ROBUSTIFY THE LATENT SPACE: OFFLINE DISTRIBUTIONALLY ROBUST REINFORCEMENT LEARNING WITH LINEAR FUNCTION APPROXIMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Among the reasons hindering the applications of reinforcement learning (RL) to real-world problems, two factors are critical: limited data and the mismatch between the test environment (real environment in which the policy is deployed) and the training environment (e.g., a simulator). This paper simultaneously addresses these issues with *offline distributionally robust RL*, where a distributionally robust policy is learned using historical data from the source environment by optimizing against a worst-case perturbation thereof. In particular, we move beyond tabular settings and design a novel linear function approximation framework that robustifies the latent space. Our framework is instantiated into two settings, one where the dataset is well-explored and the other where the dataset has weaker data coverage. In addition, we introduce a value shift algorithmic technique specifically designed to suit the distributionally robust nature, which contributes to our improved theoretical results and empirical performance. Sample complexities $\tilde{O}(d^{1/2}/N^{1/2})$ and $\tilde{O}(d^{3/2}/N^{1/2})$ are established respectively as the first non-asymptotic results in these settings, where d denotes the dimension in the linear function space and N represents the number of trajectories in the dataset. Diverse experiments are conducted to demonstrate our theoretical findings, showing the superiority of our algorithms against the non-robust one.

1 INTRODUCTION

Unlike data-driven methods in supervised learning, reinforcement learning (RL) algorithms require active interaction with the environment to learn a near-optimal policy, often involving online trial-and-error. However, this approach can be impractical in real-world scenarios with limited or prohibited data collection. To address the limitation of online RL, offline reinforcement learning (offline RL or batch RL) (Lange et al., 2012; Levine et al., 2020), focuses on policy learning with only access to some logged datasets and expert demonstrations. Due to its non-dependence on further interaction with the environment, offline RL is increasingly appealing for various applications, including autonomous driving (Yu et al., 2018; Yurtsever et al., 2020; Shi et al., 2021), healthcare (Gottesman et al., 2019; Yu et al., 2021; Tang & Wiens, 2021) and robotics (Siegel et al., 2020; Zhou et al., 2021a; Rafailov et al., 2021).

Despite the developments in the rich literature (Yu et al., 2020; Kumar et al., 2020; Yang et al., 2021b; An et al., 2021; Cheng et al., 2022), offline RL has an implicit but questionable assumption: the test environment is the same as the training one. This assumption can result in inadequate performance of offline RL in uncertain environments since the optimal policy of a Markov decision process (MDP) may be sensitive to the transition probabilities (Mannor et al., 2004; El Ghaoui & Nilim, 2005; Simester et al., 2006). Including financial trading and robotics, many domains may prefer a robust policy that remains effective in shifting distributions from the one in the training environment. Thus, robust MDPs have been proposed to address this issue (Satia & Lave Jr, 1973; Nilim & El Ghaoui, 2005; Iyengar, 2005; Wiesemann et al., 2013; Lim et al., 2013; Ho et al., 2021; Goyal & Grand-Clement, 2022). Recent studies (Zhou et al., 2021b; Yang et al., 2021a; Shi & Chi, 2022; Panaganti et al., 2022a) demonstrate the potential of robust RL in the offline setting.

In this paper, we aim to theoretically understand linear function approximation as an important component in offline distributionally (DR) robust RL. Linear function approximation (Bertsekas & Tsitsiklis, 1995; Schweitzer & Seidmann, 1985), which uses a linear combination of features to approximate the value function, is one of the most widely used and studied solutions in high-dimensional problems and serves as a cornerstone in the path toward large-scale real-world problems.

Developing a DRRL algorithm with linear function approximation is challenging. In contrast to non-robust RL where transition kernel is fixed, DRRL assumes that the transition kernel of the MDP belongs to an ambiguity set, which significantly impacts the computational feasibility of the robust value function and policy performance. A common approach in this area is to construct the ambiguity set for each state-action pair and then project the robust value function onto a lower-dimensional subspace using linear function approximation, called Robustify-then-Approximate (RTA) approach in this paper. Although RTA style algorithms (Wiesemann et al., 2013; Goyal & Grand-Clement, 2022; Panaganti et al., 2022a) have proven that their robust projected value iteration can converge to a fixed point, as pointed out by our motivating example in Section 3, linear projection may conflict with the non-linearity of the robust Bellman operator, which may consequently lead to suboptimal decisions. Furthermore, none of these algorithms have been shown to be sample efficient under weak data conditions, which is essential in real-world applications. Recently, a study by Goyal & Grand-Clement (2022) proposes constructing the ambiguity set in the latent space (called Robustify-the-Latent Space (RLS)), which is more compatible with the linear approximator. However, they assume access to the true transition kernel, whereas in practice, we can only access data sampled from some training environment. Therefore, developing a data-driven DRRL algorithm that directly robustifies the latent space is yet to be explored.

In this paper, we mainly address the question below building on insights from high-dimensional statistics (Wainwright, 2019) to provide informative insights into the impacts of salient problem parameters on the sample complexity, especially for applications with large state-action spaces:

Is it possible to design a sample-efficient algorithm using linear function approximation for offline DRRL by robustifying the latent space, even with weaker data coverage conditions?

We give a positive response to this question. Specifically, our contributions are fourfold:

1. We point out potential conflicts between linear function approximation and robustness gain in the Robustify-then-approximate (RTA) approach by constructing a motivating example. Then we instantize the idea of robustifying the latent space (RLS) into a sample-efficient **D**istributionally **R**obust **V**alue **I**teration with **L**inear function approximation (DRVI-L) algorithm for well-explored datasets.
2. We prove a state-action space independent sample complexity guarantee for our DRVI-L algorithm with a novel value shift technique to alleviate the magnification of the estimation error in the DR optimization nature. This result can almost recover to the same dependence on d and N of that from the non-robust counterpart (Yin et al., 2022), which has never been achieved by previous literature.
3. We extend our algorithm by designing the **P**essimistic **D**istributionally **R**obust **V**alue **I**teration with **L**inear function approximation (PDRVI-L) algorithm, a pessimistic variant with our DRVI-L, and prove the first sample-efficient bound beyond the well-explored condition. Such an extension is the fruit of our RLS idea with delicate non-asymptotic analysis.
4. We establish theoretical guarantees for our two algorithms even when the MDP transition is not perfectly linear, and conduct experiments to demonstrate the balance achieved by our linear function approximation algorithm between optimality and computational efficiency.

2 PRELIMINARY

2.1 MDP STRUCTURE AND NOTATIONS

Consider an episode MDP $(\mathcal{S}, \mathcal{A}, H, \mu, P, r)$ where $\mathcal{S}$ and $\mathcal{A}$ are finite state and action spaces with cardinalities S and A. $P = \{P_h\}_{h=1}^H$ are state transition probability measures and $r = \{r_h\}_{h=1}^H$ are the reward functions, respectively. We assume that r is deterministic and bounded in $[0, 1]$. A (Markovian) policy $\pi = \{\pi_h\}_{h=1}^H$ maps, for each period state-action pair (s, a) to a probability

distribution over the set of actions $\mathcal{A}$ and induce a random trajectory $s_1, a_1, r_1, \cdots, s_H, a_H, r_H, s_{H+1}$ with $s_1 \sim \mu$, $a_h \sim \pi(\cdot|s_h)$and $s_{h+1} \sim P_h(\cdot|s_h, a_h)$ for $h \in [H]$ for some initial state distribution μ. For any policy π and any stage $h \in [H]$, the value function $V_h^\pi : \mathcal{S} \to \mathbb{R}$, the action-value function $Q : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, the expected return $R(\pi, P)$ are defined as $V_h^\pi(s) \coloneqq \mathbb{E}_P^\pi[\sum_{h=1}^H r_h(s_h, a_h)|s_h = s]$, $Q_h^\pi(s, a) \coloneqq \mathbb{E}_P^\pi[\sum_{h=1}^H r_h(s_h, a_h)|s_h = s, a_h = a]$, and $R(\pi, P) \coloneqq \mathbb{E}_{s\sim\mu}[V_1^\pi(s)]$. For any function Q and any policy π, we denote $\langle Q(s, \cdot), \pi(\cdot|s)\rangle_\mathcal{A} = \sum_{a\in\mathcal{A}} Q(s, a)\pi(a|s)$. For two non-negative sequences $\{a_n\}$ and $\{b_n\}$, we denote $\{a_n\} = O(b_n)$ if $\limsup_{n\to\infty} a_n/b_n < \infty$. We also use $\tilde{O}(\cdot)$ to denote the respective meaning within multiplicative logarithmic factors in N, d, H and δ. We denote the Kullback-Leibler (KL) divergence between two discrete probability distributions P and Q over state space as $D_{\mathrm{KL}}(P\|Q) = \sum_{s\in\mathcal{S}} P(s) \log(\frac{P(s)}{Q(s)})$.

2.2 Distributionally Robust MDPs

Before we present the DRRL setting, we first introduce the Distributionally Robust MDPs (DRMDPs). DRMDPs assume that the probability P is not exactly known but lies within a so-called ambiguity set $\mathcal{P}$ induced by a distribution distance measure, such as KL divergence. The return of any given policy is the worst-case return induced by the transition model over the ambiguity set. We define the DR value function, action-value function and expected return as $V_h^{\pi,\mathrm{rob}}(s) = \inf_{P\in\mathcal{P}} V_h^\pi(s)$, $Q_h^{\pi,\mathrm{rob}}(s, a) = \inf_{P\in\mathcal{P}} Q_h^\pi(s, a)$ and $R^{\mathrm{rob}}(\pi, \mathcal{P}) = \inf_{P\in\mathcal{P}} R(\pi, P)$, respectively. The optimal DR expected return is defined as $R^{\mathrm{rob}}(\pi^*, \mathcal{P}) \coloneqq \sup_{\pi\in\Pi} R^{\mathrm{rob}}(\pi, \mathcal{P})$ over all Markovian policies. In the sequel, we omit the superscript "rob". In fact, by the work of Goyal & Grand-Clement (2022), we can restrict to the deterministic policy class to achieve the optimal DR expected return. The performance metric for any given policy π is the so-called suboptimality, which is defined as

$$\mathrm{SubOpt}(\pi; \mathcal{P}) = R(\pi^*, \mathcal{P}) - R(\pi, \mathcal{P}).$$

As the transition models and the policies are sequences corresponding to all horizons in the episode MDP, following Iyengar (2005), we assume that $\mathcal{P}$ can be decomposed as the product of the ambiguity sets in each horizon, i.e., $\mathcal{P} = \prod_{h=1}^H \mathcal{P}_h$. For stage h, each transition model P_h, lies within the ambiguity set $\mathcal{P}_h$.

2.3 Linear MDP

Our main task in this paper is to compute the optimal policy using linear function approximation with the possible lowest suboptimality. We parameterize the Q-function, value function, and optimal policy for each horizon $h \in [H]$ using $\nu_h \in \mathbb{R}^d$, given the feature map $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$, as follows:

$$Q_{\nu_h}(s, a) \coloneqq \phi(s, a)^\top \nu_h, \quad V_{\nu_h}(s) \coloneqq \max_{a\in\mathcal{A}} Q_{\nu_h}(s, a), \quad \pi_{\nu_h}(s) \coloneqq \arg\max_{a\in\mathcal{A}} Q_{\nu_h}(s, a). \tag{1}$$

To study the linear function approximation, various assumptions on the MDP have been proposed in the literature (Jiang et al., 2017; Yang & Wang, 2019; Jin et al., 2020; Modi et al., 2020; Zanette et al., 2020; Wang et al., 2021). In particular, we consider the soft state aggregation of d factors, i.e., the transition model in each stage h can be represented using a known feature map $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ over the d latent factor spaces defined by $\psi_h : \mathcal{S} \to \mathbb{R}^d$. Such a assumption has been widely adopted in the literature (Singh et al., 1994; Duan et al., 2019; Zhang & Wang, 2019; Zanette et al., 2021). We also assume the reward functions admit linear structure w.r.t. ϕ, following the linear MDP protocol from Jin et al. (2021; 2020). Formally, we have the following definition for the soft state aggregation.

Definition 2.1 (Soft State Aggregation MDP)**.** *Consider an episode MDP instance $M = (\mathcal{S}, \mathcal{A}, H, P, r)$ and a feature map $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$. We say the transition model P admit a soft state aggregation w.r.t. ϕ (denoted as $P \in \mathrm{Span}(\phi)$) if for every $s \in \mathcal{S}, a \in \mathcal{A}, s' \in \mathcal{S}$ and every $h \in [H]$, we have*

$$P_h(s'|s, a) = \phi(s, a)^\top \psi_h(s'),$$

for some factors $\psi_h : \mathcal{S} \to \mathbb{R}^d$. Moreover, ψ satisfies,

$$\int_s \psi_{h,i}(s)ds = 1, \forall i \in [d], h \in [H].$$

We say the reward functions r admit a linear representation w.r.t. ϕ (denoted as $r \in \mathrm{Span}(\phi)$) if for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$ and $h \in [H]$, there exists $\theta_h \in \mathbb{R}^d$ satisfying $\|\theta_h\| \le \sqrt{d}$ and $r_h(s, a) = \phi(s, a)^\top \theta_h$.

3 MOTIVATING EXAMPLE

Computing an optimal robust policy for a general ambiguity set is proven to be strongly NP-Hard, as demonstrated by Wiesemann et al. (2013). To ensure computational tractability, Nilim & El Ghaoui (2005) and Iyengar (2005) introduce the (s, a)-rectangular ambiguity set. This set assumes that the perturbation of the transition probability for each (s, a) pair is independent of others, denoted as $\mathcal{P}_h = \prod_{(s,a)\in\mathcal{S}\times\mathcal{A}} \mathcal{P}_h(s, a)$. However, this assumption can be computationally expensive when dealing with large state or action spaces since it requires solving a robust optimization problem for each (s, a) pair. Moreover, it may exhibit over-conservatism, particularly when the transition probabilities possess inherent structure (Wiesemann et al., 2013; Goyal & Grand-Clement, 2022).

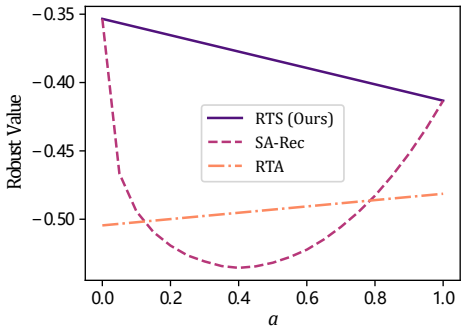


Figure 1: Motivating example. See Appendix B for the detailed experiment setup.

Before this work, the only attempts in linear function approximation for the robust RL are Tamar et al. (2014); Badrinath & Kalathil (2021). Both approaches follow the same idea: first obtaining robust values for each (s, a)-pair and then approximating robust values for the entire state-action space using a linear function. We refer to this as robustify-then-approximate (RTA). In contrast, our proposed algorithm directly robustifies the latent space (RTS). We illustrate the limitations of RTA using a continuous bandit case, which corresponds to the offline DRRL scenario with $H = 1$ and $S = 1$, where the action set is $[0, 1]$. When selecting action $a = 0$, the reward r_0 is drawn from a normal distribution with mean 1 and variance 1. If the action $a = 1$ is chosen, the reward r_1 follows a normal distribution with mean 0 and variance 0.5. When $a \in (0, 1)$, the reward distribution r_a is a mixture of r_0 and r_1, with a probability of $(1 - a)$ and a, respectively.

Given the linear structure of the problem, it is desirable to use linear function approximation to maintain a low-dimensional representation. However, as shown in Figure 1, the projected robust values using Tamar et al. (2014); Badrinath & Kalathil (2021)'s methods are irrational due to the nonlinear nature of the (s, a)-uncertainty set. Specifically, the projected algorithm behaves more pessimistically than the (s, a)-rectangular method for actions close to 0 or 1. It even **fails to preserve the order relationship between the robust values of actions** 0 **and** 1, as the robust value of action 0 is higher than 1 in the (s, a)-rectangular but becomes lower after projection. This loss of order may lead to suboptimal decisions, while our proposed algorithm, based on the d-rectangular ambiguity set (defined in Section 4), recovers the robust values for action 0 and 1 while avoiding the over-conservatism for action in between.

4 LINEAR FUNCTION APPROXIMATION: ROBUSTIFY THE LATENT SPACE

In this section, we assume the MDP enjoys the soft state-aggregation structure and introduce the concrete approach to robustify the latent space, i.e., the so-called d-rectangular ambiguity set. We then propose our first algorithm, Distributional Robust Value Iteration with Linear function approximation (DRVI-L) with corresponding sample complexity results.

Assumption 4.1 (State-Aggregation MDP)**.** *The true transition models are soft state-aggregation w.r.t. ϕ and the reward functions admit linear representation w.r.t. ϕ (Definition 2.1).*

4.1 AMBIGUITY SET STRUCTURE: ROBUSTIFY THE LATENT SPACE

To robustify the latent factor space, we assume each factor lies in an ambiguity set, which is formally stated in the Assumption 4.2.

Assumption 4.2 (d-rectangular). *For each $h \in [H]$, we assume the ambiguity set $\mathcal{P}_h$ with radius ρ admits the following structure for some probability distance $D : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_{\geq 0}$,*

$$\mathcal{P}_h(\rho) = \{(\sum_{i\in[d]} \phi_i(s,a)\psi'_{h,i}(s'))_{sas'} : \forall D(\psi'_{h,i}, \psi_{h,i}) \leq \rho\}.$$

Under Assumption 4.2, each factor $\psi_{h,i}$ is assumed to be independent, and thus can be chosen arbitrarily within the set $\mathcal{P}(\psi_h; \rho) \coloneqq \{\psi'_{h,i} : D(\psi'_{h,i}, \psi_{h,i}) \leq \rho\}$ without affecting other factors. We formulate our offline DRRL problem with d-rectangular as:

$$R(\pi; \mathcal{P}) = \inf_{P \in \mathcal{P}(\rho) = \prod_{h=1}^{H} \mathcal{P}_h(\rho)} R(\pi, P). \tag{2}$$

To facilitate algorithm design and suboptimality analysis, we choose the KL divergence as the probability distance function D. The corresponding KL ambiguity set for a horizon h and the i-th factor is denoted as $\mathcal{P}^{\mathrm{KL}}(\psi_{h,i}; \rho)$, while the ambiguity set for the horizon h is denoted as $\mathcal{P}_h^{\mathrm{KL}}(\rho) \coloneqq \prod_{i\in[d]} \mathcal{P}^{\mathrm{KL}}(\psi_{h,i}; \rho)$. To ensure computational tractability, we use Lemma 4.1, a strong duality result proven by Hu & Hong (2013), which allows us to solve the primal problem over the KL ambiguity set by solving the one-dimensional dual problem over the dual function $\sigma(Z, \beta)$.

Lemma 4.1 (Hu & Hong (2013)). *Suppose $X \sim P$ has finite moment generating function in the neighborhood of zero. We denote $Z \coloneqq \mathbb{E}_P[e^{-X/\beta}]$ and the dual function $\sigma(Z, \beta) \coloneqq -\beta \log(Z) - \beta \cdot \rho$, then,*

$$\inf_{P' : D_{\mathrm{KL}}(P' \| P) \leq \rho} \mathbb{E}_{P'}[X] = \sup_{\beta \geq 0} \sigma(Z, \beta). \tag{3}$$

As $\rho \to 0$, the LHS of the dual equation degrades to the non-robust view, i.e., $\mathbb{E}_P[X]$, and the optimum $\beta^* = \arg\sup_{\beta\geq 0} \sigma(Z, \beta)$ approaches infinity.

Based on Lemma 4.1 and Assumption 4.1, we can derive the following DR Bellman operator:

$$(\mathbb{B}_h V)(s,a) = r_h(s,a) + \inf_{P_h \in \mathcal{P}_h^{\mathrm{KL}}(\rho)} \mathbb{E}_{s' \sim P_h(\cdot|s,a)}[V(s')] = \phi(s,a)^\top(\theta_h + w_h), \tag{4}$$

where $w_{h,i} = \sup_{\beta\geq 0} \sigma(\mu_{h,i}, \beta)$ and $\mu_{h,i} \coloneqq \mathbb{E}_{\psi_{h,i}}[e^{-V(s')/\beta}] = \int_{s'} \psi_{h,i}(s') e^{-V(s')/\beta} ds'$.

The preceding result indicates that the DR Bellman operator using the d-rectangular ambiguity set can maintain ϕ representation. We formally state it in Lemma 4.2.

Lemma 4.2. *For any policy π and any epoch $h \in [H]$, the DR Q-function is linear w.r.t. ϕ. Moreover, $d(\mathbb{B}_h\mathcal{F}, \mathcal{F}) = 0$, where $d(\mathbb{B}_h\mathcal{F}, \mathcal{F}) = \sup_{g\in\mathcal{F}} \inf_{f\in\mathcal{F}} \|f - \mathbb{B}_h g\|$ is the Bellman error (Munos & Szepesvári, 2008) and $\mathcal{F} = \left\{\phi(\cdot,\cdot)^\top w : \forall w \in \mathbb{R}^d\right\}$ is a set containing all the possible function using ϕ as the feature map.*

Lemma 4.2 forms the basis for our function approximation algorithmic design. Unlike previous literature, such as Panaganti et al. (2022a), which assumes the completeness of the function class with respect to the ϕ representation without verification, our approach addresses the incompleteness issue observed in Section 3 by robustifying the feature space.

4.2 Distributionallly Robust Value Iteration with Value Shift

The key challenge in offline (DR)RL problem is that the computation of the DRRL policy is restricted to only utilize a logged dataset, rather than having access to the exact transition probability or interaction with the environment. As a result of the lack of ongoing interaction with the environment, the performance of the offline RL algorithm is adversely affected by the insufficient coverage of the offline dataset. As a starting point, we consider a robust-variant of the uniform "well-exploration" condition, which is widely adopted in many offline RL works (Jin et al., 2021; Duan et al., 2019; Xie et al., 2021).

Assumption 4.3 (Uniformly Well-explored Dataset). *Suppose $\mathcal{D}$ consists of N trajectories $\{(s_h^\tau, a_h^\tau, r_h^\tau)\}_{\tau,h=1}^{N,H}$ independently and identically induced by a fixed behavior policy $\overline{\pi}$ in the linear MDP. Meanwhile, suppose there exists an absolute constant $\underline{c} > 0$ such that at each step $h \in [H]$ and any $P \in \mathcal{P}(\rho)$*

$$\lambda_{\min}(\Sigma_h^P) \geq \underline{c}, \quad \textit{where } \Sigma_h^P = \mathbb{E}_{P,\overline{\pi}}[\phi(s_h,a_h)\phi(s_h,a_h)^\top].$$

Such an assumption requires the behavior policy to explore each feature dimension well, even in the worst-case transition model, which might need to explore some state-action pairs that the optimal policy has seldom visited. Similar assumption has appeared in (Shi & Chi, 2022) in the tabular setting.

To approximate the true Bellman operator, we construct the empirical version of Equation 4, particularly, to approximate $\mu_{h,i}$. Notably,

$$\mathbb{E}_{P_{s,a}}[e^{-V(s')/\beta}] = \int_{s'} e^{-V(s')/\beta} P(s'|s,a)ds' = \phi(s,a)^\top \mu_h,$$

where samples from $P_{s,a}$ can be obtained, motivating us to approximate μ_h by linear regression to obtain the estimator $\widehat{\mu}_h$. Note that $\mu_{h,i}$ and $\widehat{\mu}_{h,i}$ could be very close to zero, and further cause $\sigma(\widehat{\mu}_{h,i}, \beta)$ to approach infinity and damage the estimation. To address this issue, we propose a novel value shift technique by defining a new dual function $\tilde{\sigma}(Z, \beta)$ by changing Z to $Z + 1$,

$$\tilde{\sigma}(Z, \beta) = -\beta \log(Z + 1) - \beta \cdot \rho. \tag{5}$$

This ensures that $\log(Z + 1)$ remains valid even Z approaches zero. Accordingly, we adopt the shifted variant of the regression objective by subtracting 1 from the regression target, defined as

$$\begin{aligned}
&\tilde{\mathcal{E}}_h(\mu) = \sum_{\tau=1}^{N}((e^{-V(s^\tau_{h+1})/\beta} - 1) - \phi(s^\tau_h, a^\tau_h)^\top \mu)^2, \\
&\widehat{\mu}_h = \arg\min_{\mu \in \mathbb{R}^d} \tilde{\mathcal{E}}_h(\mu) + \lambda \cdot \|\mu\|^2, \quad \widehat{w}_{h,i} = \sup_{\beta \geq 0} \tilde{\sigma}(\min\{(\widehat{\mu}_{h,i})_+, 1\}, \beta).
\end{aligned} \tag{6}$$

We define $\Lambda_h = \sum_{\tau=1}^{N} \phi(s^\tau_h, a^\tau_h) + \lambda \cdot I$. $\widehat{\theta}_h$ and $\widehat{w}_h$ have the closed form as

$$\widehat{\theta}_h = \Lambda_h^{-1}(\sum_{\tau=1}^{N} \phi(s^\tau_h, a^\tau_h) r^\tau_h), \quad \widehat{\mu}_h = \Lambda_h^{-1}(\sum_{\tau=1}^{N} \phi(s^\tau_h, a^\tau_h)(e^{-V(s^\tau_{h+1}/\beta)} - 1)).$$

Our value shifting technique ensures that $\widehat{w}_h$ maintains a valid value regardless of the estimator's quality, which is essential for achieving the desired suboptimality that can nearly recover to that of the non-robust setting. We summarize our algorithm as **D**istributional **R**obust **V**alue **I**teration with **L**inear function approximation (DRVI-L) in Algorithm 1.

Prior to presenting the suboptimality analysis for our Algorithm 1, we introduce Assumption 4.4, which assumes a known, common lower bound for the optimum of the KL optimization problem in Lemma 4.1. This assumption is also necessary in the tabular case (Zhou et al., 2021b; Panaganti & Kalathil, 2022).

Assumption 4.4. *For each $h \in [H]$ and each $i \in [d]$, we denote $\beta^*_{h,i} = \arg\sup_{\beta_{h,i} \geq 0} \sigma(\mu_{h,i}, \beta_{h,i})$. We assume there exists a known $\underline{\beta}$ s.t. $0 < \underline{\beta} \leq \min_{h \in [H], i \in [d]} \beta^*_{h,i}$.*

By Proposition 2 in Hu & Hong (2013), $\beta^*_{h,i} = 0$ when the worst case happens with sufficient large probability w.r.t. ρ. In practice, it is typical to employ a small value of ρ to adapt to the problem without incurring over-conservatism (Ben-Tal & Nemirovski, 1998; 2000; Duchi & Namkoong, 2021). Thus $\beta^*_{h,i}$ would rarely be zero and enjoy a common non-zero lower bound.

Theorem 4.1. *We set $\lambda = 1$ in Algorithm 1. Under the Assumption 4.1, Assumption 4.3 and Assumption 4.4, when $N \geq 40/\underline{c} \cdot \log(4dH/\delta)$, we have the following holds with probability at least $1 - \delta$,*

$$\mathrm{SubOpt}(\widehat{\pi}; \mathcal{P}) \leq c_1 \underline{\beta}(e^{H/\underline{\beta}} - 1) d^{1/2} \zeta_1^{1/2} H / N^{1/2} + c_2 \underline{\beta}^{1/2}(e^{H/\underline{\beta}} - 1) \zeta_2^{1/2} H^{3/2} / N^{1/2}.$$

Here $\zeta_1 = \log(2N + 16Nd^{3/2}H^2 e^{H/\underline{\beta}})$, $\zeta_2 = \log(\frac{2dNH^3}{\delta\rho})$ and c_1 and c_2 are some absolute constants that only depend on $\underline{c}$.

It is worth noting that the suboptimality of Algorithm 1 primarily depends on the dimension d rather than the size of the state-action space. In contrast to tabular cases, such as Zhou et al. (2021b); Yang et al. (2021a), which focus on bounding the finite sample error for individual (s, a) pairs, Theorem 4.1 is derived by exploiting the linear structure shared by various (s, a) pairs, creating a novel ϵ-net to

Algorithm 1 DRVI-L

Input: $\underline{\beta}$, $\mathcal{D} = \{(s_h^\tau, a_h^\tau, r_h^\tau)\}_{\tau,h=1}^{N,H}$.
Init: $\widehat{V}_H = 0$, $\widehat{w}_H = 0$.
for step $h = H$ **to** 1 **do**
  $\Lambda_h = \sum_{\tau=1}^N \phi(s_h^\tau, a_h^\tau)\phi(s_h^\tau, a_h^\tau)^\top + \lambda I, \quad \widehat{\theta}_h = \Lambda_h^{-1}\left[\sum_{\tau=1}^N \phi(s_h^\tau, a_h^\tau) r_h^\tau\right]$
  if $h < H$ **then**
    Update $\widehat{w}_{h,i}$ with Equation 6.
  end if
  $\widehat{\nu}_h = \min(\widehat{\theta}_h + \widehat{w}_h, H - h + 1)_+, \quad \widehat{Q}_h(\cdot,\cdot) = \phi(\cdot,\cdot)^\top \widehat{\nu}_h$
  $\widehat{\pi}_h(\cdot \mid \cdot) = \arg\max_{\pi_h} \langle \widehat{Q}_h(\cdot,\cdot), \pi_h(\cdot \mid \cdot)\rangle_{\mathcal{A}}, \quad \widehat{V}_h(\cdot) = \langle \widehat{Q}_h(\cdot,\cdot), \widehat{\pi}_h(\cdot \mid \cdot)\rangle_{\mathcal{A}}$
end for

control the finite sample error for the entire linear function space, and utilizing the power of our value shift algorithmic ingredient. These techniques are novel compared to the non-robust counterpart.

Specifically, when $\underline{\beta}$ is relatively small, i.e., when the algorithm tends to learn a pessimistic view, we have $\mathrm{SubOpt} = \tilde{O}(d^{1/2}H^{3/2}\underline{\beta}^{1/2}e^{H/\underline{\beta}}/N^{1/2})$. When $\underline{\beta} \to \infty$, i.e., when the algorithm is learning a nearly non-robust view, our bound reduces to $\tilde{O}(d^{1/2}H^2/N^{1/2})$, which recovers the same dependence on H as the non-robust PEVI algorithm in Jin et al. (2021) and achieves the optimal dependence on N and d in Yin et al. (2022). The suboptimality in the H dependency arises as the result of our relatively simple algorithmic design to outline the first step in DRRL with linear function approximation. Adopting the advanced techniques of Yin et al. (2022) could potentially address the discrepancy, and we leave it as a future direction.

5 Extensions

5.1 Beyond Uniformly Well-explored Dataset

In practical applications, the data coverage may not satisfy Assumption 4.3, which requires the behavior policy to explore all feature dimensions with a sufficiently high exploration rate. Instead, we only need the behavior policy to adequately cover the features that the optimal policy will visit. To address this, we propose a pessimistic variant of our Algorithm 1, called **P**essimistic **D**istributionally **R**obust **V**alue **I**teration with **L**inear function approximation (PDRVI-L), inspired by the approach in Jin et al. (2021). Under a weaker data coverage condition, sample efficiency can be achieved as long as the dataset sufficiently covers the trajectory induced by the optimal policy π^*. We formalize this condition in Assumption 5.1.

Assumption 5.1 (Robust Sufficient Coverage of the Optimal Policy)**.** *Suppose there exists an absolute constant $c^\dagger > 0$ such that for any $P \in \mathcal{P}(\rho)$,*

$$\Lambda_h \geq I + c^\dagger \cdot N \cdot d \cdot \mathbb{E}_{P,\pi^*}[(\phi_i(s_h, a_h)\mathbb{1}_i)(\phi_i(s_h, a_h)\mathbb{1}_i)^\top | s_1 = s],$$

$\forall s \in \mathcal{S}, h \in [H], i \in [d]$, holds for probability at least $1 - \delta$.

Compared to the sufficient coverage condition in Jin et al. (2021), our Assumption 5.1 requires the collected samples Λ_h to cover each dimensions $i \in [d]$ uniformly well. This requirement arises from the ambiguity set constructed in the latent factor space. Moreover, we require this condition to hold uniformly across all the transition model within the ambiguity set, motivated by Blanchet et al. (2023). We summarize our algorithmic design in Appendix C and present it as Algorithm 2. In contrast to Algorithm 1, we subtract a pessimistic term $\gamma_h \sum_{i=1}^d \|\phi_i(s,a)\mathbb{1}_i\|\Lambda_h^{-1}$ from the estimated Q-value in Algorithm 2. This discourages our algorithm from selecting the action with less confidence. Compared to Jin et al. (2021), which uses $\gamma_h\|\phi(s,a)\|_{\Lambda_h^{-1}}$ as the pessimistic term in the non-robust setting, our approach provides a larger penalization and adapts to the distributionally robust nature. Under the partial coverage condition for our dataset, Algorithm 2 achieves sample efficiency, as shown in the following theorem.

Theorem 5.1. *In Algorithm 2 we set $\lambda = 1$ and*

$$\gamma_h = c_1\underline{\beta}(e^{\frac{H-h}{\underline{\beta}}}-1)d\zeta_3^{1/2} + c_2\underline{\beta}^{1/2}(e^{\frac{H-h}{\underline{\beta}}}-1)H^{1/2}\zeta_2^{1/2},$$

where ζ_2 is the same as in Theorem 4.1 and $\zeta_3 = \log(2N + 32N^2H^3d^{5/2}\zeta e^{2H/\beta})$ for some absolute constant c_1 and c_2 that are only dependent on $c^\dagger$. Then under the Assumption 4.1, 4.4 and 5.1, our algorithm 2 has the following guarantee with probability at least $1-\delta$,

$$\mathrm{SubOpt}(\widehat{\pi};\mathcal{P}) \leq c_1\underline{\beta}(e^{H/\underline{\beta}}-1)d^{3/2}H\zeta_3^{1/2}/N^{1/2} + c_2\underline{\beta}^{1/2}(e^{H/\underline{\beta}}-1)d^{1/2}H^{3/2}\zeta_2^{1/2}/N^{1/2}.$$

Our bound incurs an additional factor of d compared to Theorem 5.2 as a price of the weaker data coverage condition Specifically, the suboptimality for the Algorithm 2 is $\tilde{O}(d^{3/2}H^{3/2}\underline{\beta}^{1/2}e^{H/\underline{\beta}}/N^{1/2})$ when $\underline{\beta}$ is relatively small. When $\underline{\beta} \to \infty$, i.e., the algorithm is learning a nearly non-robust view, the suboptimality reduces to $\tilde{O}(d^{3/2}H^2/N^{1/2})$, which recovers the same dependence on d, H, and N as Jin et al. (2021). Recently, Yin et al. (2022) improves the suboptimality bound to $\tilde{O}(d^{1/2}H^{3/2}/N^{1/2})$ with a more complex algorithmic design. As our paper is the first attempt to design linear function approximation to solve the offline DRRL problem, we leave the improvement towards the optimal rate as a future direction.

5.2 Model Misspecification

The assumption of state aggregation may not be realistic when applied to real-world datasets. In this subsection, we relax the soft state-aggregation MDP assumption to allow for the possibility of a true transition kernel that is nearly a state-aggregation transition.

Assumption 5.2 (Model Misspecification in Transition Model)**.** *We assume that for all $h \in [H]$, there exists $\tilde{P}_h \in \mathrm{Span}(\phi)$ and $\xi \geq 0$ such that each (s,a), the true transition kernel $P_h(\cdot|s,a)$ satisfies $\|P_h(\cdot|s,a) - \tilde{P}_h(\cdot|s,a)\|_1 \leq \xi$. For the reward functions, we still assume that $r_h \in \mathrm{Span}(\phi)$ for all $h \in [H]$.*

Theorem 5.2 (Model Misspecification)**.** *We set $\lambda = 1$ in Algorithm 1. Under the Assumption 4.3, 4.4 and 5.2, when $N \geq 40/\underline{c} \cdot \log(4dH/\delta)$, we have the following holds with probability at least $1-\delta$,*

$$\mathrm{SubOpt}(\widehat{\pi};\mathcal{P}) \leq c_1\underline{\beta}(e^{H/\underline{\beta}}-1)(\xi d^{1/2} + d^{1/2}\zeta_1^{1/2})H/N^{1/2} + c_2\underline{\beta}^{1/2}(e^{H/\underline{\beta}}-1)H^{3/2}\zeta_2^{1/2}/N^{1/2} + \xi H^2/2.$$

Here ζ_1 and ζ_2 are the same in Theorem 4.1 and c_1 and c_2 are some absolute constants that only depend on $\underline{c}$.

Theorem 5.3 (Model Misspecification with Sufficient Coverage)**.** *In Algorithm 2 we set $\lambda = 1$ and*

$$\gamma_h = c_1\underline{\beta}(e^{\frac{H-h}{\underline{\beta}}}-1)d\zeta_3^{1/2} + c_2\underline{\beta}^{1/2}(e^{\frac{H-h}{\underline{\beta}}}-1)H^{1/2}\zeta_2^{1/2},$$

where ζ_2 and ζ_3 are the same as in Theorem 5.1 and $c_1, c_2 \geq 1$ are some absolute constants that only involve $c^\dagger$. Then based on Assumptions 4.4, 5.1 and 5.2, our Algorithm 2 has the following guarantee with probability at least $1-\delta$,

$$\mathrm{SubOpt}(\widehat{\pi};\mathcal{P}) \leq c_1\underline{\beta}(e^{H/\underline{\beta}}-1)(\xi d + d^{3/2}\zeta_3^{1/2})H/N^{1/2} + c_2\underline{\beta}^{1/2}(e^{H/\underline{\beta}}-1)d^{1/2}H^{3/2}\zeta_2^{1/2}/N^{1/2} + \xi H^2/2.$$

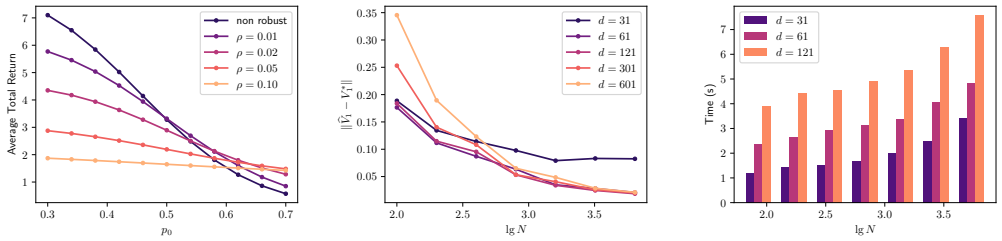
According to Theorem 5.2, when the soft-state aggregation model is inaccurate up to ξ total variation, the policy's performance incurs an approximation gap of $O(\xi \cdot (\underline{\beta}(e^{H/\underline{\beta}}-1)d^{1/2} + H^2))$ and $O(\xi \cdot (\underline{\beta}(e^{H/\underline{\beta}}-1)d + H^2))$ for our DRVI-L and PDRVI-L algorithms, respectively. The extent of degradation depends on the total-variation divergence of the empirical transition distribution from the true transition distribution, and the desired level of robustness.

6 Experiment

We evaluate the robustness and sample efficiency of our algorithms through numerical experiments in two well-known environments from the robust RL literature: the American put option environment (Tamar et al., 2014; Zhou et al., 2021b) and the CartPole environment in OpenAI Gym (Brockman et al., 2016). The American put option environment showcases the robustness and the impact of

different linear approximators, while the CartPole environment allows us to compare our algorithm with previous methods in a challenging setting with complex dynamics and a higher-dimensional, continuous state space. Additional experimental setup details can be found in Appendix D.

American Put Option: We compare Algorithm 2 with its non-robust counterpart, Pessimistic Value Iteration (PEVI) (Jin et al., 2021). Both algorithms are trained in an environment with $p_0 = 0.5$ and evaluated in a perturbed environment with varying p_0. The results, shown in Figure 2(a), demonstrate that the robust agent, particularly with a suitable radius $\rho = 0.01$, outperforms the non-robust agent in the perturbed environment with $p > 0.55$, with a slight performance degradation at $p_0 = 0.5$. Next, we investigate the impact of dimension d on suboptimality ($\|\widehat{V}_1 - V_1^*\|$) and computational time. Figure 2(b) reveals that a smaller d leads to lower estimation error and higher approximation error, given the same amount of data. The misspecification of the linear transition model introduces intrinsic bias to value estimation, but an appropriate bias reduces the estimation error with limited data, which is crucial for offline learning. Furthermore, Figure 2(c) demonstrates that the computational cost increases linearly with the dimension, rather than the size of the state-action space, indicating the potential of our algorithm for deployment in large-scale problems.



(a) Average Return (b) $\|\widehat{V}_1 - V_1^*\|$ (c) Execution time

Figure 2: Results in American Option Experiment. (a) Average total return of different KL radius ρ in the perturbed environments ($N = 1000$, $d = 61$). (b) Estimation error with different linear function dimension d's and the sizes of dataset N's ($\rho = 0.01$). (c) Execution time for different d's.

CartPole: We compare our PDRVI algorithm with several representative offline RL algorithms in the CartPole environment: (a) RFQI (Panaganti et al., 2022b), known for its capacity for non-linear approximation and superior performance; (b) RAPI (Tamar et al., 2014), further validating our message about the limitations of the RTA approach; (c) PEVI (Jin et al., 2021) as a non-robust benchmark. A summary of their features can be found in Table 3. To evaluate the algorithms' robustness, we introduce different levels of action perturbations and assess their performance in the perturbed environments. Our PDRVI algorithm demonstrate superior robust performance compared to non-robust PEVI and comparable performance to RFQI, despite using a simpler approximator. Notably, our algorithm enjoys theoretical guarantees through the DR variant of pessimism, while RFQI relies on a batch-constrained Q-learning algorithm that may not converge optimally under weak data coverage conditions. In contrast, RAPI performs poorly compared to other algorithms in all cases, supporting our claim in Section 3 that the RTA design can lead to suboptimal decisions. Despite using the more conservative R-contamination (R-con) ambiguity set, RAPI's significant performance gap compared to other algorithms confirms the ineffectiveness of the RTA approach.

Algo	Representation.	Pessimism.	Ambiguity.
PDRVI (Ours)	Linear	Yes	KL
RFQI	NN	Yes	TV
RAPI	Linear	No	R-con
PEVI	Linear	Yes	None

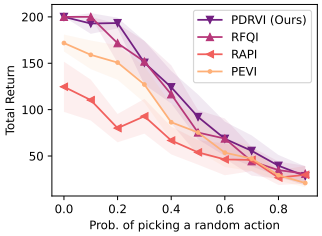


Figure 3: Experiment Results for the CartPole environment. (Left) Summary of the algorithms' features. (Right) Average return of different algorithms in the perturbed environments over 10 random seeds shadowed with standard deviation.

REFERENCES

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24, 2011.

Gaon An, Seungyong Moon, Jang-Hyun Kim, and Hyun Oh Song. Uncertainty-based offline reinforcement learning with diversified q-ensemble. *Advances in neural information processing systems*, 34:7436–7447, 2021.

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

Aharon Ben-Tal and Arkadi Nemirovski. Robust convex optimization. *Mathematics of operations research*, 23(4):769–805, 1998.

Aharon Ben-Tal and Arkadi Nemirovski. Robust solutions of linear programming problems contaminated with uncertain data. *Mathematical programming*, 88(3):411–424, 2000.

Dimitri P Bertsekas and John N Tsitsiklis. Neuro-dynamic programming: an overview. In *Proceedings of 1995 34th IEEE conference on decision and control*, volume 1, pp. 560–564. IEEE, 1995.

Dimitris Bertsimas, Vishal Gupta, and Nathan Kallus. Data-driven robust optimization. *Mathematical Programming*, 167(2):235–292, 2018.

Jose Blanchet, Miao Lu, Tong Zhang, and Han Zhong. Double pessimism is provably efficient for distributionally robust offline reinforcement learning: Generic algorithm and robust partial coverage, 2023.

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

Xinyue Chen, Zijian Zhou, Zheng Wang, Che Wang, Yanqiu Wu, and Keith Ross. Bail: Best-action imitation learning for batch deep reinforcement learning. *Advances in Neural Information Processing Systems*, 33:18353–18363, 2020.

Ching-An Cheng, Tengyang Xie, Nan Jiang, and Alekh Agarwal. Adversarially trained actor critic for offline reinforcement learning. *arXiv preprint arXiv:2202.02446*, 2022.

John C Cox, Stephen A Ross, and Mark Rubinstein. Option pricing: A simplified approach. *Journal of financial Economics*, 7(3):229–263, 1979.

Yaqi Duan, Tracy Ke, and Mengdi Wang. State aggregation learning from markov transition data. *Advances in Neural Information Processing Systems*, 32, 2019.

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

John C Duchi, Peter W Glynn, and Hongseok Namkoong. Statistics of robust optimization: A generalized empirical likelihood approach. *Mathematics of Operations Research*, 46(3):946–969, 2021.

Laurent El Ghaoui and Arnab Nilim. Robust solutions to markov decision problems with uncertain transition matrices. *Operations Research*, 53(5):780–798, 2005.

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

Scott Fujimoto, David Meger, and Doina Precup. Off-policy deep reinforcement learning without exploration. In *International conference on machine learning*, pp. 2052–2062. PMLR, 2019.

Omer Gottesman, Fredrik Johansson, Matthieu Komorowski, Aldo Faisal, David Sontag, Finale Doshi-Velez, and Leo Anthony Celi. Guidelines for reinforcement learning in healthcare. *Nature medicine*, 25(1):16–18, 2019.

Vineet Goyal and Julien Grand-Clement. Robust markov decision processes: Beyond rectangularity. *Mathematics of Operations Research*, 2022.

Chin Pang Ho, Marek Petrik, and Wolfram Wiesemann. Fast bellman updates for robust mdps. In *International Conference on Machine Learning*, pp. 1979–1988. PMLR, 2018.

Chin Pang Ho, Marek Petrik, and Wolfram Wiesemann. Partial policy iteration for l1-robust markov decision processes. *J. Mach. Learn. Res.*, 22:275–1, 2021.

Zhaolin Hu and L Jeff Hong. Kullback-leibler divergence constrained distributionally robust optimization. *Available at Optimization Online*, pp. 1695–1724, 2013.

Sandy Huang, Nicolas Papernot, Ian Goodfellow, Yan Duan, and Pieter Abbeel. Adversarial attacks on neural network policies. *arXiv preprint arXiv:1702.02284*, 2017.

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

Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E Schapire. Contextual decision processes with low bellman rank are pac-learnable. In *International Conference on Machine Learning*, pp. 1704–1713. PMLR, 2017.

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

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

David L Kaufman and Andrew J Schaefer. Robust modified policy iteration. *INFORMS Journal on Computing*, 25(3):396–410, 2013.

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

Ilya Kostrikov, Ashvin Nair, and Sergey Levine. Offline reinforcement learning with implicit q-learning. In *International Conference on Learning Representations*, 2022.

Aviral Kumar, Justin Fu, Matthew Soh, George Tucker, and Sergey Levine. Stabilizing off-policy q-learning via bootstrapping error reduction. *Advances in Neural Information Processing Systems*, 32, 2019.

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

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

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

Shiau Hong Lim, Huan Xu, and Shie Mannor. Reinforcement learning in robust markov decision processes. *Advances in Neural Information Processing Systems*, 26, 2013.

Shie Mannor, Duncan Simester, Peng Sun, and John N Tsitsiklis. Bias and variance in value function estimation. In *Proceedings of the twenty-first international conference on Machine learning*, pp. 72, 2004.

Aditya Modi, Nan Jiang, Ambuj Tewari, and Satinder Singh. Sample complexity of reinforcement learning using linearly combined model ensembles. In *International Conference on Artificial Intelligence and Statistics*, pp. 2010–2020. PMLR, 2020.

Rémi Munos and Csaba Szepesvári. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9(5), 2008.

Hongseok Namkoong and John C Duchi. Variance-based regularization with convex objectives. *Advances in neural information processing systems*, 30, 2017.

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

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

Kishan Panaganti, Zaiyan Xu, Dileep Kalathil, and Mohammad Ghavamzadeh. Robust reinforcement learning using offline data. *arXiv preprint arXiv:2208.05129*, 2022a.

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

Anay Pattanaik, Zhenyi Tang, Shuijing Liu, Gautham Bommannan, and Girish Chowdhary. Robust deep reinforcement learning with adversarial attacks. *arXiv preprint arXiv:1712.03632*, 2017.

Lerrel Pinto, James Davidson, Rahul Sukthankar, and Abhinav Gupta. Robust adversarial reinforcement learning. In *International Conference on Machine Learning*, pp. 2817–2826. PMLR, 2017.

Rafael Rafailov, Tianhe Yu, Aravind Rajeswaran, and Chelsea Finn. Offline reinforcement learning from images with latent space models. In *Learning for Dynamics and Control*, pp. 1154–1168. PMLR, 2021.

Antonin Raffin. Rl baselines3 zoo. `https://github.com/DLR-RM/rl-baselines3-zoo`, 2020.

Aravind Rajeswaran, Sarvjeet Ghotra, Balaraman Ravindran, and Sergey Levine. Epopt: Learning robust neural network policies using model ensembles. *arXiv preprint arXiv:1610.01283*, 2016.

Paria Rashidinejad, Banghua Zhu, Cong Ma, Jiantao Jiao, and Stuart Russell. Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *Advances in Neural Information Processing Systems*, 34:11702–11716, 2021.

Jay K Satia and Roy E Lave Jr. Markovian decision processes with uncertain transition probabilities. *Operations Research*, 21(3):728–740, 1973.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *ArXiv*, abs/1707.06347, 2017.

Paul J Schweitzer and Abraham Seidmann. Generalized polynomial approximations in markovian decision processes. *Journal of mathematical analysis and applications*, 110(2):568–582, 1985.

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

Tianyu Shi, Dong Chen, Kaian Chen, and Zhaojian Li. Offline reinforcement learning for autonomous driving with safety and exploration enhancement. *arXiv preprint arXiv:2110.07067*, 2021.

Noah Y Siegel, Jost Tobias Springenberg, Felix Berkenkamp, Abbas Abdolmaleki, Michael Neunert, Thomas Lampe, Roland Hafner, Nicolas Heess, and Martin Riedmiller. Keep doing what worked: Behavioral modelling priors for offline reinforcement learning. *arXiv preprint arXiv:2002.08396*, 2020.

Duncan I Simester, Peng Sun, and John N Tsitsiklis. Dynamic catalog mailing policies. *Management science*, 52(5):683–696, 2006.

Satinder Singh, Tommi Jaakkola, and Michael Jordan. Reinforcement learning with soft state aggregation. *Advances in neural information processing systems*, 7, 1994.

Aviv Tamar, Shie Mannor, and Huan Xu. Scaling up robust mdps using function approximation. In *International conference on machine learning*, pp. 181–189. PMLR, 2014.

Shengpu Tang and Jenna Wiens. Model selection for offline reinforcement learning: Practical considerations for healthcare settings. In *Machine Learning for Healthcare Conference*, pp. 2–35. PMLR, 2021.

Martin J Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*, volume 48. Cambridge University Press, 2019.

Qing Wang, Jiechao Xiong, Lei Han, Han Liu, Tong Zhang, et al. Exponentially weighted imitation learning for batched historical data. *Advances in Neural Information Processing Systems*, 31, 2018.

Ruosong Wang, Dean P Foster, and Sham M Kakade. What are the statistical limits of offline rl with linear function approximation? *arXiv preprint arXiv:2010.11895*, 2020.

Yining Wang, Ruosong Wang, Simon Shaolei Du, and Akshay Krishnamurthy. Optimism in reinforcement learning with generalized linear function approximation. In *International Conference on Learning Representations*, 2021.

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

Yifan Wu, George Tucker, and Ofir Nachum. Behavior regularized offline reinforcement learning. *arXiv preprint arXiv:1911.11361*, 2019.

Yue Wu, Shuangfei Zhai, Nitish Srivastava, Joshua M Susskind, Jian Zhang, Ruslan Salakhutdinov, and Hanlin Goh. Uncertainty weighted actor-critic for offline reinforcement learning. In *International Conference on Machine Learning*, pp. 11319–11328. PMLR, 2021.

Tengyang Xie, Ching-An Cheng, Nan Jiang, Paul Mineiro, and Alekh Agarwal. Bellman-consistent pessimism for offline reinforcement learning. *Advances in neural information processing systems*, 34, 2021.

Huan Xu and Shie Mannor. Distributionally robust markov decision processes. *Advances in Neural Information Processing Systems*, 23, 2010.

Lin Yang and Mengdi Wang. Sample-optimal parametric q-learning using linearly additive features. In *International Conference on Machine Learning*, pp. 6995–7004. PMLR, 2019.

Wenhao Yang, Liangyu Zhang, and Zhihua Zhang. Towards theoretical understandings of robust markov decision processes: Sample complexity and asymptotics. *arXiv preprint arXiv:2105.03863*, 2021a.

Yiqin Yang, Xiaoteng Ma, Li Chenghao, Zewu Zheng, Qiyuan Zhang, Gao Huang, Jun Yang, and Qianchuan Zhao. Believe what you see: Implicit constraint approach for offline multi-agent reinforcement learning. *Advances in Neural Information Processing Systems*, 34:10299–10312, 2021b.

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

Chao Yu, Jiming Liu, Shamim Nemati, and Guosheng Yin. Reinforcement learning in healthcare: A survey. *ACM Computing Surveys (CSUR)*, 55(1):1–36, 2021.

Fisher Yu, Wenqi Xian, Yingying Chen, Fangchen Liu, Mike Liao, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving video database with scalable annotation tooling. *arXiv preprint arXiv:1805.04687*, 2(5):6, 2018.

Tianhe Yu, Garrett Thomas, Lantao Yu, Stefano Ermon, James Y Zou, Sergey Levine, Chelsea Finn, and Tengyu Ma. Mopo: Model-based offline policy optimization. *Advances in Neural Information Processing Systems*, 33:14129–14142, 2020.

Ekim Yurtsever, Jacob Lambert, Alexander Carballo, and Kazuya Takeda. A survey of autonomous driving: Common practices and emerging technologies. *IEEE access*, 8:58443–58469, 2020.

Andrea Zanette. Exponential lower bounds for batch reinforcement learning: Batch rl can be exponentially harder than online rl. In *International Conference on Machine Learning*, pp. 12287–12297. PMLR, 2021.

Andrea Zanette, Alessandro Lazaric, Mykel Kochenderfer, and Emma Brunskill. Learning near optimal policies with low inherent bellman error. In *International Conference on Machine Learning*, pp. 10978–10989. PMLR, 2020.

Andrea Zanette, Martin J Wainwright, and Emma Brunskill. Provable benefits of actor-critic methods for offline reinforcement learning. *Advances in neural information processing systems*, 34:13626–13640, 2021.

Anru Zhang and Mengdi Wang. Spectral state compression of markov processes. *IEEE transactions on information theory*, 66(5):3202–3231, 2019.

Huan Zhang, Hongge Chen, Chaowei Xiao, Bo Li, Mingyan Liu, Duane Boning, and Cho-Jui Hsieh. Robust deep reinforcement learning against adversarial perturbations on state observations. *Advances in Neural Information Processing Systems*, 33:21024–21037, 2020.

Wenxuan Zhou, Sujay Bajracharya, and David Held. Plas: Latent action space for offline reinforcement learning. In *Conference on Robot Learning*, pp. 1719–1735. PMLR, 2021a.

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