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

Kishan Panaganti [1] **Zaiyan Xu** [2] **Dileep Kalathil** [2] **Mohammad Ghavamzadeh** [3]

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.

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 & 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 & Szepesvári, 2008; Farahmand et al., 2010; Chen & 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) and addressed some of the issues in offline RL, including replacing uniform concentrability with the more relaxed *single policy concentrability* assumption (Uehara & 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 & 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 distribution-

[1]California Institute of Technology [2]Texas A&M University [3]Amazon. Correspondence to: Kishan Panaganti <kpb@caltech.edu>.

Workshop on Foundations of Reinforcement Learning and Control at the 41^{st} International Conference on Machine Learning, Vienna, Austria.

ally robust learning (DRL) framework to address the issue of distribution shift in various settings (Duchi & 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 practical and scalable algorithms (Chen et al., 2020; Levy et al., 2020; Mohajerin Esfahani & 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 & Kalathil, 2021; Panaganti et al., 2022; Panaganti & Kalathil, 2022; Xu* et al., 2023; Shi & Chi, 2022; Ma et al., 2022; Wang & 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 & 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, 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 problems with large state and action spaces.

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 & 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 & 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 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 four commonly used uncertainty sets in DRL: total variation, Wasserstein, Kullback-Leibler, and chi-square 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

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

Table 1: Comparison of the offline RL algorithms in the tabular setting. The data coverage assumption is based on the single-policy concentrability $C_{\pi^*} = \max_{s,a}(d^{\pi^*}(s,a)/\mu(s,a))$ and its clipped version $C_{\pi^*,\text{clip}} = \max_{s,a}(\min\{d^{\pi^*}(s,a), 1/|\mathcal{S}|\}/\mu(s,a))$, where d^{π^*} is the discounted occupancy measure of the optimal policy π^* and μ is the state-action visitation distribution of the data generating policy. 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.

Algorithm	Algorithm-type	Data coverage assumption	Suboptimality
(Jin et al., 2021, Cor.4.5)	reward pessimism	w.h.p $\Lambda_N \geq I/N + C_{\text{sc}} \cdot \Sigma_{d^{\pi^*}}$	$\frac{d\sqrt{\text{rank}(\Sigma_{d^{\pi^*}})}}{\sqrt{C_{\text{sc}}(1-\gamma)^4N}}$
(Uehara & Sun, 2021, Th.6)	oracle model pessimism	$C_{\pi^*,\phi} < \infty$	$\sqrt{\frac{\text{rank}(\Lambda)^2 dC_{\pi^*,\phi}}{(1-\gamma)^4N}}$
LM-DRQI (this work, Th.2)	distributionally robust	$\forall i \in [d]$ w.h.p $\Lambda_N \geq I/N + C_{\text{sc}}^\dagger d \cdot \Sigma^i_{d^{\pi^*}}$	$\frac{\sqrt{\text{rank}(\Sigma_{d^{\pi^*}})d}}{\sqrt{C_{\text{sc}}^\dagger(1-\gamma)^4N}}$

Table 2: Comparison of the offline RL algorithms in the linear MDP setting. 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 $C_{\text{sc}}^\dagger$ are the sufficient coverage constants satisfying corresponding random events.

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.

(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 state 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.Lemma 11) 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 & Sun (2021). We also want to emphasize that Uehara & 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): In the final stages of working on this manuscript we came across the work by Wang et al. (2023b), who propose a similar offline RL algorithm as ours (Algorithm 1). Wang et al. (2023b) only consider the tabular setting, whereas we provide offline RL algorithms for both the tabular and linear MDP settings. Wang et al. (2023b) consider a total variation uncertainty set whereas we consider four commonly used uncertainty sets in DRL. In terms of the sample complexity guarantees, they provide a $\widetilde{\mathcal{O}}(\sqrt{(|\mathcal{S}|C^{-}_{\pi^*})/((1-\gamma)^4N)})$ 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)). To emphasize, Hoeffding's inequality (Lemma 2) 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 the analysis in Wang et al. (2023b) are sophisticated and insightful. We believe both works make interesting contributions to offline RL literature.

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 $\mathrm{Tr}(\cdot)$ 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 $f \le \mathcal{O}(g)$ to denote $f \le c \cdot g$ for some universal constants less than 100, and likewise use $f \le \widetilde{\mathcal{O}}(g)$ 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 $d^\pi_{P^o}$ 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

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

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 & Szepesvári, 2008), defined as $C_u = \sup_\pi C_\pi$. Munos & 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 & 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

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

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

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

$$V^\pi_{\mathcal{P}} = \inf_{P\in\mathcal{P}} V^\pi_P, \qquad V^*_{\mathcal{P}} = \sup_\pi \inf_{P\in\mathcal{P}} V^\pi_P. \tag{4}$$

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 & Kalathil, 2021; Panaganti et al., 2022; Panaganti & Kalathil, 2022; Xu* et al., 2023; Shi & Chi, 2022; Wang & Zou, 2021; Ma et al., 2022; Wang & Zou, 2022; Kumar et al., 2023; Li et al., 2022b; Wang et al., 2023a; Grand-Clément & 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

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

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

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

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

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

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 robust Q-function corresponding to the RMDP uncertainty set specified by the $\rho_{s,a}$ values (Xu* et al., 2023; Shi & 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 & 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 four uncertainty sets corresponding to four different distance metrics $D(\cdot,\cdot)$. We also fix a confidence level $\delta \in (0,1)$ in the following.

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

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

For the remaining uncertainty sets, the specific value of $\rho_{s,a}$ is given in appendix due to page limit.

2. Wasserstein uncertainty set ($\widehat{\mathcal{P}}^{\mathrm{w}}$): We define $\widehat{\mathcal{P}}^{\mathrm{w}} = \otimes\widehat{\mathcal{P}}^{\mathrm{w}}_{s,a}$, where $\widehat{\mathcal{P}}^{\mathrm{w}}_{s,a}$ is as in (7) with the empirical estimator $\widehat{P}^o_{s,a}$, and with the Wasserstein distance $D_{\mathrm{w}}(P, \widehat{P}^o_{s,a}) = \inf_{\nu\in\mathrm{m}(P,\widehat{P}^o_{s,a})} \int \ell(x,y)\mathrm{d}\,\nu(dx,dy)$, where the integration

Algorithm 1 Distributionally Robust Q-Iteration (DRQI) Algorithm

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

is over $(x, y) \in \mathcal{S} \times \mathcal{S}$, $\mathrm{m}(P, \widehat{P}^o_{s,a})$ denotes all probability measures on $\mathcal{S} \times \mathcal{S}$ with marginals P and $\widehat{P}^o_{s,a}$, and $\ell(\cdot, \cdot)$ is the discrete metric, $\ell(s, s') = \mathbb{1}\{s \neq s'\}$.

3. Kullback-Leibler (KL) uncertainty set ($\widehat{\mathcal{P}}^{\mathrm{kl}}$): We define $\widehat{\mathcal{P}}^{\mathrm{kl}} = \otimes \widehat{\mathcal{P}}^{\mathrm{kl}}_{s,a}$, where $\widehat{\mathcal{P}}^{\mathrm{kl}}_{s,a}$ is as in (7) with the add-$L (= 1)$ estimator $\widetilde{P}^o_{s,a}$, and with the KL distance $D_{\mathrm{KL}}(P, \widetilde{P}^o_{s,a}) = \sum_{s'} P(s') \log(P(s')/\widetilde{P}^o_{s,a}(s'))$.

4. Chi-square uncertainty set ($\widehat{\mathcal{P}}^{\mathrm{c}}$): We define $\widehat{\mathcal{P}}^{\mathrm{c}} = \otimes \widehat{\mathcal{P}}^{\mathrm{c}}_{s,a}$, where $\widehat{\mathcal{P}}^{\mathrm{c}}_{s,a}$ is as in (7) with the add-$L (= \log(1/\delta))$ estimator $\widetilde{P}^o_{s,a}$, and with the chi-square distance $D_{\mathrm{c}}(P, \widetilde{P}^o_{s,a}) = \sum_{s'} (P(s') - \widetilde{P}^o_{s,a}(s'))^2 / \widetilde{P}^o_{s,a}(s')$.

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 since we can model similar uncertainty 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.

Practical Solution to Eq. (8): The key step of our DRQI algorithm is the update $Q_{k+1} = \widehat{T} Q_k$ in line 5 of Algorithm 1. Computing this empirical robust Bellman update may seem daunting at first look: it not only involves constructing the uncertainty set $\widehat{\mathcal{P}}$ (c.f. Eq. (7) but also involves evaluating the expectation w.r.t. all models in the uncertainty set $\widehat{\mathcal{P}}$ (c.f. Eq. (8)). The DRL algorithms overcome this challenge in two different ways.

The first approach is by a dual reformulation of Eq. (8) which has been successfully applied in many works on distributionally robust supervised learning (Blanchet et al., 2019; Farnia & Tse, 2016; Duchi et al., 2022) and on distributionally robust RL(Panaganti et al., 2022; Xu* et al., 2023). For example, for the total variation uncertainty set, Eq. (8) can be rewritten as (Panaganti & Kalathil, 2022, Proposition 1)

$$(\widehat{T} Q)(s, a) = r(s, a) - \gamma \inf_{\eta \in [0, \frac{2}{\rho(1-\gamma)}]} \{\mathbb{E}_{s' \sim \widehat{P}^o_{s,a}}[(\eta - V(s'))_+] - \eta + \rho(\eta - \inf_{s''} V(s''))_+\}, \tag{10}$$

where $V(s) = \max_b Q(s, b)$. Note that the expectation in Eq. (10) above is only w.r.t. the empirical estimate $\widehat{P}^o$ which eliminates the need for constructing the uncertainty set $\widehat{\mathcal{P}}$. Moreover, the remaining optimization, $\inf_\eta$, in Eq. (10) is a scalar concave optimization over a compact real interval which makes it computationally tractable.

The second approach is to directly solve the optimization in Eq. (8) in the primal form by considering $P_{s,a}$ as an element in the $|\mathcal{S}|$-dimensional simplex. In particular, for uncertainty sets with specific structures, including the four different uncertainty sets specified before, the primal problem can be directly solved using standard convex optimization solvers such as CVXPY (Diamond & Boyd, 2016). This approach is particularly attractive for tabular settings, and hence we use this primal approach for our simulation experiments.

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 3 to 6). We defer the corresponding theorem statements and proofs to the appendix due to the page limit.

Theorem 1. *Let π_K be the DRQI policy after K iterations under the TV uncertainty set $\widehat{\mathcal{P}}^{\mathrm{tv}}$. If the total number of samples $N \geq N_{\mathrm{tv}}$, where $N_{\mathrm{tv}} = \mathcal{O}\Big(C_{\pi^*} \max\{|\mathcal{S}|^2, 2\log(2|\mathcal{S}|^2|\mathcal{A}|/\delta)\}/(\epsilon^2(1-\gamma)^4)\Big)$, 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 Sketch. Denoting $\widehat{\mathcal{P}}^{\mathrm{tv}}$ 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 Proposition 1 we show that, with the $\rho_{s,a}$ as specified above, $P^o \in \widehat{\mathcal{P}}^{\mathrm{tv}}$ with probability at least $1 - \delta$. So, by definition of the robust value function, the second term $(V^{\pi_K}_{\widehat{\mathcal{P}}}(s_0) - V^{\pi_K}_{P^o}(s_0))$ is negative.

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

Bounding $(V^{\pi^*}_{P^o}(s_0) - V^{\widehat{\pi}^*}_{\widehat{\mathcal{P}}}(s_0))$ is more technical. The key idea is to first note that $D_{\mathrm{TV}}(P_{s,\pi^*(s)}, P^o_{s,\pi^*(s)}) \leq 2\rho_{s,a}$ for any $P \in \widehat{\mathcal{P}}$, by Proposition 1 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 Lemma 1, and then use a change of measure argument to get the final bound in terms of single concentrability coefficient C_{π^*}. $\square$

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 1 (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 $\phi_i(s,a) \geq 0$ and $\|\phi(s,a)\|_2 \leq 1$ 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 $\Sigma^{(i,j)}_{d^{\pi^*}} = \mathbb{E}_{s,a\sim d^{\pi^*}}[(\phi_i(s,a)\mathbb{1}_i)(\phi_j(s,a)\mathbb{1}_j)^\top]$ 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 the 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

$$\mathcal{P} = \{P : P_{s,a}(s') = \sum_{i\in[d]} \phi_i(s,a)\nu_i(s'), \nu_i \in \mathcal{M}_i, \forall i \in [d]\},$$

$$\mathcal{M}_i = \{\nu_i : D_{\mathrm{IPM}}(\nu_i, \nu_i^o) \leq \rho_i\}, \text{ where,} \tag{11}$$

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

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

$$TQ(s,a) = r(s,a) + \gamma \sum_{i\in[d]} \phi_i(s,a) \min_{\nu_i \in \mathcal{M}_i} \mathbb{E}_{s'\sim\nu_i}(\max_b Q(s',b)). \tag{12}$$

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

$$\rho_i = \frac{c_1 \log(Nd/((1-\gamma)\delta))}{1-\gamma} \sqrt{\frac{d}{N}} \sqrt{\Lambda_N^{-1}(i,i)}. \tag{13}$$

We can now define the empirical robust Bellman operator $\widehat{T}$ exactly as in Eq. (12), 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 $\Lambda_N \geq (1/N)I + C^\dagger_{\mathrm{sc}} \cdot d \cdot \Sigma^i_{d^{\pi^*}_{P^o}}$, where $\Sigma^i_{d^{\pi^*}_{P^o}} = \mathbb{E}_{s,a\sim d^{\pi^*}}[(\phi_i(s,a)\mathbb{1}_i)(\phi_i(s,a)\mathbb{1}_i)^\top]$.*

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 ($\mathrm{rank}(\Lambda) = d$ (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 2. *Let π_K be the LM-DRQI policy after K iterations. Let Assumption 2 hold. If the total number of samples $N \geq N_{\mathrm{IPM}}$, where $N_{\mathrm{IPM}} = \widetilde{\mathcal{O}}(d \cdot \mathrm{rank}(\Sigma_{d^{\pi^*}_{P^o}})/(C^\dagger_{\mathrm{sc}}(1-\gamma)^4\epsilon^2))$, 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 statement and proofs are in the appendix due to the page limit.

5. Experiments

We evaluate the performance of our DRQI algorithm on the `FrozenLake-v1` environment ($|\mathcal{S}| = 16$, $|\mathcal{A}| = 4$) from OpenAI Gym (Brockman et al., 2016). The goal is to cross a frozen lake without falling into holes. Since the frozen lake is slippery, rather than always going in the intended direction, the agent can slip into the other directions. We implement DRQI algorithm with total variation uncertainty set using the CVXPY library (Diamond & Boyd, 2016) for the experiments. We submit our code in an anonymous Github repository: `https://github.com/aspiring-giraffe/DRQI`.

Offline Data Collection: We evaluate the algorithms using two kinds of offline datasets, *full-coverage* and *partial-*

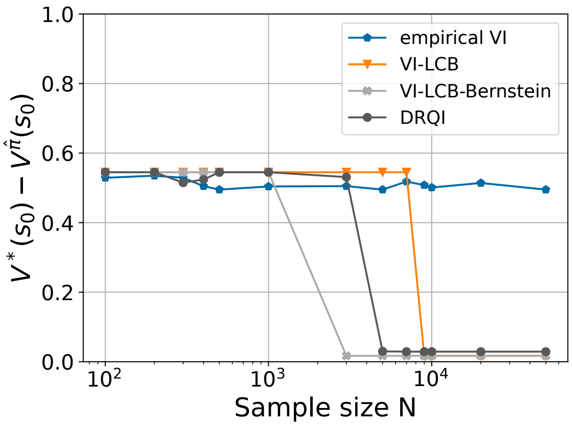

Figure 1: Convergence of DRQI algorithm under partial coverage in `FrozenLake-v1`.

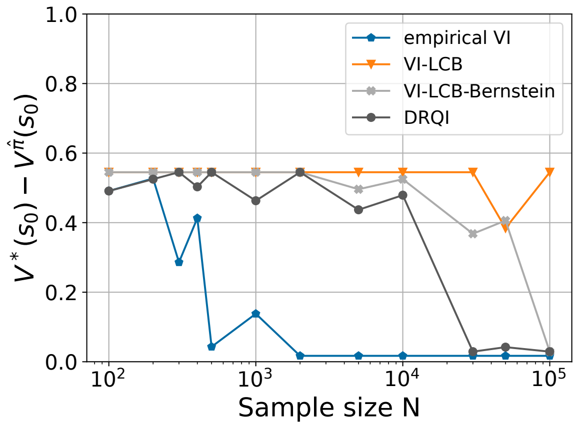

Figure 2: Convergence of DRQI algorithm under full coverage in `FrozenLake-v1`.

coverage. Full-coverage dataset is collected by using a generative model where we collect equal number of next-state samples from every (s, a) pairs. The partial-coverage dataset is generated according to the behavior policy μ, where $\mu(a \mid s) = (\frac{1}{2} + \frac{1}{2|\mathcal{A}|})\mathbb{1}_{\{a=\pi^*(s)\}} + \frac{1}{2|\mathcal{A}|}\mathbb{1}_{\{a\neq\pi^*(s)\}}$, π^* is the optimal policy for the `FrozenLake-v1` environment. It is easy to check that the single-policy concentrability coefficient C_{π^*} is bounded. Note that most of the (s, a)-pairs are un-sampled or under-sampled in the partial-coverage data set.

We compare our DRQI with three algorithms: (1) empirical value iteration (EVI) which essentially performs value iteration using the empirical model $\widehat{P}^o$, (2) VI-LCB algorithm (Rashidinejad et al., 2022), a reward pessimism-based offline RL algorithm, (3) VI-LCB-Bernstein algorithm (Li et al., 2022a), a Bernstein type reward pessimism-based offline RL algorithm. The performance metric is the value sub-optimality with respect to the optimal policy.

In the partial data coverage setting (Fig. 1), we see that the EVI algorithm does not converge even with 10^5 samples, clearly showing the inability of standard dynamic programming approaches to obtain an approximately optimal policy in such settings. On the other hand, our DRQI algorithm learns the optimal policy with roughly 4×10^3 samples. Moreover, the performance of our DRQI algorithm is on par with the state-of-the-art VI-LCB and VI-LCB-Bernstein offline RL algorithms (in fact performing better than VI-LCB but only slightly worse than VI-LCB-Bernstein). We also would like to note that both VI-LCB and VI-LCB-Bernstein algorithms require some hyperparameter tuning regarding the "universal constants" that appear in their proofs of high-probability bounds. Our DRQI algorithm, on the other hand, does not require any hyperparameter tuning and use the $\rho_{s,a}$ exactly as defined in Eq. (9).

In the full data coverage setting (Fig. 2), EVI is able to find the optimal policy since the concentration of $\widehat{P}^o$ to the true model P^o is straightforward. Our DRQI algorithm is also able to learn the optimal policy, albeit with more samples. Notably, our DRQI algorithm outperforms the two LCB-style algorithms in this setting.

Remark 2 (On the choice of the simulation settings)**.** *Our work is mainly theoretical in nature and considers the tabular and linear MDPs. We would like to emphasize that we have included the simulations for this setting, following similar theoretical works on offline RL (Li et al., 2022a). Many notable theoretical works on the sample complexity of offline RL and distributionally robust RL either do not include any simulations or consider only the tabular case as we did (Rashidinejad et al., 2022; Ma et al., 2022; Shi et al., 2022; Yin et al., 2022). Since the theory and algorithms of this paper focus on tabular and linear MDP settings, simulations in continuous state/action settings such as MuJoCo control tasks are out of the scope of the problem we address.*

6. 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 to handle large and continuous state-action space problems.

Societal Impact Statement

This paper is mainly of theoretical nature. In particular, we aim to rigorously bridge offline reinforcement learning and distributionally robust optimization framework. We hope that our work can expand the literature of theoretical offline RL. At this moment, we find no particular need for highlighting any societal consequence of our work.

References

Agarwal, A., Jiang, N., Kakade, S. M., and Sun, W. Reinforcement learning: Theory and algorithms. *CS Dept., UW Seattle, Seattle, WA, USA, Tech. Rep*, 2019.

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

Arora, V., Bhattacharyya, A., Canonne, C. L., and Yang, J. Q. Near-optimal degree testing for bayes nets. *arXiv preprint arXiv:2304.06733*, 2023.

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

Bhattacharyya, A., Gayen, S., Price, E., and Vinodchandran, N. Near-optimal learning of tree-structured distributions by chow-liu. In *Proceedings of the 53rd annual acm SIGACT symposium on theory of computing*, pp. 147–160, 2021.

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

Boucheron, S., Lugosi, G., and Massart, P. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., and Zaremba, W. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016.

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

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

Chang, J., Uehara, M., Sreenivas, D., Kidambi, R., and Sun, W. Mitigating covariate shift in imitation learning via offline data with partial coverage. *Advances in Neural Information Processing Systems*, 34:965–979, 2021.

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

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

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

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

Duchi, J., Hashimoto, T., and Namkoong, H. Distributionally robust losses for latent covariate mixtures, 2022.

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

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

Farahmand, A.-m., Szepesvári, C., and Munos, R. Error propagation for approximate policy and value iteration. *Advances in Neural Information Processing Systems*, 23, 2010.

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

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

Fujimoto, S., Meger, D., and Precup, D. Off-policy deep reinforcement learning without exploration. In *International Conference on Machine Learning*, pp. 2052–2062, 2019.

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

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

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

Jin, C., Yang, Z., Wang, Z., and Jordan, M. I. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pp. 2137–2143, 2020.

Jin, Y., Yang, Z., and Wang, Z. Is pessimism provably efficient for offline rl? In *International Conference on Machine Learning*, pp. 5084–5096. PMLR, 2021.

Kalathil, D., Borkar, V. S., and Jain, R. Empirical Q-Value Iteration. *Stochastic Systems*, 11(1):1–18, 2021.

Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R. L., Gao, I., et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pp. 5637–5664. PMLR, 2021.

Kostrikov, I., Fergus, R., Tompson, J., and Nachum, O. Offline reinforcement learning with fisher divergence critic regularization. In *International Conference on Machine Learning*, pp. 5774–5783. PMLR, 2021.

Kuhn, D., Esfahani, P. M., Nguyen, V. A., and Shafieezadeh-Abadeh, S. Wasserstein distributionally robust optimization: Theory and applications in machine learning. In *Operations research & management science in the age of analytics*, pp. 130–166. Informs, 2019.

Kumar, A., Fu, J., Soh, M., Tucker, G., and Levine, S. Stabilizing off-policy q-learning via bootstrapping error reduction. In *Advances in Neural Information Processing Systems*, pp. 11784–11794, 2019.

Kumar, A., Zhou, A., Tucker, G., and Levine, S. Conservative q-learning for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 33: 1179–1191, 2020.

Kumar, N., Derman, E., Geist, M., Levy, K. Y., and Mannor, S. Policy gradient for rectangular robust markov decision processes. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Lange, S., Gabel, T., and Riedmiller, M. Batch reinforcement learning. In *Reinforcement learning*, pp. 45–73. Springer, 2012.

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

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

Levy, D., Carmon, Y., Duchi, J. C., and Sidford, A. Large-scale methods for distributionally robust optimization. *Advances in Neural Information Processing Systems*, 33: 8847–8860, 2020.

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

Li, Y., Zhao, T., and Lan, G. First-order policy optimization for robust markov decision process. *arXiv preprint arXiv:2209.10579*, 2022b.

Liao, P., Qi, Z., Wan, R., Klasnja, P., and Murphy, S. A. Batch policy learning in average reward markov decision processes. *The Annals of Statistics*, 50(6):3364–3387, 2022.

Ma, X., Liang, Z., Xia, L., Zhang, J., Blanchet, J., Liu, M., Zhao, Q., and Zhou, Z. Distributionally robust offline reinforcement learning with linear function approximation. *arXiv preprint arXiv:2209.06620*, 2022.

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

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

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

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

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

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

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

Panaganti, K., Xu, Z., Kalathil, D., and Ghavamzadeh, M. Robust reinforcement learning using offline data. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Peng, X. B., Andrychowicz, M., Zaremba, W., and Abbeel, P. Sim-to-real transfer of robotic control with dynamics randomization. In *2018 IEEE international conference on robotics and automation (ICRA)*, pp. 3803–3810. IEEE, 2018.

Quinonero-Candela, J., Sugiyama, M., Schwaighofer, A., and Lawrence, N. D. *Dataset Shift in Machine Learning*. MIT Press, 2022.

Rashidinejad, P., Zhu, B., Ma, C., Jiao, J., and Russell, S. Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *IEEE Transactions on Information Theory*, 68(12):8156–8196, 2022.

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

Roy, A., Xu, H., and Pokutta, S. Reinforcement learning under model mismatch. In *Advances in Neural Information Processing Systems*, pp. 3043–3052, 2017.

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

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

Shi, L., Li, G., Wei, Y., Chen, Y., and Chi, Y. Pessimistic q-learning for offline reinforcement learning: Towards optimal sample complexity. In *International Conference on Machine Learning*, pp. 19967–20025. PMLR, 2022.

Si, N., Zhang, F., Zhou, Z., and Blanchet, J. Distributionally robust policy evaluation and learning in offline contextual bandits. In *International Conference on Machine Learning*, pp. 8884–8894, 2020.

Sidford, A., Wang, M., Wu, X., Yang, L., and Ye, Y. 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.

Tamar, A., Mannor, S., and Xu, H. Scaling up robust mdps using function approximation. In *International Conference on Machine Learning*, pp. 181–189, 2014.

Taori, R., Dave, A., Shankar, V., Carlini, N., Recht, B., and Schmidt, L. Measuring robustness to natural distribution shifts in image classification. *Advances in Neural Information Processing Systems*, 33:18583–18599, 2020.

Tobin, J., Fong, R., Ray, A., Schneider, J., Zaremba, W., and Abbeel, P. 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)*, pp. 23–30, 2017.

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

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

Wang, Q., Ho, C. P., and Petrik, M. Policy gradient in robust MDPs with global convergence guarantee. In *Proceedings of the 40th International Conference on Machine Learning*. PMLR, 2023a.

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

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

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

Wang, Y., Xiong, J., and Zou, S. Achieving the minimax optimal sample complexity of offline reinforcement learning: A dro-based approach, 2023b.

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

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

Xiong, W., Zhong, H., Shi, C., Shen, C., Wang, L., and Zhang, T. 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.

Xu*, Z., Panaganti*, K., and Kalathil, D. 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.

Yang, J., Zhou, K., Li, Y., and Liu, Z. Generalized out-of-distribution detection: A survey. *arXiv preprint arXiv:2110.11334*, 2021.

Yin, M., Duan, Y., Wang, M., and Wang, Y.-X. Near-optimal offline reinforcement learning with linear representation: Leveraging variance information with pessimism. In *International Conference on Learning Representations*, 2022.

Yu, T., Thomas, G., Yu, L., Ermon, S., Zou, J., Levine, S., Finn, C., and Ma, T. Mopo: Model-based offline policy optimization. In *Advances in Neural Information Processing Systems*, 2020.

Zhou, Z., Bai, Q., Zhou, Z., Qiu, L., Blanchet, J., and Glynn, P. Finite-sample regret bound for distributionally robust offline tabular reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 3331–3339, 2021.

☕ Supplementary Materials ☕

A. Useful Technical Results

Lemma 1 (Bound on binomial inverse moments (Rashidinejad et al., 2022, Lemma 14))**.** *Let $n \sim \mathrm{Binomial}(N, p)$. For any $k \geq 0$, there exists a constant c_k depending only on k such that*

$$\mathbb{E}\left[\frac{1}{(n \vee 1)^k}\right] \leq \frac{c_k}{(Np)^k},$$

where $c_k = 1 + k2^{k+1} + k^{k+1} + k\left(\frac{16(k+1)}{e}\right)^{k+1}$.

Lemma 2 (Hoeffding's inequality (Boucheron et al., 2013, see Theorem 2.8))**.** *Let $X_1, \dots, X_n$ be independent random variables such that X_i takes its values in $[a_i, b_i]$ almost surely for all $i \leq n$. Let*

$$S = \sum_{i=1}^{n} (X_i - \mathbb{E}[X_i]).$$

Then for every $t > 0$,

$$\mathbb{P}(S \geq t) \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right).$$

Furthermore, if $X_1, \dots, X_n$ are a sequence of independent, identically distributed random variables with mean μ. Let $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Suppose that $X_i \in [a, b]$, $\forall i$. Then for all $t > 0$

$$\mathbb{P}\left(\left|\overline{X}_n - \mu\right| \geq t\right) \leq 2\exp\left(-\frac{2nt^2}{(b-a)^2}\right).$$

The following lemmas characterize the sample complexity of learning discrete distributions when the accuracy is measured under four different distances, i.e., total variation, KL, chi-square, and Wasserstein.

Lemma 3 (Canonne, 2020, Theorem 1)**.** *Fix any $\delta \in (0, 1]$. Let $\widehat{P}$ be the empirical distribution constructed from N i.i.d. samples from an unknown distribution P over a finite set $\{1, \dots, k\}$. Then if the number of samples*

$$N \geq \frac{\max\{k, 2\log(2/\delta)\}}{\epsilon^2},$$

then $D_{\mathrm{TV}}(P, \widehat{P}) \leq \epsilon$ with probability at least $1 - \delta$. Moreover, this result is tight.

Lemma 4 (Bhattacharyya et al., 2021, Theorem 6.1)**.** *Fix any $\delta \in (0, 1]$. Let $\widetilde{P}$ be the empirical add-1 estimator obtained from N i.i.d. samples from an unknown distribution P over a finite set $\{1, \dots, k\}$. There exists a universal constant C such that, with probability at least $1 - \delta$,*

$$D_{\mathrm{KL}}(P, \widetilde{P}) \leq \frac{Ck\log(k/\delta)\log N}{N}.$$

Lemma 5 (Arora et al., 2023, Proposition 4.1)**.** *Fix any $\delta \in (0, 1]$ and let $L = \Theta(\log(1/\delta))$. Let $\widetilde{P}$ be the empirical add-L estimator obtained from N i.i.d. samples from an unknown distribution P over a finite set $\{1, \dots, k\}$. There exists a universal constant C such that, with probability at least $1 - \delta$,*

$$D_{\mathrm{c}}(P, \widetilde{P}) \leq \frac{Ck\log(k/\delta)}{N}.$$

Lemma 6 (Lei, 2020, Corollary 5.2)**.** *Let $p \in \mathcal{P}(\mathbb{R}^d)$ be a distribution such that $b = \mathbb{E}_{X\sim p}[\exp(a\|X\|_2)] < \infty$ for some $a > 0$. Fix $\delta \in (0, 1)$. Denote the empirical distribution from N samples of p as $\widehat{p}$. Then there exists some constant c_1 only depending on a, b such that $D_{\mathrm{w}}(p, \widehat{p}) \leq \sqrt{c_1 d \log(1/\delta)/N}$ holds at least with probability $1 - \delta$.*

Here we mention a uniform concentration result from Agarwal et al. (2019) corresponding to linear MDP transition model P^o. From Section 4, recall Λ_N and the model estimate of P^o denoted by $\widehat{P}^o$. We note that $\mathcal{O}$ notation in this result only removes dependence on universal constants.

Lemma 7 (Linear MDP Uniform Concentration Bound (Agarwal et al., 2019, Lemma 8.7))**.** *Fix $\delta \in (0,1)$ and let $\lambda = 1$. Consider $\mathcal{V} = \{V(\cdot) = \max_a \phi^\top(\cdot, a)w : w \in \mathbb{R}^d, \|w\|_2 \le 1/(1-\gamma)\}$. We have (1) $\| \sum_{t=1}^N \phi(s_t, a_t)\epsilon_t^\top V\|_{\Lambda_N^{-1}} \le \mathcal{O}(\sqrt{dN}\log(N/((1-\gamma)\delta))/1-\gamma)$ with probability at least $1-\delta$ for any $V \in \mathcal{V}$ uniformly, and it also holds (2) $\sup_{V\in\mathcal{V}} |\int_{\mathcal{S}}(P^o_{s,a} - \widehat{P}^o_{s,a})V(ds')| \le \|\phi(s,a)\|_{\Lambda_N^{-1}}\mathcal{O}\left(\sqrt{d}\log(N/\delta)/((1-\gamma)\sqrt{N})\right)$ with probability at least $1-\delta$ for any s, a.*

Here is a useful result from (Chang et al., 2021, Theorem 21).

Lemma 8. *Let $\lambda = 1$ and $c > 0$ be some universal constant. For all s, a simultaneously, with probability at least $1-\delta$ we have $\mathbb{E}_{s,a\sim\mu}[\phi(s,a)^\top\Lambda_N^{-1}\phi(s,a)] \le c^2 \cdot \mathrm{rank}(\Lambda)(\mathrm{rank}(\Lambda) + \log(c/\delta))$ where $\Lambda_N = \frac{\lambda}{N}I + \frac{1}{N}\sum_{t=1}^N \phi(s_t,a_t)\phi(s_t,a_t)^\top$, $\Lambda = \mathbb{E}_{s,a\sim\mu}\phi(s,a)\phi(s,a)^\top$.*

B. Proofs of Distributionally Robust Q-Iteration (DRQI)

We first make the observation that the true model P^o lies in the uncertainty set $\widehat{\mathcal{P}}$ with high probability. Intuitively, the empirical estimator $\widehat{P}^o$ of P^o are statistically closer which is dependent on the number of samples. We first make this observation and intuition formal in the proposition below for the TV uncertainty set.

Proposition 1. *We have $P^o \in \widehat{\mathcal{P}}^{\mathrm{tv}}$ with probability at least $1-\delta$.*

Proof. We start with the fact that $D_{\mathrm{TV}}(p, q) \le 1$ for any distributions p, q. For the case $N(s,a) < 1$, i.e., $N(s,a) = 0$, it is trivial that $P^o_{s,a} \in \widehat{\mathcal{P}}^{\mathrm{tv}}_{s,a}$, almost surely, since $\widehat{\mathcal{P}}^{\mathrm{tv}}_{s,a} = \Delta(\mathcal{S})$.

From Lemma 3, we have $D_{\mathrm{TV}}(P^o_{s,a}, \widehat{P}^o_{s,a}) \le \sqrt{\max\{|\mathcal{S}|, 2\log(2/\delta)\}/N(s,a)}$ for any s, a pair with probability at least $1-\delta/(|\mathcal{S}||\mathcal{A}|)$. Thus $\otimes_{s,a}P^o_{s,a} \in \otimes_{s,a}\widehat{\mathcal{P}}^{\mathrm{tv}}_{s,a}$ holds with probability at least $1-\delta$. □

We now provide a similar guarantee like Proposition 1 for the Wasserstein uncertainty set.

Proposition 2. *We have $P^o \in \widehat{\mathcal{P}}^{\mathrm{w}}$ with probability at least $1-\delta$.*

Proof. From Proposition 1 and Villani et al. (2009, Theorem 6.15), it follows that $D_{\mathrm{w}}(p, U) \le B$ for any distribution p and uniform distribution U, i.e., $U(s) = 1/|\mathcal{S}|$ for all $s \in \mathcal{S}$. For the case $N(s,a) < 1$, i.e., $N(s,a) = 0$, it now follows that $P^o_{s,a} \in \widehat{\mathcal{P}}^{\mathrm{w}}_{s,a}$, almost surely, since $\widehat{P}^o_{s,a} = 1/|\mathcal{S}|$.

From Lemma 6, we have $D_{\mathrm{w}}(P^o_{s,a}, \widehat{P}^o_{s,a}) \le \sqrt{C_{s,a}|\mathcal{S}|\log(1/\delta)/N(s,a)}$ for any s, a pair with probability at least $1-\delta/(|\mathcal{S}||\mathcal{A}|)$, where $C_{s,a} > 0$ is some universal constant depending only on the distribution $P^o_{s,a}$. By setting $C = \max_{s,a} C_{s,a}$, $\otimes_{s,a}P^o_{s,a} \in \otimes_{s,a}\widehat{\mathcal{P}}^{\mathrm{w}}_{s,a}$ holds with probability at least $1-\delta$. □

We now provide a similar guarantee for the KL uncertainty set.

Proposition 3. *We have $P^o \in \widehat{\mathcal{P}}^{\mathrm{kl}}$ with probability at least $1-\delta$.*

Proof. We start with the fact that $D_{\mathrm{KL}}(p, U) \le \log(|\mathcal{S}|)$ for any distribution p and uniform distribution U, i.e., $U(s) = 1/|\mathcal{S}|$ for all $s \in \mathcal{S}$. For the case $N(s,a) < 1$, i.e., $N(s,a) = 0$, it now follows that $P^o_{s,a} \in \widehat{\mathcal{P}}^{\mathrm{kl}}_{s,a}$, almost surely, since $\widetilde{P}^o_{s,a} = 1/|\mathcal{S}|$.

From Lemma 4, we have $D_{\mathrm{KL}}(P^o_{s,a}, \widetilde{P}^o_{s,a}) \le C|\mathcal{S}|\log(|\mathcal{S}|/\delta)\log(N(s,a))/N(s,a)$ for any s, a pair with probability at least $1-\delta/(|\mathcal{S}||\mathcal{A}|)$, where $C > 0$ is some universal constant. We also know that $\log(N(s,a)) \le \log(N)$. Thus $\otimes_{s,a}P^o_{s,a} \in \otimes_{s,a}\widehat{\mathcal{P}}^{\mathrm{kl}}_{s,a}$ holds with probability at least $1-\delta$. □

We now provide a similar guarantee for the chi-square uncertainty set.

Proposition 4. *We have $P^o \in \widehat{\mathcal{P}}^{\mathrm{c}}$ with probability at least $1-\delta$.*

Proof. We start with the fact that $D_{\mathrm{c}}(p, U) \le |\mathcal{S}|+1$ for any distribution p and uniform distribution U, i.e., $U(s) = 1/|\mathcal{S}|$ for all $s \in \mathcal{S}$. For the case $N(s,a) < 1$, i.e., $N(s,a) = 0$, it now follows that $P^o_{s,a} \in \widehat{\mathcal{P}}^{\mathrm{c}}_{s,a}$, almost surely, since $\widetilde{P}^o_{s,a} = 1/|\mathcal{S}|$.

From Lemma 5, we have $D_{\rm c}(P^o_{s,a}, \widetilde{P}^o_{s,a}) \leq C|\mathcal{S}|\log(|\mathcal{S}|/\delta)/N(s,a)$ for any s, a pair with probability at least $1-\delta/(|\mathcal{S}||\mathcal{A}|)$, where $C > 0$ is some universal constant. Thus $\otimes_{s,a} P^o_{s,a} \in \otimes_{s,a} \widehat{\mathcal{P}}^{\rm c}_{s,a}$ holds with probability at least $1 - \delta$. □

We are now ready to present our main results of Section 3. With the above result (Proposition 1), we now provide the offline RL suboptimality guarantee below for the TV uncertainty set.

Theorem 3. *Let π_K be the DRQI policy after K iterations under the TV uncertainty set $\widehat{\mathcal{P}}^{\rm tv}$. With probability at least $1 - \delta$ it holds that*

$$\mathbb{E}_{s_0\sim d_0}[V^{\pi^*}(s_0) - \mathbb{E}_{\mathcal{D}}[V^{\pi_K}(s_0)]] \leq \frac{64\gamma\sqrt{C_{\pi^*}|\mathcal{S}|}}{(1-\gamma)^2}\sqrt{\frac{\max\{|\mathcal{S}|, 2\log(2|\mathcal{S}||\mathcal{A}|/\delta)\}}{N}} + \frac{2\gamma^{K+1}}{(1-\gamma)^2}.$$

Proof. We first make important definitions that will be useful for our analyses. We denote the value function of policy π for the transition dynamics model P as V^π_P. We now denote the robust value function (Panaganti & Kalathil, 2022; Xu* et al., 2023; Panaganti et al., 2022) for uncertainty set $\widehat{\mathcal{P}}^{\rm tv}$ as $V^\pi_{\widehat{\mathcal{P}}} = \min_{P\in\widehat{\mathcal{P}}} V^\pi_P$ and its optimal robust policy as $\widehat{\pi}^* = \arg\max_\pi V^\pi_{\widehat{\mathcal{P}}}$. We note that for the sake of notational simplicity we drop the superscript tv going forward, that is, we denote $\widehat{\mathcal{P}}^{\rm tv}$ simply as $\widehat{\mathcal{P}}$. We let $Q^\pi_{\widehat{\mathcal{P}}}$ be its corresponding robust Q-function. From robust RL (Panaganti & Kalathil, 2022; Xu* et al., 2023; Panaganti et al., 2022) we can write the following robust Bellman equation: $Q^\pi_{\widehat{\mathcal{P}}}(s,a) = r(s,a) + \gamma\min_{P_{s,a}\in\widehat{\mathcal{P}}_{s,a}} \mathbb{E}_{s'\sim P_{s,a}}(V^\pi_{\widehat{\mathcal{P}}}(s'))$. To make it notationally easy, we write V^{π^*} (d^π) as $V^{\pi^*}_{P^o}$ ($d^\pi_{P^o}$) making the dependence on the model P^o explicit.

We now start analyzing offline RL suboptimality as:

$$\begin{aligned}
\mathbb{E}_{s_0\sim d_0}[V^{\pi^*}_{P^o}(s_0) - V^{\pi_K}_{P^o}(s_0)] &= \mathbb{E}_{s_0\sim d_0}[V^{\pi^*}_{P^o}(s_0) - V^{\pi_K}_{\widehat{\mathcal{P}}}(s_0) + V^{\pi_K}_{\widehat{\mathcal{P}}}(s_0) - V^{\pi_K}_{P^o}(s_0)] \\
&\overset{(a)}{\leq} \mathbb{E}_{s_0\sim d_0}[V^{\pi^*}_{P^o}(s_0) - V^{\pi_K}_{\widehat{\mathcal{P}}}(s_0)] \\
&= \mathbb{E}_{s_0\sim d_0}[V^{\pi^*}_{P^o}(s_0) - V^{\widehat{\pi}^*}_{\widehat{\mathcal{P}}}(s_0) + V^{\widehat{\pi}^*}_{\widehat{\mathcal{P}}}(s_0) - V^{\pi_K}_{\widehat{\mathcal{P}}}(s_0)] \\
&\leq \mathbb{E}_{s_0\sim d_0}[V^{\pi^*}_{P^o}(s_0) - V^{\widehat{\pi}^*}_{\widehat{\mathcal{P}}}(s_0)] + \left\|V^{\widehat{\pi}^*}_{\widehat{\mathcal{P}}} - V^{\pi_K}_{\widehat{\mathcal{P}}}\right\|_\infty \\
&\overset{(b)}{\leq} \mathbb{E}_{s_0\sim d_0}[V^{\pi^*}_{P^o}(s_0) - V^{\widehat{\pi}^*}_{\widehat{\mathcal{P}}}(s_0)] + \frac{2\gamma^{K+1}}{(1-\gamma)^2},
\end{aligned} \tag{14}$$

where (a) follows from Proposition 1 and definition of robust value function $V^{\pi_K}_{\widehat{\mathcal{P}}}(s_0)$ and (b) follows from robust amplification lemma (Panaganti & Kalathil, 2022, Lemma 10, eq.(28)). For the rest of the analysis, we focus on analyzing $\mathbb{E}_{s_0\sim d_0}[V^{\pi^*}_{P^o}(s_0) - V^{\widehat{\pi}^*}_{\widehat{\mathcal{P}}}(s_0)]$.

Observe that,

$$\begin{aligned}
&\mathbb{E}_{s_0\sim d_0}[V^{\pi^*}_{P^o}(s_0) - V^{\widehat{\pi}^*}_{\widehat{\mathcal{P}}}(s_0)] = \mathbb{E}_{s_0\sim d_0}[Q^{\pi^*}_{P^o}(s_0, \pi^*(s_0)) - Q^{\widehat{\pi}^*}_{\widehat{\mathcal{P}}}(s_0, \widehat{\pi}^*(s_0))] \\
&\overset{(c)}{\leq} \mathbb{E}_{s_0\sim d_0}[Q^{\pi^*}_{P^o}(s_0, \pi^*(s_0)) - Q^{\widehat{\pi}^*}_{\widehat{\mathcal{P}}}(s_0, \pi^*(s_0))] \\
&\overset{(d)}{=} \mathbb{E}_{s_0\sim d_0}[r(s_0, \pi^*(s_0)) + \gamma\mathbb{E}_{s'\sim P^o_{s_0,\pi^*(s_0)}}(V^{\pi^*}_{P^o}(s')) \\
&\qquad - r(s_0, \pi^*(s_0)) - \gamma \min_{P_{s_0,\pi^*(s_0)}\in\widehat{\mathcal{P}}_{s_0,\pi^*(s_0)}} \mathbb{E}_{s'\sim P_{s_0,\pi^*(s_0)}}(V^{\widehat{\pi}^*}_{\widehat{\mathcal{P}}}(s'))] \\
&= \mathbb{E}_{s_0\sim d_0}[\gamma\mathbb{E}_{s'\sim P^o_{s_0,\pi^*(s_0)}}(V^{\pi^*}_{P^o}(s')) - \gamma\mathbb{E}_{s'\sim P^o_{s_0,\pi^*(s_0)}}(V^{\widehat{\pi}^*}_{\widehat{\mathcal{P}}}(s'))] \\
&\qquad + \mathbb{E}_{s_0\sim d_0}[\gamma\mathbb{E}_{s'\sim P^o_{s_0,\pi^*(s_0)}}(V^{\widehat{\pi}^*}_{\widehat{\mathcal{P}}}(s')) - \gamma \min_{P_{s_0,\pi^*(s_0)}\in\widehat{\mathcal{P}}_{s_0,\pi^*(s_0)}} \mathbb{E}_{s'\sim P_{s_0,\pi^*(s_0)}}(V^{\widehat{\pi}^*}_{\widehat{\mathcal{P}}}(s'))] \\
&= \mathbb{E}_{s_0\sim d_0}[\gamma\mathbb{E}_{s'\sim P^o_{s_0,\pi^*(s_0)}}(V^{\pi^*}_{P^o}(s') - V^{\widehat{\pi}^*}_{\widehat{\mathcal{P}}}(s'))] \\
&\qquad + \underbrace{\mathbb{E}_{s_0\sim d_0}[\gamma\mathbb{E}_{s'\sim P^o_{s_0,\pi^*(s_0)}}(V^{\widehat{\pi}^*}_{\widehat{\mathcal{P}}}(s')) - \gamma \min_{P_{s_0,\pi^*(s_0)}\in\widehat{\mathcal{P}}_{s_0,\pi^*(s_0)}} \mathbb{E}_{s'\sim P_{s_0,\pi^*(s_0)}}(V^{\widehat{\pi}^*}_{\widehat{\mathcal{P}}}(s'))]}_{(I)},
\end{aligned} \tag{15}$$

where (c) follows since $\widehat{\pi}^*$ is optimal robust policy of $V^{\pi}_{\widehat{\mathcal{P}}}$ and (d) follows from classical and robust Bellman equations.

Analyzing (I) in Eq. (15) but for any $P_{s_0,\pi^*(s_0)} \in \widehat{\mathcal{P}}_{s_0,\pi^*(s_0)}$ gives us:

$$\begin{aligned}
&\mathbb{E}_{s_0\sim d_0}[\gamma \mathbb{E}_{s'\sim P^o_{s_0,\pi^*(s_0)}}(V^{\widehat{\pi}^*}_{\widehat{\mathcal{P}}}(s')) - \gamma \mathbb{E}_{s'\sim P_{s_0,\pi^*(s_0)}}(V^{\widehat{\pi}^*}_{\widehat{\mathcal{P}}}(s'))] \qquad (16)\\
&= \mathbb{E}_{s_0\sim d_0}[\gamma \mathbb{E}_{s'\sim P^o_{s_0,\pi^*(s_0)}}(V^{\widehat{\pi}^*}_{\widehat{\mathcal{P}}}(s')) - \gamma \mathbb{E}_{s'\sim \widehat{P}^o_{s_0,\pi^*(s_0)}}(V^{\widehat{\pi}^*}_{\widehat{\mathcal{P}}}(s')) \\
&\qquad + \gamma \mathbb{E}_{s'\sim \widehat{P}^o_{s_0,\pi^*(s_0)}}(V^{\widehat{\pi}^*}_{\widehat{\mathcal{P}}}(s')) - \gamma \mathbb{E}_{s'\sim P_{s_0,\pi^*(s_0)}}(V^{\widehat{\pi}^*}_{\widehat{\mathcal{P}}}(s'))] \\
&\overset{(g)}{\leq} \frac{2\gamma}{1-\gamma}\mathbb{E}_{s_0\sim d_0}\left[\min\left\{1, \sqrt{\frac{\max\{|\mathcal{S}|, 2\log(2|\mathcal{S}||\mathcal{A}|/\delta)\}}{N(s_0,\pi^*(s_0))}}\, \mathbb{1}\{N(s_0,\pi^*(s_0))\geq 1\}\right\}\right] \\
&\qquad + \gamma \mathbb{E}_{s_0\sim d_0}[\mathbb{E}_{s'\sim \widehat{P}^o_{s_0,\pi^*(s_0)}}(V^{\widehat{\pi}^*}_{\widehat{\mathcal{P}}}(s')) - \mathbb{E}_{s'\sim P_{s_0,\pi^*(s_0)}}(V^{\widehat{\pi}^*}_{\widehat{\mathcal{P}}}(s'))] \\
&\overset{(h)}{\leq} \frac{4\gamma}{1-\gamma}\mathbb{E}_{s_0\sim d_0}\left[\sqrt{\frac{\max\{|\mathcal{S}|, 2\log(2|\mathcal{S}||\mathcal{A}|/\delta)\}}{N(s_0,\pi^*(s_0))}}\, \mathbb{1}\{N(s_0,\pi^*(s_0))\geq 1\}\right], \qquad (17)
\end{aligned}$$

where (g), holds with probability at least $1-\delta$, follows from Hölder's inequality and by Proposition 1, and (h) by Hölder's inequality and the definition of uncertainty set $\widehat{\mathcal{P}}$.

Substituting Eq. (17) back in Eq. (15), we get the following recursion

$$\begin{aligned}
\mathbb{E}_{s_0\sim d_0}[V^{\pi^*}_{P^o}(s_0) - V^{\widehat{\pi}^*}_{\widehat{\mathcal{P}}}(s_0)] &\leq \gamma \mathbb{E}_{s_0\sim d_0}[\mathbb{E}_{s'\sim P^o_{s_0,\pi^*(s_0)}}(V^{\pi^*}_{P^o}(s') - V^{\widehat{\pi}^*}_{\widehat{\mathcal{P}}}(s'))] \\
&\quad + \frac{4\gamma}{1-\gamma}\mathbb{E}_{s_0\sim d_0}\left[\sqrt{\frac{\max\{|\mathcal{S}|, 2\log(2|\mathcal{S}||\mathcal{A}|/\delta)\}}{N(s_0,\pi^*(s_0))}}\, \mathbb{1}\{N(s_0,\pi^*(s_0))\geq 1\}\right] \\
&= \gamma \mathbb{E}_{s_1\sim d^{\pi^*}_{P^o,1}}[V^{\pi^*}_{P^o}(s_1) - V^{\widehat{\pi}^*}_{\widehat{\mathcal{P}}}(s_1)] \\
&\quad + \frac{4\gamma}{1-\gamma}\mathbb{E}_{s_0\sim d_0}\left[\sqrt{\frac{\max\{|\mathcal{S}|, 2\log(2|\mathcal{S}||\mathcal{A}|/\delta)\}}{N(s_0,\pi^*(s_0))}}\, \mathbb{1}\{N(s_0,\pi^*(s_0))\geq 1\}\right] \\
&\leq \gamma^2 \mathbb{E}_{s_2\sim d^{\pi^*}_{P^o,2}}[V^{\pi^*}_{P^o}(s_2) - V^{\widehat{\pi}^*}_{\widehat{\mathcal{P}}}(s_2)] \\
&\quad + \gamma\frac{4\gamma}{1-\gamma}\mathbb{E}_{s_1\sim d^{\pi^*}_{P^o,1}}\left[\sqrt{\frac{\max\{|\mathcal{S}|, 2\log(2|\mathcal{S}||\mathcal{A}|/\delta)\}}{N(s_1,\pi^*(s_1))}}\, \mathbb{1}\{N(s_1,\pi^*(s_1))\geq 1\}\right] \\
&\quad + \frac{4\gamma}{1-\gamma}\mathbb{E}_{s_0\sim d_0}\left[\sqrt{\frac{\max\{|\mathcal{S}|, 2\log(2|\mathcal{S}||\mathcal{A}|/\delta)\}}{N(s_0,\pi^*(s_0))}}\, \mathbb{1}\{N(s_0,\pi^*(s_0))\geq 1\}\right] \\
&\leq \frac{4\gamma}{1-\gamma}\sum_{t=0}^{\infty}\gamma^t \mathbb{E}_{s_t\sim d^{\pi^*}_{P^o,t}}\left[\sqrt{\frac{\max\{|\mathcal{S}|, 2\log(2|\mathcal{S}||\mathcal{A}|/\delta)\}}{N(s_t,\pi^*(s_t))}}\, \mathbb{1}\{N(s_t,\pi^*(s_t))\geq 1\}\right] \\
&= \frac{4\gamma}{(1-\gamma)^2}\mathbb{E}_{s\sim d^{\pi^*}_{P^o}}\left[\sqrt{\frac{\max\{|\mathcal{S}|, 2\log(2|\mathcal{S}||\mathcal{A}|/\delta)\}}{N(s,\pi^*(s))}}\, \mathbb{1}\{N(s,\pi^*(s))\geq 1\}\right],
\end{aligned}$$

where last equality follows by the definition of state-distribution $d^{\pi^*}_{P^o} = (1-\gamma)\sum_{t=0}^{\infty}\gamma^t d^{\pi^*}_{P^o,t}$. Now, putting this back in Eq. (14), we see that the offline RL guarantee becomes:

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}}[\mathbb{E}_{s_0\sim d_0}[V^{\pi^*}_{P^o}(s_0) - V^{\pi_K}_{P^o}(s_0)]] \\
&\leq \frac{2\gamma^{K+1}}{(1-\gamma)^2} + \frac{4\gamma}{(1-\gamma)^2}\mathbb{E}_{s\sim d^{\pi^*}_{P^o}}\mathbb{E}_{\mathcal{D}}\left[\sqrt{\frac{\max\{|\mathcal{S}|, 2\log(2|\mathcal{S}||\mathcal{A}|/\delta)\}}{N(s,\pi^*(s))}}\, \mathbb{1}\{N(s,\pi^*(s))\geq 1\}\right] \\
&\leq \frac{2\gamma^{K+1}}{(1-\gamma)^2} + \frac{4\gamma}{(1-\gamma)^2}\mathbb{E}_{s\sim d^{\pi^*}_{P^o}}\mathbb{E}_{\mathcal{D}}\left[\sqrt{\frac{\max\{|\mathcal{S}|, 2\log(2|\mathcal{S}||\mathcal{A}|/\delta)\}}{N(s,\pi^*(s))\vee 1}}\right] \\
&\overset{(i)}{\leq} \frac{2\gamma^{K+1}}{(1-\gamma)^2} + \frac{4\gamma}{(1-\gamma)^2}\mathbb{E}_{s\sim d^{\pi^*}_{P^o}}\left[\sqrt{\max\{|\mathcal{S}|, 2\log(2|\mathcal{S}||\mathcal{A}|/\delta)\}}\frac{16}{\sqrt{N\mu(s,\pi^*(s))}}\right]
\end{aligned}$$

$$
\begin{aligned}
&\overset{(j)}{\leq} \frac{2\gamma^{K+1}}{(1-\gamma)^2} + \frac{64\gamma}{(1-\gamma)^2}\mathbb{E}_{s\sim d^{\pi^*}_{P^o}}\left[\sqrt{\frac{C_{\pi^*}\max\{|\mathcal{S}|, 2\log(2|\mathcal{S}||\mathcal{A}|/\delta)\}}{N d^{\pi^*}_{P^o}(s,\pi^*(s))}}\right] \\
&= \frac{2\gamma^{K+1}}{(1-\gamma)^2} + \frac{64\gamma\sqrt{C_{\pi^*}}}{(1-\gamma)^2}\sqrt{\frac{\max\{|\mathcal{S}|, 2\log(2|\mathcal{S}||\mathcal{A}|/\delta)\}}{N}}\sum_s \sqrt{d^{\pi^*}_{P^o}(s,\pi^*(s))} \\
&\overset{(k)}{\leq} \frac{2\gamma^{K+1}}{(1-\gamma)^2} + \frac{64\gamma\sqrt{C_{\pi^*}}}{(1-\gamma)^2}\sqrt{\frac{\max\{|\mathcal{S}|, 2\log(2|\mathcal{S}||\mathcal{A}|/\delta)\}}{N}}\sqrt{|\mathcal{S}|}.
\end{aligned} \tag{18}
$$

Recall that (s_i, a_i)-pairs in $\mathcal{D}$ are i.i.d. and follow the data generating policy μ. That is, for any (s,a), $N(s,a)$ follows $\mathrm{Binomial}(N, \mu(s,a))$. Then (i) follows from Lemma 1 with $k = 1/2$. We note here that this technique of bridging two visitation distributions, μ and $d^{\pi^*}_{P^o}$, is critical and original in our paper. We have (j) by recalling the definition of single-policy concentrability with comparator policy π^*, that is,

$$
C_{\pi^*} = \max_{s,a}\frac{d^{\pi^*}_{P^o}(s,a)}{\mu(s,a)}.
$$

(k) is due to Cauchy-Schwarz inequality and by recognizing $d^{\pi^*}_{P^o}(\cdot,\pi^*(\cdot))$ as a probability distribution. This completes the proof of this main theorem.

□

We now provide a similar offline RL suboptimality guarantee below for the Wasserstein uncertainty set using Proposition 2.

Theorem 4. *Let π_K be the DRQI policy after K iterations under the Wasserstein uncertainty set $\widehat{\mathcal{P}}^{\mathrm{w}}$. With probability at least $1-\delta$ it holds that*

$$
\mathbb{E}_{s_0\sim d_0}[V^{\pi^*}(s_0) - \mathbb{E}_{\mathcal{D}}[V^{\pi_K}(s_0)]] \leq \frac{64\gamma\sqrt{C_{\pi^*}}}{(1-\gamma)^2}\sqrt{\frac{C|\mathcal{S}|^2\log(|\mathcal{S}||\mathcal{A}|/\delta)}{N}} + \frac{2\gamma^{K+1}}{(1-\gamma)^2}.
$$

Proof. The proof follows exactly as in the proof of Theorem 3. We replace the dependence on Proposition 1 with Proposition 2. We then only have to take care of step (g) in Eq. (17). We start from analyzing (I) as in Eq. (16):

$$
\begin{aligned}
&\mathbb{E}_{s_0\sim d_0}[\gamma\mathbb{E}_{s'\sim P^o_{s_0,\pi^*(s_0)}}(V^{\widehat{\pi}^*}_{\widehat{\mathcal{P}}}(s')) - \gamma\mathbb{E}_{s'\sim P_{s_0,\pi^*(s_0)}}(V^{\widehat{\pi}^*}_{\widehat{\mathcal{P}}}(s'))] \\
&= \mathbb{E}_{s_0\sim d_0}[\gamma\mathbb{E}_{s'\sim P^o_{s_0,\pi^*(s_0)}}(V^{\widehat{\pi}^*}_{\widehat{\mathcal{P}}}(s')) - \gamma\mathbb{E}_{s'\sim \widehat{P}^o_{s_0,\pi^*(s_0)}}(V^{\widehat{\pi}^*}_{\widehat{\mathcal{P}}}(s')) \\
&\qquad + \gamma\mathbb{E}_{s'\sim \widehat{P}^o_{s_0,\pi^*(s_0)}}(V^{\widehat{\pi}^*}_{\widehat{\mathcal{P}}}(s')) - \gamma\mathbb{E}_{s'\sim P_{s_0,\pi^*(s_0)}}(V^{\widehat{\pi}^*}_{\widehat{\mathcal{P}}}(s'))] \\
&\overset{(a)}{\leq} \mathbb{E}_{s_0\sim d_0}\Bigg[\frac{2\gamma}{1-\gamma}D_{\mathrm{w}}(P^o_{s_0,\pi^*(s_0)}, \widehat{P}^o_{s_0,\pi^*(s_0)}) \\
&\qquad + \gamma\mathbb{E}_{s'\sim \widehat{P}^o_{s_0,\pi^*(s_0)}}(V^{\widehat{\pi}^*}_{\widehat{\mathcal{P}}}(s')) - \gamma\mathbb{E}_{s'\sim P_{s_0,\pi^*(s_0)}}(V^{\widehat{\pi}^*}_{\widehat{\mathcal{P}}}(s'))\Bigg] \\
&\overset{(b)}{\leq} \frac{2\gamma}{1-\gamma}\mathbb{E}_{s_0\sim d_0}\left[\sqrt{\frac{C|\mathcal{S}|\log(|\mathcal{S}||\mathcal{A}|/\delta)}{N(s_0,\pi^*(s_0))}}\mathbb{1}\{N(s_0,\pi^*(s_0))\geq 1\}\right] \\
&\qquad + \gamma\mathbb{E}_{s_0\sim d_0}[\mathbb{E}_{s'\sim \widehat{P}^o_{s_0,\pi^*(s_0)}}(V^{\widehat{\pi}^*}_{\widehat{\mathcal{P}}}(s')) - \gamma\mathbb{E}_{s'\sim P_{s_0,\pi^*(s_0)}}(V^{\widehat{\pi}^*}_{\widehat{\mathcal{P}}}(s'))] \\
&\overset{(c)}{\leq} \frac{4\gamma}{1-\gamma}\mathbb{E}_{s_0\sim d_0}\left[\sqrt{\frac{C|\mathcal{S}|\log(|\mathcal{S}||\mathcal{A}|/\delta)}{N(s_0,\pi^*(s_0))}}\mathbb{1}\{N(s_0,\pi^*(s_0))\geq 1\}\right],
\end{aligned}
$$

where (a) follows by applying the Kantorovich-Rubinstein theorem (Dudley, 2002, Theorem 11.8.2) and noting the fact that the value functions are $2/(1-\gamma)$-Lipschitz in their state dimension under the discrete metric $\ell(\cdot,\cdot)$ since $\left\|V^{\widehat{\pi}^*}_{\widehat{\mathcal{P}}}\right\|_\infty \leq 1/(1-\gamma)$, (b) holds with probability at least $1-\delta$ by Proposition 2, and (c) is again by the Kantorovich-Rubinstein theorem and the definition of uncertainty set $\widehat{\mathcal{P}}$. Now combining and analyzing the rest of the steps as in the proof of Theorem 3 completes the proof. □

We now provide a similar offline RL suboptimality guarantee below for the KL uncertainty set using Proposition 3.

Theorem 5. *Let π_K be the DRQI policy after K iterations under the KL uncertainty set $\widehat{\mathcal{P}}^{\mathrm{kl}}$ (under add-1 estimator). With probability at least $1-\delta$ it holds that*

$$\mathbb{E}_{s_0\sim d_0}[V^{\pi^*}(s_0) - \mathbb{E}_{\mathcal{D}}[V^{\pi_K}(s_0)]] \le \frac{64\gamma\sqrt{C_{\pi^*}}}{(1-\gamma)^2}\sqrt{\frac{C|\mathcal{S}|^2\log(|\mathcal{S}|^2|\mathcal{A}|/\delta)\log(N)}{N}} + \frac{2\gamma^{K+1}}{(1-\gamma)^2}.$$

Proof. The proof again follows exactly as in the proof of Theorem 3. We replace the dependence on Proposition 1 with Proposition 3. We then only have to take care of step (g) in Eq. (17). We start from analyzing (I) as in Eq. (16):

$$\begin{aligned}
&\mathbb{E}_{s_0\sim d_0}[\gamma\mathbb{E}_{s'\sim P^o_{s_0,\pi^*(s_0)}}(V^{\widehat{\pi}^*}_{\widehat{\mathcal{P}}}(s')) - \gamma\mathbb{E}_{s'\sim P_{s_0,\pi^*(s_0)}}(V^{\widehat{\pi}^*}_{\widehat{\mathcal{P}}}(s'))]\\
&= \mathbb{E}_{s_0\sim d_0}[\gamma\mathbb{E}_{s'\sim P^o_{s_0,\pi^*(s_0)}}(V^{\widehat{\pi}^*}_{\widehat{\mathcal{P}}}(s')) - \gamma\mathbb{E}_{s'\sim \widetilde{P}^o_{s_0,\pi^*(s_0)}}(V^{\widehat{\pi}^*}_{\widehat{\mathcal{P}}}(s'))\\
&\qquad + \gamma\mathbb{E}_{s'\sim \widetilde{P}^o_{s_0,\pi^*(s_0)}}(V^{\widehat{\pi}^*}_{\widehat{\mathcal{P}}}(s')) - \gamma\mathbb{E}_{s'\sim P_{s_0,\pi^*(s_0)}}(V^{\widehat{\pi}^*}_{\widehat{\mathcal{P}}}(s'))]\\
&\overset{(a)}{\le} \mathbb{E}_{s_0\sim d_0}[\gamma\sqrt{2\ln(2)D_{\mathrm{KL}}(P^o_{s_0,\pi^*(s_0)},\widetilde{P}^o_{s_0,\pi^*(s_0)})}\|V^{\widehat{\pi}^*}_{\widehat{\mathcal{P}}}(s')\|_\infty]\\
&\qquad + \gamma\mathbb{E}_{s'\sim \widetilde{P}^o_{s_0,\pi^*(s_0)}}(V^{\widehat{\pi}^*}_{\widehat{\mathcal{P}}}(s')) - \gamma\mathbb{E}_{s'\sim P_{s_0,\pi^*(s_0)}}(V^{\widehat{\pi}^*}_{\widehat{\mathcal{P}}}(s'))]\\
&\overset{(b)}{\le} \frac{2\gamma}{1-\gamma}\mathbb{E}_{s_0\sim d_0}\left[\sqrt{\frac{C|\mathcal{S}|\log(|\mathcal{S}|^2|\mathcal{A}|/\delta)\log(N)}{N(s_0,\pi^*(s_0))}}\mathbb{1}\{N(s_0,\pi^*(s_0))\ge 1\}\right]\\
&\qquad + \gamma\mathbb{E}_{s_0\sim d_0}[\mathbb{E}_{s'\sim \widetilde{P}^o_{s_0,\pi^*(s_0)}}(V^{\widehat{\pi}^*}_{\widehat{\mathcal{P}}}(s')) - \gamma\mathbb{E}_{s'\sim P_{s_0,\pi^*(s_0)}}(V^{\widehat{\pi}^*}_{\widehat{\mathcal{P}}}(s'))]\\
&\overset{(c)}{\le} \frac{4\gamma}{1-\gamma}\mathbb{E}_{s_0\sim d_0}\left[\sqrt{\frac{C|\mathcal{S}|\log(|\mathcal{S}|^2|\mathcal{A}|/\delta)\log(N)}{N(s_0,\pi^*(s_0))}}\mathbb{1}\{N(s_0,\pi^*(s_0))\ge 1\}\right],
\end{aligned}$$

where (a) follows from Hölder's inequality and Pinsker's inequality (Cover & Thomas, 1991, Lemma 12.6.1), (b) holds with probability at least $1-\delta$ by Proposition 3, and (c) again follows from Hölder's inequality and Pinsker's inequality under the definition of uncertainty set $\widehat{\mathcal{P}}$. Now combining and analyzing the rest of the steps as in the proof of Theorem 3 completes the proof. □

We also provide a similar offline RL suboptimality guarantee below for the chi-square uncertainty set using Proposition 4.

Theorem 6. *Let π_K be the DRQI policy after K iterations under the chi-square uncertainty set $\widehat{\mathcal{P}}^{\mathrm{c}}$ (under add-$\log(1/\delta)$ estimator). With probability at least $1-\delta$ it holds that*

$$\mathbb{E}_{s_0\sim d_0}[V^{\pi^*}(s_0) - \mathbb{E}_{\mathcal{D}}[V^{\pi_K}(s_0)]] \le \frac{64\gamma\sqrt{C_{\pi^*}}}{(1-\gamma)^2}\sqrt{\frac{C|\mathcal{S}|^2\log(|\mathcal{S}|^2|\mathcal{A}|/\delta)}{N}} + \frac{2\gamma^{K+1}}{(1-\gamma)^2}.$$

Proof. The proof again follows exactly as in the proof of Theorem 3. We replace the dependence on Proposition 1 with Proposition 4. We again only have to take care of step (g) in Eq. (17). We start from analyzing (I) as in Eq. (16):

$$\begin{aligned}
&\mathbb{E}_{s_0\sim d_0}[\gamma\mathbb{E}_{s'\sim P^o_{s_0,\pi^*(s_0)}}(V^{\widehat{\pi}^*}_{\widehat{\mathcal{P}}}(s')) - \gamma\mathbb{E}_{s'\sim P_{s_0,\pi^*(s_0)}}(V^{\widehat{\pi}^*}_{\widehat{\mathcal{P}}}(s'))]\\
&= \mathbb{E}_{s_0\sim d_0}[\gamma\mathbb{E}_{s'\sim P^o_{s_0,\pi^*(s_0)}}(V^{\widehat{\pi}^*}_{\widehat{\mathcal{P}}}(s')) - \gamma\mathbb{E}_{s'\sim \widetilde{P}^o_{s_0,\pi^*(s_0)}}(V^{\widehat{\pi}^*}_{\widehat{\mathcal{P}}}(s'))\\
&\qquad + \gamma\mathbb{E}_{s'\sim \widetilde{P}^o_{s_0,\pi^*(s_0)}}(V^{\widehat{\pi}^*}_{\widehat{\mathcal{P}}}(s')) - \gamma\mathbb{E}_{s'\sim P_{s_0,\pi^*(s_0)}}(V^{\widehat{\pi}^*}_{\widehat{\mathcal{P}}}(s'))]\\
&\overset{(a)}{\le} \mathbb{E}_{s_0\sim d_0}[2\gamma\sqrt{D_{\mathrm{c}}(P^o_{s_0,\pi^*(s_0)},\widetilde{P}^o_{s_0,\pi^*(s_0)})}\|V^{\widehat{\pi}^*}_{\widehat{\mathcal{P}}}\|_\infty]\\
&\qquad + \gamma[\mathbb{E}_{s'\sim \widetilde{P}^o_{s_0,\pi^*(s_0)}}(V^{\widehat{\pi}^*}_{\widehat{\mathcal{P}}}(s')) - \gamma\mathbb{E}_{s'\sim P_{s_0,\pi^*(s_0)}}(V^{\widehat{\pi}^*}_{\widehat{\mathcal{P}}}(s'))]\\
&\overset{(b)}{\le} \frac{2\gamma}{1-\gamma}\mathbb{E}_{s_0\sim d_0}\left[\sqrt{\frac{C|\mathcal{S}|\log(|\mathcal{S}|^2|\mathcal{A}|/\delta)}{N(s_0,\pi^*(s_0))}}\mathbb{1}\{N(s_0,\pi^*(s_0))\ge 1\}\right]
\end{aligned}$$

$$+ \gamma \mathbb{E}_{s_0 \sim d_0}[\mathbb{E}_{s' \sim \widetilde{P}^o_{s_0,\pi^*(s_0)}}(V^{\widehat{\pi}^*}_{\widehat{\mathcal{P}}}(s')) - \gamma \mathbb{E}_{s' \sim P_{s_0,\pi^*(s_0)}}(V^{\widehat{\pi}^*}_{\widehat{\mathcal{P}}}(s'))]$$
$$\overset{(c)}{\leq} \frac{4\gamma}{1-\gamma} \mathbb{E}_{s_0 \sim d_0}\left[\sqrt{\frac{C|\mathcal{S}|\log(|\mathcal{S}|^2|\mathcal{A}|/\delta)}{N(s_0,\pi^*(s_0))}} \mathbb{1}\{N(s_0,\pi^*(s_0)) \geq 1\}\right],$$

where (a) follows from Hölder's inequality, and from Pinsker's inequality (Cover & Thomas, 1991, Lemma 12.6.1) and (Basu et al., 2011, Lemma 11.1) we have $D_{\mathrm{TV}}(p,q) \leq 2\sqrt{D_{\mathrm{c}}(p,q)}$ for any two distributions, (b) holds with probability at least $1-\delta$ by Proposition 4, and (c) follows same as (a) but under the definition of uncertainty set $\widehat{\mathcal{P}}$. Now combining and analyzing the rest of the steps as in the proof of Theorem 3 completes the proof. □

C. Results and Proofs of LM-DRQI

In the following, we always use $c > 0$ for a small universal constant whose exact value might be changing. We allow $\lambda = \Omega(1)$ but set $\lambda = 1$ for simplicity. In what follows, we use $\mathbb{1}_i \in \mathbb{R}^{d \times 1}$ to denote vector with values 0 except 1 at position i. We first make a similar observation as in Proposition 1-Proposition 4 that the true model P^o lies in the uncertainty set $\widehat{\mathcal{P}}$ with high probability. We make this formal in the proposition below.

Proposition 5. *We have $\nu^o \in \widehat{\mathcal{M}}$ with probability at least $1-\delta$. Furthermore, $P^o \in \widehat{\mathcal{P}}$ also holds with probability at least $1-\delta$.*

Proof. Let $\mathbb{1}_i \in \mathbb{R}^{d \times 1}$ denote vector with values 0 except 1 at position i and $\mathbb{1}(s'_t) \in \mathbb{R}^{|\mathcal{S}| \times 1}$ denote vector with values 0 except 1 at position s'_t. Fixing an $i \in [d]$ and $V \in \mathcal{V}$, we have the following:

$$\mathbb{E}_{\nu^o_i}[V] - \mathbb{E}_{\widehat{\nu}_i}[V] = (\nu^o_i)^\top V - (\widehat{\nu}_i)^\top V = \mathbb{1}_i^\top (\nu^o)^\top V - \mathbb{1}_i^\top (\widehat{\nu})^\top V$$
$$\overset{(a)}{=} \mathbb{1}_i^\top \Lambda_N^{-1}(\frac{\lambda}{N} I + \frac{1}{N}\sum_{t=1}^N \phi(s_t,a_t)\phi(s_t,a_t)^\top)(\nu^o)^\top V - \mathbb{1}_i^\top (\widehat{\nu})^\top V$$
$$\overset{(b)}{=} \frac{\lambda}{N}\mathbb{1}_i^\top \Lambda_N^{-1}(\nu^o)^\top V + \frac{1}{N}\mathbb{1}_i^\top \Lambda_N^{-1}\sum_{t=1}^N \phi(s_t,a_t)(P^o_{s_t,a_t})^\top V - \mathbb{1}_i^\top (\widehat{\nu})^\top V$$
$$\overset{(c)}{=} \frac{\lambda}{N}\mathbb{1}_i^\top \Lambda_N^{-1}(\nu^o)^\top V + \frac{1}{N}\mathbb{1}_i^\top \Lambda_N^{-1}\sum_{t=1}^N \phi(s_t,a_t)(P^o_{s_t,a_t})^\top V - \frac{1}{N}\mathbb{1}_i^\top \Lambda_N^{-1}\sum_{t=1}^N \phi(s_t,a_t)\mathbb{1}(s'_t)^\top V$$
$$\overset{(d)}{=} \frac{\lambda}{N}\mathbb{1}_i^\top \Lambda_N^{-1}(\nu^o)^\top V + \frac{1}{N}\mathbb{1}_i^\top \Lambda_N^{-1}\sum_{t=1}^N \phi(s_t,a_t)\epsilon_t^\top V, \tag{19}$$

where (a) is by $\Lambda_N = \frac{\lambda}{N}I + \frac{1}{N}\sum_{t=1}^N \phi(s_t,a_t)\phi(s_t,a_t)^\top$, (b) by $\phi(s_t,a_t)^\top)(\nu^o)^\top = P^o_{s_t,a_t}(\cdot)$, (c) by $\widehat{\nu}(s') = \frac{1}{N}\Lambda_N^{-1}\sum_{t=1}^N \phi(s_t,a_t)\mathbb{1}\{s' = s'_t\}$, and (d) by setting $\epsilon_t = (P^o_{s_t,a_t} - \mathbb{1}(s'_t))$.

Before proceeding, here is a consequence of Assumption 1. Consider any $(s,a) \in \mathcal{S} \times \mathcal{A}$. For any linear MDP $P_{s,a}(s') = \phi(s,a)^\top \nu(s')$, summing both sides across s', we get

$$1 = \sum_{s'} P_{s,a}(s') = \phi(s,a)^\top \sum_{s'} \nu(s') = \sum_{i \in [d]} \phi_i(s,a).$$

Since $\phi_i(s,a) \geq 0$ and $\|x\|_2 \leq \|x\|_1$ for $x \in \mathbb{R}^d$, $\|\phi(s,a)\|_2 \leq 1$ follows. Now we analyze the two terms in Eq. (19). First,

$$|\frac{\lambda}{N}\mathbb{1}_i^\top \Lambda_N^{-1}(\nu^o)^\top V| \leq \frac{\lambda}{N}\left\|\mathbb{1}_i^\top \Lambda_N^{-1}\right\|_1 \left\|(\nu^o)^\top V\right\|_\infty$$
$$\overset{(e)}{\leq} \frac{1}{1-\gamma}\frac{\lambda}{N}\left\|\mathbb{1}_i^\top \Lambda_N^{-1}\right\|_1$$
$$\overset{(f)}{\leq} \frac{\sqrt{d}}{1-\gamma}\frac{\lambda}{N}\left\|\mathbb{1}_i^\top \Lambda_N^{-1}\right\|_2$$
$$= \frac{\sqrt{d}}{1-\gamma}\frac{\lambda}{N}\sqrt{\mathbb{1}_i^\top \Lambda_N^{-1}\Lambda_N^{-1}\mathbb{1}_i}$$

$$\stackrel{(g)}{\leq} \frac{\sqrt{d}}{1-\gamma}\frac{\lambda}{N}\sqrt{\left\|\Lambda_N^{-1}\right\|_{\mathrm{op}}}\sqrt{\mathbb{1}_i^\top \Lambda_N^{-1}\mathbb{1}_i}$$
$$= \frac{\sqrt{d}}{1-\gamma}\frac{\lambda}{N}\left\|\Lambda_N^{-1/2}\right\|_{\mathrm{op}}\|\mathbb{1}_i\|_{\Lambda_N^{-1}}$$
$$\stackrel{(h)}{\leq} \frac{1}{1-\gamma}\sqrt{\frac{d\lambda}{N}}\sqrt{\Lambda_N^{-1}(i,i)},$$

where (e) follows since $V \in \mathcal{V} = \{V(\cdot) = \max_a \phi^\top(\cdot,a)w : w \in \mathbb{R}^d, \|w\|_2 \leq 1/(1-\gamma)\}$ satisfies $|V(s)| \leq \left\|\max_a \phi^\top(s,a)\right\|_2 \|w\|_2 \leq 1/(1-\gamma)$ for any $s \in \mathcal{S}$ and ν^o is a probability distribution, (f) by $\|x\|_1 \leq \sqrt{d}\|x\|_2$ for $x \in \mathbb{R}^d$, (g) by $x^\top A y \leq \|A\|_{\mathrm{op}} x^\top y$ for positive definite matrix A with maximum eigenvalue $\|A\|_{\mathrm{op}}$ and for $x, y \in \mathbb{R}^d$, and (h) follows since Λ_N's minimal absolute value is λ/N in its diagonal entries.

Second, by Cauchy-Schwarz on Λ_N^{-1}-norm,

$$|\frac{1}{N}\mathbb{1}_i^\top \Lambda_N^{-1}\sum_{t=1}^N \phi(s_t,a_t)\epsilon_t^\top V| \leq \frac{1}{N}\left\|\mathbb{1}_i^\top\right\|_{\Lambda_N^{-1}}\left\|\sum_{t=1}^N \phi(s_t,a_t)\epsilon_t^\top V\right\|_{\Lambda_N^{-1}} = \frac{1}{N}\sqrt{\Lambda_N^{-1}(i,i)}\left\|\sum_{t=1}^N \phi(s_t,a_t)\epsilon_t^\top V\right\|_{\Lambda_N^{-1}}.$$

We now get back to analyzing Eq. (19) using these intermediate steps. Fix $i \in [d]$. For all $V \in \mathcal{V}$, we have the following uniform bound:

$$|\mathbb{E}_{\nu_i^o}[V] - \mathbb{E}_{\widehat{\nu}_i}[V]| \leq \frac{1}{1-\gamma}\sqrt{\frac{d\lambda}{N}}\sqrt{\Lambda_N^{-1}(i,i)} + \frac{1}{N}\sqrt{\Lambda_N^{-1}(i,i)}\left\|\sum_{t=1}^N \phi(s_t,a_t)\epsilon_t^\top V\right\|_{\Lambda_N^{-1}}$$
$$\stackrel{(i)}{\leq} \frac{1}{1-\gamma}\sqrt{\frac{d\lambda}{N}}\sqrt{\Lambda_N^{-1}(i,i)} + \frac{1}{N}\sqrt{\Lambda_N^{-1}(i,i)}\mathcal{O}\left(\frac{\sqrt{dN}\log(N/((1-\gamma)\delta))}{1-\gamma}\right)$$
$$\leq \frac{\mathcal{O}(\log(N/((1-\gamma)\delta)))}{1-\gamma}\sqrt{\frac{d}{N}}\sqrt{\Lambda_N^{-1}(i,i)},$$

where (i) holds with probability $1-\delta$ by Lemma 7.

Let $c_1 > 0$ be some universal constant. Furthermore, with an additional uniform bound, the following holds for all $i \in [d]$ with probability at least $1-\delta$:

$$\mathrm{d}_{\mathcal{V}}(\nu_i^o, \widehat{\nu}_i) \leq \frac{c_1 \log(Nd/((1-\gamma)\delta))}{1-\gamma}\sqrt{\frac{d}{N}}\sqrt{\Lambda_N^{-1}(i,i)}. \tag{20}$$

It is now straightforward to see $\nu^o \in \widehat{\mathcal{M}}$ holds with probability at least $1-\delta$ by recalling:

$$\widehat{\mathcal{M}} = \bigotimes_{i\in[d]} \widehat{\mathcal{M}}_i \quad \text{where} \quad \widehat{\mathcal{M}}_i = \left\{\nu_i \in \Delta(\mathcal{S}) : \mathrm{d}_{\mathcal{V}}(\nu_i, \widehat{\nu}_i) \leq \frac{c_1 \log(Nd/((1-\gamma)\delta))}{1-\gamma}\sqrt{\frac{d}{N}}\sqrt{\Lambda_N^{-1}(i,i)}\right\}.$$

Furthermore, recall $P_{s,a}^o(s') = \sum_{i\in[d]} \phi_i(s,a)\nu_i^o(s')$ and $\widehat{P}_{s,a}^o(s') = \sum_{i\in[d]} \phi_i(s,a)\widehat{\nu}_i(s')$. We now have the following equations:

$$\sup_{V\in\mathcal{V}}\left|\int_{\mathcal{S}}(P_{s,a}^o - \widehat{P}_{s,a})V(ds')\right| = \sup_{V\in\mathcal{V}}\left|\int_{\mathcal{S}}\sum_{i=1}^d \phi_i(s,a)(\nu_i^o(s') - \widehat{\nu}_i(s'))V(ds')\right|$$
$$= \sup_{V\in\mathcal{V}}\left|\sum_{i=1}^d \phi_i(s,a)\int_{\mathcal{S}}(\nu_i^o(s') - \widehat{\nu}_i(s'))V(ds')\right| \leq \sup_{V\in\mathcal{V}}\sum_{i=1}^d |\phi_i(s,a)|\left|\int_{\mathcal{S}}(\nu_i^o(s') - \widehat{\nu}_i(s'))V(ds')\right|$$
$$\leq \sum_{i=1}^d |\phi_i(s,a)| \sup_{V\in\mathcal{V}}\left|\int_{\mathcal{S}}(\nu_i^o(s') - \widehat{\nu}_i(s'))V(ds')\right|$$

$$= \sum_{i=1}^{d} |\phi_i(s,a)| \cdot \mathrm{d}_{\mathcal{V}}(\nu_i^o, \widehat{\nu}_i) \le \frac{c_1 \log(Nd/((1-\gamma)\delta))}{1-\gamma} \sqrt{\frac{d}{N}} \sum_{i=1}^{d} \|\phi_i(s,a)\mathbb{1}_i\|_{\Lambda_N^{-1}}\,, \tag{21}$$

where the last inequality follows by Eq. (20). This holds with probability at least $1-\delta$ for all s, a together. Thus we have a high probability event that $P^o \in \widehat{\mathcal{P}}$ with probability at least $1-\delta$. □

Before presenting our main result we adapt (Jin et al., 2021, Corollary 4.5) to present a high probability result adhering to the sufficient coverage assumption (Assumption 2).

Lemma 9. *For any s, a, we have with probability at least $1-\delta$ that $\sum_{i\in[d]} \mathbb{E}_{s,a\sim d^{\pi^*}_{P^o}}[\|\phi_i(s,a)\mathbb{1}_i^\top\|_{\Lambda_N^{-1}}] \le \sqrt{\mathrm{rank}(\Sigma_{d^{\pi^*}_{P^o}})/C^\dagger_{\mathrm{sc}}}$ where $\Lambda_N = \frac{\lambda}{N}I + \frac{1}{N}\sum_{t=1}^{N}\phi(s_t,a_t)\phi(s_t,a_t)^\top$, $\Sigma_{d^{\pi^*}_{P^o}} = \mathbb{E}_{s,a\sim d^{\pi^*}_{P^o}}\phi(s,a)\phi(s,a)^\top$.*

Proof. This proof follows similar steps in the proof of (Jin et al., 2021, Corollary 4.5). Firstly notice,

$$\begin{aligned}
\sum_{i\in[d]} \mathbb{E}_{s,a\sim d^{\pi^*}_{P^o}}[\|\phi_i(s,a)\mathbb{1}_i\|_{\Lambda_N^{-1}}] &= \sum_{i\in[d]} \mathbb{E}_{s,a\sim d^{\pi^*}_{P^o}}[\sqrt{(\phi_i(s,a)\mathbb{1}_i)^\top \Lambda_N^{-1}(\phi_i(s,a)\mathbb{1}_i)}] \\
&= \sum_{i\in[d]} \mathbb{E}_{s,a\sim d^{\pi^*}_{P^o}}[\sqrt{\mathrm{Tr}((\phi_i(s,a)\mathbb{1}_i)(\phi_i(s,a)\mathbb{1}_i)^\top \Lambda_N^{-1})}] \\
&\overset{(a)}{\le} \sqrt{d}\sqrt{\sum_{i\in[d]} \mathrm{Tr}(\mathbb{E}_{s,a\sim d^{\pi^*}_{P^o}}[(\phi_i(s,a)\mathbb{1}_i)(\phi_i(s,a)\mathbb{1}_i)^\top]\Lambda_N^{-1})} \\
&\overset{(b)}{\le} \sqrt{d}\sqrt{\sum_{i\in[d]} \mathrm{Tr}(\Sigma^i_{d^{\pi^*}_{P^o}} \cdot ((1/N)I + C^\dagger_{\mathrm{sc}}\cdot d \cdot \Sigma^i_{d^{\pi^*}_{P^o}})^{-1})} \\
&\overset{(c)}{=} \sqrt{d}\sqrt{\sum_{i\in[d]} \frac{\lambda^i_{d^{\pi^*}_{P^o}}}{(1/N) + C^\dagger_{\mathrm{sc}}\cdot d\cdot \lambda^i_{d^{\pi^*}_{P^o}}}} \\
&\overset{(d)}{\le} \sqrt{d}\sqrt{\frac{\mathrm{rank}(\Sigma_{d^{\pi^*}_{P^o}})}{(1/(N\cdot \max_{i\in[d]}\lambda^i_{d^{\pi^*}_{P^o}})) + C^\dagger_{\mathrm{sc}}\cdot d}} \le \sqrt{\frac{\mathrm{rank}(\Sigma_{d^{\pi^*}_{P^o}})}{C^\dagger_{\mathrm{sc}}}},
\end{aligned}$$

where (a) follows by $\|x\|_1 \le \sqrt{d}\|x\|_2$ for $x \in \mathbb{R}^d$ and Jensen's inequality, (b) holds with probability at least $1-\delta$ by the sufficient coverage assumption (Assumption 2), (c) follows by denoting eigenvalues $\lambda^i_{d^{\pi^*}_{P^o}}$ of rank-1 matrices $\Sigma^i_{d^{\pi^*}_{P^o}}$. For (d), we first notice

$$\begin{aligned}
\Sigma_{d^{\pi^*}_{P^o}} = \mathbb{E}_{s,a\sim d^{\pi^*}_{P^o}}\phi(s,a)\phi(s,a)^\top &= \mathbb{E}_{s,a\sim d^{\pi^*}_{P^o}}[\sum_{i,j\in[d]}(\phi_i(s,a)\mathbb{1}_i)(\phi_j(s,a)\mathbb{1}_j)^\top] \\
&= \sum_{i\in[d]}\Sigma^i_{d^{\pi^*}_{P^o}} + \sum_{i,j\in[d]:i\neq j}\mathbb{E}_{s,a\sim d^{\pi^*}_{P^o}}[(\phi_i(s,a)\mathbb{1}_i)(\phi_j(s,a)\mathbb{1}_j)^\top].
\end{aligned}$$

For any $k \in [d]$, let λ_k denote k^{th} smallest eigenvalue of $\Sigma_{d^{\pi^*}_{P^o}}$. For any $k \in [d]$, we know from a fact of positive semidefinite matrices that λ_k is at least as any k^{th} smallest eigenvalue of any matrix summand. Moreover, since $\|\phi(s,a)\|_2 \le 1$, it follows by Jensen's inequality $\lambda_1 = \|\Sigma_{d^{\pi^*}_{P^o}}\|_{\mathrm{op}} \le \mathbb{E}_{s,a\sim d^{\pi^*}_{P^o}}\|\phi(s,a)\phi(s,a)^\top\|_{\mathrm{op}} \le 1$. Since $\Sigma_{d^{\pi^*}_{P^o}}$ is positive semidefinite, we have all $\lambda_k \in [0,1]$. Finally, step (d) is concluded by the fact that the number of non-zero eigenvalues is equal to the rank of a positive semidefinite matrix. This completes the proof. □

We are now ready to present our main result of this linear MDP problem setting. With the above result, we now provide the offline RL suboptimality guarantee below.

Theorem 7. *Let Assumption 1 hold. Let π_K be the LM-DRQI algorithm policy after K iterations. Then, under Assumption 2, the following holds with probability at least $1-\delta$*

$$\mathbb{E}_{s_0\sim d_0}[V^{\pi^*}(s_0) - \mathbb{E}_{\mathcal{D}}[V^{\pi_K}(s_0)]] \le \frac{2\gamma^{K+1}}{(1-\gamma)^2} + \frac{c_1\log(Nd/((1-\gamma)\delta))}{(1-\gamma)^2}\sqrt{\frac{d\cdot \mathrm{rank}(\Sigma_{d^{\pi^*}_{P^o}})}{C^\dagger_{\mathrm{sc}}N}}.$$

Proof. We first recall our analyses of Theorem 3. We denote the value function of policy π for the transition dynamics model P as V_P^π. We now denote the robust value function (Panaganti & Kalathil, 2022; Xu* et al., 2023; Panaganti et al., 2022) for uncertainty set $\widehat{\mathcal{P}}$ as $V_{\widehat{\mathcal{P}}}^\pi = \min_{P\in\widehat{\mathcal{P}}} V_P^\pi$ and its optimal robust policy as $\widehat{\pi}^* = \arg\max_\pi V_{\widehat{\mathcal{P}}}^\pi$. We let $Q_{\widehat{\mathcal{P}}}^\pi$ be its corresponding robust Q-function. From robust RL (Panaganti & Kalathil, 2022; Xu* et al., 2023; Panaganti et al., 2022) we can write the following robust Bellman equation: $Q_{\widehat{\mathcal{P}}}^\pi(s,a) = r(s,a) + \gamma \min_{P_{s,a}\in\widehat{\mathcal{P}}_{s,a}} \mathbb{E}_{s'\sim P_{s,a}}(V_{\widehat{\mathcal{P}}}^\pi(s'))$. To make it notationally easy, we write $V^{\pi^*}(d^\pi)$ as $V_{P^o}^{\pi^*}(d_{P^o}^\pi)$ making the dependence on the model P^o explicit.

We again recall Eq. (14) in tandem with Proposition 5:

$$\mathbb{E}_{s_0\sim d_0}[V_{P^o}^{\pi^*}(s_0) - V_{P^o}^{\pi_K}(s_0)] \le \mathbb{E}_{s_0\sim d_0}[V_{P^o}^{\pi^*}(s_0) - V_{\widehat{\mathcal{P}}}^{\widehat{\pi}^*}(s_0)] + \frac{2\gamma^{K+1}}{(1-\gamma)^2}. \tag{22}$$

Further recalling Eq. (15) we know,

$$\begin{aligned}\mathbb{E}_{s_0\sim d_0}[V_{P^o}^{\pi^*}(s_0) - V_{\widehat{\mathcal{P}}}^{\widehat{\pi}^*}(s_0)] &\le \mathbb{E}_{s_0\sim d_0}[\gamma\mathbb{E}_{s'\sim P^o_{s_0,\pi^*(s_0)}}(V_{P^o}^{\pi^*}(s') - V_{\widehat{\mathcal{P}}}^{\widehat{\pi}^*}(s'))] \\ &\quad + \underbrace{\mathbb{E}_{s_0\sim d_0}[\gamma\mathbb{E}_{s'\sim P^o_{s_0,\pi^*(s_0)}}(V_{\widehat{\mathcal{P}}}^{\widehat{\pi}^*}(s'))] - \gamma \min_{P_{s_0,\pi^*(s_0)}\in\widehat{\mathcal{P}}_{s_0,\pi^*(s_0)}} \mathbb{E}_{s'\sim P_{s_0,\pi^*(s_0)}}(V_{\widehat{\mathcal{P}}}^{\widehat{\pi}^*}(s'))]}_{(I)}.\end{aligned} \tag{23}$$

Analyzing (I) in Eq. (23) for any $P\in\widehat{\mathcal{P}}$:

$$\begin{aligned}(I) &= \mathbb{E}_{s_0\sim d_0}[\gamma\mathbb{E}_{s'\sim P^o_{s_0,\pi^*(s_0)}}(V_{\widehat{\mathcal{P}}}^{\widehat{\pi}^*}(s')) - \gamma\mathbb{E}_{s'\sim \widehat{P}^o_{s_0,\pi^*(s_0)}}(V_{\widehat{\mathcal{P}}}^{\widehat{\pi}^*}(s')) \\ &\qquad + \gamma\mathbb{E}_{s'\sim \widehat{P}^o_{s_0,\pi^*(s_0)}}(V_{\widehat{\mathcal{P}}}^{\widehat{\pi}^*}(s')) - \gamma\mathbb{E}_{s'\sim P_{s_0,\pi^*(s_0)}}(V_{\widehat{\mathcal{P}}}^{\widehat{\pi}^*}(s'))] \\ &\overset{(g)}{\le} \frac{c_1\log(Nd/((1-\gamma)\delta))}{1-\gamma}\sqrt{\frac{d}{N}}\sum_{i=1}^d \|\phi_i(s_0,\pi^*(s_0))\mathbb{1}_i\|_{\Lambda_N^{-1}} \\ &\qquad + \gamma\mathbb{E}_{s_0\sim d_0}[\mathbb{E}_{s'\sim \widehat{P}^o_{s_0,\pi^*(s_0)}}(\widehat{V}^{\widehat{\pi}^*}(s')) - \mathbb{E}_{s'\sim P_{s_0,\pi^*(s_0)}}(\widehat{V}^{\widehat{\pi}^*}(s'))] \\ &\overset{(h)}{\le} \frac{2c_1\log(Nd/((1-\gamma)\delta))}{1-\gamma}\sqrt{\frac{d}{N}}\sum_{i=1}^d \|\phi_i(s_0,\pi^*(s_0))\mathbb{1}_i\|_{\Lambda_N^{-1}},\end{aligned} \tag{24}$$

where (g) holds with probability at least $1-\delta$, which follows from Lemma 7, and (h) follows by the definition of set $\widehat{\mathcal{P}}$.

Substituting Eq. (24) back in Eq. (23) and via recursion we get,

$$\begin{aligned}&\mathbb{E}_{s_0\sim d_0}[V_{P^o}^{\pi^*}(s_0) - V_{\widehat{\mathcal{P}}}^{\widehat{\pi}^*}(s_0)] \\ &\le \sum_{t=0}^\infty \gamma^t \mathbb{E}_{s\sim d^{\pi^*}_{P^o,t}}[\frac{c_1\log(Nd/((1-\gamma)\delta))}{1-\gamma}\sqrt{\frac{d}{N}}\sum_{i=1}^d \|\phi_i(s_0,\pi^*(s_0))\mathbb{1}_i\|_{\Lambda_N^{-1}}] \\ &= \frac{c_1\log(Nd/((1-\gamma)\delta))}{(1-\gamma)^2}\sqrt{\frac{d}{N}}\sum_{i=1}^d \mathbb{E}_{s\sim d^{\pi^*}_{P^o}}[\|\phi_i(s_0,\pi^*(s_0))\mathbb{1}_i\|_{\Lambda_N^{-1}}],\end{aligned}$$

where last equality follows by the definition of state-distribution $d^{\pi^*}_{P^o} = (1-\gamma)\sum_{t=0}^\infty \gamma^t d^{\pi^*}_{P^o,t}$. Now, putting this back in Eq. (22), the offline RL guarantee becomes:

$$\begin{aligned}\mathbb{E}_{\mathcal{D}}[\mathbb{E}_{s_0\sim d_0}[V_{P^o}^{\pi^*}(s_0) - V_{P^o}^{\pi_K}(s_0)]] &\le \frac{2\gamma^{K+1}}{(1-\gamma)^2} \\ &\quad + \frac{c_1\log(Nd/((1-\gamma)\delta))}{(1-\gamma)^2}\sqrt{\frac{d}{N}}\sum_{i=1}^d \mathbb{E}_{\mathcal{D}}[\mathbb{E}_{s\sim d^{\pi^*}_{P^o}}[\|\phi_i(s_0,\pi^*(s_0))\mathbb{1}_i\|_{\Lambda_N^{-1}}]].\end{aligned} \tag{25}$$

We now assume we have sufficient coverage Assumption 2 of linear MDP P^o. Now under Lemma 9, with probability at least $1-\delta$, from Eq. (25) we have

$$\mathbb{E}_{\mathcal{D}}[\mathbb{E}_{s_0\sim d_0}[V_{P^o}^{\pi^*}(s_0) - V_{P^o}^{\pi_K}(s_0)]] \le \frac{2\gamma^{K+1}}{(1-\gamma)^2}$$

$$+ \frac{c_1 \log(Nd/((1-\gamma)\delta))}{(1-\gamma)^2} \sqrt{\frac{d}{N}} \sum_{i=1}^{d} \mathbb{E}_{\mathcal{D}}[\mathbb{E}_{s \sim d_{P^o}^{\pi^*}}[\|\phi_i(s_0, \pi^*(s_0))\mathbb{1}_i\|_{\Lambda_N^{-1}}]]$$

$$\leq \frac{2\gamma^{K+1}}{(1-\gamma)^2} + \frac{c_1 \log(Nd/((1-\gamma)\delta))}{(1-\gamma)^2} \sqrt{\frac{d \cdot \mathrm{rank}(\Sigma_{d_{P^o}^{\pi^*}})}{C_{\mathrm{sc}}^{\dagger} N}}.$$

This proves this result. □

Different from above, we now provide the offline RL suboptimality guarantee relying on the finite relative condition instead of the sufficient coverage assumption Assumption 2. Before presenting the result, here is another high probability result similar to Lemma 9 but now relies on the finite relative condition.

Lemma 10. *Let $\lambda = 1$. For any s, a, with probability at least $1-\delta$ we have $\sum_{i\in[d]} \mathbb{E}_{s,a\sim d_{P^o}^{\pi^*}}[\|\phi_i(s,a)\mathbb{1}_i\|_{\Lambda_N^{-1}} \leq c\sqrt{C_{\pi^*,\phi}^{\dagger}\mathrm{rank}(\Lambda)(\mathrm{rank}(\Lambda) + \log(c/\delta))}$ where $\Lambda_N = \frac{\lambda}{N} I + \frac{1}{N}\sum_{t=1}^{N} \phi(s_t,a_t)\phi(s_t,a_t)^\top$, $\Lambda = \mathbb{E}_{s,a\sim\mu}\phi(s,a)\phi(s,a)^\top$, $C_{\pi^*,\phi}^{\dagger} = \max_{x\in\mathbb{R}^d} \sum_{i\in[d]} d(x^\top \Sigma_{d^{\pi^*}}^{i} x)/(x^\top \Lambda x)$.*

Proof. We follow the proof in Lemma 9 but use the relative condition number to get the required bound. Firstly notice,

$$\begin{aligned}
\sum_{i\in[d]} \mathbb{E}_{s,a\sim d_{P^o}^{\pi^*}}[\|\phi_i(s,a)\mathbb{1}_i\|_{\Lambda_N^{-1}}] &= \sum_{i\in[d]} \mathbb{E}_{s,a\sim d_{P^o}^{\pi^*}}[\sqrt{(\phi_i(s,a)\mathbb{1}_i)^\top \Lambda_N^{-1}(\phi_i(s,a)\mathbb{1}_i)}] \\
&= \sum_{i\in[d]} \mathbb{E}_{s,a\sim d_{P^o}^{\pi^*}}[\sqrt{\mathrm{Tr}((\phi_i(s,a)\mathbb{1}_i)(\phi_i(s,a)\mathbb{1}_i)^\top \Lambda_N^{-1})}] \\
&\overset{(a)}{\leq} \sqrt{d}\sqrt{\sum_{i\in[d]} \mathrm{Tr}(\mathbb{E}_{s,a\sim d_{P^o}^{\pi^*}}[(\phi_i(s,a)\mathbb{1}_i)(\phi_i(s,a)\mathbb{1}_i)^\top]\Lambda_N^{-1})} \\
&= \sqrt{d}\sqrt{\sum_{i\in[d]} \mathrm{Tr}(\Sigma_{d_{P^o}^{\pi^*}}^{i}\Lambda_N^{-1})} \\
&\overset{(b)}{\leq} \sqrt{d}\sqrt{\frac{C_{\pi^*,\phi}^{\dagger}}{d}\mathrm{Tr}(\Lambda\Lambda_N^{-1})} \\
&= \sqrt{C_{\pi^*,\phi}^{\dagger}\mathbb{E}_{s,a\sim\mu}[\phi(s,a)^\top\Lambda_N^{-1}\phi(s,a)]} \\
&\overset{(c)}{\leq} c\sqrt{C_{\pi^*,\phi}^{\dagger}\mathrm{rank}(\Lambda)(\mathrm{rank}(\Lambda) + \log(c/\delta))},
\end{aligned}$$

where (a) follows by $\|x\|_1 \leq \sqrt{d}\|x\|_2$ for $x \in \mathbb{R}^d$ and Jensen's inequality, (b) follows by $C_{\pi^*,\phi}^{\dagger}$ definition, and (c) holds by Lemma 8 with probability at least $1-\delta$. □

Corollary 1. *Let Assumption 1 hold. Let π_K be the LM-DRQI algorithm policy after K iterations. Then, with $C_{\pi^*,\phi}^{\dagger} < \infty$, the following holds with probability at least $1-\delta$*

$$\mathbb{E}_{s_0\sim d_0}[V^{\pi^*}(s_0) - \mathbb{E}_{\mathcal{D}}[V^{\pi_K}(s_0)]] \leq \frac{2\gamma^{K+1}}{(1-\gamma)^2} + \frac{c_1 \log(Nd/((1-\gamma)\delta))}{(1-\gamma)^2}\sqrt{\frac{dC_{\pi^*,\phi}^{\dagger}\mathrm{rank}(\Lambda)^2\log(c/\delta)}{N}}.$$

Proof. The proof follows from Theorem 7. In this corollary, we assume finite relative condition number $C_{\pi^*,\phi}^{\dagger} < \infty$ for linear MDP P^o instead of assuming Assumption 2. We also emphasize that in this result we only need to assume $\Sigma_{d^{\pi^*}}^{i}$ for all $i \in [d]$, due to Lemma 10, instead for all $\Sigma_{d^{\pi^*}}^{(i,j)}$, $i, j \in [d]$ in Assumption 1. Thus this result is more general than Theorem 7. Now under Lemma 10, with probability at least $1-\delta$, from Eq. (25) we have

$$\mathbb{E}_{\mathcal{D}}[\mathbb{E}_{s_0\sim d_0}[V_{P^o}^{\pi^*}(s_0) - V_{P^o}^{\pi_K}(s_0)]] \leq \frac{2\gamma^{K+1}}{(1-\gamma)^2}$$

$$+ \frac{c_1 \log(Nd/((1-\gamma)\delta))}{(1-\gamma)^2} \sqrt{\frac{d}{N}} \sum_{i=1}^{d} \mathbb{E}_{\mathcal{D}}[\mathbb{E}_{s\sim d_{Po}^{\pi^*}}[\|\phi_i(s_0, \pi^*(s_0))\mathbb{1}_i\|_{\Lambda_N^{-1}}]]$$

$$\leq \frac{2\gamma^{K+1}}{(1-\gamma)^2} + \frac{c_2 \log(Nd/((1-\gamma)\delta))}{(1-\gamma)^2} \sqrt{\frac{dC^{\dagger}_{\pi^*,\phi} \text{rank}(\Lambda)^2 \log(c/\delta)}{N}},$$

where c_2 is a universal constant that only depends on c_1 and c (c is from Lemma 10). This completes the proof. □

In the following, we show that for a class of linear MDPs, the sufficient coverage assumption in Jin et al. (2021) implies our sufficient coverage assumption (Assumption 2) adapted from Ma et al. (2022).

Lemma 11. *Consider a class of linear MDPs where $\Sigma^i_{d^{\pi^*}} = \Sigma^j_{d^{\pi^*}}$ for all $i, j \in [d]$. Define the random events $\mathcal{E}_1 = \{\omega : \Lambda_N(\omega) \geq I/N + C_{\text{sc}} \cdot \Sigma_{d^{\pi^*}}\}$ and $\mathcal{E}_2 = \{\omega : \Lambda_N(\omega) \geq I/N + C_{\text{sc}} \cdot d\Sigma^i_{d^{\pi^*}}\}$. Then we have $\mathcal{E}_1 \subseteq \mathcal{E}_2$.*

Proof. We know

$$\Sigma_{d^{\pi^*}} = \mathbb{E}_{s,a\sim d^{\pi^*}} \phi(s,a)\phi(s,a)^\top = \mathbb{E}_{s,a\sim d^{\pi^*}}[\sum_{i,j\in[d]} (\phi_i(s,a)\mathbb{1}_i)(\phi_j(s,a)\mathbb{1}_j)^\top] = \sum_{i\in[d]} \Sigma^i_{d^{\pi^*}} + \sum_{i,j\in[d]:i\neq j} \Sigma^{(i,j)}_{d^{\pi^*}}.$$

Consider some non-zero $x \in \mathbb{R}^{d\times 1}$. Since $\Sigma^{(i,j)}_{d^{\pi^*}}$ are all positive semidefinite, we have

$$x^\top \Sigma_{d^{\pi^*}} x \geq \sum_{i\in[d]} x^\top \Sigma^i_{d^{\pi^*}} x = d(x^\top \Sigma^i_{d^{\pi^*}} x).$$

Noting that $\mathcal{E}_1 = \{\omega : x^\top \Lambda_N(\omega) x \geq \|x\|_2^2/N + C_{\text{sc}} \cdot x^\top \Sigma_{d^{\pi^*}} x\}$ and $\mathcal{E}_2 = \{\omega : x^\top \Lambda_N(\omega) x \geq \|x\|_2^2/N + C_{\text{sc}} \cdot dx^\top \Sigma^i_{d^{\pi^*}} x\}$ finishes the proof. □

For a different class of linear MDPs we have the following.

Lemma 12. *Consider a class of linear MDPs where $\Sigma^i_{d^{\pi^*}} = \Sigma^{(i,j)}_{d^{\pi^*}}$ for all $i, j \in [d]$. Let $C_{\pi^*,\phi} = \max_{x\in\mathbb{R}^d} (x^\top \Sigma_{d^{\pi^*}} x)/(x^\top \Lambda x)$ and $C^{\dagger}_{\pi^*,\phi} = \max_{x\in\mathbb{R}^d} \sum_{i\in[d]} d(x^\top \Sigma^i_{d^{\pi^*}} x)/(x^\top \Lambda x)$. Then we have $C^{\dagger}_{\pi^*,\phi} = C_{\pi^*,\phi}$.*

Proof. From Lemma 11, we already know $\Sigma_{d^{\pi^*}} = \sum_{i,j\in[d]} \Sigma^{(i,j)}_{d^{\pi^*}}$. Consider any non-zero $x \in \mathbb{R}^{d\times 1}$. From the class of linear MDPs, we further have

$$x^\top \Sigma_{d^{\pi^*}} x = \sum_{i\in[d]} d(x^\top \Sigma^i_{d^{\pi^*}} x).$$

Now the statement directly follows. □

From Corollary 1, we get the offline suboptimality guarantee of the order $\frac{\sqrt{dC^{\dagger}_{\pi^*,\phi}\text{rank}(\Lambda)^2}}{\sqrt{(1-\gamma)^4 N}}$ for LM-DRQI algorithm. Furthermore, under Lemma 12, it is comparable with Uehara & Sun (2021) in Table 2.