Neural Network Approximation for Pessimistic Offline Reinforcement Learning

Di Wu[1], **Yuling Jiao**[1, 2], **Li Shen**[3], **Haizhao Yang**[4*], **Xiliang Lu**[1, 2 *]

[1] School of Mathematics and Statistics, Wuhan University, China,
[2] Hubei Key Laboratory of Computational Science, Wuhan University, China,
[3] JD Explore Academy, China,
[4] Department of Mathematics and Department of Computer Science, University of Maryland, College Park, USA,
{dwu.math,yulingjiaomath,xllv.math}@whu.edu.cn, mathshenli@gmail.com, hzyang@umd.edu

Abstract

Deep reinforcement learning (RL) has shown remarkable success in specific offline decision-making scenarios, yet its theoretical guarantees are still under development. Existing works on offline RL theory primarily emphasize a few trivial settings, such as linear MDP or general function approximation with strong assumptions and independent data, which lack guidance for practical use. The coupling of deep learning and Bellman residuals makes this problem challenging, in addition to the difficulty of data dependence. In this paper, we establish a non-asymptotic estimation error of pessimistic offline RL using general neural network approximation with $\mathcal{C}$-mixing data regarding the structure of networks, the dimension of datasets, and the concentrability of data coverage, under mild assumptions. Our result shows that the estimation error consists of two parts: the first converges to zero at a desired rate on the sample size with partially controllable concentrability, and the second becomes negligible if the residual constraint is tight. This result demonstrates the explicit efficiency of deep adversarial offline RL frameworks. We utilize the empirical process tool for $\mathcal{C}$-mixing sequences and the neural network approximation theory for the Hölder class to achieve this. We also develop methods to bound the Bellman estimation error caused by function approximation with empirical Bellman constraint perturbations. Additionally, we present a result that lessens the curse of dimensionality using data with low intrinsic dimensionality and function classes with low complexity. Our estimation provides valuable insights into the development of deep offline RL and guidance for algorithm model design.

Existing Works	Assumption	
Szepesvári and Munos (2005); Munos (2007) Antos, Szepesvári, and Munos (2007, 2008) Farahmand, Szepesvári, and Munos (2010) Scherrer (2014); Liu et al. (2019) Chen and Jiang (2019); Jiang (2019) Wang et al. (2019); Feng, Li, and Liu (2019) Liao et al. (2022); Zhang et al. (2020) Uehara, Huang, and Jiang (2020) Xie and Jiang (2021)	\	F
Nguyen-Tang et al. (2022a)	NN	
Rashidinejad et al. (2021) Yin, Bai, and Wang (2021) Shi et al. (2022b); Li et al. (2022)	Tabular	P
Jin, Yang, and Wang (2021); Chang et al. (2021) Zhang et al. (2022); Nguyen-Tang et al. (2022b) Bai et al. (2022)	Linear	
Jiang and Huang (2020)	Compact	
Zhan et al. (2022)	Convex	
Uehara, Huang, and Jiang (2020) Rashidinejad et al. (2022) Zanette and Wainwright (2022) Xie et al. (2021); Cheng et al. (2022)	Finite	
Ji et al. (2023), **Our work**	NN	

Table 1: A comparison concerning assumptions related to data coverage and approximation. F (Full), P (Partial).

1 Introduction

Online RL has demonstrated significant empirical success in specific decision-making problems (Mnih et al. 2015; Silver et al. 2016). However, numerous real-world environments only allow limited interaction, presenting challenges in cost (e.g., robotics) or safety (e.g., autonomous driving). Therefore, offline RL was introduced as a paradigm that enables learning decision-making problems from pre-collected data without additional interaction (Lange, Gabel, and Riedmiller 2012; Levine et al. 2020). The primary objective of offline RL is to leverage the available data to learn near-optimal policies even when the data is insufficiently collected.

Distribution shift is a significant challenge that offline RL faces due to inconsistency between the data used for training and the data induced by the learned policy. One way to overcome the distribution shift challenge is to enforce a policy constraint (Fujimoto, Meger, and Precup 2019; Kumar et al. 2019). However, this approach can be overly conservative and is highly dependent on the accuracy of estimating the behavior policy. To avoid overestimation and focus on reliable data distribution, several studies (Kumar et al. 2020; Xie et al. 2021; Cheng et al. 2022) have modified the Q-value function, enabling the selection of actions in a pessimistic manner. These two approaches are known as regularized policy-based and pessimistic value-based methods, respectively. The efficacy of these methods has been verified in complex offline RL environments (Fu et al. 2020).

*Correspondence to Xiliang Lu or Haizhao Yang.

The empirical success of recent studies far surpasses the theory for offline RL. Early works (Szepesvári and Munos 2005; Munos 2007; Antos, Szepesvári, and Munos 2007, 2008) assume data to be fully covered, which is unrealistic. More recent studies have relaxed this assumption to partial coverage, focusing mainly on tabular and linear function approximations (Jin, Yang, and Wang 2021; Chang et al. 2021; Zhang et al. 2022; Nguyen-Tang et al. 2022b; Bai et al. 2022; Rashidinejad et al. 2021; Yin, Bai, and Wang 2021; Shi et al. 2022b; Li et al. 2022). General function approximations have been investigated in Jiang and Huang (2020); Uehara and Sun (2021); Zhan et al. (2022); Rashidinejad et al. (2022); Zanette and Wainwright (2022); Xie et al. (2021); Cheng et al. (2022), but they still rely on additional assumptions such as finiteness and convexity. See Table 1 for a comparison to prior works. Practical applications adopt deep neural network parameterization, which is highly non-convex. Moreover, the approximation of policy function is ignored even in the actor-critic framework (Xie et al. 2021; Cheng et al. 2022). Another challenge for offline RL theory is that data is sequentially dependent. In contrast, current works assume offline data to be independent and identically distributed (i.i.d.), leading to a statistical deviation not aligning with reality. In summary, existing offline RL theory faces three main challenges: overly strong assumptions regarding the value function, the inadequate inclusion of policy function approximation, and the neglect of data dependence.

This paper investigates the performance of a deep adversarial offline RL framework that uses deep neural networks to parameterize both the value and policy functions while assuming the data to be dependent. Our result indicates the estimation error consists of two parts: the first converges to zero at a desired rate on the sample size, and the second decreases if the residual constraint is tight. We demonstrate that the estimation rate depends explicitly on the network structure, the tightness of the Bellman constraint, and partially controllable data coverage. However, the curse of dimensionality presents a challenge to the result, mainly when data dimensions are enormous. To alleviate this, we propose using low-dimensional data structures or low-complexity target functions consistent with real-world scenarios.

This paper marks the first attempt to bridge the gap between theory and practice by providing an analysis of deep pessimistic offline RL. Main contributions are as follows:

- We reassess offline RL methods and present an adversarial framework that utilizes deep neural networks to parameterize both policy and value functions, where the offline data is sequentially dependent with only partial coverage.
- We establish a non-asymptotic rate for the estimation error of deep adversarial offline RL regarding the width and depth of networks, dataset dimensions, and the concentrability of distribution shift. Our results are derived under mild assumptions and explicitly illustrate how the choices of neural network structure and algorithm setting influence the efficiency of deep offline RL.
- We mitigate the curse of dimensionality by utilizing low-dimensional data structures or low-complexity target functions, providing an idealized guarantee for real-world data.

Technical Contribution. Our technical challenge stems from multiple practical considerations, which we clarify from four distinct perspectives. (1) Confining adversarial terms presents a non-trivial task, particularly when optimized within different constrained sets. We introduce an operator perturbation analysis, which holds under L_p norm with distribution shifts. (2) Part of generalization errors is rooted in constraints rather than explicit functions. We leverage a generalized version of performance difference to disentangle them from constraints. (3) Due to dependent data, parameterized value and policy functions, the induced space is complex. We tackle this by using uniform covering numbers (focusing on samples, not the entire class), ghost point analysis, and an extended Bernstein inequality. (4) We alleviate the curse of dimensionality with two considerations. Firstly, Minkowski dimension can measure the dimensionality of highly irregular sets like fractals, where we employ Whitney's extension theorem to establish a novel bound. Secondly, our work is the first theoretical attempt to consider low complexity in RL, where we develop perturbation and recursion analysis with additional connecting layers.

2 Related Works

Offline RL. Recent advances in offline RL can be categorized into two groups. The first group, called policy-based regularization, involves directly constraining the learned policy to be similar to the behavior policy (Fujimoto, Meger, and Precup 2019; Laroche, Trichelair, and Des Combes 2019; Kumar et al. 2019; Siegel et al. 2020). However, these methods suffer from two issues: (a) they may be overly conservative, similar to behavior cloning, and (b) estimating the behavior policy can be challenging. Instead of directly constraining the policy, the second group modifies the learning objective to avoid overestimating the value function (Kumar et al. 2020; Liu et al. 2020; Jin, Yang, and Wang 2021; Kostrikov, Nair, and Levine 2021; Uehara and Sun 2021; Xie et al. 2021; Cheng et al. 2022; Rigter, Lacerda, and Hawes 2022; Bhardwaj et al. 2023). For instance, Kumar et al. (2020) and Kostrikov, Nair, and Levine (2021) utilize a conservative approach to optimize the lower bound of the value function, ensuring safe improvement. On the other hand, Xie et al. (2021); Cheng et al. (2022) introduce a bilevel scheme to emphasize Bellman-consistent pessimism, while Rigter, Lacerda, and Hawes (2022); Bhardwaj et al. (2023) employ an adversarial MDP model to minimize policy performance.

Although recent methods in offline RL have demonstrated impressive empirical results, theoretical foundations are not well understood. Early studies of offline RL theory are analyzed with strong assumptions, such as full data coverage (Szepesvári and Munos 2005; Munos 2007; Antos, Szepesvári, and Munos 2007, 2008; Farahmand, Szepesvári, and Munos 2010; Scherrer 2014; Liu et al. 2019; Chen and Jiang 2019; Jiang 2019; Wang et al. 2019; Feng, Li, and Liu 2019; Liao et al. 2022; Zhang et al. 2020; Uehara, Huang, and Jiang 2020; Xie and Jiang 2021). Recent analyses relax this assumption to partial coverage. Rashidinejad et al. (2021); Yin, Bai, and Wang (2021); Shi et al. (2022b); Li et al. (2022) have studied the tabular MDP, while Jin, Yang,

Work	Method	Curse of Dim?
Nguyen-Tang et al. (2022a)	OPE/OPL	Exist
Ji et al. (2023)	OPE	Riemannian $\mathcal{M}$
This work	Adversarial	Minkowski Dim Low Complexity

Table 2: A comparison of works with NN approximation. OPE/OPL (off-policy evaluation/learning), $\mathcal{M}$ (Manifold).

and Wang (2021); Chang et al. (2021); Zhang et al. (2022); Nguyen-Tang et al. (2022b); Bai et al. (2022) explored the linear MDP. General function approximation has been studied in (Jiang and Huang 2020; Uehara and Sun 2021; Zhan et al. 2022; Rashidinejad et al. 2022; Zanette and Wainwright 2022; Xie et al. 2021; Cheng et al. 2022). Specifically, Jiang and Huang (2020) assume the value function class to be compact and explore the convex hull. Zhan et al. (2022) assume the value function to be strongly convex, while Uehara, Huang, and Jiang (2020); Rashidinejad et al. (2022); Zanette and Wainwright (2022); Xie et al. (2021) and Cheng et al. (2022) assume the function class to be finite. However, both value and policy functions are nonconvex and infinite in reality, which is the main concern of this study. Furthermore, most of theoretical works are analyzed in the i.i.d. setting, which does not reflect data dependence.

Approximation and Generalization in Deep Learning. Extensive research has examined the estimation error of deep learning (DL) and how it guides the training process. This error typically comprises two components: approximation, which evaluates the expressive power of deep neural networks for specific general functions, and generalization, which measures the deviation between finite data samples and the expectation. The approximation theory of deep learning has been studied for continuous functions (Shen, Yang, and Zhang 2019; Yarotsky 2021; Shen, Yang, and Zhang 2021) and smooth functions (Yarotsky 2017, 2018; Suzuki 2018; Lu et al. 2021; Suzuki and Nitanda 2021; Jiao, Wang, and Yang 2023). Meanwhile, the generalization theory of deep learning has been extensively explored in the context of i.i.d. data (Anthony and Bartlett 1999; Schmidt-Hieber 2020; Nakada and Imaizumi 2020; Bauer and Kohler 2019; Farrell, Liang, and Misra 2021; Jiao et al. 2023). For dependent data, statistical techniques have been the focus of research in Yu (1994); Antos, Szepesvári, and Munos (2008); Hang and Steinwart (2017); Steinwart, Hush, and Scovel (2009); Mohri and Rostamizadeh (2008, 2010); Ralaivola and Amini (2015); Roy, Balasubramanian, and Erdogdu (2021). This paper presents the first analysis in the context of pessimistic offline RL problems with dependent data and deep neural network approximation.

Additional Related Works. Recently, two studies have offered theoretical insights into deep offline reinforcement learning. Nguyen-Tang et al. (2022a) analyze the sample complexity associated with offline policy evaluation and optimization using a deep ReLU neural network approximation. However, their findings are afflicted by the curse of dimensionality and necessitate complete data coverage. Ji et al. (2023) investigate the estimation error of the fitted Q-evaluation method employing convolutional neural networks. They introduce a novel concentrability metric and mitigate the curse of dimensionality by conceptualizing the data space as a low-dimensional Riemannian manifold.

While these two studies concentrate on the variant of the fitted Q-iteration under i.i.d. setting, our work focuses on the pessimistic approach with sequentially dependent data, necessitating a more intricate analysis due to the inclusion of uncertainty quantifiers and adversarial strategies. Furthermore, we address the curse of dimensionality by utilizing a general measure of dimensionality and target functions possessing low complexity. See Table 2 for a clear comparison.

In addition, several recent works consider RL allowing time-dependence (Zou, Xu, and Liang 2019; Kallus and Uehara 2022; Shi et al. 2022a), and developing pessimistic-type algorithms (Lyu et al. 2022; Zhou et al. 2023). However, these works focus on less practical assumptions or empirical performance, which are quite different from our concerns.

3 Preliminaries

Reinforcement Learning

Markov Decision Processes (MDPs). This work considers a discounted MDP, defined with a tuple $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$, where $\mathcal{S}$ is the set of states, $\mathcal{A}$ is the set of actions, and $\gamma \in (0,1)$ is the discount factor. $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the Markov transition kernel, and $R : \mathcal{S} \times \mathcal{A} \to \Delta(\mathbb{R})$ is the immediate reward. Given a specific pair $(s,a) \in \mathcal{S} \times \mathcal{A}$, $P(\cdot|s,a)$ refers to the probability distribution of the next state, and $R(\cdot|s,a)$ refers to the probability distribution of the immediate reward. For regularity, the reward is assumed to be bounded by $R_{\max}$, and the MDP starts at the initial state s_0. A policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ is used to decide which action to take, and accordingly, a sequence is obtained as

$$a_t \sim \pi(\cdot|s_t), r_t \sim R(\cdot|s_t, a_t), s_{t+1} \sim P(\cdot|s_t, a_t).$$

RL aims to find the optimal policy π^* maximizing the value function $V^\pi(s)$, the expected cumulative discounted reward starting from s, i.e., $V^\pi(s) = \mathbb{E}\left[\sum_{t=0}^\infty \gamma^t r_t | s_0 = s\right]$. Similarly, we define the action-value function $Q^\pi(s,a) = \mathbb{E}\left[\sum_{t=0}^\infty \gamma^t r_t | s_0 = s, a_0 = a\right]$ starting from s, taking action a and then following policy π. The boundedness of rewards guarantees $V^\pi(s)$ and $Q^\pi(s,a)$ are both in $[0, R_{\max}/(1-\gamma)]$. We define the Bellman operator as follows:

$$\mathcal{T}^\pi Q(s,a) = \mathbb{E}[R(s,a)] + \gamma P^\pi Q(s,a),$$

where $P^\pi Q(s,a) := \int P(ds'|s,a)\pi(da'|s)Q(s',a')$. It has been proven that the Bellman operator is contractive concerning the sup-norm (Sutton and Barto 2018), i.e.,

$$\|\mathcal{T}^\pi Q_1 - \mathcal{T}^\pi Q_2\|_\infty \le \gamma \|Q_1 - Q_2\|_\infty$$

for any two action-value functions Q_1 and Q_2. This property guarantees the existence of a fixed point Q^π with respect to $\mathcal{T}^\pi$, which induces the value iteration algorithm.

Offline RL. Offline RL aims to learn an optimal policy using a given dataset without interacting with the environment. The fixed dataset $\mathcal{D}$ consists of tuples s, a, r, s' with s and a sampled from the state-action distribution of a behavior policy μ, r and s' induced by the environment. For any policy

π, we define the marginal state-action occupancy measure as ρ^π. We also denote $\mu = \rho^\mu$, with a slight abuse of notation.

Definition 3.1 (Concentrability Coefficient)**.** *Let μ be the behavior policy and π be a comparator policy; define the density ratio based concentrability coefficient as follows:*

$$\mathcal{C}(\pi; \mu) := \sup_{(s,a)} \frac{\rho^\pi(s,a)}{\mu(s,a)}.$$

This definition of concentrability is widely used in literature (Szepesvári and Munos 2005; Munos 2007; Chen and Jiang 2019; Xie and Jiang 2020), and Chen and Jiang (2019) also offer rich practical insights, indicating the presence of low concentrability. The definition of the concentrability coefficient varies in a few kinds of literature, such as full coverage in (Szepesvári and Munos 2005), Bellman residual-based perspective in (Xie et al. 2021; Cheng et al. 2022), and χ^2-divergence in (Ji et al. 2023). Our result could potentially be extended to a tighter metric, e.g., Bellman residual-based, involving the separation of on/off support parts. Nonetheless, this distinction does not fall within the primary scope of our study. For a comprehensive review of concentrability, refer to Uehara and Sun (2021) and the references therein.

This work assumes only partial coverage within several particular policies, which we will explain in Section 4. We extensively utilize $\mathcal{C}(\Pi; \mu) := \sup_{\pi \in \Pi} \mathcal{C}(\pi; \mu)$ to represent the concentrability of a set of policies with respect to μ.

Feed-Forward Deep Neural Networks

In this work, our primary focus is on the multi-layer feed-forward neural network (FNN) activated by the rectified linear unit (ReLU) function $\sigma(x) = \max\{0, x\}$ with $x \in \mathbb{R}^d$:

$$\begin{aligned} f_0(x) &= x, \\ f_\ell(x) &= \sigma\left(W_\ell f_{\ell-1}(x) + b_\ell\right), \quad \ell = 1, \ldots, L-1, \\ f(x) &= f_L(x) = W_L f_{L-1}(x) + b_L. \end{aligned}$$

Here $W_\ell \in \mathbb{R}^{n_\ell \times n_{\ell-1}}, n_0 = d$ and $b_\ell \in \mathbb{R}^{n_\ell}$ are the weight parameters of the ℓ layer. The activation function σ is applied entry-wise. A network with width $\mathcal{W}$ and depth $\mathcal{L}$ means $\mathcal{W} = \max\{n_\ell, \ell = 0, \ldots, L\}, \mathcal{L} = L - 1$. That is, the maximum width of the hidden layers does not exceed $\mathcal{W}$, and the number of the hidden layers does not exceed $\mathcal{L}$. The weight parameters consist of $W_\ell, b_\ell, \ell = 0, \ldots, L$, and we denote the total number of parameters as $\mathcal{P}$. For simplicity, we may use $\mathcal{NN}$ to denote ReLU FNNs in this work.

Notations in this paper are summarized in Appendix A.

4 Main Results

Adversarial Offline RL

Adversarial offline RL has been extensively studied in literature (Kumar et al. 2020; Xie et al. 2021; Cheng et al. 2022; Rigter, Lacerda, and Hawes 2022; Bhardwaj et al. 2023). We formulate the framework with relative pessimism (Cheng et al. 2022) as a maximization-minimization problem:

$$\widehat{\pi}^* \in \arg\max_{\pi \in \Pi} \min_{f \in \mathcal{F}_\mu^{\pi,\epsilon}} \mathcal{L}_\mu(\pi, f), \tag{1}$$

with $\mathcal{L}_\mu(\pi, f) := \mathbb{E}_\mu[f(s,\pi) - f(s,a)]$, $\mathcal{F}_\mu^{\pi,\epsilon} := \{f \in \mathcal{F} \mid \mathcal{E}_\mu(\pi, f) \le \epsilon\}$ with $\mathcal{E}_\mu(\pi, f) := \|f - \mathcal{T}^\pi f\|_{2,\mu}^2$. Here, $\mathcal{F}$ is the set of functions $f : \mathcal{S} \times \mathcal{A} \to [0, V_{\max}]$ and Π is the policy function class. Practically, this constrained pessimism framework is implemented by adversarial regularized algorithms as introduced in Bhardwaj et al. (2023), to approximately address a specific sub-question. These algorithms have exhibited competent performance across diverse offline scenarios owing to the robust improvement over uncertainty. In this study, we focus on the essential max-min problem, leaving the algorithm analysis as future directions.

The population-level problem (1) is intractable because the oracle distribution is not accessible. To solve the optimization problem (1), we propose an empirical scheme that can be used for computation with neural network approximation. First, we define an estimated Bellman error:

$$\begin{aligned} \mathcal{E}_D(\pi, f) :=& \mathbb{E}_\mathcal{D}\left[(f(s,a) - r - \gamma f(s', \pi))^2\right] \\ &- \min_{f' \in \mathcal{F}} \mathbb{E}_\mathcal{D}\left[(f'(s,a) - r - \gamma f(s', \pi))\right], \end{aligned}$$

which is shown to be an unbiased estimation of $\mathcal{E}_\mu(\pi, f)$ in Antos, Szepesvári, and Munos (2008). Consider Π_θ as the set of parameterized policies $\{\pi_\theta \mid \theta \in \Theta \subseteq \mathbb{R}^d\}$. Actions are selected based on the likelihood derived from the probability density function (PDF) of the policy distribution. This PDF is approximated through a ReLU FNN in $\mathcal{NN}_1$. Importantly, the approximation need not strictly adhere to being a specific density. This flexibility arises from the possibility of drawing samples directly from the density, for instance, using kernel density estimation (Rosenblatt 1956b). The following equation gives the computation scheme we consider:

$$\widehat{\pi} = \arg\max_{\pi \in \Pi_\theta} \min_{f \in \mathcal{NN}_2 \cap \mathcal{F}_\mathcal{D}^{\pi,\epsilon}} \mathcal{L}_\mathcal{D}(\pi, f) \tag{2}$$

where $\mathcal{L}_\mathcal{D}(\pi, f) := \mathbb{E}_\mathcal{D}[f(s,\pi) - f(s,a)]$, $\mathcal{F}_\mathcal{D}^{\pi,\epsilon} := \{f \in \mathcal{F} \mid \mathcal{E}_\mathcal{D}(\pi, f) \le \epsilon\}$, and $\mathcal{NN}_2$ refer to a ReLU FNN used to approximate the value function. To simplify the theoretical analysis, we also use $\mathcal{R}_\mu(\pi, f), \mathcal{R}_\mathcal{D}(\pi, f)$ to denote $-\mathcal{L}_\mu(\pi, f)$ and $-\mathcal{L}_\mathcal{D}(\pi, f)$, respectively. Thus, problems (1) and (2) can be reformulated as minimax problems:

$$\widehat{\pi}^* \in \arg\min_{\pi \in \Pi_\theta} \max_{f \in \mathcal{F}_\mu^{\pi,\epsilon}} \mathcal{R}_\mu(\pi, f), \tag{3}$$

$$\widehat{\pi} \in \arg\min_{\pi \in \Pi_\theta} \max_{f \in \mathcal{NN}_2 \cap \mathcal{F}_\mathcal{D}^{\pi,\epsilon}} \mathcal{R}_\mathcal{D}(\pi, f). \tag{4}$$

There exists a gap between $\widehat{\pi}$ and the exact solution $\widehat{\pi}^*$, due to finite sampling and imperfect approximation. We aim to explicitly measure this gap concerning the network structure and data sampling, which guides the training process. We want to emphasize that the analysis of this problem is quite challenging due to the coupling of network approximation and the empirical constraint of the Bellman error.

Technical Assumptions

In DL and RL theory, several mild assumptions are commonly utilized. We define the Hölder function class and $\mathcal{C}$-mixing process (Maume-Deschamps 2006) as follows.

Definition 4.1 (Hölder Smooth Function Class)**.** *For $\zeta = s + r$ with $s \in \mathbb{N}^+$ and $0 < r \le 1$, the Hölder smooth function class $\mathcal{H}^\zeta$ is defined as*

$$\begin{aligned} \mathcal{H}^\zeta = \Big\{ f : [0,1]^d \to \mathbb{R} \Big| & \max_{\|\alpha\|_1 \le s} \|\partial^\alpha f\|_\infty \le B, \\ & \max_{\|\alpha\|_1 = s} \sup_{x \ne y} \frac{|\partial^\alpha f(x) - \partial^\alpha f(y)|}{\|x - y\|_\infty^r} \le B \Big\}. \end{aligned}$$

Definition 4.2 ($\mathcal{C}$-mixing process)**.** *Let $(\Omega, \mathcal{A}, \mu)$ be a probability space, $(Z, \mathcal{B})$ be a measurable space, and $\mathcal{Z} := (Z_i)_{i\geq 0}$ be a Z-valued stationary process on Ω. For any $n \geq 0$, we define the $\mathcal{C}$-mixing coefficients as*

$$\psi_{\mathcal{C}}(\mathcal{Z}, n) := \sup\{\mathrm{cor}(Y, h \circ Z_{k+n}) : k \geq 0, Y \in B_{L_1(\mathcal{A}_0^k, \mu)}, h \in B_{\mathcal{C}(Z)}\},$$

where $\mathrm{cor}(\cdot,\cdot)$ denotes the correlation of two random variables, i.e., $\mathrm{cor}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$ if $X, Y, XY \in L_1(\Omega, \mathcal{A}, \mu)$. $\mathcal{A}_0^k$ is the σ-algebra generated by $(Z_0, \ldots, Z_k)$ and $\mathcal{C}(Z)$ is the bounded function space $\{f : Z \to \mathbb{R} \mid \|f\|_\infty + \|f\| < \infty\}$ where $\|\cdot\|$ is a semi-norm.

Additionally, if we have $\psi_{\mathcal{C}}(\mathcal{Z}, n) \leq d_n$ for all $n > 0$, where $(d_n)_{n\geq 0}$ is a strictly positive sequence converging to 0, then $\mathcal{Z}$ is said to be $\mathcal{C}$-mixing with rate $(d_n)_{n\geq 0}$. If $(d_n)_{n\geq 0}$ is of the form $d_n = c\exp(-bn^\eta)$ for $b > 0$, $c \geq 0$, and $\eta > 0$, then $\mathcal{Z}$ is called geometrically $\mathcal{C}$-mixing.

Assumption 4.1 (Smoothness)**.** *Without loss of generality, we assume $z_i := \{s_i, a_i\} \in [0, 1]^d$. Density functions of policies in Π and value functions in $\mathcal{F}$ are Hölder smooth.*

Assumption 4.2 (Completeness)**.** *For any $\pi \in \Pi_\theta$ and $f \in \mathcal{NN}_2$, we have $\mathcal{T}^\pi f \in \mathcal{F}$.*

Assumption 4.1 is a generalization of Lipschitz continuity. It is commonly used in theory and efficient in capturing real-world features (Fan et al. 2020). Assumption 4.2 holds when rewards and values belong to smooth function classes (Fan et al. 2020). Moreover, Chen and Jiang (2019); Wang, Foster, and Kakade (2021) verify that completeness is indispensable even in simple scenarios. While Assumption 4.1 emphasizes the smoothness, it may not sufficiently guarantee the completeness stated in Assumption 4.2.

Assumption 4.3 (Mixing)**.** *We assume the batch data $\{s_t, a_t, r_t\}_{t\geq 0}$ satisfies the definition of strictly stationary geometrically $\mathcal{C}$-mixing process with parameters $b, c, \eta > 0$.*

Assumption 4.3 describes the mixing rate for the batch data sequence and indicates that the future weakly depends on the past. This property of weak dependence is general, encompassing ϕ-mixing (Ibragimov 1962) as a particular case and overlapping with α-mixing (Rosenblatt 1956a). The quantitative distinctions between α-mixing and $\mathcal{C}$-mixing are examined in Hang et al. (2016). Experiments (Solowjow et al. 2020) illustrate that mixing captures the autocorrelation speed of dynamical systems, including Markov chains as a specific case, which characterizes the essential nature of data dependence. We notice that several studies focus on episodic MDPs (Jin, Yang, and Wang 2021), but within a linear setting rather than a universal characterization.

Loss Consistency

Recalling the problem (3), we denote the risk of π as

$$\widetilde{\mathcal{R}}_\mu(\pi, \epsilon) = \max_{f \in \mathcal{F}_\mu^{\pi,\epsilon}} \mathcal{R}_\mu(\pi, f). \tag{5}$$

Our first main theorem presents an upper bound for the excess risk, $\widetilde{\mathcal{R}}_\mu(\widehat{\pi}, \epsilon) - \widetilde{\mathcal{R}}_\mu(\widehat{\pi}^*, \epsilon)$, which quantitatively measures the difference between $\widehat{\pi}$ and $\widehat{\pi}^*$, and demonstrates the efficacy of the adversarial offline RL framework (1).

Theorem 4.1. *Under Assumptions 4.1,4.2 and 4.3, let $\widehat{\pi}$ and $\widehat{\pi}^*$ be defined in (3). Then, for $\mathcal{NN}_1$, $\mathcal{NN}_2$ with width $\mathcal{W} = \mathcal{O}(d^{s+1}|\mathcal{D}|^{\frac{d}{2d+4\zeta^*}})$ and depth $\mathcal{L} = \mathcal{O}(\log(|\mathcal{D}|))$, the following non-asymptotic error bound holds*

$$\mathbb{E}[\widetilde{\mathcal{R}}_\mu(\widehat{\pi}, \epsilon) - \widetilde{\mathcal{R}}_\mu(\widehat{\pi}^*, \epsilon)] \leq C_1 R_{\max} d^{s+(\zeta\vee 1)/2} |\mathcal{D}|^{\frac{-\zeta^*}{d+2\zeta^*}} \log(|\mathcal{D}|)^{2+\frac{1}{\eta}} + C_2\sqrt{\epsilon},$$

where $\zeta^* = \zeta(1 \wedge \zeta)$, C_1 is a constant depending on $s, B, \mathcal{C}(\widehat{\pi}; \mu), \mathcal{C}(\widehat{\pi}_\delta^*; \mu)$ and C_2 is a constant depending on $\mathcal{C}(\widehat{\pi}; \mu), \mathcal{C}(\widehat{\pi}_\delta^*; \mu)$.

A small constant C_1 is achieved when both $\widehat{\pi}$ and $\widehat{\pi}_\delta^*$ demonstrate effective controllability concerning the behavior policy μ. Specifically, $\widehat{\pi}_\delta^*$ represents a δ-neighborhood of $\widehat{\pi}^*$ in terms of their densities, with δ being the approximation error. In other words, our assumption pertains to the partial controllability of data coverage related to $\widehat{\pi}$ and a small area around $\widehat{\pi}^*$. Nevertheless, when concentrability is poor, C_1 might be very large, aligning with the unfavorable empirical results under challenging distribution shifts (Levine et al., 2020). A larger value of ζ indicates a faster order, implying that estimating a smoother target is more manageable. The non-asymptotic bound $\mathcal{O}(|\mathcal{D}|^{\frac{-\zeta^*}{d+2\zeta^*}})$ is afflicted by the curse of dimensionality, a concern we tackle in the next section.

Under mild conditions, the optimal rate in nonparametric regression is $C_d|\mathcal{D}|^{-2\zeta/(2\zeta+d)}$ (Stone 1982), which aligns with ours. Moreover, our prefactor is polynomial in d instead of exponential (Shen, Yang, and Zhang 2019). These results are tight and new in RL. The optimality is also extensively discussed in Suzuki (2018); Suzuki and Nitanda (2021).

The hyperparameter ϵ in the second term corresponds to the Bellman constraint, as introduced in (3), which is restricted by the expressive capacity of the value function class, but may still be small. Its prefactor is also related to the concentrability of $\widehat{\pi}$ and $\widehat{\pi}_\delta^*$. The constraint ϵ significantly influences the training process, as it balances the accuracy and uncertainty aspects of the acquired value function. Note that ϵ should be at least larger than the gap between the value function space and the corresponding Bellman mapping space, but still dominated by the first term in the bound.

This explicit bound has no unknown parameters involved, including the width and depth of the network, providing informative guidance for training adversarial offline RL. By selecting suitable width and depth for neural networks, the estimation error exhibits an exponential decrease as the number of data samples increases. This result matches the empirical observation of network approximation (Montufar et al. 2014) and generalization (Novak et al. 2018).

Circumvent the Curse of Dimensionality

Theorem 4.1 indicates a curse of dimensionality when the data dimension is large. According to the "no free lunch" theorem (Wolpert 1996), any method regardless of data or model conditions is susceptible to this challenge. To mitigate this, we aim to alleviate the curse of dimensionality by utilizing a priori information under two scenarios:

- Data structure with low Minkowski dimension.
- Target function combined of low-complexity elements.

Low-Dimensional Data Structure. We start with the definitions of covering numbers (Vershynin 2018) and Minkowski dimension (Bishop and Peres 2017).

Definition 4.3 (Covering Number)**.** *Let* $\mathbf{dist}$ *be a metric,* $\epsilon > 0$ *and* $K \subset \mathbb{R}^n$*. A subset* $\mathcal{N} \subset K$ *is an* ϵ*-net of* K *if*

$$\forall x \in K, \exists x_0 \in \mathcal{N} : \mathbf{dist}(x, x_0) \leq \epsilon.$$

The smallest cardinality of an ϵ*-net of* K *is called the covering number of* K*, denoted by* $\mathcal{N}(K, \mathbf{dist}, \epsilon)$.

Definition 4.4 (Minkowski Dimension)**.** *Let* $\mathbf{dist}$ *be a metric,* $\epsilon > 0$ *and* K *be a subset of* $\mathbb{R}^n$*, i.e.,* $K \subset \mathbb{R}^n$*. We define the upper and lower Minkowski dimensions as*

$$\overline{\dim}_{\mathcal{M}}(K) = \limsup_{\epsilon \to 0} \frac{\log \mathcal{N}(K, \mathbf{dist}, \epsilon)}{-\log(\epsilon)},$$
$$\underline{\dim}_{\mathcal{M}}(K) = \liminf_{\epsilon \to 0} \frac{\log \mathcal{N}(K, \mathbf{dist}, \epsilon)}{-\log(\epsilon)}.$$

Furthermore, if $\overline{\dim}_{\mathcal{M}}(K) = \underline{\dim}_{\mathcal{M}}(K)$*, this value is called the Minkowski dimension and denoted by* $\dim_{\mathcal{M}}(K)$.

Obviously, $\mathcal{N}(K, \mathbf{dist}, \epsilon) = \epsilon^{-\dim_{\mathcal{M}}(K)+o(1)}$. This indicates that the Minkowski dimension measures the decay rate in covering numbers as ϵ tends towards 0.

Remark 4.1. *For any manifold, its Minkowski dimension is equivalent to its dimension. Although the high ambient dimensions of real-world data are quite large, such as those in MNIST (LeCun et al. 1998), CIFAR (Krizhevsky 2009), ImageNet (Deng et al. 2009), the intrinsic dimensions have been estimated to be relatively low (Recanatesi et al. 2019; Pope et al. 2021). Hence, it is reasonable to assume that the data has a low-dimensional structure, indicating that it is supported by a space with a small Minkowski dimension.*

Theorem 4.2. *Suppose that the support of* $\mathcal{S} \times \mathcal{A}$ *is* $K \subset [0,1]^d$*, and its Minkowski dimension satisfies* $\dim_{\mathcal{M}}(K) \ll d$*. Assuming Assumptions 4.1,4.2 and 4.3 hold, we define* $\widehat{\pi}$ *and* $\widehat{\pi}^*$ *as in (3). Then, for* $\mathcal{NN}_1$ *and* $\mathcal{NN}_2$ *with width* $\mathcal{W} = \mathcal{O}(d_K^{s+1}|\mathcal{D}|^{\frac{d_K}{2d_K+4\zeta^*}})$ *and depth* $\mathcal{L} = \mathcal{O}(\log(|\mathcal{D}|))$*, we can establish a non-asymptotic error bound:*

$$\mathbb{E}[\widetilde{\mathcal{R}}_\mu(\widehat{\pi}, \epsilon) - \widetilde{\mathcal{R}}_\mu(\widehat{\pi}^*, \epsilon)]$$
$$\leq \frac{C_1 R_{\max}}{(1-\lambda)^{\zeta/2}} \sqrt{d} d_K^{s+(\zeta \vee 1+1)/2} |\mathcal{D}|^{\frac{-\zeta^*}{d_K+2\zeta^*}} \log(|\mathcal{D}|)^{2+\frac{1}{\eta}} + C_2\sqrt{\epsilon},$$

where $0 < \lambda < 1$*,* $d_K = \mathcal{O}(\dim_{\mathcal{M}}(K)/\lambda^2)$*,* $\zeta^* = \zeta(1 \wedge \zeta)$*,* C_1 *is a constant depending on* $s, B, \mathcal{C}(\widehat{\pi}; \mu), \mathcal{C}(\widehat{\pi}_\delta^*; \mu)$ *and* C_2 *is a constant depending on* $\mathcal{C}(\widehat{\pi}; \mu), \mathcal{C}(\widehat{\pi}_\delta^*; \mu)$*.*

When d is large, this upper bound is less susceptible to the curse of dimensionality compared to the bound in Theorem 4.1, since the intrinsic dimension $\dim_{\mathcal{M}}(K) \ll d$. Additionally, the width of neural networks has a smaller order than that in Theorem 4.1. These comparisons indicate a significant improvement in alleviating the curse of dimensionality.

Low-Complexity Target Function. We further consider a function f combining k functions as:

$$f = G^k \circ G^{k-1} \circ \cdots \circ G^1, \tag{6}$$

where $G^i : \mathbb{R}^{l_{i-1}} \to \mathbb{R}^{l_i}$ is defined by $G^i(x) = [g_1^i(W_1^i x), \ldots, g_{l_i}^i(W_{l_i}^i x)]^\top$, with $W_j^i \in \mathbb{R}^{d_i \times l_{i-1}}$ being a matrix and $g_j^i : \mathbb{R}^{d_i} \to \mathbb{R}$ being a function.

For instance, a naive additive model (Stone 1985) is given by $f(x) = g_1^1(x_1, x_2) + g_2^1(x_3, x_4) + \cdots + g_{2^{d-1}}^1(x_{2^d-1}, x_{2^d})$ with $x = (x_1, \ldots, x_{2^d})^\top \in \mathbb{R}^{2^d}$. In this case, it indicates $l_0 = 2^d$, $d_1 = 2$, $l_1 = 2^{d-1}$, $d_2 = 2^{d-1}$ and $l_2 = 1$. Without loss of generality, we assume that the component functions g_j^i are Hölder smooth with respect to the coefficient ζ_i.

Remark 4.2. *The low-complexity structure described in (6) is common in various models. Besides the additive model, similar structures can also be observed in other statistical inference models, e.g., the single index model (Hardle, Hall, and Ichimura 1993), the projection pursuit model (Friedman and Stuetzle 1981). Moreover, recent research (Chen, Wang, and Yang 2023) has shown that operators associated with well-known PDEs, including the Poisson, parabolic, and Burgers equation, exhibit this structure or its variants. These operators have the potential to represent natural images for RL tasks such as CT scans (Shen et al. 2022).*

Theorem 4.3. *Suppose the policy and value functions satisfy the condition in (6). Assuming Assumptions 4.1,4.2 and 4.3 hold. Then for* $\mathcal{NN}_1$*,* $\mathcal{NN}_2$ *with width* $\mathcal{W} = \mathcal{O}(d_*^{s+1}|\mathcal{D}|^{\frac{d_*}{2d_*+4\zeta^*}})$ *and depth* $\mathcal{L} = \mathcal{O}(\log(|\mathcal{D}|))$*, we can establish a non-asymptotic error bound*

$$\mathbb{E}[\widetilde{\mathcal{R}}_\mu(\widehat{\pi}, \epsilon) - \widetilde{\mathcal{R}}_\mu(\widehat{\pi}^*, \epsilon)]$$
$$\leq C_1 R_{\max} d_*^{s+(\zeta \vee 1)/2} |\mathcal{D}|^{\frac{-\zeta^*}{d_*+2\zeta^*}} \log(|\mathcal{D}|)^{2+\frac{1}{\eta}} + C_2\sqrt{\epsilon},$$

where $\zeta^* = \min_i(\zeta_i \prod_{l=i+1}^k (\zeta^l \wedge 1))(1 \wedge \zeta)$*,* $d_* = \max_i d_i$*,* C_1 *is a constant depending on* $B, \zeta, s, k, \mathcal{C}(\widehat{\pi}; \mu), \mathcal{C}(\widehat{\pi}_\delta^*; \mu)$ *and* C_2 *is a constant depending on* $\mathcal{C}(\widehat{\pi}; \mu), \mathcal{C}(\widehat{\pi}_\delta^*; \mu)$*.*

In Theorem 4.3, the expression for ζ^* differs from that in Theorems 4.1 and 4.2, as it is determined by the product of Hölder coefficients of each component function g_j^i. A higher degree of smoothness in each component implies a tighter bound. The definition $d_* = \max_i d_i$ indicates a significant reduction in the curse of dimensionality since d_i is consistently much smaller than d. The changes in ζ^* and d_* also result in a reduced order for the width of the neural network.

Our results can be extended to anisotropic Besov spaces (Suzuki 2018; Suzuki and Nitanda 2021), which aligns with our motivation related to low-complexity structure.

5 Proof Sketch

This section provides a proof sketch for Theorem 4.1. The complete proof is available in Appendix B.

Useful Lemmas

Lemma 5.1. *Let* $\widetilde{\mathcal{R}}_\mu(\pi, \epsilon)$ *be defined in (5), and let* $\widehat{\pi}, \widehat{\pi}^*$ *be defined in (3). For any admissible policy* $\phi \in \Pi_\theta$*, we have:*

$$\widetilde{\mathcal{R}}_\mu(\widehat{\pi}, \epsilon) - \widetilde{\mathcal{R}}_\mu(\widehat{\pi}^*, \epsilon)$$
$$\leq \underbrace{2 \sup_{\phi \in \Pi_\theta} |\widetilde{\mathcal{R}}_{\mathcal{D}}(\phi, \epsilon) - \widetilde{\mathcal{R}}_\mu(\phi, \epsilon)|}_{(A)} + \underbrace{2 \sup_{\phi \in \Pi_\theta} |\widehat{\mathcal{R}}_\mu(\phi, \epsilon) - \widetilde{\mathcal{R}}_{\mathcal{D}}(\phi, \epsilon)|}_{(B)}$$
$$+ \underbrace{\inf_{\phi \in \Pi_\theta} \Big(\Big(\widehat{\mathcal{R}}_\mu(\widehat{\pi}, \epsilon) - \widehat{\mathcal{R}}_{\mathcal{D}}(\widehat{\pi}, \epsilon) \Big) + \Big(\widehat{\mathcal{R}}_{\mathcal{D}}(\phi, \epsilon) - \widehat{\mathcal{R}}_\mu(\phi, \epsilon) \Big) + \Big(\widetilde{\mathcal{R}}_\mu(\phi, \epsilon) - \widetilde{\mathcal{R}}_\mu(\widehat{\pi}^*, \epsilon) \Big) \Big)}_{(C)}$$

where $\widetilde{\mathcal{R}}_{\mathcal{D}}(\pi,\epsilon)=\max_{f\in\mathcal{NN}_2\cap\mathcal{F}_\mu^{\pi,\epsilon}}\mathcal{R}_\mu(\pi,f),\widehat{\mathcal{R}}_\mu(\pi,\epsilon)=\max_{f\in\mathcal{NN}_2\cap\mathcal{F}_\mu^{\pi,\epsilon}}\mathcal{R}_{\mathcal{D}}(\pi,f)$ *and* $\widehat{\mathcal{R}}_{\mathcal{D}}(\pi,\epsilon)=\max_{f\in\mathcal{NN}_2\cap\mathcal{F}_{\mathcal{D}}^{\pi,\epsilon}}\mathcal{R}_{\mathcal{D}}(\pi,f)$.

The upper bound for the excess risk has three components. The first term (A) and the second term (B) capture the approximation error and generalization error, respectively. The third term (C), known as the Bellman estimation error, combines both the approximation and generalization error coupled with the Bellman residual about on/off-support data. This lemma provides a decomposition of the excess risk, forming the foundation for further derivation. We now introduce lemmas corresponding to each part.

Lemma 5.2 (Bounding (A)). *Let $\widetilde{\mathcal{R}}_{\mathcal{D}}(\pi,\epsilon),\widetilde{\mathcal{R}}_\mu(\pi,\epsilon)$ be defined in Lemma 5.1. Under Assumption 4.1, for $\mathcal{NN}_2$ with width $38(s+1)^2 3^d d^{s+1}N\lceil\log_2(8N)\rceil$ and depth $21(s+1)^2M\lceil\log_2(8M)\rceil+2d$, it holds for any $M,N\in\mathbb{N}^+$:*

$$\sup_{\phi\in\Pi_\theta}|\widetilde{\mathcal{R}}_{\mathcal{D}}(\phi,\epsilon)-\widetilde{\mathcal{R}}_\mu(\phi,\epsilon)|\le 38B(s+1)^2d^{s+(\zeta\vee1)/2}(NM)^{-2\zeta/d}$$

Lemma 5.2 provides an approximation error between ReLU FNN and Hölder functions, exhibiting a polynomial dependency on the input data dimension d.

Lemma 5.3 (Bounding (B)). *Let $\widehat{\mathcal{R}}_\mu(\pi,\epsilon)$ and $\widetilde{\mathcal{R}}_{\mathcal{D}}(\pi,\epsilon)$ be defined as in Lemma 5.1. Under Assumption 4.3, if the size of the dataset $|\mathcal{D}|$ satisfies*

$$|\mathcal{D}|\ge n_0:=\max\Big\{\min\Big\{m\ge3:m^2\ge808c,\frac{m}{(\log m)^{2/\eta}}\Big\},e^{3/b}\Big\},$$

where b,c,η are parameters in Assumption 4.3, the following holds for any $\phi\in\Pi_\theta$

$$\mathbb{E}\sup_{\phi\in\Pi_\theta}|\widehat{\mathcal{R}}_\mu(\phi,\epsilon)-\widetilde{\mathcal{R}}_{\mathcal{D}}(\phi,\epsilon)|\le\mathcal{O}\Big(R_{\max}\sqrt{\mathcal{PL}\log(\mathcal{P})}\frac{\big(\log|\mathcal{D}|\big)^{\frac{2+\eta}{2\eta}}}{\sqrt{|\mathcal{D}|}}\Big).$$

Lemma 5.3 provides a bound on the generalization error for $\mathcal{C}$-mixing data. We employ a uniform covering, to enable the measure of the infinite neural network class.

Lemma 5.4 (Bouding (C)). *Let $\widehat{\mathcal{R}}_\mu(\pi,\epsilon)$, $\widehat{\mathcal{R}}_{\mathcal{D}}(\pi,\epsilon)$, $\widehat{\mathcal{R}}_\mu(\pi,\epsilon)$ be defined in Lemma 5.1, and let $\widehat{\pi},\widehat{\pi}^*$ be defined in (3). If the size of the dataset $|\mathcal{D}|$ satisfies the requirement in Lemma 5.3, we have*

$$\begin{aligned}&\inf_{\phi\in\Pi_\theta}\Big(\Big(\widehat{\mathcal{R}}_\mu(\widehat{\pi},\epsilon)-\widehat{\mathcal{R}}_{\mathcal{D}}(\widehat{\pi},\epsilon)\Big)+\Big(\widehat{\mathcal{R}}_{\mathcal{D}}(\phi,\epsilon)-\widehat{\mathcal{R}}_\mu(\phi,\epsilon)\Big)\\&\quad+\Big(\widetilde{\mathcal{R}}_\mu(\phi,\epsilon)-\widetilde{\mathcal{R}}_\mu(\widehat{\pi}^*,\epsilon)\Big)\Big)\\&\le C_{\mathcal{C}(\widehat{\pi};\mu),\mathcal{C}(\widehat{\pi}_\delta^*;\mu)}\sqrt{\epsilon}+C_{B,\mathcal{C}(\widehat{\pi}_\delta^*;\mu)}\delta^{1\wedge\zeta}\\&\quad+C_{\mathcal{C}(\widehat{\pi};\mu),\mathcal{C}(\widehat{\pi}_\delta^*;\mu)}\mathcal{O}\Big(R_{\max}\sqrt{\mathcal{PL}\log(\mathcal{P})}\frac{\big(\log|\mathcal{D}|\big)^{\frac{2+\eta}{2\eta}}}{\sqrt{|\mathcal{D}|}}\Big).\end{aligned}$$

The Bellman estimation consists of two terms: Bellman approximation and generalization. We bound these two terms together for consistency since a shared policy ϕ is considered an infimum. The first term on the RHS is related to the Bellman residual constraint, while the second and third terms correspond to generalization and approximation. The coefficients are similar to those discussed in Theorem 4.1.

Main Proof

Proof of Theorem 4.1. Combining the bounds in Lemma 5.2, 5.3, 5.4 into the decomposition in Lemma 5.1 yields

$$\begin{aligned}&\mathbb{E}[\widetilde{\mathcal{R}}_\mu(\widehat{\pi},\epsilon)-\widetilde{\mathcal{R}}_\mu(\widehat{\pi}^*,\epsilon)]\\&\le C_{s,B,\mathcal{C}(\widehat{\pi}_\delta^*;\mu)}d^{s+(\zeta\vee1)/2}(NM)^{-2\zeta(1\wedge\zeta)/d}+C_{\mathcal{C}(\widehat{\pi};\mu),\mathcal{C}(\widehat{\pi}_\delta^*;\mu)}\sqrt{\epsilon}\\&\quad+C_{\mathcal{C}(\widehat{\pi};\mu),\mathcal{C}(\widehat{\pi}_\delta^*;\mu)}\Big(R_{\max}\sqrt{\mathcal{PL}\log(\mathcal{P})}\frac{\big(\log|\mathcal{D}|\big)^{\frac{2+\eta}{2\eta}}}{\sqrt{|\mathcal{D}|}}\Big).\end{aligned}$$

This bound indicates that as M and N grow large, the first term decreases while the second term increases. Thus we balance them by selecting appropriate M and N to obtain

$$d^{s+(\zeta\vee1)/2}(NM)^{-2\zeta(1\wedge\zeta)/d}\approx\sqrt{\mathcal{PL}\log(\mathcal{P})}\frac{\big(\log|\mathcal{D}|\big)^{\frac{2+\eta}{2\eta}}}{\sqrt{|\mathcal{D}|}}.$$

The number of parameters $\mathcal{P}$, the width $\mathcal{W}$ and the depth $\mathcal{L}$ of the network satisfy the inequality:

$$\mathcal{P}\le\mathcal{W}(d+1)+(\mathcal{W}^2+\mathcal{W})(\mathcal{L}-1)+\mathcal{W}+1\le2\mathcal{W}^2\mathcal{L}.$$

The approximation bound is established with width $\mathcal{W}=38(s+1)^2 3^d d^{s+1}N\lceil\log_2(8N)\rceil$ and depth $\mathcal{L}=21(s+1)^2M\lceil\log_2(8M)\rceil+2d$, yielding the number of parameters $\mathcal{P}\le\mathcal{O}((s+1)^6d^{2s+2}N^2\lceil\log_2(8N)\rceil^2M\lceil\log_2(8M)\rceil)$. By setting $N=\mathcal{O}(|\mathcal{D}|^{\frac{d}{2d+4\zeta^*}})$ and $M=\mathcal{O}(\log(|\mathcal{D}|))$, we can further bound $\mathcal{W}=\mathcal{O}(d^{s+1}|\mathcal{D}|^{\frac{d}{2d+4\zeta^*}})$, $\mathcal{L}=\mathcal{O}(\log(|\mathcal{D}|))$, $\mathcal{P}=\mathcal{O}(d^{2s+2}|\mathcal{D}|^{\frac{d}{d+2\zeta^*}}\log(|\mathcal{D}|))$.

$$\begin{aligned}&\mathbb{E}[\widetilde{\mathcal{R}}_\mu(\widehat{\pi},\epsilon)-\widetilde{\mathcal{R}}_\mu(\widehat{\pi}^*,\epsilon)]\\&\le C_{s,B,\mathcal{C}(\widehat{\pi}_\delta^*;\mu)}d^{s+(\zeta\vee1)/2}|\mathcal{D}|^{\frac{-\zeta^*}{d+2\zeta^*}}\log(|\mathcal{D}|)+C_{\mathcal{C}(\widehat{\pi};\mu),\mathcal{C}(\widehat{\pi}_\delta^*;\mu)}\sqrt{\epsilon}\\&\quad+C_{\mathcal{C}(\widehat{\pi};\mu),\mathcal{C}(\widehat{\pi}_\delta^*;\mu)}R_{\max}d^{s+1}|\mathcal{D}|^{\frac{-\zeta^*}{d+2\zeta^*}}\log(|\mathcal{D}|)^{2+\frac{1}{\eta}}\\&=C_1R_{\max}d^{s+(\zeta\vee1)/2}|\mathcal{D}|^{\frac{-\zeta^*}{d+2\zeta^*}}\log(|\mathcal{D}|)^{2+\frac{1}{\eta}}+C_2\sqrt{\epsilon},\end{aligned}$$

where $\zeta^*=\zeta(1\wedge\zeta)$, C_1 depends on $s,B,\mathcal{C}(\widehat{\pi};\mu),\mathcal{C}(\widehat{\pi}_\delta^*;\mu)$, and C_2 depends on $\mathcal{C}(\widehat{\pi};\mu)$ and $\mathcal{C}(\widehat{\pi}_\delta^*;\mu)$. □

6 Conclusion

This paper examines the estimation error within a deep adversarial offline RL framework under mild assumptions. Both policy and value functions are parameterized using deep neural networks, with data assumed to exhibit dependence and partial coverage. The excess risk is decomposed into three components: generalization, approximation, and Bellman estimation error. We bound these errors by adapting tools from empirical processes and approximation theory to address intricate Bellman constraints. This derived bound explicitly reveals the interplay between network architecture, dataset dimensionality, sample size, and the concentrability of distributional shifts in influencing the estimation error. Additionally, we provide two conditions to alleviate the curse of dimensionality. Our work is the first attempt to establish a non-asymptotic estimation error for deep adversarial offline RL problems.

Acknowledgements

We would like to thank the anonymous referees for their useful comments and suggestions, which have led to considerable improvements in the paper. This work is supported by the National Key Research and Development Program of China (No.2020YFA0714200), the National Nature Science Foundation of China (No.12371424, No.12371441), "the Fundamental Research Funds for the Central Universities", the research fund of KLATASDSMOE of China, and the US National Science Foundation under awards DMS2244988, DMS2206333.

References

Anthony, M.; and Bartlett, P. 1999. *Neural network learning: theoretical foundations*. Cambridge University Press.

Antos, A.; Szepesvári, C.; and Munos, R. 2007. Fitted Q-iteration in continuous action-space MDPs. In *NeurIPS*.

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. *Machine Learning*.

Bai, C.; Wang, L.; Yang, Z.; Deng, Z.; Garg, A.; Liu, P.; and Wang, Z. 2022. Pessimistic bootstrapping for uncertainty-driven offline reinforcement learning. *arXiv preprint arXiv:2202.11566*.

Bauer, B.; and Kohler, M. 2019. On deep learning as a remedy for the curse of dimensionality in nonparametric regression. *The Annals of Statistics*.

Bhardwaj, M.; Xie, T.; Boots, B.; Jiang, N.; and Cheng, C.-A. 2023. Adversarial model for offline reinforcement learning. *arXiv preprint arXiv:2302.11048*.

Bishop, C. J.; and Peres, Y. 2017. *Fractals in probability and analysis*, volume 162. Cambridge University Press.

Chang, J.; Uehara, M.; Sreenivas, D.; Kidambi, R.; and Sun, W. 2021. Mitigating covariate shift in imitation learning via offline data with partial coverage. In *NeurIPS*.

Chen, J.; and Jiang, N. 2019. Information-theoretic considerations in batch reinforcement learning. In *ICML*.

Chen, K.; Wang, C.; and Yang, H. 2023. Deep operator learning lessens the curse of dimensionality for PDEs. *arXiv preprint arXiv:2301.12227*.

Cheng, C.-A.; Xie, T.; Jiang, N.; and Agarwal, A. 2022. Adversarially trained actor critic for offline reinforcement learning. In *ICML*.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*.

Fan, J.; Wang, Z.; Xie, Y.; and Yang, Z. 2020. A theoretical analysis of deep Q-learning. In *Learning for Dynamics and Control*.

Farahmand, A.-m.; Szepesvári, C.; and Munos, R. 2010. Error propagation for approximate policy and value iteration. In *NeurIPS*.

Farrell, M. H.; Liang, T.; and Misra, S. 2021. Deep neural networks for estimation and inference. *Econometrica*.

Feng, Y.; Li, L.; and Liu, Q. 2019. A kernel loss for solving the bellman equation. In *NeurIPS*.

Friedman, J. H.; and Stuetzle, W. 1981. Projection pursuit regression. *Journal of the American statistical Association*.

Fu, J.; Kumar, A.; Nachum, O.; Tucker, G.; and Levine, S. 2020. D4rl: Datasets for deep data-driven reinforcement learning. *arXiv preprint arXiv:2004.07219*.

Fujimoto, S.; Meger, D.; and Precup, D. 2019. Off-policy deep reinforcement learning without exploration. In *ICML*.

Hang, H.; Feng, Y.; Steinwart, I.; and Suykens, J. A. 2016. Learning theory estimates with observations from general stationary stochastic processes. *Neural computation*.

Hang, H.; and Steinwart, I. 2017. A Bernstein-type inequality for some mixing processes and dynamical systems with an application to learning. *The Annals of Statistics*.

Hardle, W.; Hall, P.; and Ichimura, H. 1993. Optimal smoothing in single-index models. *The Annals of Statistics*.

Ibragimov, I. A. 1962. Some limit theorems for stationary processes. *Theory of Probability & Its Applications*.

Ji, X.; Chen, M.; Wang, M.; and Zhao, T. 2023. Sample complexity of nonparametric off-policy evaluation on low-dimensional manifolds using deep networks. In *ICLR*.

Jiang, N. 2019. On value functions and the agent-environment boundary. *arXiv preprint arXiv:1905.13341*.

Jiang, N.; and Huang, J. 2020. Minimax value interval for off-policy evaluation and policy optimization. In *NeurIPS*.

Jiao, Y.; Shen, G.; Lin, Y.; and Huang, J. 2023. Deep nonparametric regression on approximately low-dimensional manifolds. *Annals of Statistics*.

Jiao, Y.; Wang, Y.; and Yang, Y. 2023. Approximation bounds for norm constrained neural networks with applications to regression and GANs. *Applied and Computational Harmonic Analysis*.

Jin, Y.; Yang, Z.; and Wang, Z. 2021. Is pessimism provably efficient for offline rl? In *ICML*.

Kallus, N.; and Uehara, M. 2022. Efficiently breaking the curse of horizon in off-policy evaluation with double reinforcement learning. *Operations Research*.

Kostrikov, I.; Nair, A.; and Levine, S. 2021. Offline reinforcement learning with implicit Q-learning. In *Deep RL Workshop NeurIPS 2021*.

Krizhevsky, A. 2009. Learning multiple layers of features from tiny images. *Master's thesis, University of Tront*.

Kumar, A.; Fu, J.; Soh, M.; Tucker, G.; and Levine, S. 2019. Stabilizing off-policy q-learning via bootstrapping error reduction. In *NeurIPS*.

Kumar, A.; Zhou, A.; Tucker, G.; and Levine, S. 2020. Conservative q-learning for offline reinforcement learning. In *NeurIPS*.

Lange, S.; Gabel, T.; and Riedmiller, M. 2012. Batch reinforcement learning. *Reinforcement learning: State-of-the-art*.

Laroche, R.; Trichelair, P.; and Des Combes, R. T. 2019. Safe policy improvement with baseline bootstrapping. In *ICML*.

LeCun, Y.; Bottou, L.; Bengio, Y.; and Haffner, P. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*.

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*.

Li, G.; Shi, L.; Chen, Y.; Chi, Y.; and Wei, Y. 2022. Settling the sample complexity of model-based offline reinforcement learning. *arXiv preprint arXiv:2204.05275*.

Liao, P.; Qi, Z.; Wan, R.; Klasnja, P.; and Murphy, S. A. 2022. Batch policy learning in average reward markov decision processes. *The Annals of Statistics*.

Liu, B.; Cai, Q.; Yang, Z.; and Wang, Z. 2019. Neural trust region/proximal policy optimization attains globally optimal policy. In *NeurIPS*.

Liu, Y.; Swaminathan, A.; Agarwal, A.; and Brunskill, E. 2020. Provably good batch off-policy reinforcement learning without great exploration. In *NeurIPS*.

Lu, J.; Shen, Z.; Yang, H.; and Zhang, S. 2021. Deep network approximation for smooth functions. *SIAM Journal on Mathematical Analysis*.

Lyu, J.; Ma, X.; Li, X.; and Lu, Z. 2022. Mildly conservative Q-learning for offline reinforcement learning. In *NeurIPS*.

Maume-Deschamps, V. 2006. Exponential inequalities and functional estimations for weak dependent data: applications to dynamical systems. *Stochastics and Dynamics*.

Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A. A.; Veness, J.; Bellemare, M. G.; Graves, A.; Riedmiller, M.; Fidjeland, A. K.; Ostrovski, G.; et al. 2015. Human-level control through deep reinforcement learning. *Nature*.

Mohri, M.; and Rostamizadeh, A. 2008. Rademacher complexity bounds for non-iid processes. In *NeurIPS*.

Mohri, M.; and Rostamizadeh, A. 2010. Stability Bounds for Stationary φ-mixing and β-mixing Processes. *JMLR*.

Montufar, G. F.; Pascanu, R.; Cho, K.; and Bengio, Y. 2014. On the number of linear regions of deep neural networks. In *NeurIPS*.

Munos, R. 2007. Performance bounds in l_p-norm for approximate value iteration. *SIAM journal on control and optimization*.

Nakada, R.; and Imaizumi, M. 2020. Adaptive approximation and generalization of deep neural network with intrinsic dimensionality. *JMLR*.

Nguyen-Tang, T.; Gupta, S.; Tran-The, H.; and Venkatesh, S. 2022a. On sample complexity of offline reinforcement learning with deep ReLU networks in Besov spaces. *TMLR*.

Nguyen-Tang, T.; Yin, M.; Gupta, S.; Venkatesh, S.; and Arora, R. 2022b. On instance-dependent bounds for offline reinforcement learning with linear function approximation. *arXiv preprint arXiv:2211.13208*.

Novak, R.; Bahri, Y.; Abolafia, D. A.; Pennington, J.; and Sohl-Dickstein, J. 2018. Sensitivity and generalization in neural networks: an empirical study. In *ICLR*.

Pope, P.; Zhu, C.; Abdelkader, A.; Goldblum, M.; and Goldstein, T. 2021. The intrinsic dimension of images and its impact on learning. In *ICLR*.

Ralaivola, L.; and Amini, M.-R. 2015. Entropy-based concentration inequalities for dependent variables. In *ICML*.

Rashidinejad, P.; Zhu, B.; Ma, C.; Jiao, J.; and Russell, S. 2021. Bridging offline reinforcement learning and imitation learning: A tale of pessimism. In *NeurIPS*.

Rashidinejad, P.; Zhu, H.; Yang, K.; Russell, S.; and Jiao, J. 2022. Optimal conservative offline rl with general function approximation via augmented lagrangian. *arXiv preprint arXiv:2211.00716*.

Recanatesi, S.; Farrell, M.; Advani, M.; Moore, T.; Lajoie, G.; and Shea-Brown, E. 2019. Dimensionality compression and expansion in deep neural networks. *arXiv preprint arXiv:1906.00443*.

Rigter, M.; Lacerda, B.; and Hawes, N. 2022. RAMBO-RL: robust adversarial model-based offline reinforcement learning. In *NeurIPS*.

Rosenblatt, M. 1956a. A central limit theorem and a strong mixing condition. *Proceedings of the national Academy of Sciences*.

Rosenblatt, M. 1956b. Remarks on some nonparametric estimates of a density function. *The annals of mathematical statistics*.

Roy, A.; Balasubramanian, K.; and Erdogdu, M. A. 2021. On empirical risk minimization with dependent and heavy-tailed data. In *NeurIPS*.

Scherrer, B. 2014. Approximate policy iteration schemes: a comparison. In *ICML*.

Schmidt-Hieber, A. J. 2020. Nonparametric regression using deep neural networks with ReLU activation function. *Annals of statistics*.

Shen, Z.; Wang, Y.; Wu, D.; Yang, X.; and Dong, B. 2022. Learning to scan: A deep reinforcement learning approach for personalized scanning in CT imaging. *Inverse Problems & Imaging*, 16(1).

Shen, Z.; Yang, H.; and Zhang, S. 2019. Deep network approximation characterized by number of neurons. *arXiv preprint arXiv:1906.05497*.

Shen, Z.; Yang, H.; and Zhang, S. 2021. Deep network with approximation error being reciprocal of width to power of square root of depth. *Neural Computation*.

Shi, C.; Zhang, S.; Lu, W.; and Song, R. 2022a. Statistical inference of the value function for reinforcement learning in infinite-horizon settings. *Journal of the Royal Statistical Society Series B: Statistical Methodology*.

Shi, L.; Li, G.; Wei, Y.; Chen, Y.; and Chi, Y. 2022b. Pessimistic q-learning for offline reinforcement learning: Towards optimal sample complexity. In *ICML*.

Siegel, N.; Springenberg, J. T.; Berkenkamp, F.; Abdolmaleki, A.; Neunert, M.; Lampe, T.; Hafner, R.; Heess, N.; and Riedmiller, M. 2020. Keep doing what worked: Behavioral modelling priors for offline reinforcement learning. In *ICLR*.

Silver, D.; Huang, A.; Maddison, C. J.; Guez, A.; Sifre, L.; Van Den Driessche, G.; Schrittwieser, J.; Antonoglou, I.; Panneershelvam, V.; Lanctot, M.; et al. 2016. Mastering the

game of Go with deep neural networks and tree search. *Nature*.

Solowjow, F.; Baumann, D.; Fiedler, C.; Jocham, A.; Seel, T.; and Trimpe, S. 2020. A kernel two-sample test for dynamical systems. *arXiv preprint arXiv:2004.11098*.

Steinwart, I.; Hush, D.; and Scovel, C. 2009. Learning from dependent observations. *Journal of Multivariate Analysis*.

Stone, C. J. 1982. Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*.

Stone, C. J. 1985. Additive regression and other nonparametric models. *The Annals of Statistics*.

Sutton, R. S.; and Barto, A. G. 2018. *Reinforcement learning: An introduction*. MIT press.

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

Suzuki, T.; and Nitanda, A. 2021. Deep learning is adaptive to intrinsic dimensionality of model smoothness in anisotropic Besov space. In *NeurIPS*.

Szepesvári, C.; and Munos, R. 2005. Finite time bounds for sampling based fitted value iteration. In *ICML*.

Uehara, M.; Huang, J.; and Jiang, N. 2020. Minimax weight and q-function learning for off-policy evaluation. In *ICML*.

Uehara, M.; and Sun, W. 2021. Pessimistic model-based offline reinforcement learning under partial coverage. *arXiv preprint arXiv:2107.06226*.

Vershynin, R. 2018. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press.

Wang, L.; Cai, Q.; Yang, Z.; and Wang, Z. 2019. Neural policy gradient methods: Global optimality and rates of convergence. *arXiv preprint arXiv:1909.01150*.

Wang, R.; Foster, D.; and Kakade, S. M. 2021. What are the statistical limits of offline RL with linear function approximation? In *ICLR*.

Wolpert, D. H. 1996. The lack of a priori distinctions between learning algorithms. *Neural computation*.

Xie, T.; Cheng, C.-A.; Jiang, N.; Mineiro, P.; and Agarwal, A. 2021. Bellman-consistent pessimism for offline reinforcement learning. In *NeurIPS*.

Xie, T.; and Jiang, N. 2020. Q* approximation schemes for batch reinforcement learning: A theoretical comparison. In *UAI*.

Xie, T.; and Jiang, N. 2021. Batch value-function approximation with only realizability. In *ICML*.

Yarotsky, D. 2017. Error bounds for approximations with deep ReLU networks. *Neural Networks*.

Yarotsky, D. 2018. Optimal approximation of continuous functions by very deep ReLU networks. In *COLT*.

Yarotsky, D. 2021. Elementary superexpressive activations. In *ICML*.

Yin, M.; Bai, Y.; and Wang, Y.-X. 2021. Near-optimal offline reinforcement learning via double variance reduction. In *NeurIPS*.

Yu, B. 1994. Rates of convergence for empirical processes of stationary mixing sequences. *The Annals of Probability*.

Zanette, A.; and Wainwright, M. J. 2022. Bellman residual orthogonalization for offline reinforcement learning. *arXiv preprint arXiv:2203.12786*.

Zhan, W.; Huang, B.; Huang, A.; Jiang, N.; and Lee, J. 2022. Offline reinforcement learning with realizability and single-policy concentrability. In *COLT*.

Zhang, J.; Koppel, A.; Bedi, A. S.; Szepesvari, C.; and Wang, M. 2020. Variational policy gradient method for reinforcement learning with general utilities. In *NeurIPS*.

Zhang, X.; Chen, Y.; Zhu, X.; and Sun, W. 2022. Corruption-robust offline reinforcement learning. In *AISTATS*.

Zhou, Y.; Qi, Z.; Shi, C.; and Li, L. 2023. Optimizing Pessimism in Dynamic Treatment Regimes: A Bayesian Learning Approach. In *AISTATS*.

Zou, S.; Xu, T.; and Liang, Y. 2019. Finite-sample analysis for sarsa with linear function approximation. In *NeurIPS*.