Robust Markov Decision Processes on Continuous State Spaces

Mengmeng Li	Yifan Hu	Daniel Kuhn	Yan Li
EPFL	Rutgers University	EPFL	Texas A&M University
mengmeng.li@epfl.ch	yifan.hu@rutgers.edu	daniel.kuhn@epfl.ch	yan.li@tamu.edu

Abstract

We study infinite-horizon robust Markov decision processes (MDPs) on continuous state spaces with structured rectangular ambiguity set. The proposed ambiguity set falls within the convex hull of unknown generating kernels. We utilize the dynamic formulation of the corresponding robust MDPs, and subsequently introduce a stochastic first-order method for robust policy evaluation. We establish its high probability convergence to the robust value function, which in turn leads to an $\widetilde{\mathcal{O}}(1/\epsilon^2)$ sample complexity. This high probability accuracy certificate is then used in an approximate policy iteration method that finds an ϵ-optimal policy with $\widetilde{\mathcal{O}}(1/\epsilon^2)$ samples. Notably the proposed methods can be implemented independent of the size of the state space. The obtained sample complexities for both robust policy evaluation and optimization appear to be new for robust MDPs with continuous state spaces. Of independent interest, the proposed method is also directly applicable to zero-sum Markov games, which seems to strictly improve the existing sample complexities for continuous state spaces.

1 Introduction

We consider an infinite-horizon robust Markov decision process (MDP), denoted by $(\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \gamma)$, comprising a Borel measurable state space $\mathcal{S} \subseteq \mathbb{R}^S$, a finite action space $\mathcal{A}$, a continuous reward-per-stage function $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, and an ambiguity set $\mathcal{P}$ of transition kernels $P : \mathcal{S} \times \mathcal{A} \to \mathcal{P}(\mathcal{S})$, where $\mathcal{P}(\mathcal{S})$ denotes the set of probability measures over $\mathcal{S}$, and a discount factor $\gamma \in (0, 1)$. We assume that $r(s, a) \in [0, 1]$ for all $(s, a) \in \mathcal{S} \times \mathcal{A}$. Let Π denote the set of all randomized stationary policies $\pi : \mathcal{S} \to \mathcal{P}(\mathcal{A})$. For any $\pi \in \Pi$ and $P \in \mathcal{P}$, the value function $V_\pi^P : \mathcal{S} \to \mathbb{R}$ is defined through

$$V_\pi^P(s) = \mathbb{E}_\pi^P \left[\sum_{t=0}^{\infty} \gamma^t r(S_t, A_t) \mid S_0 = s \right], \tag{1}$$

where $\{(S_t, A_t)\}_{t \geq 0}$ is the Markov chain with its law induced by π and P given $S_0 = s$ (*cf.* Ionescu Tulcea Theorem [51]), and $\mathbb{E}_\pi^P$ denotes the corresponding expectation. The robust value function is then defined as

$$V_\pi(s) = \min_{P \in \mathcal{P}} V_\pi^P(s). \tag{2}$$

We are interested in finding the optimal policy π^* of

$$\max_{\pi \in \Pi} V_\pi(s), \tag{3}$$

for a given initial state $s \in \mathcal{S}$.[1] Formulation (3) above allows the decision maker to acknowledge potentially imprecise knowledge or possible deviation of the transition dynamics in MDPs, and instead seeks a robust policy that would perform reasonably well for any kernel $P \in \mathcal{P}$. The ambiguity set $\mathcal{P}$ is typically constructed to include the set of plausible transition dynamics that conform closely with historical data or expert knowledge. For instance, it can be constructed from pre-collected real-world trajectories [26, 56], or by varying key simulation parameters of digital simulators [8, 13, 50]. Of course, (3) reduces to the non-robust MDP when $\mathcal{P}$ is a singleton.

[1] As will be clarified in Section 2, for the ambiguity set $\mathcal{P}$ considered in this manuscript, there exists a policy π^* that is optimal for (3) with any initial state $s \in \mathcal{S}$.

Similar to non-robust MDPs, the solution methods of robust MDPs largely hinge upon the existence of dynamic programming equations characterizing the value functions. In particular, the existence of dynamic programming equations has been established for various classes of so-called *rectangular* ambiguity sets [10, 14, 23, 36, 56]. This, in particular, certifies the existence of optimal stationary Markovian policies, and gives tractable computational schemes for convex $\mathcal{P}$ based on value iterations. The importance of rectangularity is further emphasized in [56], which shows that evaluating V_π (defined in (2)) for a fixed policy π becomes NP-hard for non-s-rectangular ambiguity set even for finite state spaces. It is worth noting here that in (2) the kernel is selected independent of process history (*i.e.*, in a static manner). Recently, [28] propose a dynamic counterpart of (3) between the controller and nature without requiring rectangularity of $\mathcal{P}$. It is shown therein that (3) with all existing rectangular ambiguity sets is indeed equivalent to the dynamic form. In addition, for any potentially non-rectangular ambiguity set $\mathcal{P}$, whenever dynamic programming equation exists for (3), then there must exist a counterpart of (3) with the s-rectangularized enlargement $\mathcal{P}$ that has the same value function and optimal policy.

In view of the dynamic programming equations, when $\mathcal{P}$ is given a priori, or its close approximation can be constructed and efficiently stored inside the computational memory, then (3) can be solved by classical dynamic programming methods, including (approximate) policy and value iterations [10, 11, 14, 21, 36, 37, 43, 45, 56, 58]. We refer to such solution techniques as model-based methods. On the other hand, when $\mathcal{P}$ or its approximation becomes prohibitive in terms of memory budget, one has to resort to a model-free approach, in the sense that one can avoid explicitly constructing or estimating $\mathcal{P}$ for solving (3). This, for instance, includes Q-learning based method using dual representation of distributionally robust counterpart of Bellman operators [32, 38, 54], or policy gradient methods [19, 29, 53, 55, 64] equipped with robust temporal difference learning for robust policy evaluation. Note that solution methods for (3) without dynamic programming equations have also been recently studied in [20, 25, 31].

Despite the recent progress on robust MDPs, it should be noted that the aforementioned methods come with performance guarantees only for finite state spaces.[2] This can be largely attributed to two main reasons. First, the ambiguity set $\mathcal{P}$ becomes infinite dimensional for continuous state space, which prohibits the application of model-based method that requires saving $\mathcal{P}$ into the memory. Second, existing model-free methods require evaluation of robust Bellman operator for every state, which becomes clearly infeasible for continuous state spaces. It is worth further highlighting here that two approximate computation schemes of Bellman operator commonly used for non-robust MDPs, namely, regression-based methods and fixed-point iterations in parameterized function space, can easily break down for robust MDPs. In particular, regression-based approaches that minimize the robust Bellman residual need to consider a non-convex objective due to nonlinearity of robust Bellman operator [27]. On the other hand, fixed-point methods lose the contraction property associated with Bellman operator, if the fixed-point iteration is composed with $\mathcal{L}^2$-projection onto the parameterized function space.[3] In such cases these methods can exhibit divergence behavior unless with restrictive assumptions on $\mathcal{P}$ [48]. To the best of our knowledge, there seems to be no algorithm that can achieve global convergence for robust MDPs on continuous state spaces without restrictive assumptions.

In this paper, we show that by exploiting structural properties of $\mathcal{P}$, one can design globally convergent computational schemes for solving (3), even if the state space $\mathcal{S}$ is continuous. In particular, we are interested in ambiguity sets $\mathcal{P}$ generated dynamically by taking the convex-hull of K generating transition kernels at every state. The considered modelling choice of $\mathcal{P}$ is particularly convenient if the transition kernel $P \in \mathcal{P}$, despite being supported over continuous state space (*i.e.*, infinite-dimensional), characterizes dynamics that is inherently controlled by finite-dimensional parameters. This notably covers a variety of physics-based (stochastic) differential equations, and digital simulators used in supply chain management, portfolio optimization, and robotics [8, 9, 13, 30, 50]. We will discuss some potential constructions of $\mathcal{P}$ depending on the dimension of the parameter space. For the considered ambiguity set $\mathcal{P}$, we propose a model-free method that operates in continuous state spaces while attaining the optimal sample complexity of $\widetilde{\mathcal{O}}(1/\epsilon^2)$ for both robust policy evaluation (2) and policy optimization (3). Notably, the proposed method only requires sample access to generating kernels, without requiring knowing or saving $\mathcal{P}$ a priori, and can be implemented in a computational/memory budget that is independent of the size of the state space. Our

[2] It should be noted here that in [64], the ambiguity set $\mathcal{P}$ is defined over continuous state spaces, but the global convergence of the proposed method requires a small radius of $\mathcal{O}(1-\gamma)$ for $\mathcal{P}$.

[3] Unless there is no approximation error for the value function. This is often referred to as the Bellman completeness condition (see, *e.g.*, [33]).

contributions can be summarized as follows.

First, for robust policy evaluation, we propose a stochastic first-order algorithm that can estimate the robust value function (2) up to ϵ accuracy with an optimal $\mathcal{O}(1/\epsilon^2)$ number of samples drawn from the generating kernels. In contrast to existing approaches for continuous state robust MDPs, the proposed method does not require any restrictive assumptions on $\mathcal{P}$ [48, Assumption 2] or its radius [64]. Notably the accuracy certificate is stated in a high probability sense. *En route*, we also establish the high probability convergence of temporal difference learning over unbounded domain, which might be of independent interest.

Second, for robust policy optimization, we adopt the dynamic viewpoint of formulation (3) (see, *e.g.*, [28]), and present an approximate version of policy iteration method for zero-sum Markov game over continuous state spaces. Notably the presented policy iteration method takes the previously discussed robust policy evaluation method as a subroutine, and can be implemented efficiently for continuous state spaces. In particular, the high probability accuracy certificate of robust policy evaluation appears to be essential for the global convergence for a class of policy iteration-based methods. We further establish an $\widetilde{\mathcal{O}}(1/\epsilon^2)$ sample complexity for the approximate policy iteration method for finding an ϵ-optimal policy of (3).

Both the robust policy evaluation and robust policy improvement's sample and computational complexities remain independent of the size of the state space. In addition, the obtained sample complexities of $\widetilde{O}(1/\epsilon^2)$ for policy evaluation and optimization are order-wise optimal. To the best of our knowledge, this seems to be the first computationally feasible method for solving continuous state space robust MDPs, with optimal global performance guarantees. The proposed framework also directly applies to infinite-horizon zero-sum Markov game, and the obtained $\widetilde{\mathcal{O}}(1/\epsilon^2)$ sample complexity seems to strictly improve the existing development for continuous state spaces [63].

The rest of the paper is organized as follows. Section 2 introduces the mixture ambiguity set considered in this manuscript and establishes its key structural properties. Section 3 presents an algorithmic framework for robust policy evaluation and analyzes its iteration and sample complexities. Section 4 then presents a method for robust policy optimization. Finally, we conclude in Section 5 and discuss possible future directions.

Notation. For any measurable set X, we use $\mathcal{P}(X)$ to denote the set of probability measures defined over X, and use $\mathcal{L}^2(X,\nu)$ to denote the $\mathcal{L}^2$ space associated with $\nu \in \mathcal{P}(X)$. When set X is finite, we use Δ_X to denote the probability simplex defined over X. For any $n > 0$, e_i denotes the i-th standard basis vector in $\mathbb{R}^n$. Unless stated otherwise, we reserve $\|\cdot\|$ for the Euclidean norm. Finally, we denote $[k] = \{1, \ldots, k\}$ for any $k > 0$.

2 Robust MDP with Mixture Ambiguity Set

Throughout the rest of our discussions, we focus on the following structured ambiguity set $\mathcal{P}$, which we term as the mixture ambiguity set.

Definition 1 (Mixture Ambiguity Set)**.** *Given a finite set of transition kernels $\{P_k\}_{k\in\mathcal{K}}$ where $\mathcal{K} = [K]$, and a set $W \subseteq \Delta_{\mathcal{K}}$, the mixture ambiguity set $\mathcal{P}$ is defined by*

$$\mathcal{P} = \{P : \mathcal{S} \times \mathcal{A} \to \mathcal{P}(\mathcal{S}) \mid P(\cdot|s,\cdot) \in \mathcal{P}_s,\ \forall s \in \mathcal{S}\},$$

where the statewise marginal ambiguity set $\mathcal{P}_s$ is defined as

$$\mathcal{P}_s = \left\{ \sum_{k\in\mathcal{K}} v_k P_k(\cdot|s,\cdot) \mid v \in W \right\}. \tag{4}$$

We refer to $\{P_k\}_{k\in\mathcal{K}}$ as the generating kernels of $\mathcal{P}$.[4]

Definition 1 states that the transition probability of any kernel $P \in \mathcal{P}$ from any state s falls within the convex hull of the transition probabilities corresponding to generating kernels. From Definition 1, it is also

[4] In this manuscript, we assume that for any $k \in \mathcal{K}$, the conditional probability $P_k(\cdot|s,a)$ is continuous in the weak* topology with respect to s for any $a \in \mathcal{A}$.

clear that $\mathcal{P}$ is s-rectangular, and consequently the optimal policy π^* of (3) does not depend on the initial state $s \in \mathcal{S}$ [23, 56].

Mixture ambiguity sets can be convenient for modeling continuous-state MDPs where the transition kernel, despite being infinite dimensional, corresponds to physical systems that are determined by finite-dimensional parameters. Notably, this subsumes physics-based stochastic (partial) differential equations that are typically governed by a finite set of parameters with physical meaning (*e.g.*, diffusion and damping coefficients [9]). It also includes a variety of digital simulators used in robotics [8, 50], portfolio optimization [3, 13], and supply chain management [13]. For instance, in multi-echelon inventory control [13], the key simulation parameters governing the transition dynamics only involve the Poisson rate of exogenous random demand. To facilitate our discussion, let us denote parameter space $\Theta \subseteq \mathbb{R}^d$, and let P_θ be the transition kernel corresponding to parameter $\theta \in \Theta$.

Construction of mixture ambiguity set. We discuss here some potential constructions of $\mathcal{P}$ depending on the dimension d of the parameter space Θ.

- For small values of d, one can simply discretize the parameter space into $\widehat{\Theta} = \{\theta_k\}_{k\in\mathcal{K}}$, and correspondingly let $P_k = P_{\theta_k}$ and $W = \Delta_{\mathcal{K}}$. For instance, this includes the aforementioned multi-echelon inventory control [13], where $\Theta \subset \mathbb{R}$ corresponds to the possible Poisson rate of exogenous random demand.
- For large values of d, one can proceed as follows. Let $\mu_\Theta \in \mathcal{P}(\Theta)$ denote a continuous nominal distribution over *parameter space* Θ reflecting plausible parameter configuration. There are some natural scenarios for which μ_Θ can be constructed. If no prior information on θ is available, it is natural to consider $\mu_\Theta = \mathrm{Unif}(\Theta)$. This, for instance, is adopted in domain randomization for robotic simulators with many physical parameters [40, 49]. One can also adopt a Bayesian viewpoint that helps encode data-driven information on θ. In particular, let ν_Θ be a generic prior distribution of θ and ξ be a set of pre-collected trajectories sampled from the Markov process governed by P_θ, one can then set $\mu_\Theta = \nu_{\Theta|\xi}$, where the latter denotes the posterior distribution of θ given ξ.

  Given μ_Θ representing plausible parameter configurations, consider the mixture ambiguity set $\mathcal{P}$ generated by mixing $\{P_\theta : \theta \in \Theta\}$ with distribution μ close to μ_Θ,

$$\mathcal{P}_s^c = \{\mathbb{E}_{\theta\sim\mu}\left[P_\theta(\cdot|s,\cdot)\right] :\ D_\phi(\mu,\mu_\Theta) \le \tau,\ \mu \in \mathcal{P}(\Theta)\},$$

  where $D_\phi(\mu,\mu')$ denotes the ϕ-divergence between μ and μ'. From dual representation of ϕ-divergence [44], it can be readily shown that robust MDP (3) with ambiguity set $\mathcal{P} = \prod_{s\in\mathcal{S}} \mathcal{P}_s^c$ can be efficiently approximated by taking $\{\theta_k\}_{k\in\mathcal{K}}$ generated i.i.d. from μ_Θ, and subsequently forming a counterpart of (3) with ambiguity set $\mathcal{P} = \prod_{s\in\mathcal{S}} \mathcal{P}_s$, where

$$\mathcal{P}_s = \left\{\sum_{k\in\mathcal{K}} v_k P_{\theta_k}(\cdot|s,\cdot) :\ D_\phi(v,\widehat{\mu}_{\mathcal{K}}) \le \tau,\ v \in \Delta_{\mathcal{K}}\right\}, \tag{5}$$

  and $\widehat{\mu}_{\mathcal{K}}$ denotes the empirical distribution over $\{\theta_k\}_{k\in\mathcal{K}}$. Clearly, $\mathcal{P}_s$ defined in (5) takes the form of mixture ambiguity set defined in (4). In particular, it should be noted that in this case the required K typically depends only polynomially on d.

It could be worth mentioning here some modeling advantages of (4) for robust MDPs. First, by taking the convex hull of generating kernels $\{P_k\}_{k\in\mathcal{K}}$, the size of the ambiguity set $\mathcal{P}$ (and subsequently the robustness) naturally depends on the similarity of generating kernels. This contrasts with existing modeling approaches for robust MDPs, which consider a probability ball centered at some unknown nominal kernel, while the radius of the ball is often set by an trial-and-error manner or requires substantial domain expertise.[5]

Second, it is known that a major modeling benefit of s-rectangular set is to allow coupling of probability distributions $\{P(\cdot|s,a)\}_{a\in\mathcal{A}}$ across different actions through proper construction of marginal ambiguity

[5] A notable exception is the case where the nominal kernel is constructed by samples drawing from an unknown data-generating fixed kernel, and hence is known to the decision maker beforehand. In such cases, one can often set the radius to be $\mathcal{O}(1/\sqrt{N})$ to guarantee out-of-sample performance of the corresponding robust policy, where N denotes the number of samples available [56]. It should be noted that this approach can be limited to finite-state robust MDPs due to its model-based nature.

set $\mathcal{P}_s$, defined in (4). However, unless for exceptional cases where substantial knowledge on the underlying dynamics is available for constructing such coupling explicitly, one should take caution in introducing hard-encoded artificial coupling among actions. A prior approach for introducing somewhat natural coupling among actions is by constructing the marginal ambiguity set $\mathcal{P}_s$ through statistical techniques, such as the super-level set of maximum likelihood estimation [56]. Unfortunately, this approach requires the transition kernel $P \in \mathcal{P}$ to be simple enough in the sense that the analytical form of $P(\cdot|s,a)$ is available. In comparison, the proposed mixture ambiguity set (4) allows the coupling among actions to be naturally induced by the generating kernels $\{P_k\}_{k\in\mathcal{K}}$, without manually introducing any artificial coupling among actions. At the same time, the modeling approach here can be applied to arbitrarily complex generating kernels without requiring their analytical form.

2.1 Dynamic Game

We now discuss an equivalent dynamic game formulation of (3) with the considered mixture ambiguity set (4). In particular, in the ensuing sections we will utilize the alternative formulation to derive efficient solution techniques for robust policy evaluation (2) and policy optimization (3).

Consider a two-player dynamic game in which the controller adopts a randomized policy $\pi : \mathcal{S} \to \Delta_{\mathcal{A}}$, while nature adopts a non-randomized policy $\omega : \mathcal{S} \to W$. Let us denote by $\mathcal{W}$ the set of all feasible w of nature. At any state $s \in \mathcal{S}$, The controller chooses $a \in \mathcal{A}$ and nature chooses $v \in \mathcal{W}$ simultaneously and without observing each other's action. The next state s' follows distribution given by $\sum_{k\in\mathcal{K}} v_k P_k(\cdot|s,a)$, and the immediate reward (resp. cost) for the controller (resp. nature) is given by $r(s,a)$. For any given nature's policy $\omega \in \mathcal{W}$, consider $P^\omega \in \mathcal{P}$ defined as

$$P^\omega(\cdot|s,\cdot) = \sum_{k\in\mathcal{K}} \omega(k|s) P_k(\cdot|s,\cdot) \quad \forall s \in \mathcal{S}.$$

Correspondingly, let us denote V_π^ω in short for $V_\pi^{P^\omega}$ defined in (1).

We make the following immediate observations regarding the constructed dynamic game. First, the value of the game for a joint policy (π,ω) is given by V_π^ω. Second, fixing the controller's policy π, one can view the environment of nature as a standalone cost-minimizing MDP, where the value function associated with nature's policy $\omega \in \mathcal{W}$ is given by V_π^ω. It then follows from (2) that the robust value function V_π coincides with the optimal value function of nature's cost-minimizing MDP [28]. Hence from the previous observations and [28], the robust MDP problem (3) can be reformulated into the following max-min dynamic game[6]

$$\max_{\pi\in\Pi} \min_{\omega\in\mathcal{W}} V_\pi^\omega(s). \tag{6}$$

We proceed to introduce some structural properties of nature's cost-minimizing MDP, which will prove useful in our ensuing algorithmic development.

Definition 2. *For any joint policy $(\pi,\omega) \in \Pi \times \mathcal{W}$, define joint action-value function through*

$$Q_\pi^\omega(s,a,k) = r(s,a) + \gamma \int_{\mathcal{S}} V_\pi^\omega(s') P_k(\mathrm{d}s'|s,a) \quad \forall (s,a,k) \in \mathcal{S}\times\mathcal{A}\times\mathcal{K}.$$

Nature's action-value function associated with its policy $\omega \in \mathcal{W}$ is then defined by

$$H_\pi^\omega(s,v) = \sum_{a\in\mathcal{A}} \sum_{k\in\mathcal{K}} v_k \pi(a|s) Q_\pi^w(s,a,k) \quad \forall (s,v) \in \mathcal{S}\times W.$$

Let $\mathbb{P}_\pi^\omega(\cdot|S_0 = s_0)$ denote the probability law of the Markov chain $\{(S_t,A_t)\}_{t\geq 0}$ induced by π and P^ω given initial state $S_0 = s_0$ (Ionescu Tulcea Theorem [51]). We define the discounted visitation measure induced by joint policy $(\pi,\omega) \in \Pi\times\mathcal{W}$ through

$$d_\pi^\omega(\mathcal{E}|s_0) = (1-\gamma)\sum_{t=0}^{\infty} \gamma^t \mathbb{P}_\pi^\omega\left(S_t \in \mathcal{E}|S_0 = s_0\right), \tag{7}$$

[6] It should be noted that the s-rectangularity of $\mathcal{P}$ is essential in the construction of the dynamic game (*cf.* [28]). For general $\mathcal{P}$, (3) differs from the above constructed dynamic game (6).

for any Borel set $\mathcal{E} \in \mathcal{S}$. The following lemma then characterizes the difference of nature's value functions for any pair of its policies.

Lemma 2.1. *For any nature's policies $\omega, \omega' \in \mathcal{W}$, and any $s \in \mathcal{S}$, we have*

$$V_\pi^{\omega'}(s) - V_\pi^{\omega}(s) = \frac{1}{1-\gamma} \int_{\mathcal{S}} [H_\pi^{\omega}(s', \omega'(\cdot|s')) - H_\pi^{\omega}(s', \omega(\cdot|s'))] \, d_\pi^{\omega'}(\mathrm{d}s'|s). \tag{8}$$

In addition, let

$$\psi_\pi^{\omega}(s, v) = \gamma \int V_\pi^{\omega}(s')(P_\pi^{v} - P_\pi^{\omega(s)})(\mathrm{d}s'|s) \quad \forall v \in W. \tag{9}$$

We then have

$$V_\pi^{\omega'}(s) - V_\pi^{\omega}(s) = \frac{1}{1-\gamma} \int_{\mathcal{S}} \psi^{w}(s', \omega'(\cdot|s')) d_\pi^{\omega'}(\mathrm{d}s'|s). \tag{10}$$

Proof. The proof of (8) follows directly from the performance difference lemma [17] applied to nature's cost minimizing MDP and Definition 2. In addition, (10) follows from (8), together with the definitions of ψ_π^{ω} in (9) and H_π^{ω} in Definition 2. □

In the following sections, we proceed to introduce efficient solution methods for the robust policy evaluation (2) and optimization (3) problems. Notably the proposed methods only require the sampling access to the generating kernels $\{P_k\}_{k \in \mathcal{K}}$. This can be particularly convenient for implementation purposes, as simulating a random process can be considerably easier than constructing or storing its distribution explicitly.

3 Robust Policy Evaluation

In this section, we turn our attention to the robust policy evaluation problem (2). As discussed in Section 2.1, given any to-be-evaluated policy π, it suffices to compute the optimal value function of nature's cost-minimizing MDP. It should be noted that here nature's MDP belongs to the class of continuous action MDPs, which in general is intractable. In view of this, going forward we will exploit the fact that nature's action-value function (Definition 2) is indeed affine in its action. In particular, the proposed method admits the following conceptual update:

$$\omega^{(m+1)}(\cdot|s) \leftarrow \operatorname*{argmin}_{v \in W} \sum_{t=0}^{m} \alpha_t H_\pi^{\omega^{(t)}}(s, v) + \lambda_m h(v) \quad \forall s \in \mathcal{S}, \tag{11}$$

where $\{\alpha_t\}_{t \geq 0}$ denote the stepsizes, $\lambda_m \geq 0$ denotes the regularization strength, and $h(v) = \sum_{k=1}^{K} v_k \log v_k$ denotes the negative entropy function.

A few remarks are in order for update (11). First, it is worth noting that from Definition 2, the update (11) corresponds to solving a simple convex optimization problem. This simple observation will prove to be essential in constructing the estimation of $H_\pi^{\omega^{(t)}}$ and the subsequent convergence analysis of Algorithm 1. Second, from an implementation perspective, as (11) defines the access to the updated policy $\omega^{(m+1)}$ through solving (11), there is indeed no need to explicitly compute and store $\omega^{(m+1)}(\cdot|s)$ for every state $s \in \mathcal{S}$. Instead, action $\omega^{(m+1)}(\cdot|s)$ can be generated whenever the policy $\omega^{(m+1)}$ needs to be accessed at the state of interest, and consequently the size of the state space does not affect the computational complexity of Algorithm 1 explicitly. This will be further discussed within Section 3.1.

It can be seen that (11) is only conceptual, as the state-action value functions $\{H_\pi^{\omega^{(t)}}\}_{t \geq 0}$ are rarely available. We will introduce a subroutine in Section 3.1 that provides an efficient estimation of H_π^{ω} for any given $(\pi, \omega) \in \Pi \times \mathcal{W}$, by using samples drawn from the generating kernels $\{P_k\}_{k \in \mathcal{K}}$. With this in mind, we present in Algorithm 1 the details of the proposed stochastic method for robust policy evaluation.

Clearly, Algorithm 1 differs from the conceptual update (11) by replacing the unknown state-action value function $H_\pi^{\omega^{(m)}}$ by its sample estimate $\widehat{H}_m$. In a nutshell, the estimation takes a two-step procedure, defined by first estimating the joint action-value function $Q_\pi^{\omega^{(m)}}$ (Algorithm 2) given joint policy $(\pi, \omega^{(m)})$, followed

by estimating $H_\pi^{\omega^{(m)}}$ in view of Definition 2. Clearly, the constructed $\widehat{H}_m(s,v)$ in (12) is affine in v, which ensures that update (13) can be efficiently computed.

It should be noted here that the conceptual update (13) bears some similarities with the policy dual averaging method proposed in [16] for solving general state-action space MDPs. It is hence worth discussing some important differences for development herein and in [16]. First, for continuous action space MDPs, convergence to stationary solutions is established in [16]. While nature's cost-minimizing MDP herein involves a continuous action space, in this manuscript, we heavily utilize the affine structure of nature's action-value function to establish the global convergence of Algorithm 1. Second, while [16] assumes access to oracles satisfying generic error conditions required for convergence analysis, we will work with relaxed error conditions for Algorithm 1, which are in turn satisfied by Algorithm 2. *En route*, we develop new probabilistic tools for Algorithms 1 and 2 that allow us to work with the relaxed error conditions, which could be of independent interest. Third, analysis within [16] only establishes the bias of the estimated value function. On the other hand, for the purpose of evaluation one would ideally seek estimation accuracy in a high probability sense instead of only controlling the bias.[7] As will be clarified later, this seemingly simple question of boosting accuracy certificate from bias to high probability in turn creates some technical challenges that appear non-trivial to us. Fourth, and perhaps most importantly, we will demonstrate later that high probability control on robust policy evaluation appears to be essential for robust policy optimization (Remark 1).

Our subsequent discussion on the convergence of Algorithm 1 requires the following condition on the stochastic error of $\widehat{Q}_m$.

Condition 1. *For any given $\delta \in (0,1)$, there exists $\varepsilon_Q, B > 0$, such that the estimation $\widehat{Q}_m$ of $Q_\pi^{\omega^{(m)}}$ satisfies*

$$\|\mathbb{E}_\pi^{\omega^{(m)}}[\widehat{Q}_m|\mathcal{F}_{m-1}] - Q_\pi^{\omega^{(m)}}\|_\infty \leq \varepsilon_Q, \quad \|\widehat{Q}_m\|_\infty \leq B, \quad \mathbb{E}_\pi^{\omega^{(m)}}\big[\|\widehat{Q}_m - Q_\pi^{\omega^{(m)}}\|_\infty^2|\mathcal{F}_{m-1}\big] \leq J \tag{14a}$$

with probability $1-\delta/(4M)$ for every $m = 1,\ldots,M$, where $\{\omega^{(m)}\}_{m=1}^M$ are the iterates generated by Algorithm 1. Here $\mathcal{F}_{m-1}$ denotes the σ-algebra generated up to iteration $m-1$.

Condition 1 states that $\widehat{Q}_m$ has diminishing conditional bias and has bounded norm in high probability. Notably, in Section 3.1, we will establish that the proposed subroutine Algorithm 2 indeed satisfies Condition 1. Consequently from Definition 2 and Condition 1 we have for every $m = 1,\ldots,M$ with probability

Algorithm 1 Stochastic dual averaging for robust policy evaluation

Require: Controller's policy $\pi \in \Pi$, initial nature's policy $\omega^{(0)} \in \mathcal{W}$, number of iterations M, stepsizes $\{\alpha_m\}_{m=0}^M$ and regularization parameters $\{\lambda_m\}_{m=0}^M$

1: **for** $m = 0,1,2,\ldots,M$ **do**

2: Compute $\widehat{Q}_m \approx Q_\pi^{\omega^{(m)}}$ using Algorithm 2 with inputs $(\omega^{(m)},\pi)$

3: Set

$$\widehat{V}_m(s) = \sum_{a\in\mathcal{A}}\sum_{k\in\mathcal{K}} \omega^{(m)}(k|s)\pi(a|s)\widehat{Q}_m(s,a,k), \quad \text{and} \quad \widehat{H}_m(s,v) = \sum_{a\in\mathcal{A}}\sum_{k\in\mathcal{K}} v_k\pi(a|s)\widehat{Q}_m(s,a,k) \tag{12}$$

4: Update

$$\omega^{(m+1)}(\cdot|s) \leftarrow \operatorname*{argmin}_{v\in W} \sum_{t=0}^{m} \alpha_t \widehat{H}_t(s,v) + \lambda_m h(v) \tag{13}$$

5: **end for**

6: **return** $\widehat{V}_\pi = \sum_{m=1}^M \vartheta_m \widehat{V}_m$ and $\widehat{Q}_\pi = \sum_{m=1}^M \vartheta_m \widehat{Q}_m$, where $\vartheta_m = \alpha_m/(\sum_{j=1}^M \alpha_j)$ for $m = 1,2,\ldots,M$

[7] To be more precise, we will proceed to establish an $\widetilde{\mathcal{O}}(1/\varepsilon^2)$ sample complexity for $\left\|\sum_{m=1}^M \vartheta_m \widehat{V}_m - V_\pi\right\| \leq \varepsilon$ with high probability, instead of $\mathbb{E}\left[\sum_{m=1}^M \vartheta_m V_\pi^{w^{(m)}} - V_\pi\right] \leq \varepsilon$ established in [16]. Note the latter appears to be even weaker than the convergence guarantee on mean-squared error typically performed for non-robust policy evaluation [2, 18], as it only establishes the diminishing bias of $\sum_{m=1}^M \vartheta_m V_\pi^{w^{(m)}}$ as an estimator of V_π. It remains unclear whether one can boost this accuracy certificate to high probability with $\mathcal{O}(\log(1/\delta))$ sample complexity overhead for a confidence level of $1-\delta$.

$1-\delta/(4M)$ that

$$\|\mathbb{E}_\pi^{\omega^{(m)}}[\widehat{V}_m|\mathcal{F}_{m-1}] - V_\pi^{\omega^{(m)}}\|_\infty \le \varepsilon_Q, \quad \|\widehat{V}_m\|_\infty \le B, \quad \mathbb{E}_\pi^{\omega^{(m)}}\big[\|\widehat{V}_m - V_\pi^{\omega^{(m)}}\|_\infty^2|\mathcal{F}_{m-1}\big] \le J, \tag{14b}$$

and

$$\|\mathbb{E}_\pi^{\omega^{(m)}}[\widehat{H}_m|\mathcal{F}_{m-1}] - H_\pi^{\omega^{(m)}}\|_\infty \le \varepsilon_Q, \quad \|\widehat{H}_m\|_\infty \le B, \quad \mathbb{E}_\pi^{\omega^{(m)}}\big[\|\widehat{H}_m - H_\pi^{\omega^{(m)}}\|_\infty^2|\mathcal{F}_{m-1}\big] \le J. \tag{14c}$$

To facilitate our discussion, let us denote $H_m = H_\pi^{\omega^{(m)}}$, and correspondingly let

$$\begin{aligned}
\psi_m(s,v) &= H_m(s,v) - H_m(s,\omega^{(m)}(\cdot|s)),\\
\widehat{\psi}_m(s,v) &= \widehat{H}_m(s,v) - \widehat{H}_m(s,\omega^{(m)}(\cdot|s)),\\
\widehat{\Psi}_m(s,v) &= \sum_{t=0}^{m} \alpha_t \widehat{\psi}_t(s,v).
\end{aligned}$$

Note that from Definition 2 and (9), it is clear that $\psi_m = \psi_\pi^{\omega^{(m)}}$. In addition, let us define stochastic error

$$\delta_m = \widehat{\psi}_m - \psi_m.$$

A consequence of Condition 1 together with (14b) and (14c) is summarized in the lemma below.

Lemma 3.1. *Suppose that Condition 1 holds, and set $\alpha_m = \sqrt{m}$ for all $m \le M$. In addition, let $\bar{B} = B + \frac{1}{1-\gamma}$ and denote by $\omega_\pi^\star$ the optimal policy for the inner minimization problem in (6). For any $\varepsilon \ge \varepsilon_Q$, the following hold with probability at least $1-\delta/2$ for all $(s,a,k) \in \mathcal{S}\times\mathcal{A}\times\mathcal{K}$.*

(i) $$\left|\sum_{m=1}^{M} \vartheta_m(\widehat{V}_m(s) - V_\pi^{\omega^{(m)}}(s))\right| \le \varepsilon + \frac{3\bar{B}}{\sqrt{M}}\sqrt{\log\left(\frac{8M}{\delta}\right)} + \sqrt{\frac{4J\delta}{M}};$$

(ii) $$\sum_{m=1}^{M} \frac{\vartheta_m}{1-\gamma}\int_{\mathcal{S}} \delta_m(s',\omega_\pi^\star(\cdot|s'))d_\pi^{\omega_\pi^\star}(\mathrm{d}s'|s) \le \frac{2}{1-\gamma}\left(\varepsilon + \frac{3\bar{B}}{\sqrt{M}}\sqrt{\log\left(\frac{8M}{\delta}\right)} + \sqrt{\frac{4J\delta}{M}}\right);$$

(iii) $$\left|\sum_{m=1}^{M} \vartheta_m(\widehat{Q}_m(s,a,k) - Q_\pi^{\omega^{(m)}}(s,a,k))\right| \le \varepsilon + \frac{3\bar{B}}{\sqrt{M}}\sqrt{\log\left(\frac{8M}{\delta}\right)} + \sqrt{\frac{4J\delta}{M}}.$$

Proof. Observe first that the choice of α_m gives $\sum_{m=1}^M \alpha_m = \sum_{m=1}^M \sqrt{m} \ge \int_0^M \sqrt{x}\,dx = \frac{2}{3}M^{3/2}$. Hence, we have

$$\sum_{m=1}^{M} \vartheta_m^2 = \frac{\sum_{m=1}^M \alpha_m^2}{(\sum_{j=1}^M \alpha_j)^2} \le \frac{M(M+1)/2}{(\frac{2}{3}M^{3/2})^2} \le \frac{9}{4M}. \tag{15}$$

For Assertion (i), since (14b) holds with probability $1-\delta/(4M)$, we have for every fixed $s \in \mathcal{S}$ that

$$\left|\sum_{m=1}^{M} \vartheta_m(\widehat{V}_m(s) - V_\pi^{\omega^{(m)}}(s))\right| \le \varepsilon + \bar{B}\sqrt{2\sum_{m=1}^{M}\vartheta_m^2 \log\left(\frac{8M}{\delta}\right)} + \sqrt{\frac{4J\delta}{M}} \le \varepsilon + \frac{3\bar{B}}{\sqrt{M}}\sqrt{\log\left(\frac{8M}{\delta}\right)} + \sqrt{\frac{4J\delta}{M}}$$

holds with probability at least $1-\delta/2$. The first inequality in the above expression follows from Lemma A.1, and the second inequality holds because of (15). This observation concludes the proof for Assertion (i). Assertion (ii) (resp. Assertion (iii)) follows from a similar reasoning where (14c) (resp. Condition 1) is invoked instead of (14b). □

We proceed by presenting the following basic property for each update (13) of Algorithm 1. Note that the Bregman divergence induced by h corresponds to the Kullback-Leibler divergence given by

$$\mathsf{D}(v',v) = h(v') - h(v) - \langle \nabla h(v), v'-v\rangle = \sum_{k\in\mathcal{K}} v'_k \log\left(\frac{v'_k}{v_k}\right) \quad \forall v, v' \in \Delta_{\mathcal{K}}.$$

Lemma 3.2. *Define $\widehat{\Psi}_{-1} = 0$ and $\lambda_{-1} = 0$, and suppose $\lambda_m \geq \lambda_{m-1}$ for every $m \geq 1$. Then, for all $m \geq 0$ and $s \in \mathcal{S}$, the iterates $\{\omega^{(m)}\}_{m=1}^M$ generated by Algorithm 1 satisfy*

(i) $\widehat{\Psi}_m(s, \omega^{(m+1)}(\cdot|s)) + \lambda_m \mathsf{D}(v, \omega^{(m+1)}(\cdot|s)) + \lambda_m h(\omega^{(m+1)}(\cdot|s)) - \lambda_m h(v) \leq \widehat{\Psi}_m(s, v)$ *for every* $v \in W$,

(ii) $\alpha_m \widehat{\psi}_m(s, \omega^{(m+1)}(\cdot|s)) \leq \widehat{\Psi}_m(s, \omega^{(m+1)}(\cdot|s)) - \widehat{\Psi}_{m-1}(s, \omega^{(m)}(\cdot|s)) - \lambda_{m-1} \mathsf{D}(\omega^{(m+1)}(\cdot|s), \omega^{(m)}(\cdot|s)) - \lambda_{m-1} h(\omega^{(m)}(\cdot|s)) + \lambda_{m-1} h(\omega^{(m+1)}(\cdot|s))$.

Proof. For Assertion (i), note that the first-order optimality condition of (13) in Algorithm 1 implies

$$\sum_{t=0}^{m} \alpha_t \widehat{H}_t(s, v) - \sum_{t=0}^{m} \alpha_t \widehat{H}_t(s, \omega^{(m+1)}(\cdot|s)) + \lambda_m \langle \nabla h(\omega^{(m+1)}(\cdot|s)), v - \omega^{(m+1)}(\cdot|s) \rangle \geq 0.$$

The claim then follows from the definition of Bregman divergence $\mathsf{D}(\cdot, \cdot)$ and the definition of $\widehat{\Psi}_m$. For Assertion (ii), we have

$$\begin{aligned}
\alpha_m \widehat{\psi}_m(s, \omega^{(m+1)}(\cdot|s)) &= \widehat{\Psi}_m(s, \omega^{(m+1)}(\cdot|s)) - \widehat{\Psi}_{m-1}(s, \omega^{(m+1)}(\cdot|s)) \\
&\leq \widehat{\Psi}_m(s, \omega^{(m+1)}(\cdot|s)) - \widehat{\Psi}_{m-1}(s, \omega^{(m)}(\cdot|s)) - \lambda_{m-1} \mathsf{D}(\omega^{(m+1)}(\cdot|s), \omega^{(m)}(\cdot|s)) \\
&\quad - \lambda_{m-1} h(\omega^{(m)}(\cdot|s)) + \lambda_{m-1} h(\omega^{(m+1)}(\cdot|s)),
\end{aligned}$$

where the equality holds by construction of $\widehat{\Psi}_m$, and the inequality follows from Assertion (i). □

With Lemma 3.2 in place, we proceed to establish some generic convergence properties of Algorithm 1. In what follows, Lemma 3.3 provides lower and upper bounds on estimating the robust value function V_π using weighted combination of value functions $\{V_\pi^{\omega^{(m)}}\}_{m=1}^M$ generated by Algorithm 1.

Lemma 3.3. *Suppose that (14c) holds. Then, for any $\varepsilon \geq \varepsilon_Q$, with probability at least $1 - \delta/2$, we have*

$$\begin{aligned}
0 \leq \sum_{m=1}^{M} \vartheta_m V_\pi^{\omega^{(m)}}(s) - V_\pi(s) \leq \frac{2\varepsilon}{1-\gamma} + \left(\sum_{m=1}^{M} \alpha_m \right)^{-1} \left(\sum_{m=1}^{M} \frac{\alpha_m^2 B^2}{2(1-\gamma)\lambda_{m-1}} + \frac{2\lambda_M \log(|\mathcal{K}|)}{1-\gamma} \right) \\
+ \frac{6\bar{B}}{\sqrt{M}(1-\gamma)} \sqrt{\log\left(\frac{8M}{\delta}\right)} + \frac{2}{1-\gamma} \sqrt{\frac{4J\delta}{M}} \qquad \forall s \in \mathcal{S}.
\end{aligned}$$

Proof. Taking the telescoping sum of Lemma 3.2(ii) from $m = 0$ to M, we obtain

$$\begin{aligned}
\alpha_0 \widehat{\psi}_0(s, \omega^{(1)}(\cdot|s)) &\leq \widehat{\Psi}_M(s, \omega^{(M+1)}(\cdot|s)) - \sum_{m=1}^{M} \lambda_{m-1} \mathsf{D}(\omega^{(m+1)}(\cdot|s), \omega^{(m)}(\cdot|s)) - \sum_{m=1}^{M} \alpha_m \widehat{\psi}_m(s, \omega^{(m+1)}(\cdot|s)) \\
&\quad + \lambda_{M-1} h(\omega^{(M+1)}(\cdot|s)) \\
&\leq \widehat{\Psi}_M(s, \omega^{(M+1)}(\cdot|s)) - \frac{1}{2} \sum_{m=1}^{M} \sum_{a \in \mathcal{A}} \pi(a|s) \lambda_{m-1} \|\omega^{(m+1)}(\cdot|s) - \omega^{(m)}(\cdot|s)\|_1^2 \\
&\quad - \sum_{m=1}^{M} \alpha_m \sum_{k \in \mathcal{K}} \left[\widehat{H}_m(s, \omega^{(m+1)}(k|s)) - \widehat{H}_m(s, \omega^{(m)}(k|s)) \right] + \lambda_{M-1} h(\omega^{(M+1)}(\cdot|s)) \\
&\leq \widehat{\Psi}_M(s, \omega^{(M+1)}(\cdot|s)) - \frac{1}{2} \sum_{m=1}^{M} \sum_{a \in \mathcal{A}} \pi(a|s) \lambda_{m-1} \|\omega^{(m+1)}(\cdot|s) - \omega^{(m)}(\cdot|s)\|_1^2 \\
&\quad - \sum_{m=1}^{M} \alpha_m \|\widehat{H}_m\|_\infty \|\omega^{(m+1)}(\cdot|s) - \omega^{(m)}(\cdot|s)\|_1 + \lambda_{M-1} h(\omega^{(M+1)}(\cdot|s)) \\
&\leq \widehat{\Psi}_M(s, \omega^{(M+1)}(\cdot|s)) + \sum_{m=1}^{M} \frac{\alpha_m^2 \|\widehat{H}_m\|_\infty^2}{2\lambda_{m-1}} + \lambda_{M-1} \log(|\mathcal{K}|)
\end{aligned}$$

$$\leq \widehat{\Psi}_M(s,\omega_\pi^\star(s)) + \sum_{m=1}^{M} \frac{\alpha_m^2 \|\widehat{H}_m\|_\infty^2}{2\lambda_{m-1}} + 2\lambda_M \log(|\mathcal{K}|),$$

where the second inequality holds by Pinsker's inequality and the definition of $\widehat{\psi}$, the third inequality exploits Hölder's inequality, and the fourth inequality uses Young's inequality and $|h(v)| \leq \log(|\mathcal{K}|)$ for all $v \in W$. Finally, the last inequality follows from Lemma 3.2(i) where $v = \omega_\pi^\star(s)$. Since $\alpha_0 = 0$, integrating the leftmost and rightmost sides of the above inequality with respect to $d_\pi^{\omega_\pi^\star}(\cdot|s)$ and using the definition of $\widehat{\Psi}_M$ gives

$$\begin{aligned}
0 &\leq \sum_{m=1}^{M} \alpha_m \int_{\mathcal{S}} \psi_m(s,\omega_\pi^\star(\cdot|s)) d_\pi^{\omega_\pi^\star}(\mathrm{d}s'|s) \\
&\quad + \sum_{m=1}^{M} \frac{\alpha_m^2 \|\widehat{H}_m\|_\infty^2}{2\lambda_{m-1}} + \sum_{m=1}^{M} \alpha_m \int_{\mathcal{S}} \delta_m(s,\omega_\pi^\star(\cdot|s)) d_\pi^{\omega_\pi^\star}(\mathrm{d}s'|s) + 2\lambda_M \log(|\mathcal{K}|) \\
&= \sum_{m=1}^{M} \alpha_m (1-\gamma)(V_\pi^{\omega_\pi^\star}(s) - V_\pi^{\omega^{(m)}}(s)) \\
&\quad + \sum_{m=1}^{M} \frac{\alpha_m^2 \|\widehat{H}_m\|_\infty^2}{2\lambda_{m-1}} + \sum_{m=1}^{M} \alpha_m \int_{\mathcal{S}} \delta_m(s',\omega_\pi^\star(\cdot|s')) d_\pi^{\omega_\pi^\star}(\mathrm{d}s'|s) + 2\lambda_M \log(|\mathcal{K}|),
\end{aligned}$$

where the equality follows from Lemma 2.1. Noting that $V_\pi^{\omega_\pi^\star} = V_\pi$, $\vartheta_m = \alpha_m / (\sum_{j=1}^{M} \alpha_j)$ together with Lemma 3.1(ii) concludes the proof. □

We are now ready to specify concrete choices of $\{\alpha_m\}_{m\geq 0}$ and $\{\lambda_m\}_{m\geq 0}$, and correspondingly establish the high probability convergence of $\widehat{V}_\pi$ and $\widehat{Q}_\pi$ towards both robust value function V_π and the robust joint action-value function Q_π defined through

$$Q_\pi(s,a,k) = \min_{\omega \in \mathcal{W}} Q_\pi^\omega(s,a,k) \quad \forall (s,a,k) \in \mathcal{S} \times \mathcal{A} \times \mathcal{K} \tag{16}$$

under Condition 1.

Theorem 3.4. *Suppose that Condition 1 holds. Set*

$$\alpha_m = \sqrt{m}, \quad \lambda_m = \frac{(m+1)B}{2\sqrt{\log(|\mathcal{K}|)}}, \quad \forall m \leq M$$

in Algorithm 1. Then, for any $\varepsilon \geq 4\varepsilon_Q/(1-\gamma)$, the total number of iterations required by Algorithm 1 to output

$$-\varepsilon \leq \widehat{V}_\pi(s) - V_\pi(s) \leq \varepsilon, \quad -\varepsilon \leq \widehat{Q}_\pi(s,a,k) - Q_\pi(s,a,k) \leq \varepsilon \quad \forall (s,a,k) \in \mathcal{S} \times \mathcal{A} \times \mathcal{K}$$

with probability at least $1-\delta$ is bounded by

$$M = \mathcal{O}\left(\frac{\bar{B}^2 \log(|\mathcal{K}|)}{(1-\gamma)^2 \varepsilon^2} \log\left(\frac{\bar{B}^2}{\delta(1-\gamma)^2\varepsilon^2} \right) + \frac{J\delta}{\varepsilon^2} \right). \tag{17}$$

Proof. Adding $\sum_{m=1}^{M} \vartheta_m (\widehat{V}_m(s) - V_\pi^{\omega^{(m)}}(s))$ to each part in the stated inequality of Lemma 3.3 and observing that $(1-\gamma)\varepsilon/4 \geq \varepsilon_Q$, we obtain with probability at least $1-\delta/2$ that

$$\sum_{m=1}^{M} \vartheta_m \widehat{V}_m(s) - V_\pi^{\omega^{(m)}}(s) \leq \sum_{m=1}^{M} \vartheta_m \widehat{V}_m(s) - V_\pi(s) \leq \sum_{m=1}^{M} \vartheta_m \widehat{V}_m(s) - V_\pi^{\omega^{(m)}}(s) + \frac{\varepsilon}{2} + C(M,\delta), \tag{18}$$

where

$$C(M,\delta) = \left(\sum_{m=1}^{M} \alpha_m \right)^{-1} \left(\sum_{m=1}^{M} \frac{\alpha_m^2 B^2}{2(1-\gamma)\lambda_{m-1}} + \frac{2\lambda_M \log(|\mathcal{K}|)}{1-\gamma} \right) + \frac{6\bar{B}}{\sqrt{M}(1-\gamma)} \sqrt{\log\left(\frac{8M}{\delta}\right)} + \frac{2}{1-\gamma} \sqrt{\frac{4J\delta}{M}}.$$

Using Lemma 3.1(i) with the observation $(1-\gamma)\varepsilon/4 \geq \varepsilon_Q$ to lower and upper bound the term $\sum_{m=1}^{M} \vartheta_m(\widehat{V}_m(s) - V_\pi^{\omega^{(m)}}(s))$ in (18), we obtain

$$-\frac{\varepsilon}{4} - \frac{3\bar{B}}{\sqrt{M}}\sqrt{\log\left(\frac{8M}{\delta}\right)} - \sqrt{\frac{4J\delta}{M}} \leq \sum_{m=1}^{M} \vartheta_m \widehat{V}_m(s) - V_\pi^{\omega_\pi^\star}(s) \leq \frac{3\varepsilon}{4} + C(M,\delta) + \frac{3\bar{B}}{\sqrt{M}}\sqrt{\log\left(\frac{8M}{\delta}\right)} + \sqrt{\frac{4J\delta}{M}} \tag{19}$$

with probability $1-\delta$. On the other hand, integrating all three terms of the inequalities in the statement of Lemma 3.3 with respect to $P_k(\cdot|s,a)$ gives

$$0 \leq \sum_{m=1}^{M} \vartheta_m Q_\pi^{\omega^{(m)}}(s,a,k) - Q_\pi(s,a,k) \leq \frac{\varepsilon}{2} + C(M,\delta)$$

with probabitliy $1-\delta/2$. After adding $\sum_{m=1}^{M} \vartheta_m(\widehat{Q}_m(s,a,k) - Q_\pi^{\omega^{(m)}}(s,a,k))$ to each part of the above inequality and applying Lemma 3.1(iii) with the observation $(1-\gamma)\varepsilon/4 \geq \varepsilon_Q$, we obtain with probability $1-\delta$ that

$$\begin{aligned} -\frac{\varepsilon}{4} - \frac{3\bar{B}}{\sqrt{M}}\sqrt{\log\left(\frac{8M}{\delta}\right)} - \sqrt{\frac{4J\delta}{M}} &\leq \sum_{m=1}^{M} \vartheta_m \widehat{Q}_m(s,a,k) - Q_\pi(s,a,k) \\ &\leq \frac{3\varepsilon}{4} + C(M,\delta) + \frac{3\bar{B}}{\sqrt{M}}\sqrt{\log\left(\frac{8M}{\delta}\right)} + \sqrt{\frac{4J\delta}{M}}. \end{aligned} \tag{20}$$

Finally, recall that the choices $\alpha_m = \sqrt{m}$ and $\lambda_m = \frac{(m+1)B}{2\sqrt{\log(|\mathcal{K}|)}}$ gives $\left(\sum_{m=1}^{M} \alpha_m\right)^{-1} \leq \int_0^M \sqrt{x}\,\mathrm{d}x = \frac{2}{3}M^{-3/2}$ together with $\alpha_m^2/\lambda_{m-1} = 2\sqrt{\log(|\mathcal{K}|)}/B$. Hence, choosing

$$M = \mathcal{O}\left(\frac{\bar{B}^2 \log(|\mathcal{K}|)}{(1-\gamma)^2\varepsilon^2} \log\left(\frac{\bar{B}^2}{\delta(1-\gamma)^2\varepsilon^2}\right) + \frac{J\delta}{\varepsilon^2}\right)$$

ensures that $C(M,\delta) + \frac{3\bar{B}}{\sqrt{M}}\sqrt{\log\left(\frac{8M}{\delta}\right)} + \sqrt{\frac{4J\delta}{M}}$ is at most $\varepsilon/4$. Applying the previous observation to (19) and (20) together with the construction of $\widehat{V}_\pi$ and $\widehat{Q}_\pi$ concludes the proof. □

Note that the second term in the iteration complexity (17) obtained in Theorem 3.4 scales linearly with respect to δ. As the required confidence level $1-\delta$ approaches 1, this will in turn clearly be dominated by the first term. Consequently in view of Theorem 3.4, it suffices for Algorithm 1 to take $\widetilde{\mathcal{O}}(1/\varepsilon^2)$ iterations to output an ε-accurate estimation of the robust value function V_π associated with the mixture ambiguity set in Definition 1. Of course, this iteration complexity hinges upon Condition 1, which requires accurate estimation of the value function V_π^ω and the action-value function H_π^ω of nature. This in turn will be fulfilled by Algorithm 2, which we discuss in detail in Section 3.1. Notably, we will determine the corresponding sample complexity required to certify Condition 1. Finally, we conclude in Section 3.2 by establishing the total sample complexity of Algorithm 1 for estimating the robust value function V_π.

3.1 Estimating Joint Action-value Function

We now proceed to introduce a temporal difference (TD) learning-based method [47] for estimating the joint action-value function Q_π^ω associated with any $(\pi,\omega) \in \Pi \times \mathcal{W}$. In particular, we will establish that the proposed method constitutes a valid subroutine in Algorithm 1 that certifies Condition 1.

Consider the Bellman operator T_π^ω defined through

$$(T_\pi^\omega Q)(s,a,k) = r(s,a) + \gamma \int_{\mathcal{S}} \sum_{a' \in \mathcal{A}} \sum_{k' \in \mathcal{K}} \pi(a'|s')\omega(k'|s')Q(s',a',k')P_k(\mathrm{d}s'|s,a) \quad \forall (s,a,k) \in \mathcal{S}\times\mathcal{A}\times\mathcal{K}, \tag{21}$$

for any $Q : \mathcal{S}\times\mathcal{A}\times\mathcal{K} \to \mathbb{R}$. Going forward we also write $\mathcal{Z} = \mathcal{S}\times\mathcal{A}\times\mathcal{K}$, together with $z \in \mathcal{Z}$ in short for $z = (s,a,k)$ for some $s \in \mathcal{S}, a \in \mathcal{A}, k \in \mathcal{K}$. It is clear that T_π maps the space of bounded continuous functions

onto itself, and constitutes a contraction operator in $\|\cdot\|_\infty$ norm. In addition, from dynamic programming equations it can readily seen that the unique fixed point of T_π^ω is given by Q_π^ω (Banach fixed point theorem).

It should be noted that as the set $\mathcal{Z}$ is continuous, it becomes computationally prohibitive even to represent Q_π^ω with a reasonable computing budget. Instead, we consider finding the best possible approximation of Q_π^ω within the linear space spanned by some pre-determined basis functions. In particular, let $\phi_i : \mathcal{Z} \to \mathbb{R}$ denote the i-th basis function for $i = 1, \ldots, d$, and define the feature mapping $\phi(\cdot) = [\phi_1(\cdot), \ldots, \phi_d(\cdot)]$. Without loss of generality we assume that $\|\phi(\cdot)\| \leq 1$. The constructions of basis function $\{\phi_i\}_{i=1}^d$ and the choice of dimension d have been extensively discussed in the prior literature, for instance, by using random feature mappings that can uniformly approximate a function class with bounded norms in a reproducing kernel Hilbert space [42].

We proceed as follows. Let us denote by $\mathcal{Q} = \{Q_\theta(\cdot) \coloneqq \theta^\top \phi(\cdot) : \theta \in \mathbb{R}^d\}$ the linear span of basis functions. One can view $\mathcal{Q}$ as a proper subspace of $\mathcal{L}^2(\mathcal{Z}, \nu_\pi^\omega)$, where ν_π^ω corresponds to the stationary measure of Markov chain $\{Z_t \coloneqq (S_t, A_t, K_t)\}_{t\geq 0}$ with transition law defined by $A_t \sim \pi(\cdot|S_t), K_t \sim \omega(S_t), S_{t+1} \sim P_{K_t}(\cdot|S_t, A_t)$. We are interested in finding $\theta_\pi^\omega \in \mathbb{R}^d$ such that $Q_{\theta_\pi^\omega}(\cdot)$ provides, in some sense, a reasonable approximation of Q_π^ω. In particular, from (21) and $T_\pi^\omega Q_\pi^\omega = Q_\pi^\omega$, it is natural to consider finding θ_π^ω such that

$$\Pi_{\mathcal{L}^2(\mathcal{Z}, \nu_\pi^\omega)} T_\pi^\omega Q_{\theta_\pi^\omega} = Q_{\theta_\pi^\omega}, \tag{22}$$

where $\Pi_{\mathcal{L}^2(\mathcal{Z}, \nu_\pi^\omega)}$ denotes the orthogonal projection onto $\mathcal{Q}$.[8] From the optimality condition of the projection $\Pi_{\mathcal{L}^2(\mathcal{Z}, \nu_\pi^\omega)}$, it can be readily verified that (22) is equivalent to $F(\theta_\pi^\omega) = 0$, where

$$F(\theta) = \int_{\mathcal{Z}} \phi(z) \left[\phi(z)^\top \theta \; - \; r(s,a) - \gamma \int_{\mathcal{S}} \sum_{a' \in \mathcal{A}} \sum_{k' \in \mathcal{K}} \pi(a'|s') \omega(k'|s') \phi(s', a', k')^\top \theta P_k(\mathrm{d}s'|s,a) \right] \nu_\pi^\omega(\mathrm{d}z). \tag{23}$$

We present the details of the proposed TD learning method in Algorithm 2. In a nutshell, the method operates by drawing samples from the generating kernels $\{P_k\}_{k\in\mathcal{K}}$ to construct a stochastic estimation $\widehat{F}$ of operator $F(\cdot)$ defined in (23). Note that the trajectory $\{Z_t\}_{t\geq 0}$ is sampled from the generating kernels $\{P_k\}_{k\in\mathcal{K}}$ in a Markovian manner. In view of the Markovian noise, we take a similar strategy as in [18, 60] that periodically skips τ samples within the above construction of $\widehat{F}$. In particular, at t-th iteration of Algorithm 2, the algorithm evaluates the stochastic operator $\widehat{F}$ through

$$\widehat{F}(\theta_t, \xi_t^\tau) = (\langle \phi(Z_t^\tau), \theta_t \rangle - r(S_t^\tau, A_t^\tau) - \gamma \langle \phi((Z_t^\tau)'), \theta_t \rangle) \phi(Z_t^\tau), \tag{24}$$

where $\xi_t^\tau = (Z_t^\tau, (Z_t^\tau)')$, generated by Procedure 3, corresponds to t-th transition pair that is separated from $(t-1)$-th pair by τ timesteps. We will establish later that $\widehat{F}$ constructed above serves as an approximately unbiased estimation of F in (23).

Before we proceed, it is worth mentioning here that we will establish high probability convergence of Algorithm 2 over unbounded domain, which should be contrasted with existing convergence analysis in expectation [2, 18]. As will be clarified later, the interdependence of the noise associated with $\widehat{F}$ and the norm of the iterates makes the corresponding analysis for high probability results considerably more involved compared to expectation bounds. In particular, the approach we develop for establishing boundedness of iterates in high probability seems to be new for TD-type methods, and hence may be of independent interest.

For any to-be-evaluated joint policy $(\pi, \omega) \in \Pi \times \mathcal{W}$, we make the following assumption on the Markov chain $\{Z_t\}_{t\geq 0}$ and the feature mapping ϕ.

Assumption 1. *There exist constants $C > 0$ and $\rho \in (0,1)$ such that for every Borel set $\mathcal{B} \subseteq \mathcal{S}$,*

$$\left| \mathbb{P}_\pi^\omega(S_{t+\tau} \in \mathcal{B} \mid \mathcal{F}_{t-1}) - \nu_\pi^\omega(\mathcal{B}) \right| \leq C\rho^\tau \quad \forall\, t \in \mathbb{Z}_{++},\ \tau \in \mathbb{Z}_+, \tag{25}$$

where $\mathcal{F}$ denotes the filtration up to iteration $t-1$ of Algorithm 2. In addition, let $\Sigma = \int_{\mathcal{Z}} \phi(z)\, \phi(z)^\top \nu_\pi^\omega(\mathrm{d}z)$. We assume that $\Sigma \succ 0$.

[8] Clearly, the projection $\Pi_{\mathcal{L}^2(\mathcal{Z}, \nu_\pi^\omega)}$ is necessary, otherwise solution might not exist for (22). It should be also noted that projection on $\mathcal{Z}$ in $\|\cdot\|_\infty$ norm, although being conceptually appealing and naturally leads to value iteration, is in general not implementable as the space $\mathcal{Z}$ is uncountable.

Algorithm 2 Temporal difference learning

Require: $(\pi, \omega) \in \Pi \times \mathcal{W}$, η, T, τ, $S_0 = s \in \mathcal{S}$

1: $K_0 \sim \omega(S_0)$, $A_0 \sim \pi(\cdot|S_0)$, $S_1 \sim P_{K_0}(\cdot|S_0, A_0)$, $K_1 \sim \omega(S_1)$, $A_1 \sim \pi(\cdot|S_1)$. Set $\xi_{-1}^{\tau} = (Z_0, Z_1)$.
2: **for** $t = 0, \ldots, T$ **do**
3: $\quad \xi_t^{\tau} \leftarrow \textsc{MarkovSampler}(\xi_{t-1}^{\tau}, \omega, \pi, \tau)$
4: $\quad \theta_{t+1} = \theta_t - \eta \widehat{F}(\theta_t, \xi_t^{\tau})$
5: **end for**
6: **return** $\widehat{Q}(\cdot) = \phi(\cdot)^\top \theta_{T+1}$

Procedure 3 $\textsc{MarkovSampler}(\xi, \omega, \pi, \tau)$

Require: $(\pi, \omega) \in \Pi \times \mathcal{W}$, τ, $\xi \in \mathcal{Z} \times \mathcal{Z}$

1: $\mathtt{Z}_0 = \xi(0), \mathtt{Z}_1 = \xi(1)$
2: **for** $i = 0, \ldots, \tau$ **do**
3: $\quad$ Sample $\mathtt{K}_i \sim \omega(\mathtt{S}_i)$, $\mathtt{A}_i \sim \pi(\cdot|\mathtt{S}_i)$
4: $\quad$ Sample $\mathtt{S}_{i+1} \sim P_{\mathtt{K}_i}(\cdot|\mathtt{S}_i, \mathtt{A}_i)$
5: $\quad$ Set $\mathtt{Z}_i = (\mathtt{S}_i, \mathtt{A}_i, \mathtt{K}_i)$
6: **end for**
7: **return** $(\mathtt{Z}_{\tau-1}, \mathtt{Z}_{\tau})$

To proceed, let us denote

$$\varepsilon_{\text{approx}} = \sup_{\pi \in \Pi, \omega \in \mathcal{W}} \|Q_{\theta_\pi^\omega} - Q_\pi^\omega\|_\infty$$

as the approximation error of parameterizing Q_π^ω with $Q_{\theta_\pi^\omega}$. In addition, let

$$R = \sup_{\pi \in \Pi, \omega \in \mathcal{W}} \|\theta_\pi^\omega\|, \; U = 2R + 1.$$

From (22) and Assumption 1, it can be readily verified that θ_π^ω is continuous with respect to ω, consequently $R < \infty$ given $\mathcal{W}$ being compact.

We now proceed to discuss some immediate implications of Assumption 1. To facilitate our discussion, let us define $\mathcal{K} : \mathcal{Z} \to \mathbb{R}^{d \times d}$ through

$$\mathcal{K}(z) = \phi(z) \left[\phi(z)^\top - \gamma \int_{\mathcal{S}} \sum_{a' \in \mathcal{A}} \sum_{k' \in \mathcal{K}} \pi(a'|s') \omega(k'|s') \phi(s', a', k')^\top P_k(\mathrm{d}s'|s, a) \right] \quad \forall z \in \mathcal{Z}.$$

We begin by presenting some basic properties of operator $F(\cdot)$ defined in (23).

Lemma 3.5. *Suppose Assumption 1 holds. Denote $\mu = \lambda_{\min}(\Sigma)(1-\gamma)$ and $L = \lambda_{\max}(\Sigma)(1+\gamma)$. Then, we have*

(i) $\|F(\theta)\| \le L\|\theta - \theta_\pi^\omega\|$ *for any* $\theta \in \mathbb{R}^d$,

(ii) $\langle F(\theta), \theta - \theta_\pi^\omega \rangle \ge \mu \|\theta - \theta_\pi^\omega\|^2$ *for any* $\theta \in \mathbb{R}^d$.

Proof. For Assertion (i), observe first that

$$\begin{aligned} \int_{\mathcal{Z}} \mathcal{K}(z) \, \nu_\pi^\omega(\mathrm{d}z) &= \Sigma - \gamma \int_{\mathcal{Z} \times \mathcal{S}} \sum_{a' \in \mathcal{A}} \sum_{k' \in \mathcal{K}} \pi(a'|s') \omega(k'|s') \phi(z) \phi(z')^\top P_k(\mathrm{d}s'|s, a) \nu_\pi^\omega(\mathrm{d}z) \\ &= \Sigma - \gamma \int_{\mathcal{Z}} \phi(z') \phi(z')^\top \, \nu_\pi^\omega(\mathrm{d}z') = (1-\gamma)\Sigma, \end{aligned} \tag{26}$$

where the first and last equalities follow from the definition of Σ, and the second equality holds because ν_π^ω is the steady-state distribution. We then have

$$F(\theta) = F(\theta) - F(\theta_\pi^\omega) = \int_{\mathcal{Z}} \mathcal{K}(z) \, \nu_\pi^\omega(\mathrm{d}z)(\theta - \theta_\pi^\omega) = (1-\gamma)\Sigma(\theta - \theta_\pi^\omega), \tag{27}$$

where the first equality holds because $F(\theta_\pi^\omega) = 0$, the second equality follows from linearity of F in θ, and the third equality exploits (26). Assertion (i) then follows by definition of L. For Assertion (ii), we have

$$\langle F(\theta), \theta - \theta_\pi^\omega \rangle = (\theta - \theta_\pi^\omega)^\top (1-\gamma)\Sigma(\theta - \theta_\pi^\omega) \ge \mu \|\theta - \theta_\pi^\omega\|^2,$$

where the first equality follows from (27), and the inequality follows by definition of μ. □

Let us define the stochastic error associated with $\widehat{F}(\theta_t, \xi_t^\tau)$ in (24) through

$$\zeta_t = \widehat{F}(\theta_t, \xi_t^\tau) - F(\theta_t). \tag{28}$$

Lemma 3.6 below establishes that the bias of the estimator $\widehat{F}(\theta_t, \xi_t^\tau)$ decays exponentially fast in τ under Assumption 1.

Lemma 3.6. *Suppose that Assumption 1 holds. We then have*

$$\left\| F(\theta_t) - \mathbb{E}_\pi^\omega\left[\widehat{F}(\theta_t, \xi_t^\tau) \mid \mathcal{F}_{t-1}\right]\right\| \leq (1+\gamma) C \rho^\tau \|\theta_t - \theta_\pi^\omega\| \quad \mathbb{P}_\pi^\omega\text{-a.s.}.$$

Proof. Denote by $\nu_{\tau|t} \in \mathcal{P}(\mathcal{Z})$ the conditional distribution of $Z_{t+\tau}$ given $\mathcal{F}_{t-1}$, that is, for every Borel set $\mathcal{B}$, $\mathbb{P}_\pi^\omega[Z_{t+\tau} \in \mathcal{B} \mid \mathcal{F}_{t-1}] = \int_{\mathcal{B}} \nu_{\tau|t}(\mathrm{d}z)$. We then have

$$\begin{aligned}
\left\| F(\theta_t) - \mathbb{E}_\pi^\omega\left[\widehat{F}(\theta_t, \xi_t^\tau) \mid \mathcal{F}_{t-1}\right]\right\| &= \left\| \int_{\mathcal{Z}} \mathcal{K}(z) \left(\nu_\pi^\omega(\mathrm{d}z) - \nu_{\tau|t}(\mathrm{d}z)\right)(\theta_t - \theta_\pi^\omega)\right\| \\
&\leq \left\| \int_{\mathcal{Z}} \mathcal{K}(z) \left(\nu_\pi^\omega(\mathrm{d}z) - \nu_{\tau|t}(\mathrm{d}z)\right)\right\|_{\mathrm{op}} \|\theta_t - \theta_\pi^\omega\| \\
&\leq \sup_{z \in \mathcal{Z}} \|\mathcal{K}(z)\|_{\mathrm{op}} \, \|\nu_\pi^\omega - \nu_{\tau|t}\|_{\mathrm{TV}} \, \|\theta_t - \theta_\pi^\omega\| \\
&\leq \sup_{z, z' \in \mathcal{Z}} \|\phi(z)\|(\|\phi(z)\| + \gamma\|\phi(z')\|) \, \|\nu_\pi^\omega - \nu_{\tau|t}\|_{\mathrm{TV}} \, \|\theta_t - \theta_\pi^\omega\| \\
&\leq (1+\gamma)\|\nu_\pi^\omega - \nu_{\tau|t}\|_{\mathrm{TV}} \, \|\theta_t - \theta_\pi^\omega\| \leq (1+\gamma) C \rho^\tau \|\theta_t - \theta_\pi^\omega\|,
\end{aligned}$$

where the first and the third inequalities follow from Cauchy-Schwarz inequality, the second inequality uses Hölder's inequality, and the fourth inequality holds because $\|\phi(z)\| \leq 1$ for all $z \in \mathcal{Z}$. Finally, the last inequality exploits Assumption 1. Thus, the claim follows. □

We are ready to discuss the convergence behavior of Algorithm 2, which will help us specify parameters (η, T) in the TD learning procedure that in turn certifies Condition 1. We begin by first establishing that the bias of estimate θ_t exhibits linear convergence.

Lemma 3.7. *Suppose that Assumption 1 holds, set* $\tau > (\log(\mu) - \log(C(1+\gamma)))/\log(\rho)$ *and define* $\widetilde{\mu} = \mu - (1+\gamma)C\rho^\tau > 0$. *If* $\eta \leq \widetilde{\mu}/L^2$, *then*

$$\|\mathbb{E}_\pi^\omega[\theta_t] - \theta_\pi^\omega\|^2 \leq (1 - \eta\widetilde{\mu})^t \|\theta_0 - \theta_\pi^\omega\|^2.$$

Proof. We have

$$\begin{aligned}
&\|\mathbb{E}_\pi^\omega[\theta_{t+1}] - \theta_\pi^\omega\|^2 \\
&= \|\mathbb{E}_\pi^\omega[\theta_t] - \eta F(\mathbb{E}_\pi^\omega[\theta_t]) - \eta\mathbb{E}_\pi^\omega[\zeta_t] - \theta_\pi^\omega\|^2 \\
&\leq \|\mathbb{E}_\pi^\omega[\theta_t] - \theta_\pi^\omega\|^2 - 2\eta\langle F(\mathbb{E}_\pi^\omega[\theta_t]), \mathbb{E}_\pi^\omega[\theta_t] - \theta_\pi^\omega\rangle - 2\eta\langle \mathbb{E}_\pi^\omega[\zeta_t], \mathbb{E}_\pi^\omega[\theta_t] - \theta_\pi^\omega\rangle + \eta^2\|F(\mathbb{E}_\pi^\omega[\theta_t])\|^2 \\
&\leq \|\mathbb{E}_\pi^\omega[\theta_t] - \theta_\pi^\omega\|^2 - 2\mu\eta\|\mathbb{E}_\pi^\omega[\theta_t] - \theta_\pi^\omega\|^2 + 2\eta(1+\gamma)C\rho^\tau\|\mathbb{E}_\pi^\omega[\theta_t] - \theta_\pi^\omega\|^2 + L^2\eta^2\|\mathbb{E}_\pi^\omega[\theta_t] - \theta_\pi^\omega\|^2 \\
&\leq (1 - 2\eta\widetilde{\mu} + L^2\eta^2)\|\mathbb{E}_\pi^\omega[\theta_t] - \theta_\pi^\omega\|^2,
\end{aligned}$$

where the first equality follows by the update rule in Line 4 of Algorithm 2, and the second inequality holds because of Lemma 3.5(ii), Lemma 3.6, and Lemma 3.5(i). The claim then follows from the choice of η. □

As will be clarified later in Theorem 3.10, Lemma 3.7 will be useful in showing that the output of Algorithm 2 satisfies the first inequality in (14a) of Condition 1. In addition to the fast bias reduction, we proceed to establish that the distance to the optimal solution remains bounded in Algorithm 2.

Lemma 3.8. *Suppose that Assumption 1 holds, set* $\tau > (\log(\mu) - \log(C(1+\gamma)))/\log(\rho)$ *and define* $\widetilde{\mu} = \mu - (1+\gamma)C\rho^\tau$. *If* $\eta \leq \mu/8$, *then*

$$\|\theta_{t+1} - \theta_\pi^\omega\|^2 \leq \|\theta_t - \theta_\pi^\omega\|^2 + 2\eta^2 U^2 + 2\eta\left\langle \zeta_t, \theta_t - \theta_\pi^\omega\right\rangle.$$

If, in addition, $\eta \leq \widetilde{\mu}/8$, *then*

$$\mathbb{E}_\pi^\omega[\|\theta_t - \theta_\pi^\omega\|^2] \leq \|\theta_0 - \theta_\pi^\omega\|^2 + \frac{U^2\widetilde{\mu}^2}{64}.$$

Proof. Observe first that by construction of $\widehat{F}$, we have

$$\|\widehat{F}(\theta_t, \xi_t^\tau) - \widehat{F}(\theta_\pi^\omega, \xi_t^\tau)\| = \|\phi(S_{t+\tau}) - \gamma\phi(S_{t+\tau+1})\|\|\theta_t - \theta_\pi^\omega\|\|\phi(S_{t+\tau})\| \le 2\|\theta_t - \theta_\pi^\omega\|, \tag{29}$$

and

$$\|\widehat{F}(\theta_\pi^\omega, \xi_t^\tau)\| = \|(\langle\phi(S_{t+\tau}) - \gamma\phi(S_{t+\tau+1}), \theta_t\rangle - r(S_{t+\tau}, A_{t+\tau}))\phi(S_{t+\tau})\| \le 2\|\theta_\pi^\omega\| + 1 \le U. \tag{30}$$

We then have

$$\begin{aligned}\langle\widehat{F}(\theta_t, \xi_t^\tau), \theta_{t+1} - \theta_t\rangle &= \langle\widehat{F}(\theta_t, \xi_t^\tau) - \widehat{F}(\theta_\pi^\omega, \xi_t^\tau) + \widehat{F}(\theta_\pi^\omega, \xi_t^\tau), \theta_{t+1} - \theta_t\rangle \\ &\ge -2\|\theta_t - \theta_\pi^\omega\|\|\theta_{t+1} - \theta_t\| - U\|\theta_{t+1} - \theta_t\|,\end{aligned} \tag{31}$$

where the inequality follows from Cauchy-Schwarz inequality and the previous observations (29) and (30). We then obtain

$$\begin{aligned}\eta\langle\widehat{F}(\theta_t, Z_t^\tau), \theta_{t+1} - \theta_t\rangle + \frac{1}{2}\|\theta_{t+1} - \theta_t\|^2 &\ge \|\theta_{t+1} - \theta_t\|\left(-\eta(2\|\theta_t - \theta_\pi^\omega\| + U) + \frac{1}{2}\|\theta_{t+1} - \theta_t\|\right) \\ &\ge -\frac{1}{2}\eta^2(2\|\theta_t - \theta_\pi^\omega\| + U)^2 \ge -\eta^2(4\|\theta_t - \theta_\pi^\omega\|^2 + U^2),\end{aligned} \tag{32}$$

where the first inequality follows from (31), the second inequality follows from Young's inequality, and the third inequality follows from $(a+b)^2 \le 2a^2 + 2b^2$. Note that the update rule in Line 4 of Algorithm 2 implies

$$\eta\langle\widehat{F}(\theta_t, Z_t^\tau), \theta_t - \theta_\pi^\omega\rangle + \eta\langle\widehat{F}(\theta_t, Z_t^\tau), \theta_{t+1} - \theta_t\rangle + \frac{1}{2}\|\theta_{t+1} - \theta_t\|^2 = \frac{1}{2}\|\theta_\pi^\omega - \theta_t\|^2 - \frac{1}{2}\|\theta_\pi^\omega - \theta_{t+1}\|^2.$$

Lower bounding the left-hand side of the above expression using (32) then yields

$$\eta\langle\widehat{F}(\theta_t, \xi_t^\tau), \theta_t - \theta_\pi^\omega\rangle - 4\eta^2\|\theta_t - \theta_\pi^\omega\|^2 - U^2\eta^2 \le \frac{1}{2}\|\theta_\pi^\omega - \theta_t\|^2 - \frac{1}{2}\|\theta_\pi^\omega - \theta_{t+1}\|^2.$$

Rearranging terms in the above expression further shows

$$\begin{aligned}\frac{1}{2}\|\theta_\pi^\omega - \theta_{t+1}\|^2 &\le \left(\frac{1}{2} + 4\eta^2\right)\|\theta_t - \theta_\pi^\omega\|^2 + U^2\eta^2 - \eta\langle\zeta_t, \theta_t - \theta_\pi^\omega\rangle - \eta\langle F(\theta_t), \theta_t - \theta_\pi^\omega\rangle \\ &\le \left(\frac{1}{2} - \eta\mu + 4\eta^2\right)\|\theta_t - \theta_\pi^\omega\|^2 + U^2\eta^2 - \eta\langle\zeta_t, \theta_t - \theta_\pi^\omega\rangle,\end{aligned}$$

where the second inequality follows from Lemma 3.5(ii). The first claim then follows by noting that $\eta \le \widetilde{\mu}/8$. Taking expectation and multiplying by 2 of the above expression yields

$$\begin{aligned}\mathbb{E}_\pi^\omega[\|\theta_\pi^\omega - \theta_{t+1}\|^2] &\le \left(1 - 2\eta\mu + 8\eta^2\right)\|\theta_t - \theta_\pi^\omega\|^2 + 2U^2\eta^2 - 2\eta\langle\mathbb{E}_\pi^\omega[\zeta_t], \theta_t - \theta_\pi^\omega\rangle \\ &\le \left(1 - 2\eta\widetilde{\mu} + 8\eta^2\right)\|\theta_t - \theta_\pi^\omega\|^2 + 2U^2\eta^2,\end{aligned}$$

where the second inequality follows from Lemma 3.6. Thus, the second claim also follows. □

We proceed to establish the following high probability boundedness guarantee on the iterate θ_t generated by Algorithm 2. It should be noted that as the norm of the stochastic error ζ_t (defined in (28)) itself depends on the norm of iterate θ_t, one cannot simply invoke standard concentration argument to control its accumulation in a high probability sense. In the following proposition, we construct an approximate martingale sequence that in turn will help us bound the accumulated stochastic error in high probability, which might be of independent interest.

Proposition 3.9. *Suppose that Assumption 1 holds. Fix total iterations $T > 0$ and set $\theta_0 = 0$. For any $\delta \in (0,1)$, set $\tau > \max\{-\log(\sqrt{T}), \log(\mu) - \log(C(1+\gamma))\}/\log(\rho)$ and $\eta = \nu/\sqrt{T}$ with*

$$\nu = \min\left\{\frac{\widetilde{\mu}}{L^2+8}, \frac{1}{2U}, \frac{1}{4G}\right\}, \quad G = \max\left\{2(1+\gamma)C(R^2+1), 4\sqrt{\log\left(\frac{2T}{\delta}\right)}\left[(L+2)(R^2+1) + U(R+1)\right]\right\}.$$

Then with probability at least $1-\delta$, we have

$$\|\theta_t - \theta_\pi^\omega\|^2 \le R^2 + 1, \quad \forall t \le T.$$

Proof. Define random sequences $\{X_t = \langle \zeta_t, \theta_t - \theta_\pi^\omega \rangle\}, \{\widetilde{X}_t = \langle \zeta_t, \theta_t - \theta_\pi^\omega \rangle \mathbb{1}_{\mathcal{G}_t}\}$, where $\mathcal{G}_t = \left\{Y_t \le G\sqrt{t}\right\}$, and

$$Y_0 = \widetilde{Y}_0 = 0,\ Y_t = Y_{t-1} + X_{t-1},\ \widetilde{Y}_t = \widetilde{Y}_{t-1} + \widetilde{X}_{t-1}.$$

Conditioning on event $\mathcal{G}_t$ and define $M_{(\nu,G)} = R^2 + 2\nu^2 U^2 + 2\nu G$, recursively applying Lemma 3.8 and using the parameter choice of η yields

$$\|\theta_t - \theta_\pi^\omega\|^2 \le R^2 + 2t\eta^2 U^2 + 2\eta G\sqrt{t} = R^2 + 2\nu^2 U^2 + 2\nu G = M_{(\nu,G)}, \tag{33}$$

We proceed to show that the above bound happens with probability at least $1-\delta$ given a proper choice of (ν, G). First, note that

$$\begin{aligned}
\langle \zeta_t, \theta_t - \theta_\pi^\omega \rangle &= \langle \widehat{F}_t(\theta_t, \xi_t^\tau) - \widehat{F}_t(\theta_\pi^\omega, \xi_t^\tau), \theta_t - \theta_\pi^\omega \rangle - \langle F(\theta_t), \theta_t - \theta_\pi^\omega \rangle + \langle \widehat{F}_t(\theta_\pi^\omega, \xi_t^\tau), \theta_t - \theta_\pi^\omega \rangle \\
&\le \|\widehat{F}_t(\theta_t, \xi_t^\tau) - \widehat{F}_t(\theta_\pi^\omega, \xi_t^\tau)\| \|\theta_t - \theta_\pi^\omega\| + \|F(\theta_t)\| \|\theta_t - \theta_\pi^\omega\| + \|\widehat{F}_t(\theta_\pi^\omega, \xi_t^\tau)\| \|\theta_t - \theta_\pi^\omega\| \\
&\le (L+2)\|\theta_t - \theta_\pi^\omega\|^2 + U\|\theta_t - \theta_\pi^\omega\|,
\end{aligned}$$

where the first inequality follows from Cauchy-Schwarz inequality, and the second inequality uses the bounds (29), Lemma 3.5(i), and (30). We then have

$$\left|\widetilde{X}_t\right| = \left|\langle \zeta_t, \theta_t - \theta_\pi^\omega \rangle \mathbb{1}_{\mathcal{G}_t}\right| \le (L+2)M_{(\nu,G)} + U\sqrt{M_{(\nu,G)}}.$$

In addition, observe that

$$\left|\mathbb{E}_\pi^\omega[\widetilde{X}_t | \mathcal{F}_{t-1}]\right| = \left|\mathbb{E}_\pi^\omega[\langle \zeta_t, \theta_t - \theta_\pi^\omega \rangle] \mathbb{1}_{\mathcal{G}_t}\right| \le (1+\gamma)C\rho^\tau \left\|\theta_t - \theta_\pi^\omega\right\|^2 \mathbb{1}_{\mathcal{G}_t} \le (1+\gamma)C\rho^\tau M_{(\nu,G)}, \tag{34}$$

where the equality holds because $\mathbb{1}_{\mathcal{G}_t}$ is $\mathcal{F}_{t-1}$ measurable, the first inequality follows from Lemma 3.6, and the second inequality uses (33). Let $b = (L+2)M_{(\nu,G)} + U\sqrt{M_{(\nu,G)}}$. Chernoff bound then implies

$$\begin{aligned}
\mathbb{P}\left(\widetilde{Y}_t \ge x\right) &\le \min_{\lambda>0} \exp(-\lambda x) \cdot \mathbb{E}_\pi^\omega \left[\exp\left(\sum_{i=0}^{t-1} \lambda\left(\widetilde{Y}_{i+1} - \widetilde{Y}_i\right)\right)\right] \\
&= \min_{\lambda>0} \exp(-\lambda x) \cdot \mathbb{E}_\pi^\omega \left[\exp\left(\sum_{i=0}^{t-2} \lambda\left(\widetilde{Y}_{i+1} - \widetilde{Y}_i\right)\right) \cdot \mathbb{E}_\pi^\omega\left[\lambda\left(\widetilde{Y}_t - \widetilde{Y}_{t-1}\right) \mid \mathcal{F}_{t-1}\right]\right] \\
&\le \min_{\lambda>0} \exp(-\lambda x) \cdot \mathbb{E}_\pi^\omega \left[\exp\left(\sum_{i=0}^{t-2} \lambda\left(\widetilde{Y}_{i+1} - \widetilde{Y}_i\right)\right)\right] \cdot \exp\left(\lambda(1+\gamma)C\rho^\tau M_{(\nu,G)} + \frac{b^2\lambda^2}{2}\right) \\
&\le \min_{\lambda>0} \exp(-\lambda x) \cdot \exp\left(t\lambda(1+\gamma)C\rho^\tau M_{(\nu,G)} + \frac{tb^2\lambda^2}{2}\right) \\
&= \exp\left(\min_{\lambda>0} -(x - t(1+\gamma)C\rho^\tau M_{(\nu,G)})\lambda + \frac{tb^2\lambda^2}{2}\right) = \exp\left(-\frac{(x - t(1+\gamma)C\rho^\tau M_{(\nu,G)})^2}{2tb^2}\right),
\end{aligned}$$

where the second inequality follows from Hoeffding's lemma and (34), and the third inequality holds by recursive application of the second inequality. Thus, for any $\delta \in (0,1)$, by applying the union bound over $1 \le t \le T$ in the above inequality, we have

$$\widetilde{Y}_t \le t(1+\gamma)C\rho^\tau M_{(\nu,G)} + 2\left((L+2)M_{(\nu,G)} + U\sqrt{M_{(\nu,G)}}\right)\sqrt{\log\left(\frac{2T}{\delta}\right)t}, \quad \forall t \in [T],$$

with probability at least $1-\delta$. By construction of τ, G, and ν, we have $t(1+\gamma)C\rho^\tau M_{(\nu,G)} \le G\sqrt{t}/2$, $2(L+2)M_{(\nu,G)}\sqrt{\log\left(\frac{2T}{\delta}\right)t} \le G\sqrt{t}/4$, together with $2U\sqrt{M_{(\nu,G)}}\sqrt{\log\left(\frac{2T}{\delta}\right)t} \le G\sqrt{t}/4$. Thus, $\widetilde{Y}_t \le G\sqrt{t}$ for all $t \in [T]$ with probability $1-\delta$. Denote by $\mathcal{G} = \{\widetilde{Y}_t \le G\sqrt{t}\ \forall t \in [T]\}$. We now use induction to show that $Y_t = \widetilde{Y}_t$ over $\mathcal{G}$ for all $t \le T$. Note that the claim holds trivially at $t = 0$. Suppose that the claim holds at iteration $t \ge 0$, then for any $\omega \in \mathcal{G}$, we have

$$Y_{t+1}(\omega) = Y_t(\omega) + X_t(\omega) = Y_t(\omega) + X_t(\omega)\mathbb{1}_{\left\{\widetilde{Y}_t \le G\sqrt{t}\right\}}(\omega)$$

$$= Y_t(\omega) + X_t(\omega)\mathbb{1}_{\{Y_t \le G\sqrt{t}\}}(\omega) = Y_t(\omega) + \widetilde{X}_t(\omega) = \widetilde{Y}_{t+1}(\omega),$$

where the second equality follows from the definition of $\mathcal{G}$, the third equality exploits the induction hypothesis, whereas the fourth and fifth equality follow from the definitions of $\widetilde{X}_t$ and $\widetilde{Y}_{t+1}$, respectively. The induction is then complete. We may now deduce that

$$\|\theta_t - \theta_\pi^\omega\|^2 \le M_{(\nu,G)} = R^2 + 2\nu^2 U^2 + 2\nu G \le R^2 + 1 \quad \forall t \le T,$$

where the first inequality follows from the induction claim, the equality holds by definition of $M_{(\nu,G)}$, and the second inequality uses the choice of ν that $\nu \le \min\{1/(2U), 1/(4G)\}$. Hence, the claim follows. □

We are now ready to show that with proper specification of parameters (η, T), Algorithm 2 indeed produces a solution that certifies Condition 1.

Theorem 3.10. *Under the same setup of Proposition 3.9, with probability at least $1-\delta$, the following hold for all $t \le T$.*

(i) $\|\mathbb{E}_\pi^\omega[Q_{\theta_t}] - Q_\pi^\omega\|_\infty \le (1-\eta\widetilde{\mu})^{t/2} R + \varepsilon_{\text{approx}}$;

(ii) $\|Q_{\theta_t}\|_\infty \le R + 1 + \varepsilon_{\text{approx}}$;

(iii) $\mathbb{E}_\pi^\omega[\|Q_{\theta_t} - Q_\pi^\omega\|_\infty^2] \le R^2 + \dfrac{U^2\widetilde{\mu}^2}{64} + \varepsilon_{\text{approx}}^2$.

Proof. For Assertion (i), we have

$$\begin{aligned}
\|\mathbb{E}_\pi^\omega[Q_{\theta_t}] - Q_\pi^\omega\|_\infty &\le \|\mathbb{E}_\pi^\omega[Q_{\theta_t}] - Q_{\theta_\pi^\omega}\|_\infty + \|Q_{\theta_\pi^\omega} - Q_\pi^\omega\|_\infty \\
&\le \sup_{z\in\mathcal{Z}} \|\phi(z)^\top \mathbb{E}_\pi^\omega[\theta_t] - \phi(z)^\top \theta_\pi^\omega\| + \varepsilon_{\text{approx}} \\
&\le \sup_{z\in\mathcal{Z}} \|\phi(z)\| \|\mathbb{E}_\pi^\omega[\theta_t] - \theta_\pi^\omega\| + \varepsilon_{\text{approx}} \le (1-\eta\widetilde{\mu})^{t/2} R + \varepsilon_{\text{approx}},
\end{aligned}$$

where the second inequality holds by the definition of $\varepsilon_{\text{approx}}$, and the last inequality follows from the assumption that $\|\phi(z)\| \le 1$ for all $z \in \mathcal{Z}$ and Lemma 3.7. For Assertion (ii), we have

$$\begin{aligned}
\|Q_{\theta_t}\|_\infty &\le \|Q_{\theta_t} - Q_{\theta_\pi^\omega}\|_\infty + \|Q_{\theta_\pi^\omega} - Q_\pi^\omega\|_\infty \\
&\le \sup_{z\in\mathcal{Z}} \|\phi(z)^\top \theta_t - \phi(z)^\top \theta_\pi^\omega\| + \varepsilon_{\text{approx}} \\
&\le \sup_{z\in\mathcal{Z}} \|\phi(z)\| \|\theta_t - \theta_\pi^\omega\| + \varepsilon_{\text{approx}} \le \sqrt{R^2+1} + \varepsilon_{\text{approx}} \le R + 1 + \varepsilon_{\text{approx}},
\end{aligned}$$

where the fourth inequality follows from Lemma 3.9. For Assertion (iii), observe that

$$\mathbb{E}_\pi^\omega[\|Q_{\theta_t} - Q_\pi^\omega\|_\infty^2] \le 2(\mathbb{E}_\pi^\omega[\|\theta_\pi^\omega - \theta_{t+1}\|^2] + \varepsilon_{\text{approx}}^2) \le R^2 + \frac{U^2\widetilde{\mu}^2}{64} + \varepsilon_{\text{approx}}^2,$$

where the second inequality uses Lemma 3.8. Hence, the claim follows. □

3.2 Sample Complexity for Robust Policy Evaluation

With Theorem 3.10 in place, we are now ready to establish the total number of samples required by Algorithm 1 to estimate the robust value function V_π of any policy $\pi \in \Pi$.

Theorem 3.11. *Suppose that Assumption 1 holds, fix total iterations $M > 0$ in Algorithm 1 and $\delta \in (0,1)$. Denote $B = R + 1 + \varepsilon_{\text{approx}}$. For any $\varepsilon \ge 8\varepsilon_{\text{approx}}/(1-\gamma)$, set the parameters of Algorithm 1 as*

$$\alpha_m = \sqrt{m}, \quad \lambda_m = \frac{(m+1)B}{2\sqrt{\log(|\mathcal{K}|)}}, \quad \forall m \le M,$$

where Algorithm 2 is instantiated with parameters

$$\tau > \max\{-\log(\sqrt{T}), \log(\mu) - \log(C(1+\gamma))\}/\log(\rho),\ \eta = \nu/\sqrt{T},\ \theta_0 = 0,\ T = \mathcal{O}\left(\frac{\log^2(B/\varepsilon)}{\nu^2\widetilde{\mu}^2}\right),$$

with

$$\nu = \min\left\{\frac{\widetilde{\mu}}{L^2+8}, \frac{1}{2U}, \frac{1}{4G}\right\},\ G = \max\left\{2(1+\gamma)C(R^2+1), 4\sqrt{\log\left(\frac{8MT}{\delta}\right)}\left[(L+2)(R^2+1)+U(R+1)\right]\right\}.$$

The total number of samples required by Algorithm 1 to output

$$-\varepsilon \le \widehat{V}_\pi(s) - V_\pi(s) \le \varepsilon, \quad -\varepsilon \le \widehat{Q}_\pi(s,a,k) - Q_\pi(s,a,k) \le \varepsilon \quad \forall (s,a,k) \in \mathcal{S}\times\mathcal{A}\times\mathcal{K}$$

with probability at least $1-\delta$ *is then bounded by*

$$\widetilde{\mathcal{O}}\left(\left(\frac{\bar{B}^2\log(|\mathcal{K}|)}{(1-\gamma)^2}\log\left(\frac{\bar{B}^2}{\delta(1-\gamma)^2\varepsilon^2}\right) + J\delta\right)\frac{1}{\widetilde{\mu}^2\nu^2\varepsilon^2}\right).$$

Proof. Note that the parameter choice clearly satisfies the conditions of Theorem 3.10, which in turn certifies the conditions of Theorem 3.4. In addition, observe that the total number of samples required is MT, where

$$M = \mathcal{O}\left(\frac{\bar{B}^2\log(|\mathcal{K}|)}{(1-\gamma)^2\varepsilon^2}\log\left(\frac{\bar{B}^2}{\delta(1-\gamma)^2\varepsilon^2}\right) + \frac{J\delta}{\varepsilon^2}\right)$$

is given by Theorem 3.4. Thus, the claim follows. □

A few remarks are in order before we conclude this section. First, in view of Theorem 3.11, Algorithm 1 requires $\widetilde{\mathcal{O}}(1/\varepsilon^2)$ number of samples to estimate the robust value function up to ε accuracy. In addition, the obtained accuracy certificate is stated in a high probability sense. This also appears to be the first robust policy evaluation method for continuous state robust MDPs with optimal sample complexity guarantees. Notably, in contrast to [48, 64], the proposed Algorithm 1 does not require any restrictive assumption on the radius of $\mathcal{P}$ or the discount factor. Second, although we consider action space being finite within this section, results developed here can be naturally extended to continuous action space with minor changes. Finally, in Section 4, we will utilize Algorithm 1 as a subroutine, and introduce an efficient policy optimization method for (3) that obtains the optimal sample complexity up to a logarithmic factor.

4 Robust Policy Optimization

With the development of robust policy evaluation in Section 3, we are now ready to introduce the proposed method for robust policy optimization problem (3). In view of the dynamic game formulation (6) in Section 2.1, the robust policy optimization problem (3) is equivalent to a dynamic zero-sum game between the controller and nature. In Algorithm 4, we present an approximate policy iteration method for solving (6). At the n-th iteration, the proposed method takes the form of

$$(\pi_{n+1}(\cdot|s), \omega_{n+1}(\cdot|s)) \leftarrow \max_{\pi(\cdot|s)\in\Delta_\mathcal{A}} \min_{\omega(\cdot|s)\in\Delta_\mathcal{K}} (\pi(\cdot|s))^\top \widehat{Q}_{\pi_n}(s)\omega(\cdot|s) \quad \forall s \in \mathcal{S}, \tag{35}$$

Algorithm 4 Approximate policy iteration

Require: Initial controller's policy $\pi_0 \in \Pi$, number of iterations N, error tolerance parameters $\Delta, \varepsilon > 0$, confidence level $1-\delta$

1: **for** $n = 0, 1, \ldots, N$ **do**
2: Find an approximate solution $(\pi_{n+1}(\cdot|s), \omega_{n+1}(\cdot|s))$ of (35) such that $(\pi(\cdot|s))^\top \widehat{Q}_{\pi_n}(s)\,\omega_{n+1}(\cdot|s) - (\pi_{n+1}(\cdot|s))^\top \widehat{Q}_{\pi_n}(s)\,\omega(\cdot|s) \le \Delta$ for all $(\pi(\cdot|s), \omega(\cdot|s)) \in \Delta_\mathcal{A}\times\Delta_\mathcal{K}$
3: Compute $\widehat{Q}_{\pi_{n+1}}$ using Algorithm 1 with input π_{n+1} such that $|\widehat{Q}_{\pi_{n+1}}(s,a,k) - Q_{\pi_{n+1}}(s,a,k)| \le \varepsilon$ for all $(s,a,k) \in \mathcal{S}\times\mathcal{A}\times\mathcal{K}$ with probability $1-\delta/(2N)$
4: **end for**

where $\widehat{Q}_{\pi_n}$ denotes an estimate of the robust joint action-value function defined in (16), and with a slight overload of notation we use $\widehat{Q}_{\pi_n}(s) \in \mathbb{R}^{|\mathcal{A}|\times|\mathcal{K}|}$ to denote the matrix defined as $\widehat{Q}_{\pi_n}(s)[a,k] = \widehat{Q}_{\pi_n}(s,a,k)$ for $(s,a,k) \in \mathcal{S}\times\mathcal{A}\times\mathcal{K}$. It is worth noting here that the estimate $\widehat{Q}_{\pi_n}$ is precisely constructed by invoking Algorithm 1 with parameters specified as in Theorem 3.11. In addition, we assume the access to a computational oracle that can produce $(\pi_{n+1}(\cdot|s), \omega_{n+1}(\cdot|s))$ whose duality gap is upper bounded by a prespecified precision $\Delta > 0$. Such an oracle can be instantiated by various off-the-shelf methods [5, 22, 34, 35]. Note that similar to (13) in Section 3, update (35) here defines the access to the updated controller's policy π_{n+1} via a bilinear matrix game, and hence there is no need to for explicitly solving (35) for every state $s \in \mathcal{S}$. Instead, $\pi_{n+1}(\cdot|s)$ is generated only when queried at state $s \in \mathcal{S}$ by Algorithm 1 in the robust policy evaluation step.

We start by summarizing some useful properties of the iterates generated by Algorithm 4. To facilitate our discussion, let us denote $\omega_n^\star$ as the optimal policy of nature's cost-minimizing MDP given controller's policy π_n. The existence of $\omega_n^\star$ follows from the dynamic equations of (6). Clearly, $\omega_n^\star$ can be chosen independent of the initial state in (6).

Lemma 4.1. *Fix total iteration number $N > 0$ in Algorithm 4 and let $\delta \in (0,1)$. At each iteration of Algorithm 4, the following hold for all $s \in \mathcal{S}$ with probability $1 - \delta/(2N)$.*

(i) $\left|(\pi(\cdot|s))^\top \widehat{Q}_{\pi_n}(s)\,\omega(\cdot|s) - (\pi(\cdot|s))^\top Q_{\pi_n}(s)\,\omega(\cdot|s)\right| \le \varepsilon$ *for all* $(\pi,\omega) \in \Pi\times\mathcal{W}$;

(ii) $(\pi(\cdot|s))^\top Q_{\pi_n}(s)\,\omega_{n+1}(\cdot|s) - (\pi_{n+1}(\cdot|s))^\top Q_{\pi_n}(s)\,\omega(\cdot|s) \le \Delta + 2\varepsilon$ *for all* $(\pi,\omega) \in \Pi\times\mathcal{W}$;

(iii) $(\pi_{n+1}(\cdot|s))^\top Q_{\pi_n}(s)\,\omega_{n+1}^\star(\cdot|s) - (\pi_n(\cdot|s))^\top Q_{\pi_n}(s)\,\omega_n^\star(\cdot|s) + \Delta + 2\varepsilon \ge 0$.

Proof. For Assertion (i), we have with probability $1-\delta/(2N)$ that

$$\begin{aligned}&\left|(\pi(\cdot|s))^\top \widehat{Q}_{\pi_n}(s)\omega(\cdot|s) - (\pi(\cdot|s))^\top Q_{\pi_n}(s)\omega(\cdot|s)\right| \\ &= \left|\sum_{a\in\mathcal{A}}\sum_{k\in\mathcal{K}}\left(\widehat{Q}_{\pi_n}(s,a,k) - Q_{\pi_n}(s,a,k)\right)\omega(k|s)\pi(a|s)\right| \le \varepsilon \quad \forall(\pi,\omega)\in\Pi\times\mathcal{W},\ \forall s\in\mathcal{S},\end{aligned}$$

where the inequality readily follows from the condition in Line 3 of Algorithm 4, which is in turn satisfied by the output of Algorithm 1 (*cf.* Theorem 3.11). For Assertion (ii), we have with probability $1-\delta/(2N)$ that

$$\begin{aligned}&(\pi(\cdot|s))^\top Q_{\pi_n}(s)\,\omega_{n+1}(\cdot|s) - (\pi_{n+1}(\cdot|s))^\top Q_{\pi_n}(s)\,\omega(\cdot|s) \\ &\le (\pi(\cdot|s))^\top \widehat{Q}_{\pi_n}(s)\,\omega_{n+1}(\cdot|s) - (\pi_{n+1}(\cdot|s))^\top \widehat{Q}_{\pi_n}(s)\,\omega(\cdot|s) + 2\varepsilon \le \Delta + 2\varepsilon \quad \forall(\pi,\omega)\in\Pi\times\mathcal{W},\ \forall s\in\mathcal{S},\end{aligned}$$

where the first inequality follows from Assertion (i), and the second inequality applies the condition in Line 2 of Algorithm 4. For Assertion (iii), we have with probability $1-\delta/(2N)$ that

$$\begin{aligned}(\pi_n(\cdot|s))^\top Q_{\pi_n}(s)\,\omega_n^\star(\cdot|s) &\le (\pi_n(\cdot|s))^\top Q_{\pi_n}(s)\,\omega_{n+1}(\cdot|s) \\ &\le (\pi_{n+1}(\cdot|s))^\top Q_{\pi_n}(s)\,\omega_{n+1}^\star(\cdot|s) + \Delta + 2\varepsilon \quad \forall s\in\mathcal{S},\end{aligned}$$

where the first inequality holds by optimality of $\omega_n^\star$, and the second inequality follows by taking $(\pi,\omega) = (\pi_n, \omega_{n+1}^\star)$ in Assertion (ii). □

Remark 1. *It is interesting to note here that for the purpose of the convergence characterization for Algorithm 4, one has to control the estimation error $\|\widehat{Q}_{\pi_n} - Q_{\pi_n}\|_\infty$ either in expectation, or in high probability as stated in Line 3. We proceed with Theorem 3.11 that certifies high probability control, while noting that there seems to be no existing method that can control the error in expectation for continuous state spaces. It can be also readily seen that Lemma 4.1 and the ensuing Lemma 4.3 no longer hold for approximate policy iteration if one can only control the bias term $\mathbb{E}[\widehat{Q}_{\pi_n} - Q_{\pi_n}]$. This for instance is the approach taken in [1].*

We proceed by characterizing the difference of value functions between any pair of joint policies.

Lemma 4.2. *For any joint policies $(\pi,\omega),(\pi',\omega') \in \Pi\times\mathcal{W}$ and any $s\in\mathcal{S}$, we have*

$$V_\pi^\omega(s) - V_{\pi'}^{\omega'}(s) = \frac{1}{1-\gamma}\int_{\mathcal{S}}\left[(\pi(\cdot|s'))^\top Q_{\pi'}^{\omega'}(s')\,\omega(\cdot|s') - (\pi'(\cdot|s'))^\top Q_{\pi'}^{\omega'}(s')\,\omega'(\cdot|s')\right] d_\pi^\omega(\mathrm{d}s'|s).$$

Proof. We have

$$\begin{aligned}
V_\pi^\omega(s) - V_{\pi'}^{\omega'}(s) &= (\pi(\cdot|s))^\top Q_\pi^\omega(s)\,\omega(\cdot|s) - (\pi'(\cdot|s))^\top Q_{\pi'}^{\omega'}(s)\,\omega'(\cdot|s) \\
&= (\pi(\cdot|s))^\top \left(Q_\pi^\omega(s) - Q_{\pi'}^{\omega'}(s)\right)\omega(\cdot|s) + (\pi(\cdot|s))^\top Q_{\pi'}^{\omega'}(s)\,\omega(\cdot|s) - (\pi'(\cdot|s))^\top Q_{\pi'}^{\omega'}(s)\,\omega'(\cdot|s) \\
&= \gamma \sum_{a\in\mathcal{A}}\sum_{k\in\mathcal{K}} \int_{\mathcal{S}} [V_\pi^\omega(s') - V_{\pi'}^{\omega'}(s')]\pi(a|s)\omega(k|s)P_k(\mathrm{d}s'|s,a) \\
&\quad + (\pi(\cdot|s))^\top Q_{\pi'}^{\omega'}(s)\,\omega(\cdot|s) - (\pi'(\cdot|s))^\top Q_{\pi'}^{\omega'}(s)\,\omega'(\cdot|s),
\end{aligned}$$

where the third equality follows from Definition 2. The claim then follows by iteratively expanding $V_\pi^\omega(s') - V_{\pi'}^{\omega'}(s')$ and using the definition of d_π^ω in (7). □

We are now ready to establish the following convergence characterization for each update of Algorithm 4.

Lemma 4.3. *Fix total iteration number $N > 0$ in Algorithm 4 and set $\delta \in (0,1)$. Let $\rho \in \mathcal{P}(\mathcal{S})$ with $\mathrm{supp}(\rho) = \mathcal{S}$, and define $D = \sup_{(\pi,\omega)\in\Pi\times\mathcal{W}} \|\int_{\mathcal{S}} d_\pi^\omega(\cdot|s)\rho(\mathrm{d}s)/\rho\|_\infty$. Then, we have with probability at least $1-\delta/N$ that*

$$\int_{\mathcal{S}} \left[V_{\pi^\star}(s) - V_{\pi_{n+1}}(s)\right]\rho(\mathrm{d}s) \le \left(1 - \frac{1-\gamma}{D}\right)\int_{\mathcal{S}} \left[V_{\pi^\star}(s) - V_{\pi_n}(s)\right]\rho(\mathrm{d}s) + \frac{2D(\Delta+2\varepsilon)}{1-\gamma}.$$

Proof. From Lemma 4.2, we have with probability $1-\delta/(2N)$ that

$$\begin{aligned}
&V_{\pi_{n+1}}(s) - V_{\pi_n}(s) + \frac{1}{1-\gamma}(\Delta+2\varepsilon) \\
&= \frac{1}{1-\gamma}\int_{\mathcal{S}} \left[(\pi_{n+1}(\cdot|s'))^\top Q_{\pi_n}(s')\,\omega_{n+1}^\star(\cdot|s') - (\pi_n(\cdot|s'))^\top Q_{\pi_n}(s')\,\omega_n^\star(\cdot|s') + \Delta + 2\varepsilon\right] d_{\pi_{n+1}}^{\omega_{n+1}^\star}(\mathrm{d}s'|s) \\
&\ge \frac{d_{\pi_{n+1}}^{\omega_{n+1}^\star}(\{s\}|s)}{1-\gamma}\left[(\pi_{n+1}(\cdot|s'))^\top Q_{\pi_n}(s')\,\omega_{n+1}^\star(\cdot|s') - (\pi_n(\cdot|s'))^\top Q_{\pi_n}(s')\,\omega_n^\star(\cdot|s') + \Delta + 2\varepsilon\right] \\
&\ge (\pi_{n+1}(\cdot|s'))^\top Q_{\pi_n}(s')\,\omega_{n+1}^\star(\cdot|s') - (\pi_n(\cdot|s'))^\top Q_{\pi_n}(s')\,\omega_n^\star(\cdot|s') + \Delta + 2\varepsilon \ge 0,
\end{aligned} \tag{36}$$

where the first inequality applies Lemma 4.1, and the second inequality follows from the fact that $d_{\pi_{n+1}}^{\omega_{n+1}^\star}(\{s\}|s) \ge 1-\gamma$. Consequently, integrating both sides of above inequality with respect to $d_{\pi^\star}^{\omega_{n+1}}(\cdot|s)$ yields

$$\begin{aligned}
&\int_{\mathcal{S}} \left[V_{\pi_n}(s') - V_{\pi_{n+1}}(s')\right] d_{\pi^\star}^{\omega_{n+1}}(\mathrm{d}s'|s) \\
&\le \int_{\mathcal{S}} \left[(\pi_n(\cdot|s'))^\top Q_{\pi_n}(s')\,\omega_n^\star(\cdot|s') - (\pi_{n+1}(\cdot|s'))^\top Q_{\pi_n}(s')\,\omega_{n+1}^\star(\cdot|s')\right] d_{\pi^\star}^{\omega_{n+1}}(\mathrm{d}s'|s) + \frac{1}{1-\gamma}(\Delta+2\varepsilon).
\end{aligned} \tag{37}$$

On the other hand, we have

$$\begin{aligned}
(1-\gamma)(V_{\pi^\star}(s) - V_{\pi_n}(s)) &\le (1-\gamma)(V_{\pi^\star}^{\omega_{n+1}}(s) - V_{\pi_n}(s)) \\
&= \int_{\mathcal{S}} \left[(\pi^\star(\cdot|s'))^\top Q_{\pi_n}(s')\,\omega_{n+1}(\cdot|s') - (\pi_n(\cdot|s'))^\top Q_{\pi_n}(s')\,\omega_n^\star(\cdot|s')\right] d_{\pi^\star}^{\omega_{n+1}}(\mathrm{d}s'|s),
\end{aligned} \tag{38}$$

where the inequality holds from the definition of $V_{\pi^\star}(s)$, and the equality follows from Lemma 4.2. Summing up (37) and (38) and rearranging terms, we obtain with probability $1-\delta/N$ that

$$\begin{aligned}
&\int_{\mathcal{S}} \left[V_{\pi_n}(s') - V_{\pi_{n+1}}(s')\right] d_{\pi^\star}^{\omega_{n+1}}(\mathrm{d}s'|s) + (1-\gamma)(V_{\pi^\star}(s) - V_{\pi_n}(s)) \\
&\le \int_{\mathcal{S}} \left[(\pi^\star(\cdot|s'))^\top Q_{\pi_n}(s')\,\omega_{n+1}(\cdot|s') - (\pi_{n+1}(\cdot|s'))^\top Q_{\pi_n}(s')\,\omega_{n+1}^\star(\cdot|s')\right] d_{\pi^\star}^{\omega_{n+1}}(\mathrm{d}s'|s) + \frac{1}{1-\gamma}(\Delta+2\varepsilon) \\
&\le \Delta + 2\varepsilon + \frac{1}{1-\gamma}(\Delta+2\varepsilon),
\end{aligned} \tag{39}$$

where the second inequality follows from the observation that

$$(\pi^\star(\cdot|s))^\top Q_{\pi_n}(s)\,\omega_{n+1}(\cdot|s) - (\pi_{n+1}(\cdot|s))^\top Q_{\pi_n}(s)\,\omega^\star_{n+1}(\cdot|s) \le \Delta + 2\varepsilon \quad \forall s \in \mathcal{S},$$

which in turn follows from Lemma 4.1(ii) applied to $(\pi, \omega) = (\pi^\star, \omega^\star_{n+1})$. Finally, we obtain

$$\begin{aligned}
&D\int_{\mathcal{S}} \left[V_{\pi_n}(s) - V_{\pi_{n+1}}(s)\right] \rho(\mathrm{d}s) + (1-\gamma)\int_{\mathcal{S}} \left[V_{\pi^\star}(s) - V_{\pi_n}(s)\right] \rho(\mathrm{d}s) \\
&= D\int_{\mathcal{S}} \left[V_{\pi_n}(s) - V_{\pi_{n+1}}(s) - \frac{1}{1-\gamma}(\Delta + 2\varepsilon)\right] \rho(\mathrm{d}s) + \frac{D}{1-\gamma}(\Delta + 2\varepsilon) + (1-\gamma)\int_{\mathcal{S}} \left[V_{\pi^\star}(s) - V_{\pi_n}(s)\right] \rho(\mathrm{d}s) \\
&\le \iint_{\mathcal{S}\times\mathcal{S}} \left[V_{\pi_n}(s') - V_{\pi_{n+1}}(s') - \frac{1}{1-\gamma}(\Delta + 2\varepsilon)\right] d_{\pi^\star}^{\omega_{n+1}}(\mathrm{d}s'|s)\rho(\mathrm{d}s) \\
&\quad + \frac{D}{1-\gamma}(\Delta + 2\varepsilon) + (1-\gamma)\int_{\mathcal{S}} \left[V_{\pi^\star}(s) - V_{\pi_n}(s)\right] \rho(\mathrm{d}s) \\
&= \int_{\mathcal{S}} \left[\int_{\mathcal{S}} \left[V_{\pi_n}(s') - V_{\pi_{n+1}}(s')\right] d_{\pi^\star}^{\omega_{n+1}}(\mathrm{d}s'|s) + (1-\gamma)(V_{\pi^\star}(s) - V_{\pi_n}(s))\right] \rho(\mathrm{d}s) \\
&\quad - \frac{1}{1-\gamma}(\Delta + 2\varepsilon) + \frac{D}{1-\gamma}(\Delta + 2\varepsilon) \\
&= \int_{\mathcal{S}} \left[\Delta + 2\varepsilon + \frac{1}{1-\gamma}(\Delta + 2\varepsilon)\right] \rho(\mathrm{d}s) - \frac{1}{1-\gamma}(\Delta + 2\varepsilon) + \frac{D}{1-\gamma}(\Delta + 2\varepsilon) \\
&\le \Delta + 2\varepsilon + \frac{D}{1-\gamma}(\Delta + 2\varepsilon),
\end{aligned}$$

where the first inequality follows from (36) and $D \ge 1$ by construction, the second inequality holds because of (39). The claim thus follows after dividing both sides in the above expression by D and rearranging terms. □

With Lemma 4.3 in place, we are now ready to present the total sample complexity of Algorithm 4 for solving the robust policy optimization problem (3).

Theorem 4.4. *Let $\delta \in (0,1)$ and $\rho \in \mathcal{P}(\mathcal{S})$ with $\mathrm{supp}(\rho) = \mathcal{S}$. Define $D = \sup_{(\pi,\omega)\in\Pi\times\mathcal{W}} \| \int_{\mathcal{S}} d_\pi^\omega(\cdot|s)\rho(\mathrm{d}s)/\rho \|_\infty$. For any $\varepsilon \ge 8\varepsilon_{\mathrm{approx}}/(1-\gamma)$ and $\epsilon \ge 4D^2(\Delta+2\varepsilon)/(1-\gamma)^2$, run Algorithm 4 for $N = \mathcal{O}\left(\frac{D}{1-\gamma}\log\left(\frac{1}{(1-\gamma)\epsilon}\right)\right)$, where Algorithm 1 is executed with the same parameter settings as those specified in Theorem 3.11, except that δ is replaced by $\delta/(2N)$ therein. Then with probability at least $1-\delta$, the output π_N of Algorithm 4 satisfies*

$$\int_{\mathcal{S}} V_{\pi^\star}(s)\rho(\mathrm{d}s) - \int_{\mathcal{S}} V_{\pi_N}(s)\rho(\mathrm{d}s) \le \epsilon.$$

In addition, the total number of samples required by Algorithm 4 is bounded by

$$\widetilde{\mathcal{O}}\left(\left(\frac{\bar{B}^2\log(|\mathcal{K}|)}{(1-\gamma)^2}\log\left(\frac{\bar{B}^2}{\delta(1-\gamma)^2\varepsilon^2}\right) + J\delta\right)\frac{1}{\mu^2\nu^2\varepsilon^2}\right).$$

Proof. We have

$$\begin{aligned}
\int_{\mathcal{S}} \left[V_{\pi^\star}(s) - V_{\pi_N}(s)\right] \rho(\mathrm{d}s) &\le \left(1 - \frac{1-\gamma}{D}\right)^N \int_{\mathcal{S}} \left[V_{\pi^\star}(s) - V_{\pi_0}(s)\right] \rho(\mathrm{d}s) + \frac{2D^2(\Delta+2\varepsilon)}{(1-\gamma)^2}\left(1 - \left(1 - \frac{1-\gamma}{D}\right)^N\right) \\
&\le \left(1 - \frac{1-\gamma}{D}\right)^N \int_{\mathcal{S}} \left[V_{\pi^\star}(s) - V_{\pi_0}(s)\right] \rho(\mathrm{d}s) + \frac{2D^2(\Delta+2\varepsilon)}{(1-\gamma)^2} \le \epsilon,
\end{aligned}$$

where the first inequality follows by a recursive application of Lemma 4.3, and the third inequality uses the choice of N and the condition on ϵ. The claim then follows by invoking Theorem 3.11, which states that each iteration of Algorithm 4 requires at most

$$\widetilde{\mathcal{O}}\left(\left(\frac{\bar{B}^2\log(|\mathcal{K}|)}{(1-\gamma)^2}\log\left(\frac{\bar{B}^2}{\delta(1-\gamma)^2\varepsilon^2}\right) + J\delta\right)\frac{1}{\mu^2\nu^2\varepsilon^2}\right)$$

number of samples. □

In view of Theorem 4.4, the number of samples required by Algorithm 4 to find an ϵ-optimal policy for the robust policy optimization problem (3) is bounded by $\widetilde{\mathcal{O}}(1/\epsilon^2)$. Clearly, the obtained sample complexity bound is order-wise optimal up to a logarithmic factor. Notably the accuracy certificate of Algorithm 4 is also stated in a high probability sense.

Before we conclude our discussion, it might worth noting here that the proposed methods can be straightfowardly adapted for solving general infinite-horizon finite-action zero-sum Markov games with continuous state space, and all the performance guarantees we obtained extend directly to this setting. Zero-sum Markov games have received considerable recent attention. When the the model is known, dynamic programming methods and more recently policy gradient methods have been studied in [4, 39, 41, 46, 52, 61]. When the model is unknown, optimal $\mathcal{O}(1/\epsilon^2)$ sample complexities have been established with model-based methods [24, 59, 62] and model-free methods [1, 6]. On the other hand, it is unclear how the same order-wise optimal complexity bound could be obtained for continuous state MDP, as existing optimal methods do not admit direct generalization to continuous state spaces. There also exists a fruitful line of research on developing model-free methods for solving continuous state space zero-sum Markov games [63]. Yet it appears that the current development either obtains suboptimal sample complexity [63], unless for finite horizon problems [12, 15] or with additional assumption on the structure of the model [7, 57]. In this sense, the proposed method in this manuscript seems to be the first algorithm for infinite-horizon zero-sum Markov game with an order-wise optimal sample complexity of $\widetilde{\mathcal{O}}(1/\epsilon^2)$ in the continuous state spaces.

5 Concluding Remarks

In this manuscript, we study robust MDPs on continuous state spaces with structured s-rectangular ambiguity set. The proposed ambiguity set lies within the convex hull of unknown generating kernels, which are accessible only via samples. We propose a stochastic first-order method for robust policy evaluation, and establish its high probability convergence to the robust value function, from which we establish an $\widetilde{\mathcal{O}}(1/\epsilon^2)$ sample complexity. With the high probability accuracy certificate of robust policy evaluation, we introduce an approximate policy iteration method that finds an ϵ-optimal policy to the robust policy optimization problem with $\mathcal{O}(1/\epsilon^2)$ sample complexity. Notably the proposed methods operate independently on the size of the state space, and thus can be implemented efficiently for continuous state spaces. The obtained sample complexities also appear to be new for solving robust MDPs with continuous state spaces.

We discuss a few directions that might be worth future investigations. First, it is interesting to consider ambiguity sets that are, in some sense, generated nonlinearly from the generating kernels, *e.g.*, when the generation process involves a nonlinear parameterized function that is convex in its parameters. It would be also interesting to study whether similar methods could be developed for nonparametric ambiguity sets, for instance those generated by the Kullback–Leibler divergence or the Wasserstein distance. Lastly, instead of approximate policy iteration considered in this manuscript, it is also highly rewarding to directly design primal-dual based methods for robust MDPs in the stochastic setting.

References

[1] Ahmet Alacaoglu, Luca Viano, Niao He, and Volkan Cevher. A natural actor-critic framework for zero-sum Markov games. In *International Conference on Machine Learning*, pages 307–366, 2022.

[2] Jalaj Bhandari, Daniel Russo, and Raghav Singal. A finite time analysis of temporal difference learning with linear function approximation. In *Conference on Learning Theory*, pages 1691–1692, 2018.

[3] Stephen Boyd, Enzo Busseti, Steve Diamond, Ronald N Kahn, Kwangmoo Koh, Peter Nystrup, and Jan Speth. Multi-period trading via convex optimization. *Foundations and Trends® in Optimization*, 3(1):1–76, 2017.

[4] Shicong Cen, Yuejie Chi, Simon S Du, and Lin Xiao. Faster last-iterate convergence of policy optimization in zero-sum Markov games. *arXiv preprint arXiv:2210.01050*, 2022.

[5] Antonin Chambolle and Thomas Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*, 40(1):120–145, 2011.

[6] Zaiwei Chen, Kaiqing Zhang, Eric Mazumdar, Asuman Ozdaglar, and Adam Wierman. A finite-sample analysis of payoff-based independent learning in zero-sum stochastic games. In *Neural Information Processing Systems*, volume 36, pages 75826–75883, 2023.

[7] Zixiang Chen, Dongruo Zhou, and Quanquan Gu. Almost optimal algorithms for two-player zero-sum linear mixture Markov games. In *International Conference on Algorithmic Learning Theory*, pages 227–261, 2022.

[8] Erwin Coumans and Yunfei Bai. Pybullet, a python module for physics simulation for games, robotics and machine learning, 2016.

[9] Giuseppe Da Prato and Jerzy Zabczyk. *Stochastic Equations in Infinite Dimensions*. Cambridge University Press, 2014.

[10] Vineet Goyal and Julien Grand-Clement. Robust Markov decision process: Beyond rectangularity. *Mathematics of Operations Research*, 48(1):203–226, 2022.

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

[12] Baihe Huang, Jason D Lee, Zhaoran Wang, and Zhuoran Yang. Towards general function approximation in zero-sum Markov games. *arXiv preprint arXiv:2107.14702*, 2021.

[13] Christian D Hubbs, Hector D Perez, Owais Sarwar, Nikolaos V Sahinidis, Ignacio E Grossmann, and John M Wassick. OR-Gym: A reinforcement learning library for operations research problems. *arXiv preprint arXiv:2008.06319*, 2020.

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

[15] Chi Jin, Qinghua Liu, and Tiancheng Yu. The power of exploiter: Provable multi-agent RL in large state spaces. In *International Conference on Machine Learning*, pages 10251–10279, 2022.

[16] Caleb Ju and Guanghui Lan. Policy optimization over general state and action spaces. *arXiv preprint arXiv:2211.16715*, 2022.

[17] Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *International Conference on Machine Learning*, pages 267–274, 2002.

[18] Georgios Kotsalis, Guanghui Lan, and Tianjiao Li. Simple and optimal methods for stochastic variational inequalities, I: Operator extrapolation. *SIAM Journal on Optimization*, 32(3):2041–2073, 2022.

[19] Navdeep Kumar, Esther Derman, Matthieu Geist, Kfir Y Levy, and Shie Mannor. Policy gradient for rectangular robust Markov decision processes. In *Neural Information Processing Systems*, pages 59477–59501, 2023.

[20] Navdeep Kumar, Adarsh Gupta, Maxence Mohamed Elfatihi, Giorgia Ramponi, Kfir Yehuda Levy, and Shie Mannor. Dual formulation for non-rectangular L_p robust Markov decision processes. *arXiv preprint arXiv:2502.09432*, 2025.

[21] Navdeep Kumar, Kaixin Wang, Kfir Yehuda Levy, and Shie Mannor. Efficient value iteration for s-rectangular robust Markov decision processes. In *International Conference on Machine Learning*, pages 25682–25725, 2024.

[22] Guanghui Lan and Yan Li. A novel catalyst scheme for stochastic minimax optimization. *arXiv preprint arXiv:2311.02814*, 2023.

[23] Yann Le Tallec. *Robust, Risk-Sensitive, and Data-Driven Control of Markov Decision Processes*. PhD thesis, Massachusetts Institute of Technology, 2007.

[24] Gen Li, Yuejie Chi, Yuting Wei, and Yuxin Chen. Minimax-optimal multi-agent RL in Markov games with a generative model. In *Neural Information Processing Systems*, pages 15353–15367, 2022.

[25] Mengmeng Li, Daniel Kuhn, and Tobias Sutter. Policy gradient algorithms for robust MDPs with non-rectangular uncertainty sets. *arXiv preprint arXiv:2305.19004*, 2023.

[26] Mengmeng Li, Daniel Kuhn, and Tobias Sutter. Towards optimal offline reinforcement learning. *arXiv preprint arXiv:2503.12283*, 2025.

[27] Yan Li and Guanghui Lan. First-order policy optimization for robust policy evaluation. *arXiv preprint arXiv:2307.15890*, 2023.

[28] Yan Li and Alexander Shapiro. Rectangularity and duality of distributionally robust Markov decision processes. *arXiv preprint arXiv:2308.11139*, 2023.

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

[30] Zhongyu Li, Xuxin Cheng, Xue Bin Peng, Pieter Abbeel, Sergey Levine, Glen Berseth, and Koushil Sreenath. Reinforcement learning for robust parameterized locomotion control of bipedal robots. In *IEEE International Conference on Robotics and Automation*, pages 2811–2817, 2021.

[31] Zhenwei Lin, Chenyu Xue, Qi Deng, and Yinyu Ye. A single-loop robust policy gradient method for robust Markov decision processes. In *International Conference on Machine Learning*, pages 30392 – 30426, 2024.

[32] Zijian Liu, Qinxun Bai, Jose Blanchet, Perry Dong, Wei Xu, Zhengqing Zhou, and Zhengyuan Zhou. Distributionally robust Q-learning. In *International Conference on Machine Learning*, pages 13623–13643, 2022.

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

[34] Arkadi Nemirovski. Prox-method with rate of convergence $\mathcal{O}(1/t)$ for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1):229–251, 2004.

[35] Yu Nesterov. Smooth minimization of non-smooth functions. *Mathematical Programming*, 103(1):127–152, 2005.

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

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

[38] Kishan Panaganti, Zaiyan Xu, Dileep Kalathil, and Mohammad Ghavamzadeh. Robust reinforcement learning using offline data. In *Neural Information Processing Systems*, pages 32211–32224, 2022.

[39] Stephen D. Patek. *Stochastic Shortest Path Games: Theory and Algorithms*. PhD thesis, Massachusetts Institute of Technology, 1997.

[40] Xue Bin Peng, Marcin Andrychowicz, Wojciech Zaremba, and Pieter Abbeel. Sim-to-real transfer of robotic control with dynamics randomization. In *IEEE International Conference on Robotics and Automation*, pages 3803–3810, 2018.

[41] Julien Perolat, Bruno Scherrer, Bilal Piot, and Olivier Pietquin. Approximate dynamic programming for two-player zero-sum Markov games. In *International Conference on Machine Learning*, pages 1321–1329, 2015.

[42] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Neural Information Processing Systems*, 2007.

[43] Shyam Sundhar Ramesh, Pier Giuseppe Sessa, Yifan Hu, Andreas Krause, and Ilija Bogunovic. Distributionally robust model-based reinforcement learning with large state spaces. In *International Conference on Artificial Intelligence and Statistics*, pages 100–108, 2024.

[44] Alexander Shapiro. Distributionally robust stochastic programming. *SIAM Journal on Optimization*, 27(4):2258–2275, 2017.

[45] Laixi Shi, Gen Li, Yuting Wei, Yuxin Chen, Matthieu Geist, and Yuejie Chi. The curious price of distributional robustness in reinforcement learning with a generative model. In *Neural Information Processing Systems*, pages 79903–79917, 2023.

[46] Zhuoqing Song, Jason D Lee, and Zhuoran Yang. Can we find Nash equilibria at a linear rate in Markov games? *arXiv preprint arXiv:2303.03095*, 2023.

[47] Richard S Sutton. Learning to predict by the methods of temporal differences. *Machine Learning*, 3(1):9–44, 1988.

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

[49] Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 23–30, 2017.

[50] Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5026–5033, 2012.

[51] CT Ionescu Tulcea. Mesures dans les espaces produits. *Atti Accad. Naz. Lincei Rend*, 7:208–211, 1949.

[52] Jan Van der Wal. Discounted Markov games: Generalized policy iteration method. *Journal of Optimization Theory and Applications*, 25(1):125–138, 1978.

[53] Qiuhao Wang, Chin Pang Ho, and Marek Petrik. Policy gradient in robust MDPs with global convergence guarantee. In *International Conference on Machine Learning*, pages 35763–35797, 2023.

[54] Shengbo Wang, Nian Si, Jose Blanchet, and Zhengyuan Zhou. A finite sample complexity bound for distributionally robust Q-learning. In *International Conference on Artificial Intelligence and Statistics*, pages 3370–3398, 2023.

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

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

[57] Qiaomin Xie, Yudong Chen, Zhaoran Wang, and Zhuoran Yang. Learning zero-sum simultaneous-move Markov games using function approximation and correlated equilibrium. In *Conference on Learning Theory*, pages 3674–3682, 2020.

[58] Zaiyan Xu, Kishan Panaganti, and Dileep Kalathil. Improved sample complexity bounds for distributionally robust reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pages 9728–9754, 2023.

[59] Yuling Yan, Gen Li, Yuxin Chen, and Jianqing Fan. Model-based reinforcement learning for offline zero-sum Markov games. *Operations Research*, 72(6):2430–2445, 2024.

[60] Bin Yu. Rates of convergence for empirical processes of stationary mixing sequences. *The Annals of Probability*, 22(1):94–116, 1994.

[61] Sihan Zeng, Thinh Doan, and Justin Romberg. Regularized gradient descent ascent for two-player zero-sum Markov games. In *Neural Information Processing Systems*, pages 34546–34558, 2022.

[62] Kaiqing Zhang, Sham M Kakade, Tamer Basar, and Lin F Yang. Model-based multi-agent RL in zero-sum Markov games with near-optimal sample complexity. *Journal of Machine Learning Research*, 24(175):1–53, 2023.

[63] Yulai Zhao, Yuandong Tian, Jason Lee, and Simon Du. Provably efficient policy optimization for two-player zero-sum Markov games. In *International Conference on Artificial Intelligence and Statistics*, pages 2736–2761, 2022.

[64] Ruida Zhou, Tao Liu, Min Cheng, Dileep Kalathil, PR Kumar, and Chao Tian. Natural actor-critic for robust reinforcement learning with function approximation. In *Neural Information Processing Systems*, pages 97–133, 2023.

A Appendix

Lemma A.1. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $\{\mathcal{F}_m\}_{m\geq 0}$ be a filtration. Assume that the stochastic processes $\{X_m\}_{m\geq 0}$ and $\{\widehat{X}_m\}_{m\geq 0}$ with $X_m : \mathcal{S} \to [0, \bar{X}]$, $\widehat{X}_m : \mathcal{S} \to \mathbb{R}$ are such that X_m and $\widehat{X}_m$ are both $\mathcal{F}_m$-measurable for every $m \geq 0$. Fix a positive integer $M > 1$, and suppose that for any $\delta \in (0, 1)$, there exist $\varepsilon_Q, B, J > 0$, potentially dependent on δ, such that*

$$\|\mathbb{E}[\widehat{X}_m|\mathcal{F}_{m-1}] - X_m\|_\infty \leq \varepsilon_Q, \quad \|\widehat{X}_m\|_\infty \leq B, \quad \mathbb{E}\Big[\big\|\widehat{X}_m - X_m\big\|_\infty^2 \mid \mathcal{F}_{m-1}\Big] \leq J$$

hold with probability $1 - \delta/(2M)$. Then, for any $\varepsilon \geq \varepsilon_Q$, we have with probability $1 - \delta$ that

$$\left|\sum_{m=1}^{M} \vartheta_m(\widehat{X}_m(s) - X_m(s))\right| \leq \varepsilon + (B + \bar{X})\sqrt{2\sum_{m=1}^{M} \vartheta_m^2 \log\left(\frac{4M}{\delta}\right)} + \sqrt{\frac{2J\delta}{M}}.$$

Proof. Define a random sequence $\{Y_m\}_{m=0}^M$ through $Y_m = \widehat{X}_m - X_m$ for all $m \leq M$, and let $Y'_m(s) = Y_m(s)\mathbb{1}_{\{\|Y_m\|_\infty \leq B+\bar{X}\}}$ for all $m \leq M$ and $s \in \mathcal{S}$. We aim to establish an upper bound for $\mathbb{E}[Y'_m]$. To this end, observe first that

$$\mathbb{E}\Big[\big|Y_m(s)\mathbb{1}_{\{\|Y_m\|_\infty \geq B+\bar{X}\}}\big|\Big] \leq \sqrt{\mathbb{E}[\|Y_m\|^2] \cdot \mathbb{E}\Big[\big|\mathbb{1}_{\{\|Y_m\|_\infty \geq B+\bar{X}\}}\big|^2\Big]} \leq \sqrt{\frac{J\delta}{2M}}, \tag{40}$$

where the first inequality follows from Cauchy-Schwarz inequality, and the second inequality holds because of the assumptions $\mathbb{E}[\|\widehat{X}_m - X_m\|^2 \mid \mathcal{F}_{m-1}] \leq J$ together with $\mathbb{P}(\|\widehat{X}_m\|_\infty \leq B) \geq 1 - \delta/(2M)$. Next, note that

$$\begin{aligned}
\mathbb{E}[Y_m(s)] &= \mathbb{E}\Big[\big(\widehat{X}_m(s) - X_m(s)\big)\mathbb{1}_{\left\{\|\mathbb{E}[\widehat{X}_m|\mathcal{F}_{m-1}] - X_m\|_\infty \leq \varepsilon_Q\right\}} + \big(\widehat{X}_m(s) - X_m(s)\big)\mathbb{1}_{\left\{\|\mathbb{E}[\widehat{X}_m|\mathcal{F}_{m-1}] - X_m\|_\infty > \varepsilon_Q\right\}}\Big] \\
&\leq \mathbb{E}\Big[\big\|\mathbb{E}[\widehat{X}_m|\mathcal{F}_{m-1}] - X_m\big\|_\infty \mathbb{1}_{\left\{\|\mathbb{E}[\widehat{X}_m|\mathcal{F}_{m-1}] - X_m\|_\infty \leq \varepsilon_Q\right\}}\Big] \\
&\quad + \sqrt{\mathbb{E}\Big[\big\|\widehat{X}_m - X_m\big\|_\infty^2\Big]\mathbb{P}\Big(\|\mathbb{E}[\widehat{X}_m|\mathcal{F}_{m-1}] - X_m\|_\infty > \varepsilon_Q\Big)} \\
&\leq \varepsilon_Q + \sqrt{\frac{J\delta}{2M}},
\end{aligned} \tag{41}$$

where the first inequality holds by the law of iterated expectations and the Cauchy-Schwarz inequality, and the second inequality follows from the assumptions $\mathbb{E}[\|\widehat{X}_m - X_m\|^2 \mid \mathcal{F}_{m-1}] \leq J$ together with $\mathbb{P}(\|\widehat{X}_m\|_\infty \leq B) \geq 1 - \delta/(2M)$. We then have for every $m \leq M$ that

$$\mathbb{E}[Y'_m(s)] = \mathbb{E}\Big[Y_m(s) - Y_m(s)\mathbb{1}_{\{\|Y_m\|_\infty \geq B + \bar{X}\}}\Big] \leq \mathbb{E}\Big[Y_m(s)\Big] + \mathbb{E}\Big[\big|Y_m(s)\mathbb{1}_{\{\|Y_m\|_\infty \geq B + \bar{X}\}}\big|\Big] \leq \varepsilon_Q + \sqrt{\frac{2J\delta}{M}},$$

where the second inequality follows by combining the previous two observations (40) and (41). Applying Azuma's inequality to $\{\vartheta_m Y'_m(s)\}_{m \geq 0}$, we then obtain

$$\left|\sum_{m=1}^{M} \vartheta_m Y'_m(s)\right| \leq (B + \bar{X})\sqrt{2\sum_{m=1}^{M} \vartheta_m^2 \log\left(\frac{4M}{\delta}\right)} + \varepsilon_Q + \sqrt{\frac{2J\delta}{M}},$$

which holds with probability $1 - \delta/2$. Finally, since $\mathbb{P}(\|\widehat{X}_m\|_\infty \leq B) \geq 1 - \delta/(2M)$, we obtain

$$\mathbb{P}\left(\sum_{m=1}^{M} Y_m(s) \neq \sum_{m=1}^{M} Y'_m(s)\right) \leq \frac{\delta}{2}.$$

Hence, the claim follows. □