Efficient Duple Perturbation Robustness in Low-rank MDPs*

Yang Hu YANGHU@G.HARVARD.EDU
Haitong Ma HAITONGMA@G.HARVARD.EDU
Na Li NALI@SEAS.HARVARD.EDU
School of Engineering and Applied Sciences, Harvard University

Bo Dai BODAI@CC.GATECH.EDU
Google DeepMind & Georgia Institute of Technology

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

Abstract

The pursuit of robustness has recently been a popular topic in reinforcement learning (RL) research, yet the existing methods generally suffer from computation issues that obstruct their real-world implementation. In this paper, we consider MDPs with low-rank structures, where the transition kernel can be written as a linear product of feature map and factors. We introduce *duple perturbation* robustness, i.e. perturbation on both the feature map and the factors, via a novel characterization of (ξ, η)-ambiguity sets featuring computational efficiency. Our novel low-rank robust MDP formulation is compatible with the low-rank function representation view, and therefore, is naturally applicable to practical RL problems with large or even continuous state-action spaces. Meanwhile, it also gives rise to a provably efficient and practical algorithm with theoretical convergence rate guarantee. Lastly, the robustness of our proposed approach is justified by numerical experiments, including classical control tasks with continuous state-action spaces.

Keywords: Reinforcement learning, duple perturbation robustness, low-rank MDPs.

1. Introduction

The recent years have witnessed the rapid development of reinforcement learning (RL), a discipline that has shown its power in various areas, ranging from gaming (Mnih et al., 2015; Silver et al., 2017; Vinyals et al., 2019), robotics (Dalal et al., 2021; Levine et al., 2016), to large language models (Ouyang et al., 2022; Ziegler et al., 2019). These successes of RL are achieved with huge amounts of data, yet for applications like robotics and autonomous driving, the real-world data collection is expensive and dangerous. As a result, RL agents are usually trained (and even tested) in simulated environments. However, simulators are typically not accurate for revealing the intrinsic uncertainty and ambiguity in the real-world dynamics. Therefore, RL agents trained in simulated environments generally suffer from a *sim-to-real* performance degradation (Peng et al., 2018; Salvato et al., 2021), which is the cost for overlooking the distributional shift from simulation to real-world environments.

Robust MDPs are proposed to mitigate such performance degradation (Satia and Lave Jr, 1973; Silver, 1963), in which the value of a policy is evaluated with respect to the worst possible realization within a prescribed ambiguity set of dynamics. The modern formulation of robust MDPs shall be attributed to Iyengar (2005) and El Ghaoui and Nilim (2005), with a whole series of theoretical efforts aiming at solving the robust MDP (Blanchet et al., 2023; Goyal and Grand-Clement, 2023; Ho et al., 2018, 2021; Yang et al., 2022; Zhou et al., 2021) and various empirical successes thereafter (Pattanaik et al., 2017; Pinto et al., 2017; Rajeswaran et al., 2016; Zhang et al., 2020). Nevertheless, most of these theoretical works consider *tabular MDPs*, and the computational complexity and/or

* The appendices can be found attached to the full paper at https://arxiv.org/abs/2404.08089.

sample complexity are, at best, polynomial in the size of state and action spaces. Therefore, these algorithms are bound to suffer from the *curse of dimensionality*, which makes these algorithms impossible to deploy in environments with large or even infinite state-action spaces, and thus restricts their modeling ability of real-world environments.

Towards alleviation of computational burden, people have been working to identify structures that reduce the intrinsic dimensionality of the problem. Among these attempts, MDPs with low-rank structures (Jin et al., 2020; Yang and Wang, 2020) turn out to be a representative model family with practical implications (Ren et al., 2022c, 2023). Specifically, in this paper we consider MDPs with such low-rank structures that the transition probability kernels can be represented by an inner product of the *feature map* $\boldsymbol{\phi}(s,a) \in \mathbb{R}^d$ and the *factor* $\boldsymbol{\mu}(s') \in \mathbb{R}^d$ (i.e., $\mathbb{P}(s'|s,a) = \langle \boldsymbol{\phi}(s,a), \boldsymbol{\mu}(s') \rangle$), and the reward functions can also be decomposed in a similar way. We will stick to the term "*low-rank* MDPs"[1] to highlight the fact that both the feature map and the factors may be unknown or subject to uncertainty. Along this line of work, provably efficient algorithms have been developed such that the sample complexity is polynomial in d, the dimension of the feature space (Agarwal et al., 2020; Jin et al., 2020; Ren et al., 2022b; Uehara et al., 2021; Yang and Wang, 2020).

Despite these promising results for low-rank MDPs, this line of work in general only deals with the non-robust case. Even though studies on misspecified low-rank MDPs do take modelling errors into consideration, the focus of that line of work is to bound the regret of learning in terms of the misspecification error, assuming online interaction with a fixed MDP. On the contrary, the robust MDP formulation aims to handle the model discrepancy between the planning/learning phase (using simulators or estimated models) and the deployment phase (where the uncertainty is intrinsic and we should prepare for the worst-case scenario). Recently, there are some emerging efforts to introduce function approximation into the realm of robust MDPs (Badrinath and Kalathil, 2021; Tamar et al., 2014), yet they only treat function approximation of value functions as a computational tool, largely ignoring its connection to the potentially beneficial restructuring of ambiguity sets. Ma et al. (2023) and Goyal and Grand-Clement (2023) are the first to exploit such structures by introducing d-rectangular ambiguity sets in terms of the factor $\boldsymbol{\mu}(s')$, yet the ambiguity is *single-sided* in that they assume a known fixed feature map $\boldsymbol{\phi}(s,a)$, and thus the transition kernels and value functions always lie in a fixed function space induced by $\boldsymbol{\phi}(s,a)$. Besides, they also require the ambiguity sets to have specific soft state-aggregate and convexity structures. These restrict the real-world application of their robust MDP formulation since there is no reason to believe that the feature map is accurate and structured in simulators.

To the best of our knowledge, none of the existing methods handle the uncertainty of feature map $\boldsymbol{\phi}(s,a)$ and factor $\boldsymbol{\mu}(s')$ together. However, in real-world systems, feature maps may be truncated (Ren et al., 2023), kernel-induced (Ren et al., 2022c) or generated by latent variables (Ren et al., 2022a), and are thus subject to uncertainty as well, just like the factors. One major computational challenge of handling duple perturbation robustness in low-rank MDPs is that the naive formulation (see Section 3.1) would require finding the worst-case feature map $\boldsymbol{\phi}(s,a)$ and factor $\boldsymbol{\mu}(s')$, leading to $|\mathcal{S}||\mathcal{A}|$ non-convex problems with at least $d|\mathcal{S}|$ decision variables, and therefore losing the advantage brought by the low-rank representation. These observations motivate the following research question:

How to design a ***provably efficient*** *algorithm for* ***duple*** *robustness in low-rank MDPs?*

1. Different names are used to refer to such MDPs with low-rank structures in literature, including factorized linear MDPs (Yao et al., 2014), factor matrix MDPs (Goyal and Grand-Clement, 2023), (soft) state-aggregate MDPs (Ma et al., 2023; Singh et al., 1994), linear MDPs (Jin et al., 2020; Yang and Wang, 2020), etc., which all leverage the idea to represent the transition kernel as an inner product in the feature space.

Here by "provably efficient" we mean the algorithm is computationally efficient with theoretical guarantees, and by "duple" we refer to model uncertainty on both the feature map and the factors.

Contribution. In this paper, we provide an **affirmative** answer to this question. To handle the aforementioned challenge of the large number of decision variables in the naive low-rank robustness formulation, we propose a novel formulation of *robust low-rank MDPs* with (ξ, η)-*rectangular* ambiguity sets. The new formulation features compatibility and scalability for low-rank MDPs, in the sense that it performs duple perturbation around the nominal model using only $\Theta(d)$ variables. We justify the proposed formulation by examples and numerical results, and also present key properties that relate it to existing robustness concepts.

Then we design an algorithm, **R**epresentation **R**obust **P**olicy **G**radient ($\mathsf{R^2PG}$), to solve the proposed robust low-rank MDPs with (ξ, η)-rectangularity. The algorithm is tractable as the optimization involved can be reduced to a $\Theta(d)$-dimensional semi-definite programming (SDP) problem, and is thus potentially scalable to work with large state-action spaces. We then show the quasi-contraction property and the extended performance difference lemma for our robust low-rank MDP formulation, based on which we establish a theoretical guarantee for the proposed $\mathsf{R^2PG}$ algorithm that ensures provably efficient convergence to the optimal low-rank robust policy with bounded suboptimality. In addition, we evaluate the numerical performance of our $\mathsf{R^2PG}$ algorithm on multiple continuous control benchmarks, where the robust algorithm leveraging the proposed low-rank robustness significantly outperforms its non-robust counterpart.

Due to limited space, a detailed literature review is provided in Appendix A: Related Works. Appendices can be found attached to the full paper at https://arxiv.org/abs/2404.08089.

2. Preliminaries

In this section, we introduce some basic notations and definitions in RL.

Notations. Denote by $\|\cdot\|$ the Euclidean 2-norm, and by $\langle \cdot, \cdot \rangle$ the standard vector inner product. Write $[n]$ for the set $\{1, 2, \ldots, n\}$. For any set S, let $\Delta(S)$ denote the probability simplex over S.

Markov Decision Processes (MDPs). We consider a *finite-horizon MDP*, which is described by a tuple $\mathsf{M} = (H, \mathcal{S}, \mathcal{A}, \{\mathbb{P}_h\}, \{r_h\}, \rho)$. Here H is the *horizon* or the length of each episode, $\mathcal{S}$ is the *state space*, and $\mathcal{A}$ is the *action space*; at step $h \in [H]$, $\mathbb{P}_h : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the *transition probability kernel*, and $r_h : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is the *reward function*; $\rho \in \Delta(\mathcal{S})$ denotes the initial state distribution. A *policy* $\pi = (\pi_1, \ldots, \pi_H)$ is composed of stage policies $\pi_h : \mathcal{S} \to \Delta(\mathcal{A})$, which determines a distribution over actions for each observed state at step $h \in [H]$.

For an MDP with transition probability kernel $\mathbb{P}$, given policy π, let $\mathbb{E}_{\pi,\mathbb{P}}$ denote the expectation over a trajectory with initial distribution ρ and evolving by $a_\tau \sim \pi_\tau(\cdot|s_\tau)$, $s_{\tau+1} \sim \mathbb{P}_\tau(\cdot|s_\tau, a_\tau)$ (the domain of time τ shall be inferred from context). Define $\Pr_{\pi,\mathbb{P}}$ similarly for the probability.

For any policy π, the standard V- and Q-functions starting from step $h \in [H]$ are defined as

$$V_h^\pi(s) := \mathbb{E}_{\pi,\mathbb{P}}\Big[\sum_{\tau=h}^H r_\tau(s_\tau, a_\tau) \;\Big|\; s_h = s\Big], \tag{1a}$$

$$Q_h^\pi(s, a) := \mathbb{E}_{\pi,\mathbb{P}}\Big[\sum_{\tau=h}^H r_\tau(s_\tau, a_\tau) \;\Big|\; s_h = s,\ a_h = a\Big]. \tag{1b}$$

For simplicity, we abuse the above notation a bit and write $V_1^\pi(\rho) := \mathbb{E}_{s_1\sim\rho}[V_1^\pi(s)]$. The (recursive) Bellman equation with respect to policy π, as well as the *Bellman update operator* $\mathcal{B}_h$, are defined by

$$Q_h^\pi(s, a) = r_h(s, a) + \mathbb{E}_{s'\sim\mathbb{P}_{h+1}(\cdot|s,a)}\big[V_{h+1}^\pi(s')\big] =: [\mathcal{B}_h V_{h+1}^\pi](s, a), \tag{2}$$

where $V_h^\pi = \big\langle Q_h^\pi(s, \cdot), \pi_h(\cdot|s)\big\rangle$. Note that here the Bellman operator performs an update from V_{h+1} to Q_h, which agrees with the definition in Ma et al. (2023), but may be different from other literature.

Given a policy π, the *state occupancy measure* is defined by $\rho_h^\pi(s) := \Pr_{\pi,\mathbb{P}}[s_h = s \mid s_1 \sim \rho]$, and the *state-action occupancy measure* is defined by $d_h^\pi(s,a) := \Pr_{\pi,\mathbb{P}}[s_h = s,\ a_h = a \mid s_1 \sim \rho]$ at step $h \in [H]$. Note that we always have $d_h^\pi(s,a) = \rho_h^\pi(s)\pi(a|s)$ and $\rho_h^\pi(s') = \sum_{s'} d_h^\pi(s,a)\mathbb{P}_h(s'|s,a)$.

Low-rank MDPs. An MDP $\mathcal{M}$ is said to have a low-rank structure (i.e., being a *low-rank MDP*), if there exists a *feature map* $\boldsymbol{\phi}_h : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ and two *factors* $\boldsymbol{\mu}_h : \mathcal{S} \to \mathbb{R}^d$, $\boldsymbol{\nu}_h \in \mathbb{R}^d$ for each step $h \in [H]$, such that for any $h \in [H]$, $s, s' \in \mathcal{S}$ and $a \in \mathcal{A}$, we have

$$\mathbb{P}_h(s'|s,a) = \langle \boldsymbol{\phi}_h(s,a), \boldsymbol{\mu}_h(s') \rangle, \quad r_h(s,a) = \langle \boldsymbol{\phi}_h(s,a), \boldsymbol{\nu}_h \rangle. \tag{3}$$

We point out that both the feature map and the factors are allowed to be different across steps. As discussed in Section 1, different names are used in literature to refer to MDPs with such decomposition structure. In this paper, though we mainly focus on the robust MDP planning problem with known nominal feature map and factors, we adopt the term "*low-rank* MDPs" to highlight the fact that the robustness we study will involve duple perturbations on both the feature $\boldsymbol{\phi}$ and the factor $\boldsymbol{\mu}$.

It is well-known that, in low-rank MDPs, Q_h^π is also representable by the same feature map as

$$Q_h^\pi(s,a) = [r_h(s,a) + \mathbb{P}_h V_{h+1}^\pi](s,a) = \Big\langle \boldsymbol{\phi}_h(s,a), \underbrace{\boldsymbol{\nu}_h + \textstyle\sum_{s'} V_{h+1}^\pi(s')\boldsymbol{\mu}_h(s')}_{\boldsymbol{\omega}_h^\pi} \Big\rangle, \tag{4}$$

where $\boldsymbol{\omega}_h^\pi$ is called the *factor* of Q-function that is irrelevant of s, a and s'.

We make the following assumption that is standard (Jin et al., 2020; Uehara et al., 2021).

Assumption 2.1 (Bounded norms of feature maps and factors) *$\|\boldsymbol{\phi}_h(s,a)\| \le 1$, $\|\boldsymbol{\nu}_h\| \le \sqrt{d}$, $\|\sum_s V(s)\boldsymbol{\mu}_h(s)\| \le \sqrt{d}$ for any $s \in \mathcal{S}$, $a \in \mathcal{A}$ and $V : \mathcal{S} \to [0,H]$.*

Low-rank structures of MDPs are favored in that it enables the design of efficient planning and learning algorithms (Agarwal et al., 2020; Jin et al., 2020; Ren et al., 2022b; Uehara et al., 2021; Yang and Wang, 2020) by leveraging the intrinsic low-rank structure, the sample/computational complexity of which only depends on the intrinsic dimension d rather than $|\mathcal{S}|$ and $|\mathcal{A}|$. We emphasize that the low-rank structure should not be seen as a restrictive assumption in general. Low-rank MDPs naturally arise from various latent structures, such as latent variable representations (Agarwal et al., 2020; Ren et al., 2022a) and block MDPs (Dann et al., 2018; Du et al., 2019). Additionally, generic transition kernels can be effectively approximated by low-rank kernels through techniques like spectral decomposition (Duan et al., 2019; Ren et al., 2022b, 2023). However, approximation or truncation of feature maps and factors is often necessary to obtain low-rank structures, which introduces intrinsic modeling errors in both feature maps and factors. This necessitates the use of low-rank robust MDPs. For instance, to fit the continuous control benchmarks discussed in Section 6.2 into the low-rank MDP framework, we follow the approach of Ren et al. (2023) to truncate the infinite-dimensional Gaussian kernel representation, resulting in finite-dimensional approximate representations that inherently induce modeling errors.

3. Robust Low-Rank MDP with Duple Perturbation

As discussed earlier, when formulating problems as low-rank MDPs, modeling errors often occur in both feature maps and factors. These errors can be understood as *perturbations* centered around a nominal model — this is a natural equivalent understanding of ambiguity sets in the standard formulation of (s,a)-rectangular (tabular) robust MDPs (Dong et al., 2022; Panaganti et al., 2022). However, existing work like Goyal and Grand-Clement (2023); Ma et al. (2023) only considers single-sided perturbation, assuming that features are known or fixed; even though misspecified low-rank MDPs take one step further to consider modeling errors, as discussed in Section 1, the misspecified setting is fundamentally different from the robust setting in terms of the interaction

pattern with the environment. Later in in this section, we will see that there exists some intrinsic difficulty in developing computationally efficient algorithms for robust low-rank MDP with duple perturbations by following the naive way of formulating the ambiguity sets. Therefore, in this section, we propose a new robust low-rank MDP formulation that enables the design of efficient robust RL algorithms for large state-action spaces with duple perturbation robustness (i.e., perturbations are applied to both the feature map and the factors).

3.1. The Proposed Robust Low-Rank MDP

When formulating robust low-rank MDPs, in parallel to the tabular robust MDP with (s,a)-rectangular ambiguity set, it is tempting to consider the following *naive robust low-rank MDP* formulation with (ϕ,μ,ν)*-rectangular ambiguity set* $\widetilde{\mathcal{M}} = \bigotimes_{h\in[H]} \widetilde{\mathcal{M}}_h$, where $\widetilde{\mathcal{M}}_h$ is the stage ambiguity set at step $h \in [H]$ centered around the nominal model $(\boldsymbol{\phi}_h^\circ, \boldsymbol{\mu}_h^\circ, \boldsymbol{\nu}_h^\circ)$, defined by[2]

$$\widetilde{\mathcal{M}}_h := \left\{ (\boldsymbol{\phi}_h, \boldsymbol{\mu}_h, \boldsymbol{\nu}_h) \;\middle|\; \begin{array}{l} \|\boldsymbol{\phi}_h(s,a) - \boldsymbol{\phi}_h^\circ(s,a)\| \le R_{\phi,h},\ \forall (s,a) \in \mathcal{S}\times\mathcal{A}; \\ \|\boldsymbol{\mu}_h(s') - \boldsymbol{\mu}_h^\circ(s')\| \le R_{\mu,h},\ \langle \boldsymbol{\phi}_h(s,a), \boldsymbol{\mu}_h(\cdot)\rangle \in \Delta(\mathcal{A}),\ \forall s' \in \mathcal{S}; \\ \|\boldsymbol{\nu}_h - \boldsymbol{\nu}_h^\circ\| \le R_{\nu,h}. \end{array} \right\}. \quad (5)$$

The objective of robust low-rank MDP is to find an optimal policy $\widetilde{\pi}^*$ to optimize the *worst-case* performance under this ambiguity set. Similarly, given a policy π, the robust value function $\widetilde{V}_h^\pi$ is defined as the worst-case performance under the set, and we can also define corresponding robust Bellman operators $\widetilde{\mathcal{B}}_h$ to perform robust policy evaluations. Detailed definitions of $\widetilde{\pi}^*$, $\widetilde{V}_h^\pi$, and $\widetilde{\mathcal{B}}_h$ in the naive robust MDP setting can be found in Appendix B.1.

Unfortunately, the naive approach faces a major computational challenge of handling duple perturbation robustness in low-rank MDPs, since to apply planning algorithms like value iteration and policy iteration, we still need to solve $|\mathcal{S}||\mathcal{A}|$ independent optimizations, each involving $d|\mathcal{S}|$ optimization variables, and even coming with additional simplex constraints (i.e., $\langle \boldsymbol{\phi}_h(s,a), \boldsymbol{\mu}_h(\cdot)\rangle \in \Delta(\mathcal{S})$) that makes the feasible region highly non-convex. As a result, the naive formulation fails to exploit the advantage of introducing low-rank representations, but rather, only adds to the complexity of optimization and magnifies the drawbacks of representations.

To consider an alternative formulation that could address the complexity issues discussed above, we first note that, from a high-level perspective, the key to ensuring robustness is to introduce perturbation in the nominal process. In standard tabular robust MDPs and naive low-rank robust MDPs, the Bellman updates are component-wise independently perturbed across the entire ambiguity set, resulting in prohibitively high computational complexity. On the contrary, our proposed low-rank robustness formulation takes advantage of the structure in Bellman update operators to "merge" the independent perturbations into "effective perturbations", leading to reduced complexity.

Robust Low-rank MDP with $(\boldsymbol{\xi},\boldsymbol{\eta})$-rectangularity. As discussed above, in our proposed robust low-rank MDP, robustness is achieved by perturbing the policy evaluation scheme. Note that in the nominal MDP, the objective $V_1^\pi(\rho)$ can be expanded in terms of $Q_h^\pi(\cdot,\cdot)$ as

$$V_1^\pi(\rho) = \textstyle\sum_{\tau<h} \mathbb{E}_{(s_\tau,a_\tau)\sim d_\tau^\pi}[r_\tau(s_\tau,a_\tau)] + \mathbb{E}_{(s_h,a_h)\sim d_h^\pi}[Q_h^\pi(s_h,a_h)].$$

Therefore, the way Q_h^π appears in V_1^π is only through the weighted average Q-function

$$\mathbb{E}_{(s_h,a_h)\sim d_h^\pi}[Q_h^\pi(s_h,a_h)] = \mathbb{E}_{(s_h,a_h)\sim d_h^\pi}\left[\left\langle \boldsymbol{\phi}_h^\circ(s_h,a_h), \boldsymbol{\omega}_h^\circ\right\rangle\right] = \left\langle \mathbb{E}_{(s_h,a_h)\sim d_h^\pi}[\boldsymbol{\phi}_h^\circ(s_h,a_h)], \boldsymbol{\omega}_h^\circ\right\rangle.$$

Now we may perturb the above "effective" appearance of Q_h^π for robust evaluation of a given policy π. Formally, given policy π, we recursively define the following robust policy evaluation scheme: for the boundary conditions, set $\widehat{V}_{H+1}^\pi(\cdot) \equiv 0$; for the recursive update at time step $h \in [H]$, given

2. Here we extend the concept of ambiguity sets to allow any proper parametrization of the model. Throughout the paper, different types of ambiguity sets will be marked with different diacritics.

$\widehat{V}^{\pi}_{h+1}(s')$ for step $h+1$, we first compute the nominal Q-factor $\boldsymbol{\omega}^{\circ}_h := \boldsymbol{\nu}^{\circ}_h + \sum_{s'} \widehat{V}^{\pi}_{h+1}(s')\boldsymbol{\mu}^{\circ}_h(s')$ using $\widehat{V}^{\pi}_{h+1}(\cdot)$, and then solve the optimization problem

$$\min_{(\boldsymbol{\xi}_h,\boldsymbol{\eta}_h)\in\widehat{\mathcal{M}}_h} \big\langle \mathbb{E}_{(s_h,a_h)\sim d^{\pi}_h}[\boldsymbol{\phi}^{\circ}_h(s_h,a_h)] + \boldsymbol{\eta}_h, \boldsymbol{\omega}^{\circ}_h + \boldsymbol{\xi}_h \big\rangle, \tag{6}$$

which can also be viewed as a certain type of perturbation around the nominal dynamics. Here d^{π}_h still refers to the state-action occupancy measure in the *nominal* model. We say $\widehat{\mathcal{M}}_h$ is a (ξ,η)-*rectangular ambiguity set*, if it is rectangular in terms of $(\boldsymbol{\xi}_h,\boldsymbol{\eta}_h)$ as

$$\widehat{\mathcal{M}}_h := \big\{(\boldsymbol{\xi}_h,\boldsymbol{\eta}_h) \;\big|\; \|\boldsymbol{\xi}_h\| \le R_{\xi,h}, \|\boldsymbol{\eta}_h\| \le R_{\eta,h}\big\}, \tag{7}$$

where $(R_{\xi,h}, R_{\eta,h})$ are the radii of perturbation. Further, we define $\widehat{\mathcal{M}} := \bigotimes_{h\in[H]} \widehat{M}_h$. Note that compared to the (ϕ,μ,ν)-rectangular ambiguity set $\widetilde{\mathcal{M}}_h$ in the naive robust low-rank MDP (defined in (5)), the (ξ,η)-rectangular ambiguity set $\widehat{\mathcal{M}}_h$ contains much fewer components, which contributes to greatly reducing the computational complexity as we will see later.

With the solution $(\boldsymbol{\xi}^*_h,\boldsymbol{\eta}^*_h)$, we proceed to calculate the robust Q-functions under the new robustness concept. Although the optimization problem (6) does not produce individual perturbed features, we may simply perturb each feature $\boldsymbol{\phi}^{\circ}_h(\cdot,\cdot)$ by the same amount $\boldsymbol{\eta}^*_h$ so that the weighted average Q-function is not affected. In this way, the new Bellman update can be written as

$$[\widehat{\mathcal{B}}^{\pi}_h \widehat{V}^{\pi}_{h+1}](s,a) := \langle \boldsymbol{\phi}^{\circ}_h(s,a) + \boldsymbol{\eta}^*_h, \boldsymbol{\omega}^{\circ}_h + \boldsymbol{\xi}^*_h \rangle, \; \forall (s,a) \in \mathcal{S}\times\mathcal{A}. \tag{8}$$

Note that here we explicitly mark the policy behind a Bellman update operator in its superscript, since the policy is implicitly invoked when we solve (6). Finaly, to complete the Bellman update, the low-rank robust Q- and V-functions can be recovered by

$$\widehat{Q}^{\pi}_h(s,a) := [\widehat{\mathcal{B}}_h \widehat{V}^{\pi}_{h+1}](s,a), \quad \widehat{V}^{\pi}_h(s) = \big\langle \pi_h(\cdot|s), \widehat{Q}^{\pi}_h(s,\cdot)\big\rangle. \tag{9}$$

The objective of the robust planning problem is to find the optimal policy that maximizes the robust value at the initial step, namely $\widehat{\pi}^* := \arg\max_{\pi} \widehat{V}^{\pi}_1(\rho)$.

3.2. Rationale of the Proposed Low-rank Robustness

In this section, we provide some additional rationale for the proposed low-rank robustness concept. We first show that the low-rank robust MDP can be interpreted as an *implicit* step-wise pseudo-MDP[3] perturbation around the nominal MDP via $2H$ *effectively equivalent* perturbation vectors $\boldsymbol{\xi}_{1:H}$ and $\boldsymbol{\eta}_{1:H}$. We then proceed to present a few key properties and examples to promote the understanding of the proposed robustness concept.

An Alternative Interpretation of the Proposed Low-rank Robust MDPs. We show that the proposed low-rank robust MDPs can be viewed as a relaxed version of the naive low-rank robust MDPs. For this purpose, define the perturbation vectors

$$\begin{aligned}
\boldsymbol{\delta}_{\phi,h}(s,a) &:= \boldsymbol{\phi}_h(s,a) - \boldsymbol{\phi}^{\circ}_h(s,a), \; \forall (s,a)\in\mathcal{S}\times\mathcal{A};\\
\boldsymbol{\delta}_{\mu,h}(s') &:= \boldsymbol{\mu}_h(s') - \boldsymbol{\mu}^{\circ}_h(s'), \; \forall s'\in\mathcal{S};\\
\boldsymbol{\delta}_{\nu,h} &:= \boldsymbol{\nu}_h - \boldsymbol{\nu}^{\circ}_h,
\end{aligned}$$

so that for any $(s,a)\in\mathcal{S}\times\mathcal{A}$, the naive robust Bellman update (20) can be equivalently rewritten as

$$\min_{\boldsymbol{\delta}_{\phi,h},\boldsymbol{\delta}_{\mu,h},\boldsymbol{\delta}_{\nu,h}} \left\langle \boldsymbol{\phi}^{\circ}_h(s,a) + \boldsymbol{\delta}_{\phi,h}(s,a), \boldsymbol{\nu}^{\circ}_h + \boldsymbol{\delta}_{\nu,h} + \sum_{s'} \widetilde{V}^{\pi}_{h+1}(s')\big(\boldsymbol{\mu}^{\circ}_h(s') + \boldsymbol{\delta}_{\mu,h}(s')\big) \right\rangle. \tag{10}$$

Comparing the forms of (6) and (10), we shall regard

$$\boldsymbol{\xi}_h := \boldsymbol{\delta}_{\nu,h} + \sum_{s'} \widetilde{V}^{\pi}_{h+1}(s')\boldsymbol{\delta}_{\mu,h}(s'), \; \forall h\in[H] \tag{11}$$

3. Pseudo-MDPs are MDP-like processes that allow transition probabilities to lie out of the probability simplex. See Yao et al. (2014) for detailed definitions and properties.

as a collection of all perturbation terms in the second component of the inner product. On the other hand, following the above argument, to formulate an effectively equivalent perturbation of the feature map, we shall focus on the weighted average of (10) over $(s,a) \sim d_h^\pi$ and define

$$\boldsymbol{\eta}_h := \mathbb{E}_{(s,a)\sim d_h^\pi}[\boldsymbol{\delta}_{\phi,h}(s,a)],\ \forall h \in [H]. \tag{12}$$

Plugging $\boldsymbol{\xi}_h$ and $\boldsymbol{\eta}_h$ back, we have recovered the form of (6). This interpretation reveals the influence of duple perturbations over feature maps and factors on the policy evaluation scheme, justifying the idea to consider the "effective" perturbation over $\boldsymbol{\xi}_h$ and $\boldsymbol{\eta}_h$ for the sake of computational efficiency.

Remark 1 (pseudo-MDP perturbation) *As equations (11) and (12) suggest, in theory we may write out a set $\widehat{\mathcal{M}}$ that contains exactly the (ξ,η)-pairs corresponding to valid perturbed MDPs around the nominal model, at the cost of additional constraints. However, for computational efficiency, we choose to relax the ambiguity set by replacing the simplex constraints on $\boldsymbol{\xi}_h$ with a simple ℓ_2-ball constraint. In this way, we are also including perturbations corresponding to pseudo-MDPs.*

Relationship with Nominal Updates and Naive Robust Updates. With the above interpretation, one may wonder how to relate the proposed low-rank robust value $\widehat{V}_h^\pi$ to the naive low-rank robust value $\widetilde{V}_h^\pi$ (defined in (18)) and the standard nominal value V_h^π, assuming that $R_{\xi,h}$ and $R_{\eta,h}$ are selected such that the (ξ,η)-ambiguity set induced by $\widetilde{\mathcal{M}}$ is a subset of $\widehat{\mathcal{M}}$ (see Appendix B.2 for details). It turns out that we can prove the following relationship.

Theorem 2 *Suppose the (ξ,η)-ambiguity set induced by the (ϕ,μ,ν)-rectangular ambiguity set $\widetilde{\mathcal{M}}$ is a subset of the (ξ,η)-rectangular ambiguity set $\widehat{\mathcal{M}}$. Then for any step $h \in [H]$ we have:*

1. *Ordinal relation: $\widehat{Q}_h^\pi(s,a) \le \widetilde{Q}_h^\pi(s,a) \le Q_h^\pi(s,a)$, and $\widehat{V}_h^\pi(s) \le \widetilde{V}_h^\pi(s) \le V_h^\pi(s)$;*
2. *Bounded gap: $\widetilde{V}_h^\pi(s) - \widehat{V}_h^\pi(s) \le V_h^\pi(s) - \widehat{V}_h^\pi(s) \le \sum_{\tau\ge h}\big(2R_{\eta,\tau}\sqrt{d} + (1+R_{\eta,\tau})R_{\xi,\tau}\big)$.*

The above theorem again justifies that the proposed robust policy evaluation scheme can be viewed as a relaxation of the naive robust evaluation scheme (given appropriate radii). Indeed, the proposed scheme evaluates the policy more pessimistically, and is thus expected to yield more conservative policies. We would also like to mention that the above bound seems to be loose in that the gap between V_1^π, $\widetilde{V}_1^\pi$ and $\widehat{V}_1^\pi$ is in the order of $\Theta(H)$. However, examples can be constructed to show that this is actually the best bound we can expect in the worst case.

Details of the proof and the example can be found in Appendix B.3.

4. R²PG: Representation Robust Policy Gradient

With the new robustness concept in hand, now we shall present our algorithm that iteratively solves for the optimal robust policy $\widehat{\pi}^*$ as defined in Section 3.1.

4.1. Algorithm Design

The proposed algorithm, **R**epresentation **R**obust **P**olicy **G**radient (R²PG), can be found in Algorithm 1 below. Overall, it follows the iterative evaluation-improvement protocol of policy iteration methods, except that it uses the proposed low-rank robust policy evaluation scheme defined in Section 3.1. Specifically, R²PG consists of two main components for each step in each iteration:

1. *Robust policy evaluation (line 6-8).* The algorithm follows the effective robust policy evaluation scheme formulated in Section 3.1. Prior to the recursion, we first recursively compute the state occupancy measure $d_h^{\pi^k}$ for the current policy π^k *(line 4)*. Then, at each time step h, it computes the nominal Q-factor $\boldsymbol{\omega}_h^{\circ,k}$ and solves (13) to get the optimal $(\boldsymbol{\xi}_h^*, \boldsymbol{\eta}_h^*)$ *(line 6)*, and then updates the robust Q- and V-functions according to (8) and (9) with identical perturbations $\boldsymbol{\eta}_h^*$ around each nominal feature $\boldsymbol{\phi}_h^\circ(\cdot,\cdot)$ *(line 7-8)*.

2. *Policy improvement (line 9).* To update the policy, we apply the Natural Policy Gradient (NPG) algorithm that is widely-used in literature (Agarwal et al., 2021; Cen et al., 2022; Mei et al., 2021). Given the robust Q-function $\widehat{Q}_h^{\pi^k}$ of policy π^k, the policy update rule is given by $\pi_h^{k+1}(a|s) \propto \pi_h^k(a|s) \cdot \exp\big(\alpha \widehat{Q}_h^{\pi^k}(s,a)\big)$ for some step size $\alpha > 0$. Detailed explanations of the intuitions behind NPG can be found in Appendix D.3.

Algorithm 1 R^{2}PG: **R**epresentation **R**obust **P**olicy **G**radient

1 Initialize $\pi_h^1(\cdot|s) \leftarrow \mathsf{Unif}(\mathcal{A})$
2 **for** $k = 1, 2, \cdots, K$ **do**
3 Initialize $\widehat{V}_{H+1}^k(s) \leftarrow 0$.
4 Compute $d_1^{\pi^k}(s,a) \leftarrow \rho(s)\pi(a|s)$, $d_{h+1}^{\pi^k}(s',a') \leftarrow \sum_{s,a} d_h^{\pi^k}(s,a)\mathbb{P}_h^\circ(s'|s,a)\pi(a'|s')$.
5 **for** $h = H, H-1, \cdots, 1$ **do**
6 Compute $\boldsymbol{\omega}_h^{\circ,k} \leftarrow \boldsymbol{\nu}_h^\circ + \sum_{s'} \widehat{V}_{h+1}^k(s')\boldsymbol{\mu}_h^\circ(s')$, and solve the optimization:
$$(\boldsymbol{\xi}_h^k, \boldsymbol{\eta}_h^k) \leftarrow \mathop{\arg\min}_{\substack{\|\boldsymbol{\xi}_h^k\| \le R_{\xi,h}, \\ \|\boldsymbol{\eta}_h^k\| \le R_{\eta,h}}} \Big\langle \mathbb{E}_{(s,a)\sim d_h^{\pi^k}}[\boldsymbol{\phi}_h^\circ(s,a)] + \boldsymbol{\eta}_h^k, \boldsymbol{\omega}_h^{\circ,k} + \boldsymbol{\xi}_h^k \Big\rangle. \tag{13}$$
7 Perform feature update: $\boldsymbol{\phi}_h^k(s,a) \leftarrow \boldsymbol{\phi}_h^\circ(s,a) + \boldsymbol{\eta}_h^k$, $\boldsymbol{\omega}_h^k \leftarrow \boldsymbol{\omega}_h^{\circ,k} + \boldsymbol{\xi}_h^k$.
8 Update value functions: $\widehat{Q}_h^k(s,a) \leftarrow \langle \boldsymbol{\phi}_h^k(s,a), \boldsymbol{\omega}_h^k \rangle$, $\widehat{V}_h^k(s) \leftarrow \sum_a \pi_h^k(a|s)\widehat{Q}_h^k(s,a)$.
9 Update policy by Natural Policy Gradient: $\pi_h^{k+1}(a|s) \propto \pi_h^k(a|s) \cdot \exp\big(\alpha \widehat{Q}_h^\pi(s,a)\big)$.
10 **return** $\pi^{\mathrm{out}} \sim \mathsf{Unif}(\pi^{1:K})$

Remark 3 (computational considerations) *For efficient implementation, a few computational challenges in Algorithm 1 need to be settled, including solving the non-convex optimization problem (13) (which can be reduced to a constrained semi-definite program (SDP)) and estimating parameters including $d_h^{\pi^k}$ and $\boldsymbol{\omega}_h^{\circ,k}$, details of which can be found in Appendix C.*

5. Theoretical Analysis

In this section, we present the convergence guarantee for our R^{2}PG algorithm. Specifically, we would like to show that the robust value of the output policy π^{out} is close to that of the optimal robust policy $\widehat{\pi}^*$ defined in Section 3.1. To this end, we have the following bound.

Theorem 4 (Convergence) *Under Assumption 2.1, by running R^{2}PG (Algorithm 1) with step size $\alpha = \sqrt{2\log A/(KH^2)}$, the low-rank robust V-function of π^{out} satisfies*

$$\mathbb{E}_{\pi^{\mathrm{out}}}\Big[\widehat{V}_1^*(\rho) - \widehat{V}_1^{\pi^{\mathrm{out}}}(\rho)\Big] \le \sqrt{\frac{2H^4 \log A}{K}} + 2\sum_{h=1}^H R_{\eta,h}\sqrt{d}. \tag{14}$$

Proof sketch. The proof is similar to that in Liu et al. (2023), which consists of three main steps.

Step 1: Quasi-contraction property. The first key observation is that the new robust Bellman operator satisfies a *quasi*-contraction property, which is similar to the contraction of nominal and naive robust Bellman operators (Iyengar, 2005), but also different in that it only holds in expectation over d_h^π and comes with an additional constant bias term due to the misspecification of pseudo-MDPs.

Lemma 5 *For any V-functions $V, V' : \mathcal{S} \to \mathbb{R}$ and any policy π, we have*

$$\mathbb{E}_{(s,a)\sim d_h^\pi}\Big[[\widehat{\mathcal{B}}_h^\pi V](s,a) - [\widehat{\mathcal{B}}_h^\pi V'](s,a)\Big] \le \mathbb{E}_{s'\sim\rho_{h+1}^\pi}\big[V(s') - V'(s')\big] + 2R_{\eta,h}\sqrt{d}. \tag{15}$$

Step 2: Extended performance difference lemma. We proceed to prove an extended performance difference lemma that is similar to its counterpart in Efroni et al. (2020).

Lemma 6 *For any policies π and π', we have*

$$\widehat{V}_1^{\pi}(\rho) - \widehat{V}_1^{\pi'}(\rho) \leq \sum_{h=1}^{H} \mathbb{E}_{\pi,\mathbb{P}^\circ}\left[\left\langle \widehat{Q}_h^{\pi'}(s_h,\cdot), \pi_h(\cdot|s_h) - \pi'_h(\cdot|s_h)\right\rangle + \left\langle [\widehat{\mathcal{B}}_h^{\pi}\widehat{V}_{h+1}^{\pi'}](s_h,\cdot) - \widehat{Q}_h^{\pi'}(s_h,\cdot), \pi_h(\cdot|s_h)\right\rangle\right] + 2\sum_{h=1}^{H} R_{\eta,h}\sqrt{d}. \tag{16}$$

Step 3: Plugging in the upper bounds. To complete the proof, we average (16) over the K policies obtained in K episodes. To bound the first term in the expectation in (16), we plug in the regret bound for NPG (Corollary 16); the second term can be bounded by a technical lemma (Lemma 17) that characterizes the difference between robust Bellman updates with respect to different policies.

Details of the proof can be found in Appendix D.[4]

Remark 7 (Eliminating the constant bias) *We point out that the constant bias term $2\sum_{h=1}^{H} R_{\eta,h}\sqrt{d}$ scaling in the order of $\Theta(H)$ in the worst case is accounted for by the intrinsic misspecification error of pseudo-MDPs from MDPs, as mentioned in Theorem 1, which is common in psuedo-MDP literature (Yao et al., 2014). A potential way to eliminate this constant bias is to add more structure to the (ξ,η)-ambiguity set such that each perturbation corresponds to at least one valid MDP around the nominal model. Appendix E provides some preliminary insights that consider latent-variable-style low-rank structures, though suffering from additional assumptions and computational challenges.*

6. Numerical Simulation

In this section, we study the robust behavior of policies induced by solving the proposed duple perturbation low-rank robustness concept via numerical simulations. We first show a direct implementation of the R^{2}PG algorithm in a toy model, and then move on to a modified implementation to handle a series of classical control tasks with continuous state-action spaces.

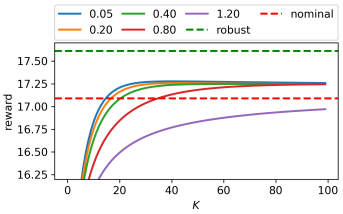

Figure 1: Robust evaluation of policies in a toy model.

6.1. A Toy Model

We first show the numerical performance of our R^{2}PG algorithm via simulations in a 4-state toy model, where the states are manually designed to represent different "uncertainty levels". Details of the setup are deferred to Appendix F.1.

The proposed R^{2}PG algorithm is run with different perturbation radii $R_{\eta,h} = 0.01$ and $R_{\xi,h} \in \{0.05, 0.2, 0.4, 0.8, 1.2\}$, and the policies obtained in all the episodes are evaluated by the *minimum* cumulative reward in several perturbed MDPs, the results of which are plotted in Figure 1. It can be observed that policies converge in all executions, and a larger perturbation radius generally leads to more conservative behavior. This phenomenon is largely expected in that, as perturbation radius increases, the misspecification error induced by the worst-case pseudo-MDP also increases, which leads to an intrinsically pessimistic estimation of policy values. Nevertheless, we point out the output policies still perform better than the nominal optimal policy when the MDP is appropriately perturbed, again highlighting the need for robustness in environments with uncertainty.

Analyzing the output policies in details, we shall further find that, as time elapses, all output policies gradually lean towards the safe state by increasing the transition probabilities to it. However,

4. Proofs and additional theoretical results can be found at https://arxiv.org/abs/2404.08089.

since the algorithm is designed to optimize over the average performance for policy evaluation, it is also reasonable that it does not fully converge to a policy that yields optimal worst-case performance.

6.2. Continuous Control Tasks

In this section, we show the numerical performance of our robust policy evaluation scheme in a series of classical continuous control tasks, including Pendulum (Brockman et al., 2016), CartPendulum (Brockman et al., 2016), Pendubot (Fantoni et al., 2000), and Quadrotor drone control (Hoffmann et al., 2007). To handle planning in MDPs with continuous state-action spaces, we adapt the Spectral Dynamics Embedding Control (SDEC) algorithm (Ren et al., 2023) to incorporate our robust policy evaluation scheme. We also use the truncated Gaussian kernel representation proposed in (Ren et al., 2023) as the nominal model, avoiding the complication of representation learning. Details of the setup and the implementation are deferred to Appendix F.2.

We compare the performance of Robust-SDEC-R (which incorporates the robust policy evaluation scheme with perturbation radius $R_{\eta,h} = R_{\xi,h} = R$) against the SDEC baseline (the vanilla algorithm without robust policy evaluation). The output policies are evaluated with respect to perturbed physical parameters, as shown in Figure 2. It can be observed that, in presence of perturbed physical parameters, the performance of Robust-SDEC suffers much less degradation as compared to the baseline SDEC, and the advantage is more significant with larger perturbations. These numerical results justify the effectiveness of the duple perturbation low-rank robustness concept.

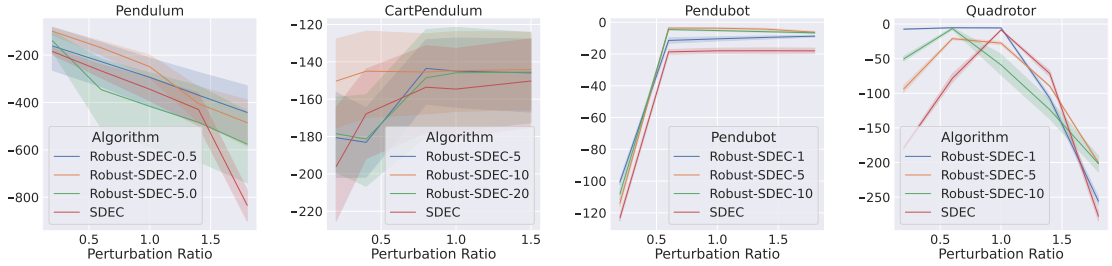

Figure 2: Evaluation results against different perturbation ratios of the physical parameters. *(curve: average cumulative reward; shade: 95% confidence region; both over 200 evaluation episodes)*

7. Conclusion

In this paper, we propose a novel low-rank robust MDP with duple perturbations, which efficiently and effectively achieves robust behavior in low-rank MDPs. The new robustness concept features computational efficiency, scalability and compatability with low-rank representation structure. Based on the new robustness concept, we design an algorithm (R^{2}PG) to solve the proposed robust low-rank MDP that provably converges to the optimal robust policy with bounded suboptimality gap.

Despite the theoretical results and empirical success, algorithms solving the proposed low-rank robust MDPs generally suffer from constant performance gaps induced by the pseudo-MDP relaxation and the nominal occupancy measure approximation, while the initial attempts to resolve these issues result in new computational challenges. In addition, sample-based learning methods remain challenging for robust low-rank MDP. Therefore, future work along this line includes designing algorithms that solve for the robust policy in an asymptotically accurate and/or more computationally efficient way, incorporating sample-based methods to estimate the nominal MDP and use the estimated MDP to generate robust policies, and further, discovering other robustness concepts that are compatible with different low-rankness concepts for more scalable robust RL.

Acknowledgments

This paper is supported by NSF AI Institute: 2112085, NSF CNS: 2003111, NSF ECCS: 2328241, ONR: N000142512173; NSF ECCS: 2401391, NSF IIS: 2403240.

References

Alekh Agarwal, Sham Kakade, Akshay Krishnamurthy, and Wen Sun. Flambe: Structural complexity and representation learning of low rank MDPs. *Advances in neural information processing systems*, 33:20095–20107, 2020.

Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *The Journal of Machine Learning Research*, 22(1):4431–4506, 2021.

Wenbao Ai and Shuzhong Zhang. Strong duality for the CDT subproblem: a necessary and sufficient condition. *SIAM Journal on Optimization*, 19(4):1735–1756, 2009.

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

Amir Beck and Yonina C Eldar. Strong duality in nonconvex quadratic optimization with two quadratic constraints. *SIAM Journal on optimization*, 17(3):844–860, 2006.

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. *arXiv preprint arXiv:2305.09659*, 2023.

Steven J Bradtke and Andrew G Barto. Linear least-squares algorithms for temporal difference learning. *Machine learning*, 22:33–57, 1996.

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

Shicong Cen, Chen Cheng, Yuxin Chen, Yuting Wei, and Yuejie Chi. Fast global convergence of natural policy gradient methods with entropy regularization. *Operations Research*, 70(4): 2563–2578, 2022.

Sheng Cheng and Nuno C. Martins. An optimality gap test for a semidefinite relaxation of a quadratic program with two quadratic constraints. *SIAM Journal on Optimization*, 31(1):866–886, January 2021.

Murtaza Dalal, Deepak Pathak, and Russ R Salakhutdinov. Accelerating robotic reinforcement learning via parameterized action primitives. *Advances in Neural Information Processing Systems*, 34:21847–21859, 2021.

Christoph Dann, Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E Schapire. On oracle-efficient PAC RL with rich observations. *Advances in neural information processing systems*, 31, 2018.

A Devinatz. Integral representations of positive definite functions. *Transactions of the American Mathematical Society*, 74(1):56–77, 1953.

Jing Dong, Jingwei Li, Baoxiang Wang, and Jingzhao Zhang. Online policy optimization for robust MDP. *arXiv preprint arXiv:2209.13841*, 2022.

Simon Du, Akshay Krishnamurthy, Nan Jiang, Alekh Agarwal, Miroslav Dudik, and John Langford. Provably efficient RL with rich observations via latent state decoding. In *International Conference on Machine Learning*, pages 1665–1674. PMLR, 2019.

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

Yonathan Efroni, Lior Shani, Aviv Rosenberg, and Shie Mannor. Optimistic policy optimization with bandit feedback. *arXiv preprint arXiv: 2002.08243*, 2020.

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

Isabelle Fantoni, Rogelio Lozano, and Mark W Spong. Energy based control of the pendubot. *IEEE transactions on automatic control*, 45(4):725–729, 2000.

Bernd Gärtner and Jiri Matousek. *Approximation algorithms and semidefinite programming*. Springer Science & Business Media, 2012.

Vineet Goyal and Julien Grand-Clement. Robust Markov decision processes: Beyond rectangularity. *Mathematics of Operations Research*, 48(1):203–226, 2023.

Elad Hazan. Sparse approximate solutions to semidefinite programs. In *Latin American symposium on theoretical informatics*, pages 306–316. Springer, 2008.

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

Chin Pang Ho, Marek Petrik, and Wolfram Wiesemann. Partial policy iteration for ℓ_1-robust Markov decision processes. *The Journal of Machine Learning Research*, 22(1):12612–12657, 2021.

Gabriel Hoffmann, Haomiao Huang, Steven Waslander, and Claire Tomlin. Quadrotor helicopter flight dynamics and control: Theory and experiment. In *AIAA guidance, navigation and control conference and exhibit*, page 6461, 2007.

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

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

Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. End-to-end training of deep visuomotor policies. *The Journal of Machine Learning Research*, 17(1):1334–1373, 2016.

Yunfan Li and Lin Yang. On the model-misspecification in reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pages 2764–2772. PMLR, 2024.

Qinghua Liu, Gellért Weisz, András György, Chi Jin, and Csaba Szepesvári. Optimistic natural policy gradient: A simple efficient policy optimization framework for online RL. *arXiv preprint arXiv: 2305.11032*, 2023.

Xiaoteng Ma, Zhipeng Liang, Jose Blanchet, Mingwen Liu, Li Xia, Jiheng Zhang, Qianchuan Zhao, and Zhengyuan Zhou. Distributionally robust offline reinforcement learning with linear function approximation. *arXiv preprint arXiv: 2209.06620*, 2023.

Jincheng Mei, Yue Gao, Bo Dai, Csaba Szepesvari, and Dale Schuurmans. Leveraging non-uniformity in first-order non-convex optimization. In *International Conference on Machine Learning*, pages 7555–7564. PMLR, 2021.

Kohei Miyaguchi. Asymptotically exact error characterization of offline policy evaluation with misspecified linear models. *Advances in Neural Information Processing Systems*, 34:28573–28584, 2021.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

Francesco Orabona. A modern introduction to online learning. *arXiv preprint arXiv:1912.13213*, 2019.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744, 2022.

Kishan Panaganti, Zaiyan Xu, Dileep Kalathil, and Mohammad Ghavamzadeh. Robust reinforcement learning using offline data. *Advances in neural information processing systems*, 35:32211–32224, 2022.

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

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

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

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

Tongzheng Ren, Chenjun Xiao, Tianjun Zhang, Na Li, Zhaoran Wang, Sujay Sanghavi, Dale Schuurmans, and Bo Dai. Latent variable representation for reinforcement learning. *arXiv preprint arXiv:2212.08765*, 2022a.

Tongzheng Ren, Tianjun Zhang, Lisa Lee, Joseph E Gonzalez, Dale Schuurmans, and Bo Dai. Spectral decomposition representation for reinforcement learning. *arXiv preprint arXiv:2208.09515*, 2022b.

Tongzheng Ren, Tianjun Zhang, Csaba Szepesvári, and Bo Dai. A free lunch from the noise: Provable and practical exploration for representation learning. In *Uncertainty in Artificial Intelligence*, pages 1686–1696. PMLR, 2022c.

Tongzheng Ren, Zhaolin Ren, Na Li, and Bo Dai. Stochastic nonlinear control via finite-dimensional spectral dynamic embedding. *arXiv preprint arXiv:2304.03907*, 2023.

Erica Salvato, Gianfranco Fenu, Eric Medvet, and Felice Andrea Pellegrino. Crossing the reality gap: A survey on sim-to-real transferability of robot controllers in reinforcement learning. *IEEE Access*, 9:153171–153187, 2021.

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

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *nature*, 550(7676):354–359, 2017.

Edward Allan Silver. *arkovian decision processes with uncertain transition probabilities or rewards*. PhD thesis, Massachusetts Institute of Technology, Dept. of Civil Engineering, 1963.

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*, pages 181–189. PMLR, 2014.

Masatoshi Uehara, Xuezhou Zhang, and Wen Sun. Representation learning for online and offline RL in low-rank MDPs. *arXiv preprint arXiv:2110.04652*, 2021.

Daniel Vial, Advait Parulekar, Sanjay Shakkottai, and R Srikant. Improved algorithms for misspecified linear Markov decision processes. In *International Conference on Artificial Intelligence and Statistics*, pages 4723–4746. PMLR, 2022.

Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *Nature*, 575(7782):350–354, 2019.

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

Lin Yang and Mengdi Wang. Reinforcement learning in feature space: Matrix bandit, kernels, and regret bound. In *International Conference on Machine Learning*, pages 10746–10756. PMLR, 2020.

Wenhao Yang, Liangyu Zhang, and Zhihua Zhang. Toward theoretical understandings of robust Markov decision processes: Sample complexity and asymptotics. *The Annals of Statistics*, 50(6): 3223–3248, 2022.

Hengshuai Yao, Csaba Szepesvári, Bernardo Avila Pires, and Xinhua Zhang. Pseudo-MDPs and factored linear action models. In *2014 IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning (ADPRL)*, pages 1–9. IEEE, 2014.

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.

Tianjun Zhang, Tongzheng Ren, Mengjiao Yang, Joseph Gonzalez, Dale Schuurmans, and Bo Dai. Making linear MDPs practical via contrastive representation learning. In *International Conference on Machine Learning*, pages 26447–26466. PMLR, 2022.

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*, pages 3331–3339. PMLR, 2021.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.