathcal{X})$. The right-linear operator $P^\pi \cdot : \mathcal{B}(\mathcal{X}; V_{\max}) \to \mathcal{B}(\mathcal{X}; V_{\max})$ is defined by

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

We denote by $a \vee b$ the max operation $\max\{a, b\}$. We denote by $C^n(\mathcal{X})$ the set of all n-times continuously differential function on domain $\mathcal{X}$.

2.2 Related Literature

A direct approach to OPE estimates a model of the system and uses this model to estimate the performance of the evaluation policy. This has been studied in the tabular case in (Mannor et al., 2004). In practice, the state space of MDPs is however often infinite or continuous, thus function approximations are often deployed in approximate dynamic programming such as fitted Q-iteration and 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).

A popular approach to OPE uses importance sampling (IS) to obtain an unbiased value estimate of new policies by reweighing sample rewards (Precup et al., 2000). Doubly robust estimations combine IS with model-based estimators to reduce the high variance (Dudík et al., 2011; Jiang and Li, 2015; Thomas and Brunskill, 2016; Farajtabar et al., 2018; Kallus and Uehara, 2019). (Liu et al., 2018) proposed to directly estimate the stationary state distribution instead of the cumulative importance ratio to break the curse of horizon. Following this, many works reformulated the problem of estimating the stationary state distribution as a density ratio estimation problem and proposed different estimators (Nachum et al., 2019a; Zhang et al., 2020a,b; Nachum et al., 2019b).

Theoretically, (Xie et al., 2019; Yin and Wang, 2020) provided the probably sharpest OPE error bound for tabular MDPs. (Jiang and Li, 2016) showed a Cramer-Rao lower bound for discrete-tree MDP. While most existing theoretical results apply only to tabular MDP without function approximation, (Duan and Wang, 2020) provided the probably shapest error bound for OPE with linear function approximation. (Le et al., 2019) provided the error bound of OPE with general function approximation and only hinted an extension to deep neural networks. To our best knowledge, the error bound for OPE with deep neural networks has not been studied and our results appear to be the first and sharpest error bounds for OPE with deep ReLU networks. The most related work to our work is perhaps (Yang et al., 2019) which also considers deep neural network approximation; however, (Yang et al., 2019) focused on analyzing deep Q-learning instead of the OPE problem.

2.3 Deep ReLU Networks

We study fitted Q evaluation method with deep networks for OPE. We consider deep networks with rectified linear unit (ReLU) activation function $\sigma(x) = \max\{x, 0\}$ and with bounded and quantized weights. This is motivated by practical settings as one only has a fixed number of bits for storing each weight of a neural network on a typical computer and ReLU is probably the most widely used activation in practice. For any integer L and d, and $\{d_i\}_{i=0}^L \subset \mathbb{N}$ where $d_0 = d$, a neural network Φ with input dimension d and L layers is a sequence of matrix-vector tuples

$$\Phi = ((W_1, b_1), ..., (W_L, b_L)),$$

where each $W_l \in \mathbb{R}^{d_l \times d_{l-1}}$, and $b_l \in \mathbb{R}^{d_l}$. The realization of the neural network Φ is defined as $f_\Phi : \mathbb{R}^d \to \mathbb{R}^{d_L}, x \mapsto x_L$ where

$$\begin{cases} x_0 &= x, \\ x_l &= \sigma(W_l x_{l-1} + b_l), \forall l \in \{1, 2, ..., L-1\}, \\ x_L &= W_L x_{L-1} + b_L. \end{cases}$$

For any $\epsilon > 0$ and $s \in \mathbb{N}$, a neural network $\Phi = ((W_1, b_1), ..., (W_L, b_L))$ is said to have (s, ϵ)-quantized weights if all entries of $W_1, ..., W_L, b_1, ..., b_L$ are elements of $[-\frac{1}{\epsilon^s}, \frac{1}{\epsilon^s}] \cap 2^{-s\lceil \log_2(1/\epsilon) \rceil}\mathbb{Z}$ (Petersen and Voigtlaender, 2018). We denote by W the total number of nonzero weights in the ReLU network functions. Denote by $\mathcal{F}(d, B)$ the set of all ReLU network functions $f_\Phi : \mathbb{R}^d \to \mathbb{R}$ with (s, ϵ)-quantized weights for some $\epsilon > 0$ and $s \in \mathbb{N}$ and $\|f_\Phi\|_\infty \leq B$.

2.4 Regularity of the regression functions

To study the rate of convergence for OPE with deep ReLU networks, we need assumptions on the regularity of the regression function. To this end, we consider the notion of smoothness as follows. For $d \in \mathbb{N}$ and for any $\beta \in (0, \infty)$ with $\beta = m + \zeta$ where $m \in \mathbb{N}_0$ and $\zeta = (0, 1]$, we define the norm for $f \in C^m([-1/2, 1/2]^d)$ as

$$\|f\|_{C^{0,\beta}} := \max_{|\alpha| = m} \|\partial^\alpha f\|_\infty \vee \max_{|\alpha| = n} Lip_\zeta(\partial f),$$

where $Lip_\zeta(g) := \sup_{x, y \in \Omega, x \neq y} \frac{|g(x) - g(y)|}{|x - y|^\zeta}$ for $g : \Omega \subset \mathbb{R}^d \to \mathbb{R}$, and $\alpha = (\alpha_1, ..., \alpha_d)$ denotes multi-index with $|\alpha| = \sum_{i=1}^d \alpha_i$. Then for $B > 0$, we define the following class of smooth functions

$$\mathcal{G}(\beta, d, B) := \{f \in C^m([-1/2, 1/2]^d) : \|f\|_{C^{0,\beta}} \leq B\}.$$

In words, $\mathcal{G}(\beta, d, B)$ consists of the functions from $C^m([-1/2, 1/2]^d)$ where all derivatives of order m are Lipschitz continuous. Note that this is equivalent to the Sobolev spaces of order $m + 1$ (Yarotsky, 2017). Also note that this assumption in the case $\beta = m + 1$ is slightly weaker than assuming $f \in C^{m+1}$ as we do not require $f \in \mathcal{G}(m + 1, d, B)$ to be $m + 1$ times continuously differentiable.

3 Fitted Q-Evaluation with Deep ReLU Networks

This section presents the FQE with deep ReLU networks followed by our finite-sample error bound analysis.

3.1 Algorithm

Given an offline data $\mathcal{D}_n = \{(s_i, a_i, r_i, s_i')\}_{i=1}^n$ and an evaluation policy π, our goal is to estimate from $\mathcal{D}_n$ the expected discounted reward of the policy π defined as $v^\pi = \int Q^\pi(s, a)\rho(ds, da)$. We consider a fitted Q-evaluation (FQE) method for this off-policy evaluation problem. The algorithm details for FQE are presented Algorithm 1.

Algorithm 1: Fitted Q-Evaluation (FQE)

```
Require: MDP$(\mathcal{S}, \mathcal{A}, P, R, \gamma, \rho)$, function class
         $\mathcal{F}$, number of iterations $K$, evaluation
         policy $\pi$, offline data
         $\mathcal{D}_n = \{(s_i, a_i, r_i, s_i')\}_{i=1}^n$ where
         $(s_i, a_i)_{i=1}^n \overset{i.i.d.}{\sim} \mu$, $s_i' \sim P(\cdot|s_i, a_i)$ and
         $r_i \sim R(s_i, a_i)$.
Initialization: $Q_0 \in \mathcal{F}$
for $k=1,2,...,K$ do
    Compute $y_i = r_i + \gamma \int_{\mathcal{A}} Q_{k-1}(s_i', a)\pi(da|s_i')$
    $Q_k \leftarrow \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n (f(s_i, a_i) - y_i)^2$
end
Output: $v_K = \|Q_K\|_\rho$
```

The algorithm initializes $Q_0 \in \mathcal{F}$ arbitrarily and iteratively computes Q_k as follows: at each iteration k, the algorithm computes the Bellman targets $\{y_i\}_{i=1}^n$ using the previous estimate Q_{k-1} to construct a new regression data $\{(x_i, y_i)\}_{i=1}^n$ where $x_i = (s_i, a_i)$. It then fits the function class $\mathcal{F}$ to the constructed regression data by minimizing the mean squared error. Formally,

$$y_i = r_i + \gamma \int Q_{k-1}(s_i', a_i)\pi(da|s_i'),$$

$$Q_k = \arg\inf_{f \in \mathcal{F}} \frac{1}{n}\sum_{i=1}^n (f(s_i, a_i) - y_i)^2.$$

In this paper, we restrict $\mathcal{F}$ to be the class of deep ReLU networks $\mathcal{F}(d, V_{\max})$ as described in Section 2.3 and analyze the resulting OPE estimator.

3.2 Finite-Sample Error Bound

In this section, we establish finite-sample error bound for FQE with deep ReLU networks. Specifically, we are interested in upper bounding the quantity $\|Q_K^\pi - Q^\pi\|_\rho$ where $Q_K^\pi(s,a)$ is returned by Algorithm 1 and $\rho \in \mathcal{P}(\mathcal{S} \times \mathcal{A})$ is the initial state-action distribution. We consider a continuous action space [1] and assume that the state-action space $\mathcal{X} = \mathcal{S} \times \mathcal{A}$ is a compact space in $\mathbb{R}^d$. Without loss of generality, we assume that $\mathcal{X} = [-1/2, 1/2]^d$. We specify the function class $\mathcal{F}$ in Algorithm 1 to be the class of ReLU network functions $\mathcal{F}(d, V_{\max})$. In what follows, we present all the assumptions necessary to establish our result.

Assumption 3.1 (Completeness). There exists some $\beta \in (0, \infty)$ such that any $f \in \mathcal{F}(d, V_{\max})$, $T^\pi f \in \mathcal{G}(\beta, d, V_{\max})$.

This completeness assumption specifies that the Bellman operator T^π applied on any ReLU network function in $\mathcal{F}(d, V_{\max})$ results in a function that is smooth enough and stays in the function class $\mathcal{G}(d, V_{\max})$. This holds when e.g., both the expected reward function $r(s,a)$ and the transition density $P(s'|s,a)$ for each fixed s' are sufficiently smooth. The completeness assumption is common in the batch RL literature (Chen and Jiang, 2019).

Next, we make an assumption about the concentration coefficients which measure the distributional shift from a sampling state-action distribution to any state-action distribution *admissible* under the MDP.

Assumption 3.2 (Concentration coefficients of future state-action distribution). Consider a sampling state-action distribution μ and initial state-action distribution ρ. Assume that $\rho \ll \mu$ and $\rho P^{\pi_1} ... P^{\pi_m} \ll \mu$ for any sequence of policies $\{\pi_t\}$, and denote

$$\kappa_m(\rho\|\mu) := \begin{cases} \left\| \frac{d\rho}{d\mu} \right\|_\infty & \text{if } m = 0, \\ \sup_{\pi_1, \dots, \pi_m} \left\| \frac{d(\rho P^{\pi_1} ... P^{\pi_m})}{d\mu} \right\|_\infty & \text{if } m \geq 1. \end{cases}$$

Further assume that

$$\kappa(\rho\|\mu) := (1-\gamma)\sum_{m \geq 0} \gamma^m \kappa_m(\rho\|\mu) < \infty,$$

where $(1-\gamma)$ is a normalization factor as $\sum_{m\geq 0}\gamma^m = (1-\gamma)^{-1}$.

The existence of $\kappa_m(\rho\|\mu), \forall m \geq 0$ requires that the sampling distribution μ has sufficient coverage over $\mathcal{S} \times \mathcal{A}$ which is a common requirement in offline RL, e.g., (Jiang and Li, 2015; Chen and Jiang, 2019). The finite $\kappa(\rho\|\mu)$ assumption is valid 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). Note that Assumption 3.2 is in fact slightly weaker than that in (Munos and Szepesvári, 2008) which we modify to FQE context.

Before characterizing our main finite-sample error bound, we present some relevant results which our main result build upon. To increase the readability, we only briefly state the proof ideas here and defer all detailed proofs to the appendix. The general procedure for establishing our main result is similar to that for the fitted Q-iteration with general function classes (Munos and Szepesvári, 2008; Le et al., 2019). The main difference is that we use deep ReLU networks as function approximation and take into account the network's approximation error for the estimation error. We also employ a different technique using local Rademacher complexity to establish our result.

Proposition 1 (Error propagation through iterations). *Let $\epsilon_{k-1} := Q_k - T^\pi Q_{k-1}, \forall k \in \{1, ..., K\}$ where each Q_k is found at iteration k by Algorithm 1. We have*

$$|v_K - v^\pi| \leq \frac{\sqrt{\kappa(\rho\|\mu)}}{1-\gamma} \max_{0 \leq k \leq K-1} \|\epsilon_k\|_\mu + \frac{2V_{max}\gamma^{K/2}}{(1-\gamma)^{1/2}}.$$

Proposition 1 states that the estimation error $|v_K - v^\pi|$ can be decomposed into the statistical error (the first term) and the algorithmic error (the second term). While the algorithmic error converges to 0 at linear rate, the statistical error reflects the fundamental difficulty of the problem. In particular, the statistical error depends on the concentration coefficient and the

[1] Our analysis can easily apply to the finite action space but we focus on the continuous action space for simplicity.

estimation error for the regression problem at each iteration k. Note that this result is agnostic to function approximation and its similar variants are also established in (Munos and Szepesvári, 2008; Le et al., 2019). We employ similar techniques using the contraction of the induced Markov processes to prove this result.

Proposition 1 allows us to turn our focus on studying the estimation error $\|\epsilon_k\|_\mu$ for the regression problem of each iteration k. Consider a fixed $Q_{k-1} \in \mathcal{F}$. We construct the dataset $D_n = \{(x_i, y_i) : 1 \le i \le n\}$ where $x_i = (s_i, a_i)$ and $y_i = r_i + \gamma \int_{\mathcal{A}} \pi(da|s_i')Q_{k-1}(s_i', a)$. Let $l(x, y) = (x - y)^2$ be the squared loss function. We have $f_* := T^\pi Q_{k-1}$ is the regression function, i.e., $f_* = \arg\inf_f \mathbb{E}[l(f(X), Y)] = \mathbb{E}[Y|X]$. Denote the the empirical risk minimizer over the function class $\mathcal{F} \subseteq \mathcal{B}(\mathcal{X}; V_{\max})$ as

$$\hat{f} := Q_k = \arg\inf_{f \in \mathcal{F}} \sum_{i=1}^{n} l(f(X_i), Y_i) \qquad (1)$$

where Q_k is founded at line 3 in Algorithm 1. Denote $d_{\mathcal{F}} := \inf_{f \in \mathcal{F}} \|f - f_*\|_\mu$ be the approximation error induced by the function class $\mathcal{F}$ w.r.t. the regression function f_*. Let $Pdim(\mathcal{F})$ be the pseudo-dimension of $\mathcal{F}$.

Proposition 2. *If $n \ge Pdim(\mathcal{F})$, for any $\lambda > 0$, with probability at least $1 - e^{-\lambda}$, we have*

$$\|\hat{f} - f_*\|_\mu \le C\Bigg(d_{\mathcal{F}} + \sqrt{\frac{\lambda + \log\log_2(n/Pdim(\mathcal{F}))}{n}} + \sqrt{\frac{Pdim(\mathcal{F})}{n}\log\frac{n\sqrt{n}}{Pdim(\mathcal{F})}}\Bigg),$$

for some universal constant C independent of n.

Proposition 2 states that the estimation error $\|\hat{f} - f_*\|_\mu$ is bounded by the approximation error $d_{\mathcal{F}}$ and the complexity of the general function class $\mathcal{F}$. We leverage local Rademacher complexity (Bartlett et al., 2005) and the localization argument (Farrell et al., 2018) to obtain this proposition. This result is similar to the one established in (Farrell et al., 2018) but the main difference is that we further simplify the analysis and explicitly use the sub-root function argument to obtain the bound.

The next proposition presents a finite-sample error bound on the estimation error $\|\hat{f} - f_*\|_\mu$ when the function class $\mathcal{F}$ is specifically chosen to be the ReLU network functions $\mathcal{F}(d, V_{\max})$ under Assumption 3.1.

Proposition 3. *If $\mathcal{F}$ in Proposition 2 is restricted to the ReLU network functions $\mathcal{F}(d, V_{\max})$, and $f_* \in \mathcal{G}(\beta, d, V_{\max})$ (under Assumption 3.1), with $\hat{f}$ solving Equation (1) for a ReLU network with $L \le (2 + \lceil\log_2(\beta)\rceil)(11 + \beta/d)$ layers and $W_n \asymp n^{\frac{d}{d+2\beta}}$ nonzero, quantized weights, then for any $\lambda \ge 0$, with probability at least $1 - e^{-\lambda}$,*

$$\|\hat{f} - f_*\|_\mu \le C\left(n^{-\frac{\beta}{d+2\beta}}\log n + \sqrt{\frac{\lambda + \log\log n}{n}}\right),$$

for some universal constant C independent of n.

The above result enjoys the rate of convergence $O(n^{\frac{-\beta}{d+2\beta}}\log n)$ which is the minimax-optimal statistical rate of convergence within the class of sufficiently smooth functions $\mathcal{G}(d, V_{\max})$ defined on $[-1/2, 1/2]^d$ (Stone, 1982), thus it cannot be improved further. In addition, this result does not require the number of layers of the ReLU networks to grow with n, as opposed to the result in (Yarotsky, 2017; Farrell et al., 2018) where L grows at $O(\log n)$. The bound in Proposition 3 is also slightly better than that in Corollary 1 of (Farrell et al., 2018) where the first term in our bound depends on $\log n$ instead of $\log^2 n$.

We are now ready to proceed with our main finite-sample error bound.

Theorem 1. *Under Assumption 3.1 and 3.2, for any $\lambda \ge 0$, with probability at least $1 - \exp(-\lambda)$, we have*

$$|v_K - v^\pi| \le \frac{2\gamma^{K/2}V_{\max}}{(1-\gamma)^{1/2}} + \frac{C\sqrt{\kappa(\rho\|\mu)}}{1-\gamma}\left(n^{-\frac{\beta}{d+2\beta}}\log n + \sqrt{\frac{\lambda + \log K + \log\log n}{n}}\right)$$

where C is a universal constant independent of n.

Proof. Theorem 1 is a direct consequence of Proposition 1 and 3 under Assumption 3.1 and 3.2. □

The error bound in Theorem 1 explicitly characterizes how different factors contribute to the estimation error of OPE. In addition, using ReLU network functions, we are able to remove the explicit dependence of the error bound on the *inherent Bellman error* for general function approximation presented in (Le et al., 2019).

4 Discussion

In this paper, we analyze off-policy evaluation using deep ReLU networks. In particular, we establish a finite-sample error bound of the method which appears to be the first and sharpest bound of this kind in the setting. Our analysis highlights the promising role of deep learning in batch reinforcement learning if one can leverage the expressiveness of deep neural networks. A future direction for this work is to take into account the optimization error for the analysis.

References

Anthony, M. and Bartlett, P. L. (2002). *Neural Network Learning - Theoretical Foundations.* Cambridge University Press.

Antos, A., Szepesvári, C., and Munos, R. (2008). Learning near-optimal policies with bellman-residual minimization based fitted policy iteration and a single sample path. *Mach. Learn.*, 71(1):89–129.

Bartlett, P. L., Bousquet, O., and Mendelson, S. (2005). Local rademacher complexities. *Ann. Statist.*, 33(4):1497–1537.

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

Bertsekas, D. P. and Tsitsiklis, J. N. (1995). Neuro-dynamic programming: an overview. In *Proceedings of 1995 34th IEEE Conference on Decision and Control*, volume 1, pages 560–564. IEEE.

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

Duan, Y. and Wang, M. (2020). Minimax-optimal off-policy evaluation with linear function approximation. *CoRR*, abs/2002.09516.

Dudík, M., Langford, J., and Li, L. (2011). Doubly robust policy evaluation and learning. *arXiv preprint arXiv:1103.4601.*

Farajtabar, M., Chow, Y., and Ghavamzadeh, M. (2018). More robust doubly robust off-policy evaluation. *arXiv preprint arXiv:1802.03493.*

Farrell, M. H., Liang, T., and Misra, S. (2018). Deep neural networks for estimation and inference: Application to causal effects and other semiparametric estimands. *arXiv preprint arXiv:1809.09953.*

Grünewälder, S., Lever, G., Baldassarre, L., Pontil, M., and Gretton, A. (2012). Modelling transition dynamics in mdps with RKHS embeddings. In *ICML*. icml.cc / Omnipress.

Harvey, N., Liaw, C., and Mehrabian, A. (2017). Nearly-tight vc-dimension bounds for piecewise linear neural networks. In *Conference on Learning Theory*, pages 1064–1068.

Jiang, N. and Li, L. (2015). Doubly robust off-policy value evaluation for reinforcement learning.

Jiang, N. and Li, L. (2016). Doubly robust off-policy value evaluation for reinforcement learning. In *ICML*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 652–661. JMLR.org.

Jong, N. K. and Stone, P. (2007). Model-based function approximation in reinforcement learning. In *AAMAS*, page 95. IFAAMAS.

Kallus, N. and Uehara, M. (2019). Double reinforcement learning for efficient off-policy evaluation in markov decision processes.

Lagoudakis, M. G. and Parr, R. (2003). Least-squares policy iteration. *J. Mach. Learn. Res.*, 4:1107–1149.

Le, H. M., Voloshin, C., and Yue, Y. (2019). Batch policy learning under constraints. In *ICML*, volume 97 of *Proceedings of Machine Learning Research*, pages 3703–3712. PMLR.

Lei, Y., Ding, L., and Bi, Y. (2016). Local rademacher complexity bounds based on covering numbers. *Neurocomputing*, 218:320–330.

Levine, S., Kumar, A., Tucker, G., and Fu, J. (2020). Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643.*

Liu, Q., Li, L., Tang, Z., and Zhou, D. (2018). Breaking the curse of horizon: Infinite-horizon off-policy estimation. In *NeurIPS*, pages 5361–5371.

Mannor, S., Simester, D., Sun, P., and Tsitsiklis, J. N. (2004). Bias and variance in value function estimation. In *ICML*, volume 69 of *ACM International Conference Proceeding Series.* ACM.

Munos, R. (2003). Error bounds for approximate policy iteration. In *ICML*, pages 560–567. AAAI Press.

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

Nachum, O., Chow, Y., Dai, B., and Li, L. (2019a). Dualdice: Behavior-agnostic estimation of discounted stationary distribution corrections.

Nachum, O., Dai, B., Kostrikov, I., Chow, Y., Li, L., and Schuurmans, D. (2019b). Algaedice: Policy gradient from arbitrary experience. *ArXiv*, abs/1912.02074.

Petersen, P. and Voigtlaender, F. (2018). Optimal approximation of piecewise smooth functions using deep relu neural networks. *Neural Networks*, 108:296–330.

Precup, D., Sutton, R. S., and Singh, S. P. (2000). 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. Morgan Kaufmann Publishers Inc.

Rebeschini, P. (2019). Oxford Algorithmic Foundations of Learning, Lecture Notes: Sub-Gaussian Concentration Inequalities. Bounds in Probability.

URL: http://www.stats.ox.ac.uk/~rebeschi/teaching/AFoL/20/material/lecture06.pdf. Last visited on Sep. 14, 2020.

Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *The annals of statistics*, pages 1040–1053.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement Learning: An Introduction*. A Bradford Book, Cambridge, MA, USA.

Thomas, P. and Brunskill, E. (2016). Data-efficient off-policy policy evaluation for reinforcement learning. In *International Conference on Machine Learning*, pages 2139–2148.

Tosatto, S., Pirotta, M., D'Eramo, C., and Restelli, M. (2017). Boosted fitted q-iteration. In *ICML*, volume 70 of *Proceedings of Machine Learning Research*, pages 3434–3443. PMLR.

Xie, T., Ma, Y., and Wang, Y.-X. (2019). Towards optimal off-policy evaluation for reinforcement learning with marginalized importance sampling.

Yang, Z., Xie, Y., and Wang, Z. (2019). A theoretical analysis of deep q-learning. *CoRR*, abs/1901.00137.

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

Yin, M. and Wang, Y. (2020). Asymptotically efficient off-policy evaluation for tabular reinforcement learning. In *AISTATS*, volume 108 of *Proceedings of Machine Learning Research*, pages 3948–3958. PMLR.

Zhang, R., Dai, B., Li, L., and Schuurmans, D. (2020a). Gendice: Generalized offline estimation of stationary values. *ArXiv*, abs/2002.09072.

Zhang, S., Liu, B., and Whiteson, S. (2020b). Gradientdice: Rethinking generalized offline estimation of stationary values. *ArXiv*, abs/2001.11113.

Appendix A. Proofs

In this appendix, we present the proofs to all the claims stated in our main paper.

A.1. Proof of Proposition 1

Proof. Let $\epsilon_{k-1} := Q_k - T^\pi Q_{k-1}, \forall k \in \{1, ..., K\}$. Since Q^π is the (unique) fixed point of T^π, for all $k \in \{1, ..., K\}$, 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

$$\begin{aligned} Q_K - Q^\pi &= \gamma^K (P^\pi)^K (Q_0 - Q^\pi) + \sum_{k=0}^{K-1} \gamma^k (P^\pi)^k \epsilon_{K-1-k} \\ &= \frac{1-\gamma^{K+1}}{1-\gamma} \left(\frac{(1-\gamma)\gamma^K}{1-\gamma^{K+1}} (P^\pi)^K (Q_0 - Q^\pi) + \sum_{k=0}^{K-1} \frac{(1-\gamma)\gamma^k}{1-\gamma^{K+1}} (P^\pi)^k \epsilon_{K-1-k} \right) \\ &= \frac{1-\gamma^{K+1}}{1-\gamma} \sum_{k=0}^{K} \alpha_k A_k \xi_k, \end{aligned}$$

where

$$\begin{aligned} \alpha_k &:= \frac{(1-\gamma)\gamma^k}{1-\gamma^{K+1}}, \forall k \in \{0, ..., K\}, \\ A_k &:= (P^\pi)^k, \forall k \in \{0, ..., K\}, \\ \xi_k &:= \begin{cases} \epsilon_{K-1-k} & \text{for } k \in \{0, ..., K-1\} \\ Q_0 - Q^\pi & \text{for } k = K. \end{cases} \end{aligned}$$

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:

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

Now we are ready to bound $\|Q_K - Q^\pi\|_\rho^2$. We have

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

$$
\begin{aligned}
&= \frac{1-\gamma^{K+1}}{(1-\gamma)^2}(1-\gamma)\sum_{k=0}^{K-1}\gamma^k\kappa_k(\rho\|\mu)\|\xi_k\|_\mu^2 + \frac{4(1-\gamma^{K+1})V_{max}^2\gamma^K}{1-\gamma} \\
&\leq \frac{1-\gamma^{K+1}}{(1-\gamma)^2}\kappa(\rho\|\mu)\max_{0\leq k\leq K-1}\|\xi_k\|_\mu^2 + \frac{4(1-\gamma^{K+1})V_{max}^2\gamma^K}{1-\gamma} \\
&\leq \frac{1}{(1-\gamma)^2}\kappa(\rho\|\mu)\max_{0\leq k\leq K-1}\|\xi_k\|_\mu^2 + \frac{4V_{max}^2\gamma^K}{1-\gamma}.
\end{aligned}
$$

The inequalities (a) and (b) follow from Jensen's inequality, (c) follows from $Q_0, Q^\pi \in \mathcal{B}(\mathcal{X}; V_{\max})$, and (d) follows from Assumption 3.2 that $\rho A_k = \rho(P^\pi)^k \leq \kappa_k(\rho\|\mu)\mu$. Thus we have

$$
\begin{aligned}
|v_K - v^\pi| &= \Big|\mathbb{E}_\rho[Q_K(s,a)] - \mathbb{E}_\rho[Q^\pi(s,a)]\Big| \\
&\leq \mathbb{E}_\rho\left[|Q_K(s,a) - Q^\pi(s,a)|\right] \\
&\leq \sqrt{\mathbb{E}_\rho\left[(Q_K(s,a) - Q^\pi(s,a))^2\right]} \\
&= \|Q_K - Q^\pi\|_\rho \\
&\leq \frac{1-\gamma^{K+1}}{(1-\gamma)^2}\kappa(\rho\|\mu)\max_{0\leq k\leq K-1}\|\xi_k\|_\mu^2 + \frac{4(1-\gamma^{K+1})V_{max}^2\gamma^K}{1-\gamma} \\
&\leq \frac{1}{1-\gamma}\sqrt{\kappa(\rho\|\mu)}\max_{0\leq k\leq K-1}\|\xi_k\|_\mu + \frac{2V_{max}\gamma^{K/2}}{(1-\gamma)^{1/2}}.
\end{aligned}
$$

□

A.2. Proof of Proposition 2

Notations. Consider a measurable space $(\mathcal{X}, \Sigma)$ where $\mathcal{X} = \mathcal{S} \times \mathcal{A}$ and all the following quantities are defined on this space. Denote $X_i = (s_i, a_i)$ where $\mathcal{X}_1, ..., X_n \overset{i.i.d.}{\sim} \mu$. Let $P_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ be the associated empirical measure. Let $L_2(\mu) := \{f : \mathcal{X} \to \mathbb{R} : \int f(x)^2\mu(dx) \leq \infty\}$ be the class of μ-integrable measurable functions. We denote $\|f\|_\mu^2 = \mathbb{E}[f^2] = \int f(x)^2\mu(dx)$ the norm of the space $L_2(\mu)$. Similarly, we denote $\|f\|_n^2 - \|f\|_{P_n}^2 = \mathbb{E}_n[f^2] = \int f(x)^2 P_n(dx) = \frac{1}{n}\sum_{i=1}^N f(X_i)^2$.

Let $\sigma_1, ..., \sigma_n$ be n independent Rademacher random variables, i.e., $\sigma_i \in \{-1, 1\}$ and $P(\sigma_i = 1) = P(\sigma_i = -1) = 1/2$. For any function class $\mathcal{F} \subseteq L_2(P)$ and any $f \in \mathcal{F}$, denote $R_n f = \frac{1}{n}\sum_{i=1}^n \sigma_i f(X_i)$, and $R_n\mathcal{F} = \sup_{f\in\mathcal{F}} R_n f$. The empirical Rademacher average of $\mathcal{F}$ is $\mathbb{E}_\sigma R_n\mathcal{F}$ and the Rademacher average of $\mathcal{F}$ is $\mathbb{E}R_n\mathcal{F} = \mathbb{E}\mathbb{E}_\sigma R_n\mathcal{F}$ where the leftmost expectation is taken over the random empirical data $\{X_i\}_{i=1}^n$.

Setup. Consider random vectors $Z = (X, Y)$ and n i.i.d. samples $\{(X_i, Y_i)\}$. Assume that $|Y| \leq M := V_{\max}$. Let $l(x, y) = (x - y)^2$ be the squared loss function. Consider a class of uniformly bounded functions $\mathcal{F} \subseteq \mathcal{B}(\mathcal{X}; M) \cap L_2(\mu)$. Denote $l_f(z) = l(f(x), y)$ where $z = (x, y)$. We have

$$
|l_{f_1}(z) - l_{f_2}(z)| = |l(f_1(x), y) - l(f_2(x), y)| \leq L|f_1(x) - f_2(x)|, \forall f_1, f_2 \in \mathcal{F}, \forall x \in \mathcal{X}, |y| \leq M, \tag{2}
$$

where $L := 4M$. Let f^* be the regression function, i.e., $f_* = \arg\inf_f \mathbb{E}\left[l(f(X), Y)\right] = \mathbb{E}[Y|X]$. Then, for any function f, we also have $\|f - f_*\|_\mu^2 = \mathbb{E}[(l_f - l_{f_*})]$. Let $\hat{f} := \arg\inf_{f\in\mathcal{F}} \mathbb{E}_n\left[l(f(X), Y)\right]$ be the empirical risk minimizer over the function class $\mathcal{F}$. We are interested in bounding $\|\hat{f} - f_*\|_\mu$.

To this end, we decompose it as follows:

$$
\|\hat{f} - f_*\|_\mu^2 = \mathbb{E}(l_{\hat{f}} - l_{f_*}) \leq \mathbb{E}(l_{\hat{f}} - l_{f_*}) + \mathbb{E}_n(l_{f_\perp} - l_{\hat{f}}) = (\mathbb{E} - \mathbb{E}_n)(l_{\hat{f}} - l_{f_*}) + \mathbb{E}_n(l_{f_\perp} - l_{f_*}), \tag{3}
$$

where $f_\perp := \arg\inf_{f\in\mathcal{F}} \|f - f_*\|_\mu$ be the projection of the regression function f_* onto the function class $\mathcal{F}$.

Step 1: Bounding the bias term. First, we bound the bias term $\mathbb{E}_n(l_{f_\perp} - l_{f_*})$ using Bernstein's inequality. Let $h = l_{f_\perp} - l_{f_{f_*}}$ and denote by $d_\mathcal{F} := \|f_\perp - f_*\|_\mu$ the projection distance. We have

$$
\begin{aligned}
Var[h] &\leq 16M^2 d_\mathcal{F}^2, \\
h - \mathbb{E}[h] &\leq 8M^2.
\end{aligned}
$$

Thus, for any $\lambda > 0$, with probability at least $1 - e^{-\lambda}$,

$$\mathbb{E}_n(l_{f_\perp} - l_{f_*}) \leq d_{\mathcal{F}}^2 + \frac{8M^2\lambda}{3n} + 4Md_{\mathcal{F}}\sqrt{\frac{2\lambda}{n}} \leq (d_{\mathcal{F}} + 2M\sqrt{\frac{2\lambda}{n}})^2. \tag{4}$$

Step 2: Bounding $(\mathbb{E} - \mathbb{E}_n)(l_f - l_{f_*})$ using local Rademacher complexity and sub-root function.

For any $f \in \mathcal{F}$, we have

$$|l_f - l_{f^*}| \leq L|f - f_*| \leq 8M^2,$$
$$Var[l_f - l_{f^*}] \leq \mathbb{E}[(l_f - l_{f^*})^2] \leq 16M^2\mathbb{E}(f - f_*)^2.$$

For any $r > 0$, any $f \in \mathcal{F}$ such that $\|f - f_*\|_\mu^2 \leq r$, and any $\lambda > 0$, with probability at least $1 - e^{-\lambda}$, we have

$$\begin{aligned}(\mathbb{E} - \mathbb{E}_n)(l_f - l_{f_*}) &\leq 3\mathbb{E}R_n \left\{l_f - l_{f_*} : f \in \mathcal{F}, \|f - f_*\|_\mu^2 \leq r\right\} + 4M\sqrt{\frac{2r\lambda}{n}} + \frac{112M^2\lambda}{3n} \\ &\leq 12M\mathbb{E}R_n \left\{f - f_* : f \in \mathcal{F}, \|f - f_*\|_\mu^2 \leq r\right\} + 4M\sqrt{\frac{2r\lambda}{n}} + \frac{112M^2\lambda}{3n} \end{aligned} \tag{5}$$

where the first inequality follows from Lemma 1.1 with $\alpha = 1/2$ and the second one follows from the contraction of Rademacher complexity.

Step 3. Bounding $\|\hat{f} - f_*\|_\mu$ using a sub-root function and localization argument.

Let ψ be a sub-root function (Definition 3.1 in (Bartlett et al., 2005)) with the fixed point r_* and assume that for any $r \geq r_*$, we have

$$\psi(r) \geq 6M\mathbb{E}R_n\{f - f_* : f \in \mathcal{F}, \|f - f_*\|_\mu^2 \leq r\}. \tag{6}$$

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. Note that $\psi(r) \leq \sqrt{rr_*}, \forall r \geq r_*$. Combining (3), (4), (5), and (6), we have: for any $r \geq r_*$ and any $\lambda \geq 0$, if $\|\hat{f} - f_*\|_\mu \leq r$, with probability at least $1 - 2e^{-\lambda}$,

$$\begin{aligned}\|\hat{f} - f_*\|_\mu^2 &\leq 2\psi(r) + 4M\sqrt{\frac{2r\lambda}{n}} + \frac{112M^2\lambda}{3n} + d_{\mathcal{F}}^2 + \frac{8M^2\lambda}{3n} + 4Md_{\mathcal{F}}\sqrt{\frac{2\lambda}{n}} \\ &\leq 2\sqrt{rr_*} + 4M\sqrt{\frac{2r\lambda}{n}} + \frac{112M^2\lambda}{3n} + (d_{\mathcal{F}} + 2\sqrt{2}\sqrt{\frac{\lambda}{n}})^2\end{aligned}$$

Consider $r_0 \geq r_*$ (to be chosen later) and let

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

where $l = \log_2(\frac{4M^2}{r_0}) \leq \log_2(\frac{4M^2}{r_*})$. We have $B_0 \subseteq B_1 \subseteq ... \subseteq B_l$ and since $\|f - f_*\|_\mu^2 \leq 4M^2, \forall |f|_\infty \leq M$, we have $P(A_l) = 1$. Now fix $\lambda > 0$. If $\|\hat{f} - f_*\|_\mu^2 \leq 2^i r_0$ for some $i \leq l$, then with probability at least $1 - 2e^{-\lambda}$, we have

$$\begin{aligned}\|\hat{f} - f_*\|_\mu^2 &\leq 2\sqrt{2^i r_0 r_*} + 4M\sqrt{\frac{2^{i+1}r_0\lambda}{n}} + \frac{112M\lambda}{3n} + (d_{\mathcal{F}} + 2\sqrt{2}\sqrt{\frac{\lambda}{n}})^2 \\ &\leq 2^{i-1}r_0,\end{aligned}$$

if the following inequalities hold

$$2\sqrt{2^i r_*} + 4M\sqrt{\frac{2^{i+1}\lambda}{n}} \leq \frac{1}{2}2^{i-1}\sqrt{r_0}$$
$$\frac{112M^2\lambda}{3n} + (d_{\mathcal{F}} + 2\sqrt{2}\sqrt{\frac{\lambda}{n}})^2 \leq \frac{1}{2}2^{i-1}r_0.$$

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

$$\begin{aligned}\sqrt{r_0} &= \frac{2}{2^{i-1}}\left(2\sqrt{2^i r_*} + 4M\sqrt{\frac{2^{i+1}\lambda}{n}}\right)\Bigg|_{i=0} + \sqrt{\frac{2}{2^{i-1}}\left(\frac{112M^2\lambda}{3n} + (d_{\mathcal{F}} + 2\sqrt{2}\sqrt{\frac{\lambda}{n}})^2\right)}\Bigg|_{i=0} + \sqrt{r_*} \\ &\leq C_1\left(d_{\mathcal{F}} + \sqrt{\frac{\lambda}{n}} + \sqrt{r_*}\right),\end{aligned}$$

for some constant $C_1 = C_1(M) > 0$ depending only on M.

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

$$\begin{aligned}P(B_0) &= P(B_1) - P(B_1 \cap B_0^c) = P(B_2) - P(B_2 \cap B_1^c) - P(B_1 \cap B_0^c) \\ &= P(B_l) - \sum_{i=0}^{l-1} P(B_{i+1} \cap B_i^c) \\ &\geq 1 - 2le^{-\lambda}.\end{aligned}$$

By changing λ to $\log(2l) + \lambda$, we can equivalently rewrite the above inequality as: for any $\lambda > 0$, with probability at least $1 - e^{-\lambda}$, we have

$$\|\hat{f} - f_*\|_\mu \leq C_1\left(d_{\mathcal{F}} + \sqrt{\frac{\lambda + \log(2l)}{n}} + r_*\right). \tag{7}$$

Step 4: Finding a sub-root bound of a local Rademacher complexity.

It remains to find a sub-root function $\psi(r)$ that satisfies (6) and the fixed point of that sub-root function. The main idea is to bound the local Rademacher complexity $\mathbb{E}R_n\{f - f_* : f \in \mathcal{F}, \|f - f_*\|_\mu^2 \leq r\}$ by its empirical local Rademacher complexity $\mathbb{E}_\sigma R_n\{f - f_* : f \in \mathcal{F}, \|f - f_*\|_n \leq r'\}$ for some r' is a function of r as the latter can then be bounded by a sub-root function represented by covering numbers and pseudo-dimension of $\mathcal{F}$.

Step 4.1: Bounding the empirical local Rademacher complexity. We first bound the empirical local Rademacher complexity $\mathbb{E}_\sigma R_n\{f - f_* : f \in \mathcal{F}, \|f - f_*\|_n \leq r\}$ by covering numbers and pseudo-dimension of $\mathcal{F}$. We denote that

$$\mathcal{F}_* := \{f - f_* : f \in \mathcal{F}\}.$$

It follows from Lemma 5 and Lemma 8 that for any $n > Pdim(\mathcal{F})$ and any $\epsilon > 0$,

$$\log \mathcal{N}(\epsilon, \mathcal{F}_*, L_2(P_n)) = \log \mathcal{N}(\epsilon, \mathcal{F}, L_2(P_n)) \leq \log \mathcal{N}(\epsilon, \mathcal{F}|_{x_1,\dots,x_n}, L_\infty) \leq Pdim(\mathcal{F}) \log(\beta_n \epsilon^{-1}), \tag{8}$$

where we denote $\beta_n = \frac{2eMn}{Pdim(\mathcal{F})} \geq 2eM$. For any $\epsilon > 0$ and $n \geq Pdim(\mathcal{F})$, with $\{\epsilon_k = \frac{\epsilon}{2^k}\}_{k=0}^\infty$, we have

$$\begin{aligned}\mathbb{E}_\sigma R_n\{f \in \mathcal{F}_* - \mathcal{F}_* : \|f\|_n \leq \epsilon\} &\leq 4\sum_{k=1}^N \epsilon_{k-1}\sqrt{\frac{\log \mathcal{N}(\epsilon_k, \mathcal{F}_* - \mathcal{F}_*, L_2(P_n))}{n}} + \epsilon_N \\ &\leq 4\sum_{k=1}^N \epsilon_{k-1}\sqrt{\frac{\log \mathcal{N}(\epsilon_k/2, \mathcal{F}_*, L_2(P_n))}{n}} + \epsilon_N \\ &\leq 4\sqrt{\frac{Pdim(\mathcal{F})}{n}}\sum_{i=1}^N \frac{\epsilon}{2^{k-1}}\sqrt{(k+1)\log 2 + \log(\beta_n \epsilon^{-1})} + \frac{\epsilon}{2^N} \\ &\leq 4\sqrt{\frac{Pdim(\mathcal{F})}{n}}\sum_{i=1}^N \frac{\epsilon}{2^{k-1}}(\sqrt{(k+1)\log 2} + \sqrt{\log(\beta_n \epsilon^{-1})}) + \frac{\epsilon}{2^N} \\ &= 8\epsilon\sqrt{\frac{Pdim(\mathcal{F})}{n}\log(\beta_n/\epsilon)}(1 - \frac{1}{2^N}) + 4\epsilon\sqrt{\frac{Pdim(\mathcal{F})}{n}}\log 2\sum_{k=1}^N \frac{\sqrt{k+1}}{2^{k-1}} + \frac{\epsilon}{2^N}.\end{aligned} \tag{9}$$

Here, the first inequality follows from Lemma 6, the second inequality follows from Lemma 4, the third inequality follows from (8), and the final inequality follows from that $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}, \forall a, b \geq 0$. Now, take $N \to \infty$ both sides of (9) and choose $\epsilon \leq \beta_n/2$, noting that $\sum_{k=1}^{\infty} \frac{\sqrt{k+1}}{2^{k-1}} < \infty$, we have

$$\mathbb{E}_\sigma R_n\{f \in \mathcal{F}_* - \mathcal{F}_* : \|f\|_\mu \leq \epsilon\} \leq C_2 \epsilon \sqrt{\frac{Pdim(\mathcal{F})}{n} \log(\beta_n/\epsilon)},$$

for some constant $C_2 > 0$. Combining the inequality above with Lemma 5, we have

$$\begin{aligned}
\mathbb{E}_\sigma R_n\{f \in \mathcal{F}_* : \|f\|_n^2 \leq r\} &\leq \inf_{\epsilon>0} \left[\mathbb{E}_\sigma R_n\{f \in \mathcal{F}_* - \mathcal{F}_* : \|f\|_\mu \leq \epsilon\} + \sqrt{\frac{2r \log \mathcal{N}(\epsilon/2, \mathcal{F}_*, L_2(P_n))}{n}}\right] \\
&\leq \left[C_2 \epsilon \sqrt{\frac{Pdim(\mathcal{F})}{n} \log(\beta_n/\epsilon)} + \sqrt{2r \frac{Pdim(\mathcal{F})}{n} \log(2\beta_n/\epsilon)}\right]\Bigg|_{\epsilon=1/\sqrt{n}} \\
&\leq C_3 \left[\frac{\sqrt{Pdim(\mathcal{F}) \log \frac{n\sqrt{n}}{Pdim(\mathcal{F})}}}{n} + \sqrt{\frac{r Pdim(\mathcal{F}) \log \frac{n\sqrt{n}}{Pdim(\mathcal{F})}}{n}}\right] \\
&=: \psi_1(r),
\end{aligned}$$

for some constant $C_3 = C_3(M) > 0$ depending only on M and for any $n \geq \max\{Pdim(\mathcal{F}), \left(\frac{Pdim(\mathcal{F})}{eM}\right)^{2/3}\}$.

Step 4.2: Bounding the local Rademacher complexity by the empirical local Rademacher complexity.

Now, we bound $\mathbb{E}_\sigma R_n\{f - f_* : f \in \mathcal{F}, \|f\|_P^2 \leq r\}$ by $\mathbb{E}_\sigma R_n\{f \in \mathcal{F}_* : \|f\|_n^2 \leq r'\}$ for some $r' = r'(r)$. For $g = (f - f_*)^2 = (f - f_*)(f + f_*)$ for $f \in \mathcal{F}$, we have $Var[g] \leq \mathbb{E}[g^2] \leq 4M^2 \mathbb{E}(f - f_*)^2$ and $|g| \leq 4M^2$. Note that since $\psi(r)$ is a sub-root function, we have $\psi(r) \leq r, \forall r \geq r_*$. Thus, consider $r \geq r_*$, it follows from Lemma 1.1 that with probability at least $1 - \frac{1}{n}$, for any $f \in \mathcal{F}$ such that $\|f - f_*\|_\mu^2 \leq r$, we have

$$\begin{aligned}
\|f - f_*\|_n^2 &\leq \|f - f_*\|_\mu^2 + 3\mathbb{E} R_n\{(f - f_*)^2 : \|f - f_*\|_\mu^2 \leq r\} + 2M\sqrt{\frac{2r \log n}{n}} + \frac{56}{3}\frac{\log n}{n} \\
&\leq \|f - f_*\|_\mu^2 + 6M\mathbb{E} R_n\{f - f_*^2 : \|f - f_*\|_\mu^2 \leq r\} + 2M\sqrt{\frac{2r \log n}{n}} + \frac{56}{3}\frac{\log n}{n} \\
&\leq r + \psi(r) + r + r \\
&\leq 4r,
\end{aligned}$$

if r satisfies that

$$\begin{cases}
r & \geq r_* \\
\sqrt{r} & \geq 2M\sqrt{\frac{2 \log n}{n}} \\
r & \geq \frac{56}{3}\frac{\log n}{n}.
\end{cases}$$

Thus, for any $r \geq r_* \vee 8M^2 \frac{logn}{n} \vee \frac{56}{3}\frac{logn}{n}$, denoting $E_r = \{\|f - f_*\|_n^2 \leq 4r\} \cap \{\|f - f_*\|_P^2 \leq r\}$, we have

$$\begin{aligned}
6M\mathbb{E} R_n\{f - f_* : f \in \mathcal{F}, \|f - f\|_\mu^2 \leq r\} &= 6M\mathbb{E}\mathbb{E}_\sigma R_n\{f - f_* : f \in \mathcal{F}, \|f - f\|_\mu^2 \leq r\} \\
&\leq 6M\mathbb{E}\left[1_E \mathbb{E}_\sigma R_n\{f - f_* : f \in \mathcal{F}, \|f - f\|_\mu^2 \leq r\} + 2M(1 - 1_E)\right] \\
&\leq 6M\mathbb{E}\left[\mathbb{E}_\sigma R_n\{f - f_* : f \in \mathcal{F}, \|f - f\|_n^2 \leq 4r\} + 2M(1 - 1_E)\right] \\
&\leq 6M(\psi_1(4r) + \frac{2M}{n}) \\
&= 6MC_3\left[\frac{\sqrt{Pdim(\mathcal{F}) \log \frac{n\sqrt{n}}{Pdim(\mathcal{F})}}}{n} + \sqrt{\frac{4r Pdim(\mathcal{F}) \log \frac{n\sqrt{n}}{Pdim(\mathcal{F})}}{n}}\right] + \frac{12M^2}{n} \\
&\leq C_4\left[\frac{\sqrt{Pdim(\mathcal{F}) \log \frac{n\sqrt{n}}{Pdim(\mathcal{F})}}}{n} + \sqrt{\frac{r Pdim(\mathcal{F}) \log \frac{n\sqrt{n}}{Pdim(\mathcal{F})}}{n}}\right] \\
&=: \psi(r),
\end{aligned}$$

for some constant $C_4 = C_4(M)$ depending only on M. It is easy to check that $\psi(r)$ is a sub-root function. The fixed point r_* of ψ, i.e., $r_* = \psi(r_*)$ can be solved analytically as $r_* = \psi(r_*)$ is a simple quadratic equation. Solving that yields

$$C_4 \sqrt{\frac{Pdim(\mathcal{F})}{n}} \leq \sqrt{r_*} \leq C_5 \sqrt{\frac{Pdim(\mathcal{F})}{n} \log \frac{n\sqrt{n}}{Pdim(\mathcal{F})}}, \tag{10}$$

for some constant $C_5 > 0$. Now, combining (7) and (10), we have, for any $\lambda > 0$, with probability at least $1 - e^{-\lambda}$,

$$\|\hat{f} - f_*\|_P \leq C_6 \left(d_{\mathcal{F}} + \sqrt{\frac{\lambda + \log \log_2 (n/Pdim(\mathcal{F}))}{n}} + \sqrt{\frac{Pdim(\mathcal{F})}{n} \log \frac{n\sqrt{n}}{Pdim(\mathcal{F})}} \right),$$

for some constant $C_6 > 0$.

A.3. Proof of Proposition 3

The result of Proposition 3 is constructed from Proposition 2, Lemma 9 and Lemma 10. Using the same notations as in the proof of Proposition 2, when $\mathcal{F}$ is restricted to $\mathcal{F}(d, M)$ and $f_* \in \mathcal{G}(\beta, d, M)$, it follows from Lemma 10 that for any $\epsilon_n \in (0, 1/2)$, there is a neural network Φ with $L \leq (2 + \lceil \log_2 \beta \rceil)(11 + \beta/d)$ layers and $W_n \leq c\epsilon_n^{-d/\beta}$ nonzero, quantized weights such that

$$\|f_\Phi - f_*\|_\mu \leq \epsilon_n.$$

Thus, we have

$$d_{\mathcal{F}(d,M)} = \|f_\perp - f_*\|_\mu \leq \|f_\Phi - f_*\|_\mu \leq \epsilon_n.$$

Note that the upper bound in Proposition 2 can be further simplified into

$$C \left(\epsilon_n + \sqrt{\frac{\lambda + \log \log n}{n}} + \sqrt{\frac{W_n L \log W_n}{n} \log n} \right),$$

with some universal constant $C > 0$ independent of n. It now remains to find optimal ϵ_n to minimize the upper bound above. It is easy to see that such optimal value is $\epsilon_n \asymp n^{\frac{-\beta}{d+2\beta}}$. At this value, the upper bound above becomes

$$C \left(n^{-\frac{\beta}{d+2\beta}} \log n + \sqrt{\frac{\lambda + \log \log n}{n}} \right),$$

with $W_n \asymp n^{\frac{d}{d+2\beta}}$. This conludes our proof.

Appendix B. Supporting lemmas

Lemma 1 ((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 2 (**Bernstein's inequality** (Rebeschini, 2019))**.** *Let $X_1, ..., X_n \sim X$ be i.i.d. real-valued random variables that satisfy the one-sided Bernstein's condition with parameter $b > 0$. Then, for any $\epsilon > 0$ and $\delta \in [0,1]$, we have*

$$P\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}X < \frac{b}{n}\log(1/\delta) + \sqrt{\frac{2(VarX)\log(1/\delta)}{n}}\right) \geq 1-\delta.$$

Lemma 3 (**Contraction property**)**.** *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 4. *Denote $\mathcal{F} - \mathcal{F} := \{f - g : f, g \in \mathcal{F}\}$. We have*

$$\log \mathcal{N}(\epsilon, \mathcal{F} - \mathcal{F}, L_2(P_n)) \leq 2\log \mathcal{N}(\epsilon/2, \mathcal{F}, L_2(P_n)).$$

Lemma 5 (Lemma 1 in (Lei et al., 2016))**.** *Let $\mathcal{F}$ be a function class and P_n be the empirical measure supported on $X_1, ..., X_n$, then for any $r > 0$ (which can be stochastic w.r.t X_i), we have*

$$\mathbb{E}_\sigma R_n\{f \in \mathcal{F} : \|f\|^2_{L_2(P_n)} \leq r\} \leq \inf_{\epsilon>0}\left[\mathbb{E}_\sigma R_n\{f \in \mathcal{F} - \mathcal{F} : \|f\|_{L_2(P)} \leq \epsilon\} + \sqrt{\frac{2r\log \mathcal{N}(\epsilon/2, \mathcal{F}, L_2(P_n))}{n}}\right]$$

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

$$\mathbb{E}_\sigma R_n\{f \in \mathcal{F} : \|f\|_{L_2(P_n)} \leq \epsilon_0\} \leq 4\sum_{k=1}^{N}\epsilon_{k-1}\sqrt{\frac{\log \mathcal{N}(\epsilon_k, \mathcal{F}, L_2(P_n))}{n}} + \epsilon_N.$$

Lemma 7. *On a probability space $(\mathcal{X}, P)$, for any class of measurable functions $\mathcal{F} \subseteq \{\mathcal{X} \to \mathbb{R}\}$ and any $\{x_1, ..., x_n\} \subset \mathcal{X}$ with the associated empirical measure P_n, we have*

$$\mathcal{N}(\epsilon, \mathcal{F}, L_2(P_n)) \leq \mathcal{N}(\epsilon, \mathcal{F}|_{x_1,...,x_n}, L_\infty), \forall \epsilon > 0.$$

Lemma 8 (Theorem 12.2 in (Anthony and Bartlett, 2002))**.** *Let $\mathcal{F} \subseteq \{f : \mathcal{X} \to [-M, M]\}$. Denote by $Pdim(\mathcal{F})$ the pseudo-dimension of $\mathcal{F}$. For $n \geq Pdim(\mathcal{F})$ and any $\epsilon > 0$, we have*

$$\sup\{\mathcal{N}(\epsilon, \mathcal{F}|_{x_1,...,x_n}, L_\infty) : x_i \in \mathcal{X}\} \leq \left(\frac{2eMn}{\epsilon Pdim(\mathcal{F})}\right)^{Pdim(\mathcal{F})}.$$

Lemma 9 (Theorem 6 in (Harvey et al., 2017))**.** *Let $\mathcal{F}$ be a ReLU network architecture with W non-zero weights and L layers, then the VC-dimension and pseudo-dimension satisfy*

$$cWL\log(W/L) \leq VCdim(\mathcal{F}) \leq CWL\log(W),$$

with some universal constants $c, C > 0$. The same result also holds for $Pdim(\mathcal{F})$.

Lemma 10 (Theorem 3.1 in (Petersen and Voigtlaender, 2018))**.** *For any $d \in \mathbb{N}$ and $\beta, B, p > 0$, there exists a constant $s = s(d, \beta, B, p) \in \mathbb{N}$ and $c = c(d, \beta, B) > 0$ such that for any function $f_* \in \mathcal{G}(\beta, d, B)$ and any $\epsilon \in (0, 1/2)$, there is a neural network Φ_ϵ with $L \leq (2 + \lceil\log_2 \beta\rceil)(11 + \beta/d)$ layers and $W \leq c\epsilon^{-d/\beta}$ nonzero, (s, ϵ)-quantized weights such that*

$$\|f_\Phi - f_*\|_{L_p([-1/2,1/2]^d)} \leq \epsilon \text{ and } \|f_\Phi\|_\infty \leq \lceil B\rceil.$$