Bridging Distributionally Robust Learning and Offline RL: An Approach to Mitigate Distribution Shift and Partial Data Coverage

Kishan Panaganti KPB@CALTECH.EDU
California Institute of Technology

Zaiyan Xu ZXU43@TAMU.EDU
Texas A&M University

Dileep Kalathil DILEEP.KALATHIL@TAMU.EDU
Texas A&M University

Mohammad Ghavamzadeh GHAVAMZA@AMAZON.COM
Amazon AGI

Editors: N. Ozay, L. Balzano, D. Panagou, A. Abate

Abstract

The goal of an offline reinforcement learning (RL) algorithm is to learn the optimal policy using offline data, without access to the environment for online exploration. One of the main challenges in offline RL is the distribution shift which refers to the difference between the state-action visitation distribution of the data generating policy and the learning policy. Many recent works have used the idea of pessimism for developing offline RL algorithms and characterizing their sample complexity under a relatively weak assumption of single policy concentrability. Different from the offline RL literature, the area of distributionally robust learning (DRL) offers a principled framework that uses a minimax formulation to tackle model mismatch between training and testing environments. In this work, we aim to bridge these two areas by showing that the DRL approach can tackle the distributional shift problem in offline RL. In particular, we propose two offline RL algorithms using the DRL framework, for the tabular and linear function approximation settings, and characterize their sample complexity under the single policy concentrability assumption. We also demonstrate the performance of our algorithm through simulation experiments and by comparing it with other state-of-the-art tabular offline RL algorithms.

Keywords: Offline Reinforcement Learning, Distributionally Robust Reinforcement Learning, Statistical Learning Theory.

1. Introduction

The goal of an offline RL algorithm is to learn an approximately optimal policy using minimal amount of offline data collected according to a behavior policy Lange et al. (2012); Levine et al. (2020). The lack of online exploration makes the offline RL problem particularly challenging due to *distribution shift* and *partial data coverage*. Distribution shift refers to the difference between the state-action visitation distribution of the behavior policy and that of the learned policy. Partial data coverage refers to the fact that the data generated according to the behavior policy may only contain samples from parts of the state-action spaces. While these two issues are not the same, in effect, they both cause the problem of out-of-distribution (OOD) data (Yang et al., 2021; Robey et al., 2020), i.e., distributions of training and testing data being different.

In the past few years, many works have developed deep offline RL algorithms mitigating distribution shift and partial data coverage, but have been mainly focused on the algorithmic and empirical aspects (Fujimoto et al., 2019; Kumar et al., 2019, 2020; Fujimoto and Gu, 2021; Kostrikov et al., 2021). Most of the early theoretical works on offline RL, however, analyzed the performance of their algorithms by making the strong assumption of *uniformly bounded concentrability* which requires that the ratio of the state-action occupancy distribution induced by *any* policy and the data generating distribution being bounded uniformly over all states and actions (Munos, 2007; Antos et al., 2008; Munos and Szepesvári, 2008; Farahmand et al., 2010; Chen and Jiang, 2019; Liao et al., 2022). The more recent theoretical results have used the principle of pessimism or conservatism (Yu et al., 2020; Buckman et al., 2021; Jin et al., 2021) for the offline RL problem, including replacing uniform concentrability with the more relaxed *single policy concentrability* assumption (Uehara and Sun, 2021; Rashidinejad et al., 2022; Li et al., 2022a).

1.1. Motivation: Why Distributionally Robust Learning for Offline RL?

Classical supervised learning is based on empirical risk minimization (ERM), which assumes that the train and test data are drawn from the same distribution (Shalev-Shwartz and Ben-David, 2014). However, this assumption is hardly satisfied in many real-world applications (Quinonero-Candela et al., 2022), and the performance of supervised learning algorithms degrade significantly in the out-of-distribution setting (Taori et al., 2020; Koh et al., 2021). A large body of work has been recently developed that uses the distributionally robust learning (DRL) framework to address the issue of distribution shift in various settings (Duchi and Namkoong, 2021; Kuhn et al., 2019; Chen et al., 2020). The DRL framework considers an uncertainty set of data distributions around a nominal distribution (typically the training data distribution), and solves a minimax optimization problem to find a function that minimizes the expected loss, where the expectation is taken w.r.t. the distribution in the uncertainty set that maximizes the loss. DRL is a principled framework that provides generalization guarantees, accommodates ways of constructing domain-specific uncertainty sets (e.g., using f-divergence and Wasserstein distance), and offers scalable algorithms (Chen et al., 2020; Levy et al., 2020; Mohajerin Esfahani and Kuhn, 2018).

The issue of out-of-distribution data arises in real-world RL applications because of the mismatch between the train and test environments (MDP models). This issue is also known as the simulation-to-reality (sim-to-real) gap (Tobin et al., 2017). Modeling errors and changes in the real-world system parameters are inevitable in RL applications, and standard RL policies can fail dramatically even when they face a mild mismatch between the train and test environments (Tobin et al., 2017; Peng et al., 2018). Many works have used the heuristic of domain randomization (Weng, 2019) to make the learned RL policy robust against the sim-to-real gap. More recently, several works have proposed to use the DRL framework in RL to mitigate the sim-to-real gap problem (Tamar et al., 2014; Roy et al., 2017; Panaganti and Kalathil, 2021; Panaganti et al., 2022; Panaganti and Kalathil, 2022; Xu* et al., 2023; Ma et al., 2022; Wang and Zou, 2022; Kumar et al., 2023; Li et al., 2022b; Wang et al., 2023a), building on the formalism of robust Markov decision processes (RMDPs) (Iyengar, 2005; Nilim and El Ghaoui, 2005). However, these works do not consider the offline RL setting in which the out-of-distribution issues are due to the distribution shift and partial data coverage.

Offline RL closely resembles supervised learning because its goal is to learn a policy from an offline dataset, as opposed to the conventional RL goal of learning through online exploration. So,

Algorithm	Algorithm-type	Data coverage assumption	Suboptimality
Lower bound (Rashidinejad et al., 2022, Th.7)	-	single-policy	$\widetilde{\mathcal{O}}\left(\sqrt{\frac{\lvert\mathcal{S}\rvert(C_{\pi^*}-1)}{(1-\gamma)^3 N}}\right)$
(Rashidinejad et al., 2022, Th.6)	reward pessimism	single-policy	$\widetilde{\mathcal{O}}\left(\sqrt{\frac{\lvert\mathcal{S}\rvert C_{\pi^*}}{(1-\gamma)^5 N}}\right)$
(Uehara and Sun, 2021, Cor.1)	oracle model pessimism	single-policy	$\widetilde{\mathcal{O}}\left(\sqrt{\frac{\lvert\mathcal{S}\rvert^2\lvert\mathcal{A}\rvert C_{\pi^*}}{(1-\gamma)^4 N}}\right)$
DRQI (this work, Th.1, Th.4-7)	distributionally robust	single-policy	$\widetilde{\mathcal{O}}\left(\sqrt{\frac{\lvert\mathcal{S}\rvert^2 C_{\pi^*}}{(1-\gamma)^4 N}}\right)$

Table 1: Comparison of the offline RL algorithms in the tabular setting. The data coverage assumption is based on the single-policy concentrability $C_{\pi^*} = \max_{s,a}(d^{\pi^*}(s,a)/\mu(s,a))$, where d^{π^*}, μ is the state-action visitation distributions of the optimal and data generating policies respectively. The suboptimality column is the statistical bounds for the offline RL objective (Eq. (1)), where $|\mathcal{S}|$ and $|\mathcal{A}|$ are the number of states and actions, γ is the discount factor, and N is the size of the offline data.

offline RL faces similar out-of-distribution issues as in supervised learning. As mentioned above, DRL has shown to be an attractive framework to address the out-of-distribution issues arising in supervised learning, offering practical algorithms with provable guarantees. These observations motivate us to ask the following questions:

> *Can we address the distributional shift issues in offline RL using distributionally robust learning as a principled approach? What kind of theoretical performance guarantees can we provide and under what kind of assumptions?*

In this work, we answer these questions affirmatively. In particular, we propose offline RL algorithms using the framework of DRL for the tabular and linear MDP settings, and characterize their sample complexity. Moreover, we show that our approach enables the relaxation of the strong assumption of uniform concentrability to single policy concentrability. Apart from the technical contributions, we believe that establishing this connection, the proverbial bridge, between the DRL literature and offline RL literature is an interesting contribution itself, as it will enable door to bring the state-of-the-art algorithms from the active area of DRL to the offline RL, especially for high-dimensional problems. *We refer to our supplement Panaganti et al. (2023), due to the page limit, for detailed proofs of our technical results and empirical evaluations.*

1.2. Comparisons and Contributions

We outline our contributions and compare our theoretical results with several recent works that, similar to ours, only use the single concentrability assumption.

Uehara and Sun (2021) propose a pessimistic model-based offline RL algorithm, which we refer to as *oracle model pessimism* in Table 1 and Table 2. While their proposed algorithm is similar to the max-min formulation of DRL, they do not offer a computationally tractable implementation for it. It is known in the RMDP literature (Iyengar, 2005; Nilim and El Ghaoui, 2005; Wiesemann et al., 2013) that solving the max-min objective (Eq. (4)) can be NP-hard without additional structural assumptions, such as *rectangularity*. Rashidinejad et al. (2022) propose a lower confidence bound algorithm based on the idea of pessimism in the face of uncertainty. The algorithm subtracts a

Algorithm	Algorithm-type	Data coverage assumption	Suboptimality
(Jin et al., 2021, Cor.4.5)	reward pessimism	$C_{\text{sc}} < \infty$	$\frac{d\sqrt{\text{rank}(\Sigma_{d^{\pi^*}})}}{\sqrt{C_{\text{sc}}(1-\gamma)^4 N}}$
(Uehara and Sun, 2021, Th.6)	oracle model pessimism	$C_{\pi^*,\phi} < \infty$	$\sqrt{\frac{\text{rank}(\Lambda)^2 d C_{\pi^*,\phi}}{(1-\gamma)^4 N}}$
LM-DRQI (this work, Th.2)	distributionally robust	$C^{\dagger}_{\text{sc}} < \infty$	$\frac{\sqrt{\text{rank}(\Sigma_{d^{\pi^*}})d}}{\sqrt{C^{\dagger}_{\text{sc}}(1-\gamma)^4 N}}$

Table 2: Comparison of the offline RL algorithms in linear MDP setting. With dataset size N, here $\Sigma_{d^{\pi^*}} = \mathbb{E}_{s,a\sim d^{\pi^*}}[\phi(s,a)\phi(s,a)^\top]$, $\Lambda = \mathbb{E}_{s,a\sim\mu}[\phi(s,a)\phi(s,a)^\top]$, Λ_N is an estimate of Λ, $C_{\pi^*,\phi} = \max_{x\in\mathbb{R}^d} (x^\top \Sigma_{d^{\pi^*}} x)/(x^\top \Lambda x)$, $\Sigma^i_{d^{\pi^*}} = \mathbb{E}_{s,a\sim d^{\pi^*}}[(\phi_i(s,a)\mathbb{1}_i)(\phi_i(s,a)\mathbb{1}_i)^\top]$, $\mathbb{1}_i$ is the unit vector in ith dimension, $\phi(s,a) \in \mathbb{R}^d$ is d-dimensional feature vector, and C_{sc} and $C^{\dagger}_{\text{sc}}$ are the sufficient coverage constants satisfying the following random events: C_{sc} is such that it holds with high probability $\Lambda_N \geq I/N + C_{\text{sc}} \cdot \Sigma_{d^{\pi^*}}$ and $C^{\dagger}_{\text{sc}}$ is such that it holds, $\forall i \in [d]$, with high probability $\Lambda_N \geq I/N + C^{\dagger}_{\text{sc}} d \cdot \Sigma^i_{d^{\pi^*}}$.

pessimistic term from the reward estimate, and hence we call it *reward pessimism* in Table 1. They also provide a lower bound on the sample complexity of offline RL algorithms. Li et al. (2022a) also propose a reward pessimism-based offline RL algorithm. They are also able to use an improved clipped concentrability coefficient which is less than the single policy concentrability used in other works. We note that Rashidinejad et al. (2022) and Li et al. (2022a) only study the tabular setting. In the linear function approximation setting, the state-of-the-art algorithms are based on *reward pessimism*, and their sample complexity guarantees depend on the linear feature dimension (Jin et al., 2021; Yin et al., 2022; Xiong et al., 2022).

Our Contributions: (i) We propose a novel offline RL algorithm using the DRL framework, called Distributionally Robust Q-Iteration (DRQI), for the tabular setting. We show that our approach is able to relax the strong assumption of uniform concentrability to a weaker single policy concentrability assumption. We also provide detailed analysis and sample complexity results for DRQI with five commonly used uncertainty sets in DRL: total variation, Wasserstein, Kullback-Leibler, chi-square, and Hellinger uncertainty sets. The comparison with the relevant works is given in Table 1.
(ii) We extend our distributionally robust approach for offline RL to the linear MDP setting and propose the Linear MDP DRQI (LM-DRQI) algorithm. We characterize its sample complexity using only the *sufficient coverage* assumption (Jin et al., 2021) which only requires that the trajectory induced by the optimal policy π^* is covered by the offline data sufficiently well. In particular, we do not require the uniform concentrability assumption. The comparison with the relevant works is given in Table 2.
(iii) We demonstrate the superior performance of our DRQI algorithm through simulation experiments and by comparing it with other state-of-the-art tabular offline RL algorithms. In the partial data coverage setting, DRQI algorithm performs better than the standard dynamic programming approach and performs at par with the state-of-the-art reward pessimism-based offline RL algorithms. In the full coverage setting, DRQI algorithm outperforms the reward pessimism-based offline RL algorithms. We present these in our supplement Panaganti et al. (2023) due to the page limit.
(iv) We believe that establishing a connection between the DRL and offline RL literature is also a contribution of this work. It provides the opportunity to bring the ideas and algorithms from the

DRL to solve the offline RL problem. In particular, we expect that the offline RL problems with large states and action spaces could greatly benefit from this.

We note that our sample complexity result is $\mathcal{O}(\sqrt{|\mathcal{S}|/(1-\gamma)})$ away from the state-of-the-art lower-bound (and the matching upper-bound) in the tabular setting (c.f. Table 1). In the linear MDP setting, our result is comparable to Jin et al. (2021) as long as $C_{\text{sc}} \leq dC_{\text{sc}}^{\dagger}$. However, for a certain class of linear MDPs Jin et al. (2021)'s data coverage assumption implies ours (c.f.Theorem 29) and hence $C_{\text{sc}} = C_{\text{sc}}^{\dagger}$, our result improves over Jin et al. (2021) by $\sqrt{d}$. Our result is not directly comparable with that of Uehara and Sun (2021). We also want to emphasize that Uehara and Sun (2021) do not provide a tractable implementation. However, our LM-DRQI algorithm can use the least squares regression approach Ma et al. (2022) for implementation.

Comparison with Wang et al. (2023b): The closest and recent work Wang et al. (2023b) proposes a similar offline RL algorithm as ours (Algorithm 1). *Wang et al. (2023b) consider a total variation uncertainty set whereas we consider five commonly used uncertainty sets in DRL.* In terms of the sample complexity guarantees, they provide a $\widetilde{\mathcal{O}}(\sqrt{(|\mathcal{S}|C_{\pi^*})/((1-\gamma)^4 N)})$ bound. However, we want to point out that there is a technical error in their application of Hoeffding's inequality to L^1-norm (Wang et al., 2023b, Eq.(10)) (version 1). To emphasize, Hoeffding's inequality (Theorem 6) gives a concentration result for *single-valued random variables*, hence we incur an additional $|\mathcal{S}|$ factor in the concentration of total variation distance (equivalently for L^1-norm) between two *random vectors*. This observation matches the tightness of concentration of empirical distributions under total variation distance (Canonne, 2020, Theorem 1). This technical error makes their bound appear $\sqrt{|\mathcal{S}|}$ better than it should be. If this error is fixed, then their sample complexity results will match ours. Wang et al. (2023b) also derive an improved bound using the Bernstein-based analysis techniques (Li et al., 2022a). Although this bound is optimal, it is only when the sample size N exceeds $\widetilde{\mathcal{O}}(1/((1-\gamma)\mu_{\min}^2))$, where μ is the data generating distribution and $\mu_{\min}$ is its minimal positive value. Hence they get quadratic dependence on $|\mathcal{S}|$ and $|\mathcal{A}|$ for sample complexity when μ is a uniform distribution. Nonetheless, we want to emphasize that their analysis is sophisticated and insightful. Moreover, they only consider the tabular setting, whereas we provide *offline RL algorithms for both the tabular and linear MDP settings* to account for high-dimensional problem settings using function approximation.

2. Preliminaries

Notations: For a set $\mathcal{X}$, we denote its cardinality as $|\mathcal{X}|$. The set of probability distributions over $\mathcal{X}$ is denoted as $\Delta(\mathcal{X})$. For any vector x and positive semidefinite matrix A, $\|x\|_A = \sqrt{x^\top A x}$. Let $\text{Tr}(\cdot)$ denote the trace operator. Denote $\mathbb{1}_i \in \{0,1\}^{d\times 1}$ as a zero-vector but with value 1 at index i. We use $f \leq \mathcal{O}(g)$ to denote $f \leq c \cdot g$ for some universal constants less than 100, and likewise use $f \leq \widetilde{\mathcal{O}}(g)$ to absorb all the universal constants less than 100 and the polylog terms depending on d, N and $1/(1-\gamma)$.

Markov Decision Process (MDP): An MDP is a tuple $(\mathcal{S}, \mathcal{A}, r, P^o, \gamma, d_0)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ the action space, $r : \mathcal{S}\times\mathcal{A} \to [0,1]$ is the reward function, $P^o : \mathcal{S}\times\mathcal{A} \to \Delta(\mathcal{S})$ is the probability transition function (model), γ is the discount factor, and d_0 is the initial state distribution. A stationary (stochastic) policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ specifies a distribution over actions for each state. Each policy $\pi \in \Pi$ induces a discounted occupancy distribution over state-action pairs, denoted as $d^\pi : \mathcal{S}\times\mathcal{A} \to [0,1]$, where $d^\pi(s,a) = (1-\gamma)\sum_{t=0}^{\infty}\gamma^t P_t(s_t = s, a_t = a; \pi)$, and $P_t(s_t = s, a_t = a; \pi)$ denotes the visitation probability of state-action pair (s,a) at time step t, starting at

$s_0 \sim d_0(\cdot)$ and following π on the model P^o. For simplicity, we denote $P_t(s_t = s, a_t = a; \pi)$ by $d_t^\pi(s,a)$. The value of a policy π at state $s \in \mathcal{S}$ is $V_{P^o}^\pi(s) = \mathbb{E}_{\pi,P^o}[\sum_{t=0}^\infty \gamma^t r(s_t,a_t) \mid s_0 = s]$, where $a_t \sim \pi(\cdot|s_t)$ and $s_{t+1} \sim P^o_{s_t,a_t}$. Similarly, we define the Q-value of a policy as $Q_{P^o}^\pi(s,a) = \mathbb{E}_{\pi,P^o}\left[\sum_{t=0}^\infty \gamma^t r_t \mid s_0 = s, a_0 = a\right]$. We sometimes denote d^π as $d_{P^o}^\pi$ making its dependence on the model P^o clearer.

Offline RL: In offline RL, we only have access to a pre-collected offline dataset consisting of N samples: $\mathcal{D} = \{(s_i, a_i, r_i, s_i')\}_{i=1}^N$, where $r_i = r(s_i,a_i)$ and $s_i' \sim P^o_{s_i,a_i}$. We assume that (s_i,a_i) pairs are generated i.i.d. by following a data generating (behavior) distribution $\mu \in \Delta(\mathcal{S} \times \mathcal{A})$. The goal of offline RL is to learn a *good* policy $\widehat{\pi}$ close to an optimal policy π^* of MDP M^o based on the offline data $\mathcal{D}$. More formally, for a prescribed accuracy level ϵ, we seek to find an ϵ-optimal policy $\widehat{\pi}$ satisfying

$$\mathbb{E}_{s_0\sim d_0}[V^{\pi^*}(s_0) - \mathbb{E}_{\mathcal{D}}[V^{\widehat{\pi}}(s_0)]] \leq \epsilon, \tag{1}$$

with high probability using an offline dataset $\mathcal{D}$ containing as few samples as possible.

Analysis of offline RL algorithm crucially depends on the *data coverage* assumption, which is quantified using the *concentrability coefficient*. For a given policy π, the concentrability coefficient C_π is defined as $C_\pi = \max_{(s,a)\in\mathcal{S}\times\mathcal{A}} d^\pi(s,a)/\,\mu(s,a)$. Most of the past theoretical works on offline RL use the strong assumption of bounded *uniform concentrability* (Munos and Szepesvári, 2008), defined as $C_u = \sup_\pi C_\pi$. Munos and Szepesvári (2008) proposed the fitted Q-iteration algorithm and gave offline RL guarantees under uniform concentrability. Recently, some works have proposed offline RL algorithms using the idea of pessimism and showed that the uniform concentrability can be relaxed to a *single concentrability* assumption, i.e., C_{π^*} is bounded (Uehara and Sun, 2021; Rashidinejad et al., 2022; Li et al., 2022a). *We also make only the same single concentrability assumption in this work.*

Robust Markov Decision Process (RMDP): The RMDP formulation considers a set of models called uncertainty set, denoted as $\mathcal{P}$. We assume that $\mathcal{P}$ satisfies the standard (s,a)*-rectangularity condition* (Iyengar, 2005). An RMDP can be specified as $(\mathcal{S}, \mathcal{A}, r, \mathcal{P}, \gamma, d_0)$ in which

$$\mathcal{P} = \otimes_{(s,a)\in\mathcal{S}\times\mathcal{A}} \mathcal{P}_{s,a}, \tag{2}$$

$$\mathcal{P}_{s,a} = \{P_{s,a} \in \Delta(\mathcal{S}) \;:\; D(P_{s,a}, P^o_{s,a}) \leq \rho_{s,a}\}, \tag{3}$$

where $D(\cdot,\cdot)$ is a distance metric between two probability distributions and $\rho_{s,a} > 0$ is the radius of the uncertainty set. In other words, $\mathcal{P}$ is the set of all models around P^o within a particular distance.

The *robust value function* $V_\mathcal{P}^\pi$ corresponding to a policy π and the *optimal robust value function* $V_\mathcal{P}^*$ are defined as (Iyengar, 2005; Nilim and El Ghaoui, 2005)

$$V_\mathcal{P}^\pi = \inf_{P\in\mathcal{P}} V_P^\pi, \qquad V_\mathcal{P}^* = \sup_\pi \inf_{P\in\mathcal{P}} V_P^\pi. \tag{4}$$

An *optimal robust policy* $\pi_\mathcal{P}^*$ is such that $V^{\pi_\mathcal{P}^*} = V_\mathcal{P}^*$. It is known that there exists a stationary and deterministic optimal policy (Iyengar, 2005) for the RMDP. The *robust Bellman operator* T is defined as (Iyengar, 2005) $(TQ)(s,a) =$

$$r(s,a) + \gamma \inf_{P_{s,a}\in\mathcal{P}_{s,a}} \mathbb{E}_{s'\sim P_{s,a}}[\max_b Q(s',b)]. \tag{5}$$

It is known that T is a contraction mapping in the infinity norm and hence it has a unique fixed point $Q_\mathcal{P}^*$ with $V_\mathcal{P}^*(s) = \max_a Q_\mathcal{P}^*(s,a)$ and $\pi_\mathcal{P}^*(s) = \arg\max_a Q_\mathcal{P}^*(s,a)$ (Iyengar, 2005). The robust

Q-Iteration can now be defined using the robust Bellman operator as $Q_{k+1} = TQ_k$. Since T is a contraction, it follows that $Q_k \to Q^*_{\mathcal{P}}$. So, robust Q-Iteration can be used to compute $Q^*_{\mathcal{P}}$ and $\pi^*_{\mathcal{P}}$ in the tabular setting with a known uncertainty set $\mathcal{P}$.

Recently, many works have proposed robust RL algorithms for solving the RMDP problem using only the data from the nominal model P^o around which $\mathcal{P}$ is defined (Tamar et al., 2014; Roy et al., 2017; Panaganti and Kalathil, 2021; Panaganti et al., 2022; Panaganti and Kalathil, 2022; Xu* et al., 2023; Shi and Chi, 2022; Wang and Zou, 2021; Ma et al., 2022; Wang and Zou, 2022; Kumar et al., 2023; Li et al., 2022b; Wang et al., 2023a; Grand-Clément and Kroer, 2021).

Remark 1 (Difference between the offline RL objective and robust RL objective) *We want to emphasize that the mathematical objectives of offline RL and robust RL are fundamentally different. More precisely, the goal of offline RL is to learn the optimal value/policy for the model P^o, i.e. $\max_\pi V^\pi_{P^o}$. In contrast, the goal of the robust RL is to learn the optimal robust policy as a max-min solution w.r.t. uncertainty set $\mathcal{P}$, i.e., $\max_\pi \min_{P\in\mathcal{P}} V^\pi_{P^o}$. Since the goal of this work is to develop an offline RL algorithm with provable guarantees, we compare our theoretical and empirical results only with the state-of-the-art offline RL algorithms, not with the robust RL algorithms works mentioned above.*

3. Distributionally Robust Q-Iteration (DRQI) Algorithm

In this section, we propose our DRQI algorithm to solve the offline RL problem in the tabular setting and provide its theoretical guarantees.

First denote $N(s,a) = \sum_{i=1}^N \mathbb{1}\{(s_i,a_i) = (s,a)\}$ and $N(s,a,s') = \sum_{i=1}^N \mathbb{1}\{(s_i,a_i,s'_i) = (s,a,s')\}$. We then construct an empirical estimate of P^o as

$$\widehat{P}^o_{s,a}(s') = \frac{N(s,a,s')\mathbb{1}\{N(s,a)\geq 1\}}{N(s,a)} + \frac{\mathbb{1}\{N(s,a)=0\}}{|\mathcal{S}|}. \tag{6}$$

We also consider the add-L estimate (Bhattacharyya et al., 2021; Arora et al., 2023) of P^o given by $\widetilde{P}^o_{s,a}(s') = (N(s,a,s')+L)/(N(s,a)+L|\mathcal{S}|)$, where the value of L is defined later. Following the uncertainty set definition (c.f. Eq. (2)-Eq. (3)), we construct the empirical uncertainty set $\widehat{\mathcal{P}}$ around $\widehat{P}^o$ or $\widetilde{P}^o$ as, $\widehat{\mathcal{P}} = \bigotimes_{s,a} \widehat{\mathcal{P}}_{s,a}$, where

$$\widehat{\mathcal{P}}_{s,a} = \{P \in \Delta(\mathcal{S}) : D(P, \widehat{P}^o_{s,a} \text{ or } \widetilde{P}^o_{s,a}) \leq \rho_{s,a}\}. \tag{7}$$

Similarly (c.f. Eq. (5)), we can define the *empirical robust Bellman operator* $\widehat{T}$ as $(\widehat{T}Q)(s,a) =$

$$r(s,a) + \gamma \inf_{P_{s,a}\in\widehat{\mathcal{P}}_{s,a}} \mathbb{E}_{s'\sim P_{s,a}}[\max_b Q(s',b)]. \tag{8}$$

Note that for $\rho_{s,a} = 0$, $\widehat{T}$ is the same as the standard (non-robust) empirical Bellman operator. Thus, the empirical Q-value iteration $Q_{k+1} = \widehat{T}Q_k$ will give an approximately optimal Q-value function under the standard generative model assumption where there are $N(s,a) = N$ next-state samples from each (s,a) pairs (Haskell et al., 2016; Kalathil et al., 2021; Sidford et al., 2018). However, since the data is generated according to a behavior policy in the offline RL, the generative model assumption is not valid here. On the other hand, for a fixed $\rho_{s,a} > 0$, the update $Q_{k+1} = \widehat{T}Q_k$ is exactly equal to empirical robust Q-iteration, and it will converge to an approximately optimal

Algorithm 1 Distributionally Robust Q-Iteration (DRQI)

Input: Offline data $\mathcal{D} = (s_i, a_i, r_i, s_i')_{i=1}^N$, Confidence level $\delta \in (0,1)$
Initialize: $Q_0 \equiv 0$
Construct the empirical estimate $\widehat{P}^o$ as in Eq. (6)
for $k = 0, \cdots, K-1$: Compute $Q_{k+1} = \widehat{T} Q_k$ (8)
Output: $\pi_K = \arg\max_a Q_K(s,a)$

robust Q-function corresponding to the RMDP uncertainty set specified by the $\rho_{s,a}$ values (Xu* et al., 2023; Shi and Chi, 2022).

The key idea behind our algorithm is to use the update $Q_{k+1} = \widehat{T} Q_k$ as a DRL style approximate Q-iteration. To see this, recall the standard DRL problem (Duchi and Namkoong, 2021; Chen et al., 2020): $\max_\theta \min_{q \in \mathcal{Q}} \mathbb{E}_{x \sim q}[f(x;\theta)]$, where f is a function to be maximized w.r.t. a parameter θ and $\mathcal{Q}$ is an uncertainty set for the probability distribution. The nomenclature 'distributionally robust' is due to the term $\min_{q \in \mathcal{Q}}$ in the objective. Now, in our case, the minimization over the uncertainty set $\widehat{\mathcal{P}}$ in the definition of $\widehat{T}$, i.e., $\inf_{P_{s,a} \in \widehat{\mathcal{P}}_{s,a}}$, also represents this distributionally robust objective. Observing that the degree of the robustness depends on the radius of the uncertainty set $\rho_{s,a}$, we propose to control this robustness by choosing an appropriate value for $\rho_{s,a}$ depending on the offline data $\mathcal{D}$. In particular, we will choose $\rho_{s,a} = \min\left(c_1, c_2/\sqrt{N(s,a)}\right)$ (where c_1 and c_2 are problem-dependent constants to be specified later) that quantify the radius of the uncertainty set caused by the insufficiency in samples. Moreover, this idea also allows us to bring algorithms from the DRL and robust RL literature to solve offline RL problems, hence bridging these areas.

In this work, we consider five uncertainty sets corresponding to five different distance metrics $D(\cdot,\cdot)$. We also fix a confidence level $\delta \in (0,1)$ in the following.

Total variation (TV) uncertainty set ($\widehat{\mathcal{P}}^{\text{tv}}$) : We define $\widehat{\mathcal{P}}^{\text{tv}} = \otimes \widehat{\mathcal{P}}^{\text{tv}}_{s,a}$, where $\widehat{\mathcal{P}}^{\text{tv}}_{s,a}$ is as in (7) with the empirical estimator $\widehat{P}^o_{s,a}$, the total variation distance $D_{\text{TV}}(P, \widehat{P}^o_{s,a}) = (1/2)\|P - \widehat{P}^o_{s,a}\|_1$, and radius

$$\rho_{s,a} = 1 \wedge \sqrt{\frac{\max\{|\mathcal{S}|, 2\log(2|\mathcal{S}||\mathcal{A}|/\delta)\}}{N(s,a)}} \mathbb{1}\{N(s,a) \geq 1\}. \tag{9}$$

We describe the remaining uncertainty sets and the corresponding results in our supplement Panaganti et al. (2023) due to the page limit. We assume that the reward function is known to focus on the key idea of distributional robustness. This simplification is made without loss of generality, as we can model similar sets $\mathcal{P}$ or $\widehat{\mathcal{P}}$ for the reward also (Si et al., 2020; Zhou et al., 2021).

Our DRQI algorithm is summarized in Algorithm 1. We discuss practical solutions and experiments in our supplement Panaganti et al. (2023). We now present the sample complexity of DRQI with TV uncertainty set, and a proof sketch. Please note that we do obtain sample complexities of the same order for all other uncertainty sets (see Theorems 19 to 22 in the supplement).

Theorem 2 *Let π_K be the DRQI policy after K iterations under the TV uncertainty set $\widehat{\mathcal{P}}^{\text{tv}}$. If the total number of samples $N \geq \mathcal{O}\left(C_{\pi^*} \max\{|\mathcal{S}|^2, 2\log(2|\mathcal{S}|^2|\mathcal{A}|/\delta)\}/(\epsilon^2(1-\gamma)^4)\right)$, then $\mathbb{E}_{s_0 \sim d_0}[V^{\pi^*}(s_0) - \mathbb{E}_{\mathcal{D}}[V^{\pi_K}(s_0)]] \leq \epsilon$ with probability at least $1-\delta$ and a sufficiently large K.*

Proof [Proof Sketch] Denoting $\widehat{\mathcal{P}}^{\text{tv}}$ simply as $\widehat{\mathcal{P}}$, we first write $V^{\pi^*}_{P^o}(s_0) - V^{\pi_K}_{P^o}(s_0) = (V^{\pi^*}_{P^o}(s_0) - V^{\pi_K}_{\widehat{\mathcal{P}}}(s_0)) + (V^{\pi_K}_{\widehat{\mathcal{P}}}(s_0) - V^{\pi_K}_{P^o}(s_0))$, where $V^{\pi_K}_{\widehat{\mathcal{P}}} = \inf_{P \in \widehat{\mathcal{P}}} V^{\pi_K}_P$ is the robust value of policy π_K

corresponding to the uncertainty set $\widehat{\mathcal{P}}$. In Theorem 14 we show that, with the $\rho_{s,a}$ as specified above, $P^o \in \widehat{\mathcal{P}}^{\text{tv}}$ with probability at least $1-\delta$. So, by definition of the robust value function, the second term $(V_{\widehat{\mathcal{P}}}^{\pi_K}(s_0) - V_{P^o}^{\pi_K}(s_0))$ is negative.

To bound the first term, we decompose it as $(V_{P^o}^{\pi^*}(s_0) - V_{\widehat{\mathcal{P}}}^{\pi_K}(s_0)) = (V_{P^o}^{\pi^*}(s_0) - V_{\widehat{\mathcal{P}}}^{\widehat{\pi}^*}(s_0)) + (V_{\widehat{\mathcal{P}}}^{\widehat{\pi}^*}(s_0) - V_{\widehat{\mathcal{P}}}^{\pi_K}(s_0))$, where $\widehat{\pi}^* = \arg\max_\pi V_{\widehat{\mathcal{P}}}^{\pi}$ is the optimal robust policy w.r.t. $\widehat{\mathcal{P}}$. Then, due to the contraction property of the robust Bellman operator, $(V_{\widehat{\mathcal{P}}}^{\widehat{\pi}^*}(s_0) - V_{\widehat{\mathcal{P}}}^{\pi_K}(s_0))$ will converge to zero exponentially in K.

Bounding $(V_{P^o}^{\pi^*}(s_0) - V_{\widehat{\mathcal{P}}}^{\widehat{\pi}^*}(s_0))$ is more technical. The key idea is to first note that $D_{\text{TV}}(P_{s,\pi^*(s)}, P^o_{s,\pi^*(s)}) \le 2\rho_{s,a}$ for any $P \in \widehat{\mathcal{P}}$, by Theorem 14 and definition of $\widehat{\mathcal{P}}$. Now, unrolling along the trajectory generated by π^* on P^o and using the form of $\rho_{s,a}$, we can get an upper bound in terms of $\mathbb{E}_{s\sim d^{\pi^*}}[1/\sqrt{N(s,\pi^*(s))}]$. We will then express $N(s,\pi^*(s))$ in terms of $N\mu(s,\pi^*(s))$ using Theorem 5, and then use a change of measure argument to get the final bound in terms of single concentrability coefficient C_{π^*}. ■

4. Linear-MDP Distributionally Robust Q-Iteration (LM-DRQI) Algorithm

In this section, we propose our LM-DRQI algorithm to solve the offline RL problem in the linear MDP setting, and give its sample complexity guarantees.

Definition 3 (Linear MDP (Jin et al., 2020)) *We say an MDP $M = (\mathcal{S}, \mathcal{A}, r, P, \gamma)$ is a linear MDP with a known feature map $\phi : \mathcal{S}\times\mathcal{A} \to \mathbb{R}^d$, if there exists d unknown probability measures $\nu = (\nu_1(\cdot), \dots, \nu_d(\cdot))$ over $\mathcal{S}$ and an unknown vector $\theta \in \mathbb{R}^d$, such that for any $(s,a) \in \mathcal{S}\times\mathcal{A}$, we have $P_{s,a} = \langle \phi(s,a), \nu(\cdot)\rangle$, $r(s,a) = \langle \phi(s,a), \theta\rangle$.*

Similar to the tabular setting, here also we assume that the reward function (equivalently θ) is known. We make the following assumption.

Assumption 1 *Let $M = (\mathcal{S}, \mathcal{A}, r, P^o, \gamma)$ be a linear MDP with a known feature map ϕ and unknown measure ν^o. We assume that $\phi_i(s,a) \ge 0$ and $\|\phi(s,a)\|_2 \le 1$ for all $(s,a) \in \mathcal{S}\times\mathcal{A}$ and $i \in [d]$. We also assume that $\Lambda = \mathbb{E}_{s,a\sim\mu}[\phi(s,a)\phi(s,a)^\top]$ and $\Sigma_{d^{\pi^*}}^{(i,j)} = \mathbb{E}_{s,a\sim d^{\pi^*}}[(\phi_i(s,a)\mathbb{1}_i)(\phi_j(s,a)\mathbb{1}_j)^\top]$ for all $i,j \in [d]$ are positive semi-definite matrices.*

We use the d-rectangularity uncertainty set construction which exploits the linear structure (Ma et al., 2022). Instead of focusing on the set of all models around P^o, we consider only the set of linear models around P^o. This is achieved indirectly by considering an uncertainty set around ν^o using an integral probability metric (IPM) (Müller, 1997) and translating that to an uncertainty set around P^o through the known feature vector ϕ. More precisely, the d-rectangularity uncertainty set $\mathcal{P}$ is defined as

$$\begin{aligned}
\mathcal{P} &= \{P : P_{s,a}(s') = \sum_{i\in[d]} \phi_i(s,a)\nu_i(s'), \nu_i \in \mathcal{M}_i, \forall i \in [d]\},\\
\mathcal{M}_i &= \{\nu_i : D_{\text{IPM}}(\nu_i, \nu_i^o) \le \rho_i\}, \text{ where,}
\end{aligned} \tag{10}$$

$D_{\text{IPM}}(p,q) = \sup_{V\in\mathcal{V}} |\int_s (p(s) - q(s))V(s)\mathrm{d}\,s|$, and $\mathcal{V} = \{V(\cdot) = \max_a \phi^\top(\cdot,a)w : w \in \mathbb{R}^d, \|w\|_2 \le 1/(1-\gamma)\}$.

It is straight forward to show that the optimal robust value function is linear w.r.t. ϕ under the d-rectangularity uncertainty set. Moreover, we can also show that the robust Bellman operator (Eq. (5)) can be written as

$$TQ(s,a) = r(s,a) + \gamma \sum_{i\in[d]} \phi_i(s,a) \min_{\nu_i \in \mathcal{M}_i} \mathbb{E}_{s'\sim\nu_i}(\max_b Q(s',b)). \tag{11}$$

We can get an empirical estimate $\widehat{P}^o$ of P^o with ridge linear regression using the offline data (Agarwal et al., 2019, Section 8.3) as $\widehat{P}^o_{s,a}(s') = \phi(s,a)^\top \widehat{\nu}^o(s')$, where $\widehat{\nu}^o(s') = \frac{1}{N}\sum_{i=1}^N \Lambda_N^{-1} \phi(s_i,a_i)\mathbb{1}\{s' = s'_i\}$, $\Lambda_N = \frac{\lambda}{N} I + \frac{1}{N}\sum_{i=1}^N \phi(s_i,a_i)\phi(s_i,a_i)^\top$ and λ is a constant. We construct an estimate $\widehat{\mathcal{M}}_i$ of $\mathcal{M}_i$ by replacing unknown ν_i^o with its estimate $\widehat{\nu}_i^o$. Similarly, we construct the empirical uncertainty set $\widehat{\mathcal{P}}$ by replacing $\mathcal{M}_i$ by $\widehat{\mathcal{M}}_i$. We fix the radius

$$\rho_i = \frac{c_1 \log(Nd/((1-\gamma)\delta))}{1-\gamma} \sqrt{\frac{d}{N}} \sqrt{\Lambda_N^{-1}(i,i)}. \tag{12}$$

We can now define the empirical robust Bellman operator $\widehat{T}$ exactly as in Eq. (11), but by replacing $\mathcal{M}_i$ by its estimate $\widehat{\mathcal{M}}_i$. Our LM-DRQI algorithm then follows the same procedure as our DRQI algorithm using this $\widehat{T}$. We omit rewriting the algorithm procedure due to page limitation.

We make the following assumption that specifies coverage requirements to provide offline RL guarantees.

Assumption 2 (Sufficient coverage assumption) *For all $i \in [d]$, with probability $1-\delta$, it holds $\Lambda_N \geq (1/N)I + C^\dagger_{\text{sc}} \cdot d \cdot \Sigma^i_{d^{\pi^*}_{Po}}$, where $\Sigma^i_{d^{\pi^*}_{Po}} = \mathbb{E}_{s,a\sim d^{\pi^*}}[(\phi_i(s,a)\mathbb{1}_i)(\phi_i(s,a)\mathbb{1}_i)^\top]$.*

The sufficient coverage assumption was originally used by Jin et al. (2021) for showing that pessimism-based offline RL algorithms can learn optimal policy without assuming the uniform concentrability ($\text{rank}(\Lambda) = d$ (Wang et al., 2021) in linear MDPs). The *sufficient coverage* assumption only requires that the trajectory induced by the optimal policy π^* is covered by the offline data sufficiently well. The assumption we use is from Ma et al. (2022), which addressed the robust RL problem using offline data. This assumption stipulate sufficient coverage in each dimension $i \in [d]$. We now give the sample complexity of our LM-DRQI algorithm.

Theorem 4 *Let π_K be the LM-DRQI policy after K iterations. Let Assumption 2 hold. If the total number of samples $N \geq N_{\text{IPM}}$, where $N_{\text{IPM}} = \widetilde{\mathcal{O}}(d \cdot \text{rank}(\Sigma_{d^{\pi^*}_{Po}})/(C^\dagger_{\text{sc}}(1-\gamma)^4\epsilon^2))$, then $\mathbb{E}_{s_0\sim d_0}[V^{\pi^*}(s_0) - \mathbb{E}_{\mathcal{D}}[V^{\pi_K}(s_0)]] \leq \epsilon$ with probability at least $1-\delta$.*

More detailed theorem statements and proofs are in Panaganti et al. (2023) due to the page limit.

5. Conclusion

In this work, we presented offline RL algorithms for the tabular and linear MDP setting using the framework of DRL. We characterized the sample complexity of these algorithms only using the single policy concentrability assumption. We also demonstrated the superior performance of our proposed algorithm through simulation experiments. In the future, we plan to extend these results to general function approximation settings (linear and general function approximations to the value function space, moving beyond linear MDPs) to handle continuous state-action space problems.

Acknowledgments

This work was supported in part by the National Science Foundation (NSF) grants NSF-CAREER-EPCN-2045783 and NSF ECCS 2038963. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the sponsoring agencies.

References

Alekh Agarwal, Nan Jiang, Sham M Kakade, and Wen Sun. Reinforcement learning: Theory and algorithms. *CS Dept., UW Seattle, Seattle, WA, USA, Tech. Rep*, 2019.

András Antos, Csaba Szepesvári, and Rémi Munos. Learning near-optimal policies with Bellman-residual minimization based fitted policy iteration and a single sample path. *Machine Learning*, 71(1):89–129, 2008.

Vipul Arora, Arnab Bhattacharyya, Clément L Canonne, and Joy Qiping Yang. Near-optimal degree testing for bayes nets. *arXiv preprint arXiv:2304.06733*, 2023.

A. Basu, H. Shioya, and C. Park. *Statistical Inference: The Minimum Distance Approach*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. CRC Press, 2011. ISBN 9781420099669.

Arnab Bhattacharyya, Sutanu Gayen, Eric Price, and NV Vinodchandran. Near-optimal learning of tree-structured distributions by chow-liu. In *Proceedings of the 53rd annual acm SIGACT symposium on theory of computing*, pages 147–160, 2021.

Jose Blanchet, Yang Kang, and Karthyek Murthy. Robust wasserstein profile inference and applications to machine learning. *Journal of Applied Probability*, 56(3):830–857, 2019.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016.

Jacob Buckman, Carles Gelada, and Marc G Bellemare. The importance of pessimism in fixed-dataset policy optimization. In *International Conference on Learning Representations*, 2021.

Clément L Canonne. A short note on learning discrete distributions. *arXiv preprint arXiv:2002.11457*, 2020.

Jonathan Chang, Masatoshi Uehara, Dhruv Sreenivas, Rahul Kidambi, and Wen Sun. Mitigating covariate shift in imitation learning via offline data with partial coverage. *Advances in Neural Information Processing Systems*, 34:965–979, 2021.

Jinglin Chen and Nan Jiang. Information-theoretic considerations in batch reinforcement learning. In *International Conference on Machine Learning*, pages 1042–1051, 2019.

Ruidi Chen, Ioannis Ch Paschalidis, et al. Distributionally robust learning. *Foundations and Trends® in Optimization*, 4(1-2):1–243, 2020.

Thomas M Cover and Joy A Thomas. Information theory and the stock market. *Elements of Information Theory. Wiley Inc., New York*, pages 543–556, 1991.

Steven Diamond and Stephen Boyd. Cvxpy: A python-embedded modeling language for convex optimization. *The Journal of Machine Learning Research*, 17(1):2909–2913, 2016.

John Duchi, Tatsunori Hashimoto, and Hongseok Namkoong. Distributionally robust losses for latent covariate mixtures, 2022.

John C. Duchi and Hongseok Namkoong. Learning models with uniform performance via distributionally robust optimization. *The Annals of Statistics*, 49(3):1378 – 1406, 2021. doi: 10.1214/20-AOS2004.

Richard M Dudley. *Real analysis and Probability*. Cambridge University Press, 2002.

Amir-massoud Farahmand, Csaba Szepesvári, and Rémi Munos. Error propagation for approximate policy and value iteration. *Advances in Neural Information Processing Systems*, 23, 2010.

Farzan Farnia and David Tse. A minimax approach to supervised learning. In *Advances in Neural Information Processing Systems*, 2016.

Scott Fujimoto and Shixiang Shane Gu. A minimalist approach to offline reinforcement learning. *Advances in neural information processing systems*, 34:20132–20145, 2021.

Scott Fujimoto, David Meger, and Doina Precup. Off-policy deep reinforcement learning without exploration. In *International Conference on Machine Learning*, pages 2052–2062, 2019.

Julien Grand-Clément and Christian Kroer. Scalable first-order methods for robust mdps. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12086–12094, 2021.

William B Haskell, Rahul Jain, and Dileep Kalathil. Empirical dynamic programming. *Mathematics of Operations Research*, 41(2):402–429, 2016.

Garud N Iyengar. Robust dynamic programming. *Mathematics of Operations Research*, 30(2): 257–280, 2005.

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

Ying Jin, Zhuoran Yang, and Zhaoran Wang. Is pessimism provably efficient for offline rl? In *International Conference on Machine Learning*, pages 5084–5096. PMLR, 2021.

Dileep Kalathil, Vivek S Borkar, and Rahul Jain. Empirical Q-Value Iteration. *Stochastic Systems*, 11(1):1–18, 2021.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pages 5637–5664. PMLR, 2021.

Ilya Kostrikov, Rob Fergus, Jonathan Tompson, and Ofir Nachum. Offline reinforcement learning with fisher divergence critic regularization. In *International Conference on Machine Learning*, pages 5774–5783. PMLR, 2021.

Daniel Kuhn, Peyman Esfahani, Viet Anh Nguyen, and Soroosh Shafieezadeh-Abadeh. Wasserstein distributionally robust optimization: Theory and applications in machine learning. In *Operations research & management science in the age of analytics*, pages 130–166. Informs, 2019.

Aviral Kumar, Justin Fu, Matthew Soh, George Tucker, and Sergey Levine. Stabilizing off-policy q-learning via bootstrapping error reduction. In *Advances in Neural Information Processing Systems*, pages 11784–11794, 2019.

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 33:1179–1191, 2020.

Navdeep Kumar, Esther Derman, Matthieu Geist, Kfir Yehuda Levy, and Shie Mannor. Policy gradient for rectangular robust markov decision processes. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

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

Jing Lei. Convergence and concentration of empirical measures under Wasserstein distance in unbounded functional spaces. *Bernoulli*, 26(1):767 – 798, 2020.

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

Daniel Levy, Yair Carmon, John C Duchi, and Aaron Sidford. Large-scale methods for distributionally robust optimization. *Advances in Neural Information Processing Systems*, 33:8847–8860, 2020.

Gen Li, Laixi Shi, Yuxin Chen, Yuejie Chi, and Yuting Wei. Settling the sample complexity of model-based offline reinforcement learning. *arXiv preprint arXiv:2204.05275*, 2022a.

Yan Li, Tuo Zhao, and Guanghui Lan. First-order policy optimization for robust markov decision process. *arXiv preprint arXiv:2209.10579*, 2022b.

Peng Liao, Zhengling Qi, Runzhe Wan, Predrag Klasnja, and Susan A Murphy. Batch policy learning in average reward markov decision processes. *The Annals of Statistics*, 50(6):3364–3387, 2022.

Xiaoteng Ma, Zhipeng Liang, Li Xia, Jiheng Zhang, Jose Blanchet, Mingwen Liu, Qianchuan Zhao, and Zhengyuan Zhou. Distributionally robust offline reinforcement learning with linear function approximation. *arXiv preprint arXiv:2209.06620*, 2022.

Peyman Mohajerin Esfahani and Daniel Kuhn. Data-driven distributionally robust optimization using the wasserstein metric: performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1-2):115–166, 2018.

Alfred Müller. Integral probability metrics and their generating classes of functions. *Advances in applied probability*, 29(2):429–443, 1997.

Rémi Munos. Performance bounds in l_p-norm for approximate value iteration. *SIAM journal on control and optimization*, 46(2):541–561, 2007.

Rémi Munos and Csaba Szepesvári. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9(27):815–857, 2008.

Arnab Nilim and Laurent El Ghaoui. Robust control of Markov decision processes with uncertain transition matrices. *Operations Research*, 53(5):780–798, 2005.

Kishan Panaganti and Dileep Kalathil. Robust reinforcement learning using least squares policy iteration with provable performance guarantees. In *International Conference on Machine Learning (ICML)*, pages 511–520, 2021.

Kishan Panaganti and Dileep Kalathil. Sample complexity of robust reinforcement learning with a generative model. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 9582–9602, 2022.

Kishan Panaganti, Zaiyan Xu, Dileep Kalathil, and Mohammad Ghavamzadeh. Robust reinforcement learning using offline data. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Kishan Panaganti, Zaiyan Xu, Dileep Kalathil, and Mohammad Ghavamzadeh. Bridging distributionally robust learning and offline rl: An approach to mitigate distribution shift and partial data coverage. *arXiv preprint arXiv:2310.18434*, 2023.

Xue Bin Peng, Marcin Andrychowicz, Wojciech Zaremba, and Pieter Abbeel. Sim-to-real transfer of robotic control with dynamics randomization. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 3803–3810. IEEE, 2018.

Joaquin Quinonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. *Dataset Shift in Machine Learning*. MIT Press, 2022.

Paria Rashidinejad, Banghua Zhu, Cong Ma, Jiantao Jiao, and Stuart Russell. Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *IEEE Transactions on Information Theory*, 68(12):8156–8196, 2022.

Alexander Robey, Hamed Hassani, and George J Pappas. Model-based robust deep learning: Generalizing to natural, out-of-distribution data. *arXiv preprint arXiv:2005.10247*, 2020.

Aurko Roy, Huan Xu, and Sebastian Pokutta. Reinforcement learning under model mismatch. In *Advances in Neural Information Processing Systems*, pages 3043–3052, 2017.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Laixi Shi and Yuejie Chi. Distributionally robust model-based offline reinforcement learning with near-optimal sample complexity. *arXiv preprint arXiv:2208.05767*, 2022.

Laixi Shi, Gen Li, Yuting Wei, Yuxin Chen, and Yuejie Chi. Pessimistic q-learning for offline reinforcement learning: Towards optimal sample complexity. In *International Conference on Machine Learning*, pages 19967–20025. PMLR, 2022.

Nian Si, Fan Zhang, Zhengyuan Zhou, and Jose Blanchet. Distributionally robust policy evaluation and learning in offline contextual bandits. In *International Conference on Machine Learning*, pages 8884–8894, 2020.

Aaron Sidford, Mengdi Wang, Xian Wu, Lin Yang, and Yinyu Ye. Near-optimal time and sample complexities for solving markov decision processes with a generative model. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2018.

Aviv Tamar, Shie Mannor, and Huan Xu. Scaling up robust mdps using function approximation. In *International Conference on Machine Learning*, pages 181–189, 2014.

Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. *Advances in Neural Information Processing Systems*, 33:18583–18599, 2020.

Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 23–30, 2017.

Masatoshi Uehara and Wen Sun. Pessimistic model-based offline reinforcement learning under partial coverage. In *International Conference on Learning Representations*, 2021.

Cédric Villani et al. *Optimal transport: old and new*, volume 338. Springer, 2009.

Qiuhao Wang, Chin Pang Ho, and Marek Petrik. Policy gradient in robust MDPs with global convergence guarantee. In *Proceedings of the 40th International Conference on Machine Learning*. PMLR, 2023a.

Ruosong Wang, Dean Foster, and Sham M. Kakade. What are the statistical limits of offline RL with linear function approximation? In *International Conference on Learning Representations*, 2021.

Yue Wang and Shaofeng Zou. Online robust reinforcement learning with model uncertainty. *Advances in Neural Information Processing Systems*, 34:7193–7206, 2021.

Yue Wang and Shaofeng Zou. Policy gradient method for robust reinforcement learning. In *International Conference on Machine Learning*, pages 23484–23526. PMLR, 2022.

Yue Wang, Jinjun Xiong, and Shaofeng Zou. Achieving the asymptotically optimal sample complexity of offline reinforcement learning: A dro-based approach. *arXiv preprint arXiv:2305.13289v1*, 2023b.

Lilian Weng. Domain randomization for sim2real transfer. *lilianweng.github.io*, 2019.

Wolfram Wiesemann, Daniel Kuhn, and Berç Rustem. Robust Markov decision processes. *Mathematics of Operations Research*, 38(1):153–183, 2013.

Wei Xiong, Han Zhong, Chengshuai Shi, Cong Shen, Liwei Wang, and Tong Zhang. Nearly minimax optimal offline reinforcement learning with linear function approximation: Single-agent mdp and markov game. In *The Eleventh International Conference on Learning Representations*, 2022.

Zaiyan Xu*, Kishan Panaganti*, and Dileep Kalathil. Improved sample complexity bounds for distributionally robust reinforcement learning. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*. Conference on Artificial Intelligence and Statistics, 2023.

Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *arXiv preprint arXiv:2110.11334*, 2021.

Ming Yin, Yaqi Duan, Mengdi Wang, and Yu-Xiang Wang. Near-optimal offline reinforcement learning with linear representation: Leveraging variance information with pessimism. In *International Conference on Learning Representations*, 2022.

Tianhe Yu, Garrett Thomas, Lantao Yu, Stefano Ermon, James Zou, Sergey Levine, Chelsea Finn, and Tengyu Ma. Mopo: Model-based offline policy optimization. In *Advances in Neural Information Processing Systems*, 2020.

Zhengqing Zhou, Qinxun Bai, Zhengyuan Zhou, Linhai Qiu, Jose Blanchet, and Peter Glynn. Finite-sample regret bound for distributionally robust offline tabular reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pages 3331–3339, 2021.