Robust Policy Mirror Descent for Controlling Uncertain Markov Decision Process [†]

Yan Li [‡] Tuo Zhao [§] Guanghui Lan [¶]

August 12, 2022

Abstract

We consider the problem of solving robust Markov decision process (MDP), which involves a set of discounted, finite state, finite action space MDPs with uncertain transition kernels. The goal of planning is to find a robust policy that optimizes the worst-case values against the transition uncertainties, and thus encompasses the standard MDP planning as a special case. For $(\mathbf{s}, \mathbf{a})$-rectangular uncertainty sets, we develop a policy-based first-order method, namely the robust policy mirror descent (RPMD), and establish an $\mathcal{O}(\log(1/\epsilon))$ and $\mathcal{O}(1/\epsilon)$ iteration complexity for finding an ϵ-optimal policy, with two increasing-stepsize schemes. The prior convergence of RPMD is applicable to any Bregman divergence, provided the policy space has bounded radius measured by the divergence when centering at the initial policy. Moreover, when the Bregman divergence corresponds to the squared euclidean distance, we establish an $\mathcal{O}\left(\max\left\{1/\epsilon, 1/(\eta\epsilon^2)\right\}\right)$ complexity of RPMD with any constant stepsize η. For a general class of Bregman divergences, a similar complexity is also established for RPMD with constant stepsizes, provided the uncertainty set satisfies the relative strong convexity. We further develop a stochastic variant of the robust policy mirror descent method, named SRPMD, when the first-order information is only available through online interactions with the nominal environment. For general Bregman divergences, we establish an $\mathcal{O}(1/\epsilon^2)$ and $\mathcal{O}(1/\epsilon^3)$ sample complexity with two increasing-stepsize schemes. For the euclidean Bregman divergence, we establish an $\mathcal{O}(1/\epsilon^3)$ sample complexity with constant stepsizes. To the best of our knowledge, all the aforementioned results appear to be new for policy-based first-order methods applied to the robust MDP problem.

1 Introduction

We consider the problem of solving the robust Markov decision process (MDP) where the transition kernel is uncertain, and one seeks to learn a policy that behaves robustly against such uncertainties. Specifically, a robust MDP $\mathcal{M}_{\mathcal{U}} := \{\mathcal{M}_u = (\mathcal{S}, \mathcal{A}, c, \mathbb{P}_u, \gamma) : u \in \mathcal{U}\}$ consists of a set of MDPs, where $\mathcal{S}$ and $\mathcal{A}$ denote the state and action space, respectively; $\mathbb{P}_u : \mathcal{S} \times \mathcal{A} \to [0, 1]$ denotes the transition kernel, indexed by $u \in \mathcal{U}$; $c : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ denotes the cost function, which we assume with loss of generality that $0 < c(s, a) \leq 1$ for all (s, a); γ denotes the discount factor. The standard value function $V_u^\pi : \mathcal{S} \to \mathbb{R}$ of a policy π with respect to MDP $\mathcal{M}_u$, is defined as

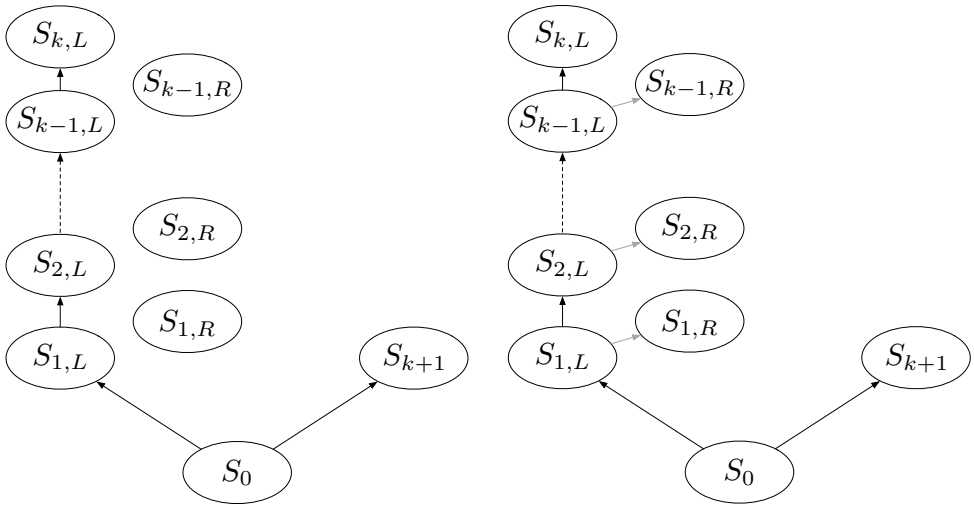
$$V_u^\pi(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t c(s_t, a_t) \big| s_0 = s, a_t \sim \pi(\cdot|s_t), s_{t+1} \sim \mathbb{P}_u(\cdot|s_t, a_t)\right], \quad \forall s \in \mathcal{S}.$$

[†] Work in progress. Initial version was released at https://gzliyan113.github.io/pdfs/papers/rpmd.pdf on August 12, 2022.

[‡] H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA, 30332. (E-mail: yli939@gatech.edu).

[§] H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA, 30332. (E-mail: tourzhao@gatech.edu).

[¶] H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA, 30332. (E-mail: george.lan@isye.gatech.edu).



(a) Nominal $\mathcal{M}$ with transition kernel $\mathbb{P}$. (b) $\mathcal{M}_u$ with transition kernel $\mathbb{P}_u \approx \mathbb{P}$.

Figure 1: A nominal MDP, and its approximate clone with small changes to the transition kernel.

Our end goal is to learn a policy π^* that is the solution of the following problem

$$\pi^* \in \cap_{s\in\mathcal{S}} \operatorname*{Argmin}_{\pi\in\Pi} \max_{u\in\mathcal{U}} V_u^\pi(s), \tag{1.1}$$

where Π denotes the set of all stationary and randomized policies. That is, (1.1) aims to learn a policy that minimizes the worst-case value simultaneously for every state. Clearly, when $\mathcal{U}$ is a singleton, the problem of finding robust policy (1.1) reduces to solving a standard MDP planning problem.

Before any technical discussions, we first construct a simple example motivating our study of finding a robust policy in the sense of (1.1), when facing transition uncertainty. Specifically, we will construct a pair of MDP $\mathcal{M}$ and $\mathcal{M}_u$, with the same $(\mathcal{S}, \mathcal{A}, c, \gamma)$, and the transition kernels of two are close to each other, yet the optimal policy for $\mathcal{M}$ achieves highly suboptimal value in $\mathcal{M}_u$. In contrast, we show that there exists a policy that achieves close-to-optimal performances in both $\mathcal{M}$ and $\mathcal{M}_u$.

Example 1.1 (Tradeoff between planning efficiency and robustness)**.** Consider the nominal MDP $\mathcal{M}$ illustrated in Figure 1a. Starting at S_0, the agent has two actions $\{L, R\}$, consisting of going either left or right. Going right incurs a cost of -1, going left incurs a cost of 0. We assume the cost occurs immediately after the action is made. Whenever the agent arrives at $S_{i,R}$ for $1 \le i \le k-1$ and S_{k+1}, the agent stays at the same place going forward.

Additionally, if the agent goes left from S_0, then in the ensuing rounds the agent has only one available action, which is to transit to the next state following the arrows. No cost is incurred until the agent transits from $S_{k-1,L}$ to $S_{k,L}$, which incurs a cost of $-(1+\epsilon)\gamma^{-k+1}$ for some small positive number $\epsilon \ll 1$ (e.g., $\epsilon = 0.01$).

Now consider another MDP $\mathcal{M}_u$ that is exact the same as $\mathcal{M}$, except in the transition kernel, illustrated in Figure 1b. In particular, for any $1 \le m \le k-1$, the transition changes to $\mathbb{P}_u(S_{m+1,L}|S_{m,L}) = p$, $\mathbb{P}_u(S_{m,R}|S_{m,L}) = 1-p$, for some $p \in (0,1)$ close to 1 (e.g., $p = 0.99$). Thus $\mathcal{M}_u$ can be viewed as an approximate copy of $\mathcal{M}$.

It should be clear that for MDP $\mathcal{M}$, going left at S_0 incurs a value of $-(1+\epsilon)$, which is close, but still strictly small than the value -1 of going right. Thus the optimal policy π^* for $\mathcal{M}$ is to always go left.

However, when deploying π^* in the slightly changed environment $\mathcal{M}_u$, one can clearly see that going left at S_0 now incurs a value of $-(1+\epsilon)p^{k-1}$, while going right still has the value of -1. For k large enough, we obverse that the value of going left approaches 0, which is significantly worse than the value of going right.

In conclusion, we see that going left at S_0 serves as the optimal policy in $\mathcal{M}$, but its performance degrades significantly despite being deployed at a similar environment $\mathcal{M}_u$. In fact, such a policy is even worse than the policy of randomly going left or right with equal probability. In contrast, going right at S_0 is an ϵ-optimal policy in $\mathcal{M}$, and also the optimal policy in $\mathcal{M}_u$, thus being a much more desirable choice in terms of robustness.

We remark that tensions between robustness and accuracy have been discussed extensively in the supervised learning literature [22,39,45,46]. Example 1.1 demonstrates that similar tensions between the planning efficiency and robustness also exists in the control of uncertain Markov decision process. It should be noted that the key ingredient for the lack of robustness in the MDP demonstrated in Figure 1 is the *reward (cost) sparseness*. This feature has been widely observed for practical MDP applications [24,28], and we believe the same mechanism can be one of the few important factors that lead to brittle robustness observed in existing reinforcement learning applications.

To proceed, we will focus on the case of $(\mathbf{s},\mathbf{a})$-rectangular uncertainty set, defined below.

Definition 1.1 ($(\mathbf{s},\mathbf{a})$-Rectangular Uncertainty)**.** We assume the transition kernel $\mathbb{P}_u$ for the MDP $\mathcal{M}_u$ takes the form of

$$\mathbb{P}_u(\cdot|s,a) = \mathbb{P}(\cdot|s,a) + u(\cdot|s,a),\ \forall (s,a) \in \mathcal{S}\times\mathcal{A}, \tag{1.2}$$

where $\mathbb{P}(\cdot|s,a)$ denotes the nominal transition kernel, and $u \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|\times|\mathcal{A}|}$ denotes the perturbation to the nominal transition kernel. The uncertainty is said to be rectangular if $\mathcal{U}$ satisfies

$$\mathcal{U} = \Pi_{s\in\mathcal{S},a\in\mathcal{A}}\mathcal{U}_{s,a}, \quad \text{where} \ \ \mathcal{U}_{s,a} = \{u(\cdot|s,a) : u \in \mathcal{U}\}.$$

We assume $\mathcal{U}$ is compact. In addition, we let

$$\mathcal{P}_{s,a} = \mathbb{P}(\cdot|s,a) + \mathcal{U}_{s,a} \tag{1.3}$$

denote the set of possible transition probabilities at $(s,a) \in \mathcal{S}\times\mathcal{A}$.

Remark 1.1. While one can also consider uncertain cost function c_u when modeling, our motivation to focus on modeling uncertain transitions is due to the observation that cost function is mostly an endogenous user choice, and thus seems less suitable to be modeled as an uncertainty.

From Definition 1.1, it should be also clear that Example 1.1 can be readily modeled into a robust MDP with a rectangular uncertainty set. We will also define the nominal environment of the robust MDP $\mathcal{M}_\mathcal{U}$ as follows, which will be useful for our discussions in the stochastic settings.

Definition 1.2 (Nominal Environment)**.** The nominal environment $\mathcal{M}_\mathrm{N}$ for a robust MDP problem $\mathcal{M}_\mathcal{U}$, is the MDP with transition kernel $\mathbb{P}_u$ specified in (1.2) with $u = \mathbf{0}$. We denote the transition kernel of the $\mathcal{M}_\mathrm{N}$ by $\mathbb{P}_\mathrm{N}$.

Given a policy π, we define its robust value function $V_r^\pi : \mathcal{S} \to \mathbb{R}$ as $V_r^\pi(s) = \max_{u\in\mathcal{U}} V_u^\pi(s)$ for all $s \in \mathcal{S}$. Consequently, solving (1.1) is equivalently to minimizing the robust value function:

$$\pi \in \cap_{s\in\mathcal{S}} \operatorname*{Argmin}_{\pi\in\Pi} V_r^\pi(s). \tag{1.4}$$

The existence of an optimal policy π^* solving (1.4) is well known in the literature [11,26], and we denote the set of optimal policies as $\Pi^* \subseteq \Pi$. Hence, we can succinctly reformulate (1.4) into a single objective optimization problem

$$\min_{\pi\in\Pi} \{f_\rho(\pi) \coloneqq \mathbb{E}_{s\sim\rho} V_r^\pi(s)\}, \tag{1.5}$$

where ρ is a nonnegative measure defined over the state space $\mathcal{S}$.

For any $u \in \mathcal{U}$, we also define the state-action value function of policy π with respect to $\mathcal{M}_u$ as

$$Q_u^\pi(s,a) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t c(s_t,a_t) \big| s_0 = s, a_0 = a, , a_t \sim \pi(\cdot|s_t), s_t \sim \mathbb{P}_u(\cdot|s_{t-1},a_{t-1}), t \geq 1\right], \quad \forall (s,a) \in \mathcal{S} \times \mathcal{A}.$$

Accordingly, the robust state-action value function $Q_r^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ as $Q_r^\pi(s,a) = \max_{u \in \mathcal{U}} Q_u^\pi(s,a)$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$. Note that from the definition of Q_u^π and V_u^π, we have

$$V_u^\pi(s) = \langle Q_u^\pi(s,\cdot), \pi(\cdot|s)\rangle \coloneqq \langle Q_u^\pi, \pi\rangle_s. \tag{1.6}$$

Given a policy π, and an uncertainty $u \in \mathcal{U}$, the discounted state visitation measure jointly induced by (π, u) is defined as

$$d_s^{\pi,u}(s') = (1-\gamma)\sum_{t=0}^{\infty} \gamma^t \mathbb{P}_u^\pi(s_t = s'|s_0 = s), \tag{1.7}$$

where $\mathbb{P}_u^\pi(s_t = s'|s_0 = s) \coloneqq \mathbb{P}\left(s_t = s'|s_0 = s, a_t = \pi(\cdot|s_t), s_{t+1} \sim \mathbb{P}_u(\cdot|s_t,a_t)\right)$ denotes the probability of reaching state s' at timestep t, given initial state s, and following policy π within MDP $\mathcal{M}_u$. Given any distribution ρ over $\mathcal{S}$, we define distribution $d_\rho^{\pi,u}$ over $\mathcal{S}$ as $d_\rho^{\pi,u}(s') = \mathbb{E}_{s\sim\rho} d_s^{\pi,u}(s')$. It is worth mentioning that the worst-case environment u_π defined in (2.2) can be non-unique. In this case, one can choose any of them by an arbitrary deterministic rule. For a finite set $\mathcal{X}$, we will denote $\Delta_{\mathcal{X}}$ as the $(|\mathcal{X}|-1)$-dimensional simplex.

Related Literature. Solving robust MDP (1.1) with rectangular uncertainty sets has been extensively studied in the dynamic programming literature. Among value-based methods, value iteration (VI) is known to achieve linear convergence to the optimal robust values [11, 26]. When the environment is unknown, sample-based value based methods [27, 32, 41], including robust Q-learning, have also been developed to directly learn the optimal value function. Policy-based methods, including the (modified) policy iteration (PI), have been studied in [10, 11, 14, 33, 43]. Approximate dynamic programming (ADP) techniques [29] for both type of methods have also been developed, which allow approximate computation of policy update and evaluation in PI [2, 38], or approximate Bellman update of VI [38]. The application of ADP to policy-based methods also enables function approximation to be used to handle the curse of dimensionality [2, 38], and using simulation-based methods to conduct approximate policy evaluation (e.g., robust TD learning [16, 32]). It should be noted there also exists another complementary line of research, on studying robust MDP with uncertainty set beyond $(\mathbf{s}, \mathbf{a})$-rectangularity [6, 8, 18, 43].

In addition to the prior developments in the context of dynamic programming, there has been a rising interests in developing first-order methods for solving the special case of (1.5), where there is no uncertainty in the transition (e.g., $\mathcal{U} = \{\mathbf{0}\}$). By using first-order information of objective (1.5) to update the policy, these policy-based methods are thus termed policy gradient methods (PGM), with their convergence behavior extensively studied in the literature. Sublinear convergence of the optimality gap for constant stepsize PGMs have been established in [1, 20], and linearly converging variants have been proposed in [3, 15, 20, 44], with local superlinear convergence studied in [15, 23]. [23] recently further characterizes the policy convergence of a PGM variant. Moreover, stochastic PGMs, which utilize sample to estimate the first-order information, have also been proposed in [20, 34, 47], and both sample and iteration complexity have been studied therein. Complementary to the policy-based first-order methods, [7] propose an accelerated first-order value-based method, and establish improved dependence on discount factor compared to value iteration.

In contrast to the aforementioned developments of PGMs for solving non-robust MDP, solving robust MDP (1.5) with first-order methods has been largely under-explored. Specifically, [9] propose a first-order value-based method derived based on value iteration, while [42] seems to be the only PGM variant to

date that directly aims to solve (1.5), which focuses on a subclass of polyhedral uncertainty. Given the abundant empirical observations on the unsatisfactory performance of PGM-trained RL agents when the deployment environment differs from the training environment [35,48], there seems to be a practical need to develop first-order policy-based methods that can learn a policy with robustness guarantees.

Our contributions mainly exist in the the following aspects. First, we develop a first-order policy-based method, named robust policy mirror descent (RPMD), for solving the robust MDP problem (1.5). En route, we establish some new structural results for the robust Markov decision process, which might be of independent interests for other algorithmic developments (e.g., natural robust policy gradient, see Section 3). Despite the non-convex and non-smooth structure of the objective (see [42]), RPMD achieves linear and sublinear convergence in the optimality gap, with two different increasing-stepsize schemes. Specifically, RPMD takes $\mathcal{O}(\log(1/\epsilon))$ (resp. $\mathcal{O}(1/\epsilon)$) iterations for the linearly (resp. sublinearly) converging variant to find an ϵ-optimal policy. The established convergence results hold for any Bregman divergence, as long as the policy space has a bounded distance to the initial policy measured in the same divergence, which holds for both the squared euclidean distance and the KL-divergence. To the best of our knowledge, no existing PGM can attain the obtained iteration complexity when solving (1.5).

Second, we establish sublinear convergence of constant-stepsize RPMD. Specifically, for RPMD with euclidean Bregman divergence and a constant stepsize of η, we establish a $\mathcal{O}\left(\min\left\{1/\epsilon, 1/(\eta\epsilon^2)\right\}\right)$ iteration complexity for finding an ϵ-optimal policy, which improves the rate of existing PGMs applied to (1.5) by orders of magnitude, and is applicable to general $(\mathbf{s},\mathbf{a})$-rectangular uncertainty sets. For a more general class of Bregman divergence and any relatively strongly convex uncertainty set (see Definition 3.2), we establish an $\mathcal{O}\left(\min\left\{1/\epsilon, R^2/(\eta\mu\epsilon^2)\right\}\right)$ iteration complexity for finding an ϵ-optimal policy, where R denotes the relative strong convexity constant for the uncertainty set, and μ denotes the strong convexity modulus of the distance-generating function with respect to ℓ_1-norm.

Third, we develop stochastic variants of the RPMD method, named SRPMD, when the first-order information is only available through online interactions with the nominal environment. For general Bregman divergences, we show an $\mathcal{O}(1/\epsilon^2)$ (resp. $\mathcal{O}(1/\epsilon^3)$) sample complexity for the linearly (resp. sublinearly) converging SRPMD variants, using different increasing-stepsize schemes. For euclidean Bregman divergence, we show an $\mathcal{O}(1/\epsilon^3)$ sample complexity with a properly chosen constant stepsize. To the best of our knowledge, all the developed sample complexity results of RPMD appear to be new for PGM methods applied to the robust MDP problem.

The rest of this manuscript is organized as follows. Section 2 makes some structural observations on the robust Markov decision process that will prove useful in the ensuing algorithmic developments. Section 3 introduces the deterministic RPMD method and establish its convergence properties. Section 4 develops the stochastic variants of RPMD when only stochastic first-order information is available. Section 5 then establishes the sample complexity for the proposed stochastic RPMD methods. Finally, concluding remarks are made in Section 6.

2 Structural Properties of Robust MDP

In this section, we develop some important observations on the structural properties of robust MDP, which will prove to be useful in our ensuing algorithmic developments.

2.1 Structure of Robust Value Functions

We first characterize the robust value function of any stochastic policy, following similar arguments for deterministic policies in [26].

Proposition 2.1. For robust MDP $\mathcal{M}_{\mathcal{U}}$ with a compact rectangular uncertainty set $\mathcal{U}$, defined in Defi-

nition 1.1, the robust value function satisfies the following nonlinear Bellman equation

$$V_r^\pi(s) = \sum_{a\in\mathcal{A}} r(s,a)\pi(a|s) + \gamma \sum_{a\in\mathcal{A}} \pi(a|s) \max_{u\in\mathcal{U}} \sum_{s'\in\mathcal{S}} \mathbb{P}_u(s'|s,a) V_r^\pi(s'), \quad \forall s \in \mathcal{S}. \tag{2.1}$$

In addition, a worst-case transition kernel $\mathbb{P}_{u_\pi}$ for the policy π is given by

$$u_\pi(\cdot|s,a) \in \operatorname*{Argmax}_{u(\cdot|s,a)\in\mathcal{U}_{s,a}} \sum_{s'\in\mathcal{S}} \mathbb{P}_u(s'|s,a) V_r^\pi(s'), \quad \forall (s,a) \in \mathcal{S}\times\mathcal{A}, \tag{2.2}$$

or equivalently,

$$V_r^\pi(s) = \sum_{a\in\mathcal{A}} r(s,a)\pi(a|s) + \gamma \sum_{a\in\mathcal{A}} \pi(a|s) \sum_{s'\in\mathcal{S}} \mathbb{P}_{u_\pi}(s'|s,a) V_r^\pi(s'), \quad \forall s \in \mathcal{S}.$$

The proof of Proposition 2.1 is deferred to Appendix A. Note that the last relation in Proposition 2.1 also shows that V_r^π is the solution of the standard Bellman equation for standard value function with uncertainty u_π, denoted by $V_{u_\pi}^\pi$. Hence from the uniqueness of the solution for the Bellman equation, we obtain

$$V_r^\pi = V_{u_\pi}^\pi. \tag{2.3}$$

Following similar lines as in Proposition 2.1, we can establish the following properties of Q_r^π.

Proposition 2.2. The robust state-action value function Q_r^π satisfies

$$Q_r^\pi(s,a) = c(s,a) + \max_{u\in\mathcal{U}} \gamma \sum_{s'\in\mathcal{S}} \mathbb{P}_u(s'|s,a) \sum_{a'\in\mathcal{A}} \pi(a'|s') Q_r^\pi(s',a'), \quad \forall (s,a) \in \mathcal{S}\times\mathcal{A}. \tag{2.4}$$

Moreover, Q_r^π and V_r^π satisfies the following relation

$$Q_r^\pi(s,a) = r(s,a) + \gamma \max_{u\in\mathcal{U}} \sum_{s'\in\mathcal{S}} \mathbb{P}_u(s'|s,a) V_r^\pi(s'), \quad \forall (s,a) \in \mathcal{S}\times\mathcal{A}, \tag{2.5}$$

Finally, we also have

$$V_r^\pi(s) = \langle Q_r^\pi(s,\cdot), \pi(\cdot|s)\rangle := \langle Q_r^\pi, \pi\rangle_s, \quad \forall s \in \mathcal{S}, \tag{2.6}$$

$$Q_r^\pi = Q_{u_\pi}^\pi, \tag{2.7}$$

where u_π is defined as in (2.2).

Proof. Property (2.4) follows from similar lines as the proof of Proposition 2.1. To show (2.5), note that

$$Q_r^\pi(s,a) = \max_{u\in\mathcal{U}} Q_u^\pi(s,a) = r(s,a) + \max_{u\in\mathcal{U}} \sum_{s'\in\mathcal{S}} \mathbb{P}_u(s'|s,a) V_u^\pi(s'),$$

where the last inequality follows from the standard relation between V_u^π and Q_u^π. It suffices to note that by taking $u = u_\pi$ defined in (2.2), we have $V_{u_\pi}^\pi = V_r^\pi$ and thus $\max_{u\in\mathcal{U}} \sum_{s'\in\mathcal{S}} \mathbb{P}_u(s'|s,a) V_u^\pi(s') \geq \sum_{s'\in\mathcal{S}} \mathbb{P}_{u_\pi}(s'|s,a) V_r^\pi(s') = \max_{u\in\mathcal{U}} \sum_{s'\in\mathcal{S}} \mathbb{P}_u(s'|s,a) V_r^\pi(s')$, where the last equality follows from (2.2). On the other hand, we have

$$\max_{u\in\mathcal{U}} \sum_{s'\in\mathcal{S}} \mathbb{P}_u(s'|s,a) V_u^\pi(s') \leq \max_{u\in\mathcal{U}} \max_{u'\in\mathcal{U}} \sum_{s'\in\mathcal{S}} \mathbb{P}_u(s'|s,a) V_{u'}^\pi(s') = \max_{u\in\mathcal{U}} \sum_{s'\in\mathcal{S}} \mathbb{P}_u(s'|s,a) V_r^\pi(s'),$$

hence we obtain

$$Q_r^\pi(s,a) = r(s,a) + \max_{u\in\mathcal{U}} \sum_{s'\in\mathcal{S}} \mathbb{P}_u(s'|s,a) V_u^\pi(s') = r(s,a) + \max_{u\in\mathcal{U}} \sum_{s'\in\mathcal{S}} \mathbb{P}_u(s'|s,a) V_r^\pi(s').$$

Thus (2.5) is proved. Moreover, (2.6) follows from taking expectation with respect to $a \sim \pi(\cdot|s)$ on both sides of (2.5) and making use of (2.1) and the rectangularity of the uncertainty set. Finally, from (2.5) and (2.6), it is also clear that

$$Q_r^\pi(s,a) = c(s,a) + \gamma \sum_{s' \in \mathcal{S}} \mathbb{P}_{u_\pi}(s'|s,a) \sum_{a' \in \mathcal{A}} \pi(a'|s') Q_r^\pi(s',a'), \quad \forall (s,a) \in \mathcal{S} \times \mathcal{A},$$

where u_π is defined as in (2.2), then $Q_r^\pi = Q_{u_\pi}^\pi$ since $Q_{u_\pi}^\pi$ is the unique solution of the previous system. □

2.2 Differentiability of Robust Values

A seemingly natural concept to solve (1.5) is to iteratively update the policy by following its negative gradient direction. At the same time, it should be noted that even for a fixed uncertainty (i.e., non-robust MDP), for any state $s \in \mathcal{S}$, the value function $V_u^\pi(s)$ is only well defined over the set of randomized policies Π, for which any $\pi \in \Pi$ must satisfy $\mathbf{1}^\top \pi(\cdot|s) = 1$ for all $s \in \mathcal{S}$. Hence $\Pi = \mathrm{dom}(f_\rho)$ belongs to a lower-dimensional subspace in $\mathbb{R}^{|\mathcal{A}||\mathcal{S}|}$ with dimension $(|\mathcal{A}|-1)|\mathcal{S}|$. Consequently, the implicitly assumed gradient ∇f_ρ, given by $\lim_{\delta \to \mathbf{0}, \delta \in \mathbb{R}^d} |f_\rho(\pi+\delta) - f_\rho(\pi) - \langle \nabla f_\rho(\pi), \delta \rangle| / \|\delta\| \to 0$, is not well defined. Given this observation, we then adopt the following definition of policy gradient for objective (1.5) when considering direct policy parameterization.

Definition 2.1 (Policy Gradient with Direct Parameterization)**.** For any function of policy $f : \Pi \to \mathbb{R}$, the policy gradient of f with respect to π, denoted by $\nabla f(\pi)$, is the vector satisfying the following,

$$\lim_{\delta \to \mathbf{0}, \pi + \delta \in \Pi} |f(\pi+\delta) - f(\pi) - \langle \nabla f(\pi), \delta \rangle| / \|\delta\|_2 \to 0. \tag{2.8}$$

Given Definition 2.1, it should be clear that for any policy $\pi \in \Pi$, if $\nabla f(\pi)$ exists, then it is unique. Moreover, Definition 2.1 slightly generalizes the notion of Fréchet derivative of objective (1.5) as Π is a closed set in its affine span. We then proceed to derive the policy gradient of $V_u^\pi(s)$ for a given uncertainty $u \in \mathcal{U}$.

Lemma 2.1 (Policy Gradient for Fixed Uncertainty with Direct Parameterization)**.** Given $u \in \mathcal{U}$ and a state $s \in \mathcal{S}$, then the policy gradient of $V_u^\pi(s)$ with respect to π is given by

$$\nabla V_u^\pi(s)[s',a] = \tfrac{1}{1-\gamma} d_s^{\pi,u}(s') Q_u^\pi(s',a), \quad \forall (s,a) \in \mathcal{S} \times \mathcal{A},$$

where $\nabla V_u^\pi(s)[s',a]$ denotes the entry of $\nabla V_u^\pi(s)$ corresponding to the (s',a) state-action pair.

Proof. For MDP $\mathcal{M}_u$, and any pair of policies (π, π') with $\pi' = \pi + \delta$, we have

$$\begin{aligned}
V_u^{\pi'}(s) - V_u^\pi(s) &\overset{(a)}{=} \tfrac{1}{1-\gamma} \mathbb{E}_{s' \sim d_s^{\pi',u}} \langle Q_u^\pi(s',\cdot), \pi'(\cdot|s') - \pi(\cdot|s') \rangle \\
&= \tfrac{1}{1-\gamma} \sum_{s' \in \mathcal{S}} \sum_{a \in \mathcal{A}} d_s^{\pi',u}(s') Q_u^\pi(s',a) \left(\pi'(a|s') - \pi(a|s')\right) \\
&= \tfrac{1}{1-\gamma} \sum_{s' \in \mathcal{S}} \sum_{a \in \mathcal{A}} d_s^{\pi,u}(s') Q_u^\pi(s',a) \left(\pi'(a|s') - \pi(a|s')\right) \\
&\quad + \tfrac{1}{1-\gamma} \sum_{s' \in \mathcal{S}} \sum_{a \in \mathcal{A}} (d_s^{\pi',u}(s') - d_s^{\pi,u}(s')) Q_u^\pi(s',a) \left(\pi'(a|s') - \pi(a|s')\right) \\
&= \underbrace{\tfrac{1}{1-\gamma} \sum_{s' \in \mathcal{S}} \sum_{a \in \mathcal{A}} d_s^{\pi,u}(s') Q_u^\pi(s',a) \delta(a|s')}_{(A)} \\
&\quad + \underbrace{\tfrac{1}{1-\gamma} \sum_{s' \in \mathcal{S}} \sum_{a \in \mathcal{A}} (d_s^{\pi',u}(s') - d_s^{\pi,u}(s')) Q_u^\pi(s',a) \left(\pi'(a|s') - \pi(a|s')\right)}_{(B)},
\end{aligned}$$

where equality (a) follows directly from the performance difference lemma for standard MDPs [12,20]. It is clear that term $(A) = \langle g, \delta \rangle$, where the entry of g associated with (s', a) state-action pair, denoted by $g(a|s')$, is given by $g(a|s') = d_s^{\pi,u}(s')Q_u^\pi(s', a)$. It remains to show that term $(B) = \mathcal{O}(\|\pi - \pi'\|_2)$ where we identify π as a matrix in $\mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$. To this end, is suffices to show $|d_s^{\pi',u}(s') - d_s^{\pi,u}(s')| = \mathcal{O}(\|\pi - \pi'\|_2)$ for any $s' \in \mathcal{S}$.

Let us define $\mathbb{P}_u^\pi : \mathcal{S} \times \mathcal{S} \to [0, 1]$ by $\mathbb{P}_u^\pi(s', s) = \sum_{a\in\mathcal{A}} \mathbb{P}_u(s'|s, a)\pi(a|s)$, then for any $\pi \in \Pi$, we obtain

$$d_s^{\pi,u} = (1-\gamma)\sum_{t=0}^{\infty} \gamma^t (\mathbb{P}_u^\pi)^t e_s = (1-\gamma)\left(1-\gamma\mathbb{P}_u^\pi\right)^{-1} e_s, \tag{2.9}$$

where in the last inequality we use the fact that $\rho(\gamma\mathbb{P}_u^\pi) \le \gamma < 1$. Hence we have

$$\begin{aligned} d_s^{\pi',u} - d_s^{\pi,u} &= (1-\gamma)\left(\left(I-\gamma\mathbb{P}_u^{\pi'}\right)^{-1} - (I-\gamma\mathbb{P}_u^\pi)^{-1}\right) e_s \\ &= (1-\gamma)\gamma\left(I-\gamma\mathbb{P}_u^{\pi'}\right)^{-1}\left(\mathbb{P}_u^\pi - \mathbb{P}_u^{\pi'}\right)(I-\gamma\mathbb{P}_u^\pi)^{-1} e_s, \end{aligned}$$

where the last equality uses the matrix identity $A^{-1} - B^{-1} = A^{-1}(B - A)B^{-1}$ for any invertible matrix A, B. Note that

$$\left\|(I-\gamma\mathbb{P}_u^\pi)^{-1}\right\|_1 = \left(\min\left\{\|(I-\gamma\mathbb{P}_u^\pi)\,x\|_1 : \|x\|_1 = 1\right\}\right)^{-1} \le (1-\gamma)^{-1} \tag{2.10}$$

for any $\pi \in \Pi$ and $u \in \mathcal{U}$, we can then further obtain

$$\left\|d_s^{\pi',u} - d_s^{\pi,u}\right\|_1 \le \tfrac{\gamma}{1-\gamma}\left\|\mathbb{P}_u^{\pi'} - \mathbb{P}_u^\pi\right\|_1 \overset{(a)}{=} \mathcal{O}(\left\|\pi - \pi'\right\|_\infty) \overset{(b)}{=} \mathcal{O}(\left\|\pi - \pi'\right\|_2), \tag{2.11}$$

where the equality (a) simply follows from the definition of $\mathbb{P}_u^\pi$, and (b) follows from the equivalence of norm. Hence the proof is completed. □

Lemma 2.1 also serves as an important stepping stone for establishing the existence of Fréchet subdifferential for policy objective (1.5), when viewing it as an extended real-valued function with domain Π. Specifically, we have the following lemma.

Lemma 2.2 (Fréchet Subgradient of Robust MDP)**.** Define $\overline{f}_\rho : \mathbb{R}^{|\mathcal{S}||\mathcal{A}|} \to \overline{R} = \mathbb{R} \cup \{+\infty\}$ as the extended-value version of function f_ρ, that is, $\overline{f}_\rho(\pi) = f_\rho(\pi)$ for any $\pi \in \Pi$ and $\overline{f}_\rho(\pi) = \infty$ otherwise. For any $\pi \in \Pi$, let $\nabla f_\rho(\pi) \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ with the (s, a)-entry specified as

$$\nabla f_\rho(\pi)[s, a] = \tfrac{1}{1-\gamma} d_\rho^{\pi,u_\pi}(s) Q_{u_\pi}^\pi(s, a), \quad \forall (s, a) \in \mathcal{S} \times \mathcal{A}. \tag{2.12}$$

Then $\nabla f_\rho(\pi)$ is a Fréchet subgradient of $\overline{f}_\rho$ at π.

Proof. Fix π, it suffices to consider any π' also in Π. Let $\delta = \pi' - \pi$, we have

$$\begin{aligned} V_{u_{\pi'}}^{\pi'}(s) - V_{u_\pi}^\pi(s) &= V_{u_{\pi'}}^{\pi'}(s) - V_{u_\pi}^{\pi'}(s) + V_{u_\pi}^{\pi'}(s) - V_{u_\pi}^\pi(s) \\ &\overset{(a)}{\ge} V_{u_\pi}^{\pi'}(s) - V_{u_\pi}^\pi(s) \\ &\overset{(b)}{=} \tfrac{1}{1-\gamma}\sum_{s'\in\mathcal{S}}\sum_{a\in\mathcal{A}} d_s^{\pi,u_\pi}(s') Q_{u_\pi}^\pi(s', a)\delta(a|s') + \mathcal{O}(\left\|\pi - \pi'\right\|_2), \end{aligned}$$

where (a) follows from the definition of $u_{\pi'}$, which guarantees $V_{u_{\pi'}}^{\pi'}(s) - V_{u_\pi}^{\pi'}(s) \ge 0$ for any $s \in \mathcal{S}$, and (b) follows directly from the proof of Lemma 2.1. Thus we obtain from the definition of f_ρ that

$$f_\rho(\pi') - f_\rho(\pi) \ge \tfrac{1}{1-\gamma}\sum_{s\in\mathcal{S}} \rho(s) \sum_{s'\in\mathcal{S}}\sum_{a\in\mathcal{A}} d_s^{\pi,u_\pi}(s') Q_{u_\pi}^\pi(s', a)\delta(a|s') + \mathcal{O}(\left\|\pi - \pi'\right\|_2)$$

$$= \tfrac{1}{1-\gamma} \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} d_\rho^{\pi,u_\pi}(s) Q_{u_\pi}^{\pi}(s,a)\delta(a|s) + \mathcal{O}(\left\|\pi - \pi'\right\|_2),$$

Dividing both sides of the previous relation by $\|\delta\|$, and taking infimum over $\pi' \neq \pi$, and further taking $\|\delta\| \downarrow 0$, we obtain

$$\liminf_{\pi'\to\pi, \pi'\neq\pi} \left(f_\rho(\pi') - f_\rho(\pi) - \tfrac{1}{1-\gamma} \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} d_\rho^{\pi,u_\pi}(s) Q_{u_\pi}^{\pi}(s,a)\delta(a|s) \right) / \left\|\pi' - \pi\right\|_2 \geq 0.$$

Hence we conclude from the prior relation, and the definition of Fréchet subdifferential [17] that $\nabla f_\rho(\pi) \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ with the (s,a)-entry specified as

$$\nabla f_\rho(\pi)[s,a] = \tfrac{1}{1-\gamma} d_\rho^{\pi,u_\pi}(s) Q_{u_\pi}^{\pi}(s,a), \quad \forall (s,a) \in \mathcal{S}\times\mathcal{A}$$

is a Fréchet subgradient of $\overline{f}_\rho$ at $\pi \in \Pi$. □

It is also worth pointing out that the objective (1.5) is indeed almost everywhere differentiable (in the sense of Definition 2.1), when taking the measure to be $(|\mathcal{A}|-1)|\mathcal{S}|$-dimensional Hausdorff measure. We remark that Hausdorff measure is a natural choice for our discussion of differentiability over Π, as it adapts to the low-dimensional nature of Π. On the other hand, choosing Lebesgue measure yields a trivial almost-everywhere-differentiable claim, as Π itself takes a Lebesgue measure of zero in $\mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$. In particular, we have the following lemma.

Lemma 2.3 (Almost-everywhere Differentiability of Robust MDP)**.** The policy optimization objective (1.5) is everywhere differentiable in its domain Π except in a zero-measure set, where the measure is taken to be the $(|\mathcal{A}|-1)|\mathcal{S}|$-dimensional Hausdorff measure, and the differentiability is defined as in Definition 2.1.

Proof. We defer the proof to Appendix A given its technical nature. □

Combining Lemma 2.2 and 2.3, we are now ready to show that for any $\pi \in \mathrm{ReInt}(\Pi)$, whenever f_ρ is differentiable at π, then the policy gradient defined in the sense of Definition 2.1, is given exactly by (2.12) in Lemma 2.2.

Lemma 2.4 (Policy Gradient for Robust MDP with Direct Parameterization)**.** At any policy $\pi \in \mathrm{ReInt}(\Pi)$, except for a zero measure set (measured with the $(|\mathcal{A}|-1)|\mathcal{S}|$-dimensional Hausdorff measure), the policy optimization objective f_ρ defined in (1.5) is differentiable, and its policy gradient ∇f_ρ defined in the sense of Definition 2.1, is given by (2.12).

Proof. Let us define $e = \mathbf{1}/|\mathcal{A}| \in \mathbb{R}^{|\mathcal{A}|}$, and let $\mathbf{e} = e \otimes \mathbf{1} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$. For any policy $\pi \in \Pi$, we can accordingly define π_e where $\pi_e(\cdot|s) = \pi(\cdot|s) - e$. Conversely, let $\Pi_e = \{\pi_e : \pi \in \Pi\}$, note that the mapping $\mathcal{M} : \Pi \to \Pi_e$ from π to π_e is one-to-one and onto. Let us define function g with domain Π_e as

$$g(\pi_e) = f_\rho(\pi), \quad \forall \pi_e \in \Pi_e.$$

It should be clear that $\Pi_e \subset \mathcal{U}$ where $\mathcal{U}$ is a $(|\mathcal{A}|-1)|\mathcal{S}|$-dimensional subspace in $\mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$. Let $\|\cdot\|$ denote the euclidean-norm on $\mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, then it is immediate that $(\mathcal{U}, \|\cdot\|)$ is a Banach space, and $g : \Pi_e \subset \mathcal{U} \to \mathbb{R}$.

For any $\pi \in \mathrm{ReInt}(\Pi)$ where f_ρ is differentiable in the sense of Definition 2.1, we have $f_\rho(\pi') - f_\rho(\pi) - \langle \nabla f_\rho(\pi), \pi' - \pi\rangle = \mathcal{O}(\|\pi - \pi'\|_2)$. Thus, by letting $\mathrm{Int}_{\mathcal{U}}(X)$ denote the interior of set X inside the Banach space $(\mathcal{U}, \|\cdot\|)$, we obtain for $\pi_e = \mathcal{M}^{-1}(\pi) \in \mathrm{Int}_{\mathcal{U}}(\Pi_e)$ that

$$g(\pi_e') - g(\pi_e) - \langle \nabla f(\pi), \pi_e' - \pi_e\rangle = \mathcal{O}(\left\|\pi_e - \pi_e'\right\|), \ \forall \pi_e' \in \Pi_e.$$

That is, g is Fréchet differentiable at π_e with gradient $\nabla g(\pi_e) = \nabla f(\pi)$. Given Lemma 2.3, it should be clear that the set of π_e in Π_e where g is not Fréchet differentiable has a measure of zero in the $(|\mathcal{A}|-1)|\mathcal{S}|$-dimensional Hausdorff measure.

We next show that (2.12) defines a Fréchet subgradient of g. This is due to the fact that for any $\pi_e \in \mathrm{Int}_{\mathcal{U}}(\Pi_e)$,

$$\begin{aligned}
&\liminf_{\pi_e' \to \pi_e, \pi_e' \neq \pi_e, \pi_e' \in \Pi_e} \left(g(\pi_e') - g(\pi_e) - \tfrac{1}{1-\gamma} \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} d_\rho^{\pi,u_\pi}(s) Q_{u_\pi}^{\pi}(s,a) \left(\pi_e'(a|s) - \pi_e(a|s)\right) \right) / \left\|\pi_e' - \pi_e\right\| \\
=&\liminf_{\pi' \to \pi, \pi' \neq \pi, \pi' \in \Pi} \left(f_\rho(\pi') - f_\rho(\pi) - \tfrac{1}{1-\gamma} \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} d_\rho^{\pi,u_\pi}(s) Q_{u_\pi}^{\pi}(s,a) \left(\pi'(a|s) - \pi(a|s)\right) \right) / \left\|\pi' - \pi\right\|_2 \geq 0,
\end{aligned}$$

where the equality follows from the construction of g and π_e, and the inequality follows from Lemma 2.2.

By combining observations from the previous two paragraphs, it is clear that the set of non-Fréchet differentiable points of g form a zero-measure set in $\mathcal{U}$, when taking the $(|\mathcal{A}|-1)|\mathcal{S}|$-dimensional Hausdorff measure. The Fréchet subgradient at any point π_e is given by (2.12). Moreover, at any Fréchet differentiable point π_e of g, we conclude that its Fréchet gradient $\nabla g(\pi)$ is the only element in its Fréchet subdifferential (Proposition 1.1, [17]), which is given by (2.12).

It remains to show that for any Fréchet differentiable point $\pi_e \in \mathrm{Int}_{\mathcal{U}}(\Pi_e)$ of g, with its derivative denoted by $\nabla g(\pi)$, the corresponding $\pi \in \mathrm{ReInt}(\Pi)$ is also a differentiable point of f_ρ in the sense of Definition 2.1. To see this, note that

$$\begin{aligned}
&f_\rho(\pi') - f_\rho(\pi) - \langle \nabla g(\pi_e), \pi' - \pi \rangle \\
=&g(\pi_e') - g(\pi_e) - \langle \nabla g(\pi_e), \pi_e' - \pi_e \rangle = \mathcal{O}(\left\|\pi_e' - \pi_e\right\|) = \mathcal{O}(\left\|\pi' - \pi\right\|_2), \quad \forall \pi' \in \mathrm{ReInt}(\Pi).
\end{aligned}$$

Thus we conclude that f_ρ is also differentiable at π, and the gradient $\nabla f_\rho(\pi)$ defined in Definition 2.1 coincides with the Fréchet gradient of g at π_e, given by (2.12). The proof is then completed. □

As the last result in this subsection, we show that if for any π, the corresponding worst-case uncertainty u_π is unique, then the robust policy optimization objective (1.5) is indeed differentiable everywhere inside $\mathrm{ReInt}(\Pi)$. In Section 3, we will also propose a sufficient condition (see Lemma 3.6) which guarantees this condition to hold.

Lemma 2.5. If for any $\pi \in \mathrm{ReInt}(\Pi)$, the corresponding worst-case environment defined in (2.2) is unique. Then the policy optimization objective (1.5) is differentiable everywhere in $\mathrm{ReInt}(\Pi)$ in the sense of Definition 2.1, and the policy gradient in given as (2.12).

Proof. The proof is deferred to Appendix A. □

Give our prior discussions on the differentiability of objective (1.5), one can perhaps directly update the policy using the policy gradient specified in (2.12). Moreover, the results developed in this subsection seems to be readily extensible for computing policy gradient when function approximation is adopted.

On the other hand, it should be noted that the task of calculating or estimating the gradient for the robust MDP seems difficult when one can only access information through interactions with the nominal environment, except for some special subclasses of uncertainty sets. In particular, although one can sample directly from the worst-case environment $\mathcal{M}_{u_\pi}$ to perform estimation, in practice we believe this is against the principle of pursuing robustness. Instead, a much more desirable alternative is to train the policy in a fixed (nominal) environment, without actually deploying it in its worst-case environment. More importantly, due to the nonconvex and potentially nonsmooth landscape, one typically could only get a sublinear convergence to a stationary point in the best-iterate sense, which would further require additional landscape analysis for showing approximate optimality of the learned policy [42].

In the next subsection, we introduce an alternative viewpoint of solving the robust MDP based on variational inequalities. As we will demonstrate in Section 3, solving the variational inequality allows us to bypass the aforementioned difficulties. At the same time, such a viewpoint also enables us to develop a policy mirror descent method with fast global convergence guarantees.

2.3 A Variational Inequality Perspective

To proceed, we introduce a characterization that upper bounds the difference of robust values for two policies, and serves as a keystone in the ensuing algorithmic developments.

Lemma 2.6. For any pair of policies π, π', we have

$$V_r^{\pi'}(s) - V_r^{\pi}(s) \leq \tfrac{1}{1-\gamma}\mathbb{E}_{s'\sim d_s^{\pi', u_{\pi'}}} \langle Q_r^{\pi}, \pi' - \pi\rangle_{s'}. \tag{2.13}$$

Proof. Let $\xi_u^{\pi'}(s) \coloneqq \{(s_t, a_t)\}_{t\geq 0}$ denote the trajectory generated from by π' within MDP $\mathcal{M}_{u_{\pi'}}$, with initial state set as s. That is, $a_t \sim \pi'(\cdot|s_t)$, $s_{t+1} \sim \mathbb{P}_{u'_\pi}(\cdot|s_t, a_t)$ for all $t \geq 0$ and $s_0 = s$. We know that

$$\begin{aligned}
V_r^{\pi'}(s) - V_r^{\pi}(s) &\overset{(a)}{=} V_{u_{\pi'}}^{\pi'}(s) - V_r^{\pi}(s) \\
&= \mathbb{E}_{\xi_u^{\pi'}(s)}\left[\sum_{t=0}^{\infty}\gamma^t c(s_t, a_t)\right] - V_r^{\pi}(s) \\
&= \mathbb{E}_{\xi_u^{\pi'}(s)}\left[\sum_{t=0}^{\infty}\gamma^t \left[c(s_t, a_t) + V_r^{\pi}(s_t) - V_r^{\pi}(s_t)\right]\right] - V_r^{\pi}(s) \\
&\overset{(b)}{=} \mathbb{E}_{\xi_u^{\pi'}(s)}\left[\sum_{t=0}^{\infty}\gamma^t \left[c(s_t, a_t) + \gamma V_r^{\pi}(s_{t+1}) - V_r^{\pi}(s_t)\right]\right] + \mathbb{E}_{\xi_u^{\pi'}(s)}\left[V_r^{\pi}(s_0)\right] - V_r^{\pi}(s) \\
&\overset{(c)}{=} \mathbb{E}_{\xi_u^{\pi'}(s)}\left[\sum_{t=0}^{\infty}\gamma^t \left[c(s_t, a_t) + \gamma V_r^{\pi}(s_{t+1}) - V_r^{\pi}(s_t)\right]\right] \\
&\overset{(d)}{\leq} \mathbb{E}_{\xi_u^{\pi'}(s)}\left[\sum_{t=0}^{\infty}\gamma^t \left[Q_r^{\pi}(s_t, a_t) - V_r^{\pi}(s_t)\right]\right] \\
&\overset{(e)}{=} \tfrac{1}{1-\gamma}\sum_{s'\in\mathcal{S}} d_s^{\pi', u_{\pi'}}(s')\sum_{a'\in\mathcal{A}}\pi'(a'|s')Q_r^{\pi}(s', a') - \tfrac{1}{1-\gamma}\sum_{s'\in\mathcal{S}} d_s^{\pi', u_{\pi'}}(s')\sum_{a'\in\mathcal{A}}\pi(a'|s')Q_r^{\pi}(s', a') \\
&= \tfrac{1}{1-\gamma}\mathbb{E}_{s'\sim d_s^{\pi', u_{\pi'}}}\langle Q_r^{\pi}, \pi' - \pi\rangle_{s'},
\end{aligned}$$

where (a) follows from (2.3); (b) follows from moving $V_r^{\pi}(s_0)$ outside the summation; (c) follows from definition that $s_0 = s$; (e) follows from the definition of d_s^{π, u_π} in (1.7), and the relation between V_r^{π} and Q_r^{π} in (2.6). It remains to show that (d) holds.

For any $t \geq 0$, we have

$$\begin{aligned}
\mathbb{E}_{\xi_u^{\pi'}(s)}\left[c(s_t, a_t) + \gamma V_r^{\pi}(s_{t+1})\right] &\overset{(a')}{=} \mathbb{E}_{\xi_u^{\pi'}(s)}\left[c(s_t, a_t) + \gamma\sum_{s'\in\mathcal{S}}\mathbb{P}_{u_{\pi'}}(s'|s_t, a_t)V_r^{\pi}(s')\right] \\
&\leq \mathbb{E}_{\xi_u^{\pi'}(s)}\left[c(s_t, a_t) + \gamma\max_{u(\cdot|s_t, a_t)}\sum_{s'\in\mathcal{S}}\mathbb{P}_u(s'|s_t, a_t)V_r^{\pi}(s')\right] \\
&\overset{(b')}{=} \mathbb{E}_{\xi_u^{\pi'}(s)}\left[Q_r^{\pi}(s_t, a_t)\right],
\end{aligned}$$

where (a') follows from the definition of $\xi_u^{\pi'}(s)$, and (b') follows from the property of Q_r^{π} (2.5). Thus relation (d) follows immediately from the prior observation and the linearity of expectation. □

By applying Lemma 2.6, we know that for any policy $\pi \in \Pi$ and an optimal robust policy π^*, we have

$$0 \leq V_r^{\pi}(s) - V_r^{\pi^*}(s) \leq \tfrac{1}{1-\gamma}\mathbb{E}_{s' \sim d_s^{\pi,u_\pi}} \langle Q_r^{\pi^*}, \pi - \pi^* \rangle_s = \langle \mathcal{F}_s(\pi, \pi^*), \pi - \pi^* \rangle,$$

where $\mathcal{F}_s : \Pi \times \Pi \to \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ is defined by

$$\mathcal{F}_s(\pi, \pi')[s', a] = \tfrac{1}{1-\gamma} d_s^{\pi,u_\pi}(s') Q_r^{\pi'}(s', a) \quad \forall (s', a) \in \mathcal{S} \times \mathcal{A}.$$

Given the optimality of π^*, we then know that $\langle \mathcal{F}(\pi, \pi^*), \pi - \pi^* \rangle \geq 0$ for all $\pi \in \Pi$. Now for any $\pi \in \Pi$, let $\alpha \in (0, 1)$, we define $\pi_\alpha = \alpha\pi + (1-\alpha)\pi^*$. Substituting π_α into the prior relation, we obtain $\langle \mathcal{F}_s(\pi_\alpha, \pi^*), \pi_\alpha - \pi^* \rangle = \alpha \langle \mathcal{F}_s(\pi_\alpha, \pi^*), \pi - \pi^* \rangle \geq 0$. Consequently,

$$\langle \mathcal{F}_s(\pi_\alpha, \pi^*), \pi - \pi^* \rangle \geq 0 \tag{2.14}$$

for any $\alpha \in (0, 1)$. It is straightforward to verify that the visitation measure $d_s^{\pi,u}$ is a continuous function of (π, u) (see proof of Lemma 2.1). On the other hand, note that given (2.2), the worst-case uncertainty $u_\pi(\cdot|s, a) \in \partial \sigma_{\mathcal{U}_{s,a}}(V_r^\pi)$, where $\sigma_X(\cdot)$ denotes the support function of set X, and ∂f denotes the subdifferential of function f. For simplicity, let us assume that $\partial \sigma_{\mathcal{U}_{s,a}}(V_r^\pi)$ is always a singleton for any π (See Lemma 3.6 for example).

Now let $\alpha \downarrow 0$, we have $\pi_\alpha \to \pi^*$. Since V_r^π is Lipschitz continuous in π (see Lemma 3.8 for an elementary proof), then $V_r^{\pi_\alpha} \to V_r^{\pi^*}$. Thus given that $\sigma_{\mathcal{U}_{s,a}}$ is a closed and proper convex function, and the fact that the subdifferential map of a closed convex function is closed (Theorem 24.4, [31]), we know that any limit point of $u_{\pi_\alpha}(\cdot|s, a)$ is also a subgradient of $\sigma_{\mathcal{U}_{s,a}}(V_r^{\pi^*})$, and hence the limit point is indeed unique and the worst-case uncertainty for π^*. Denoting this limit point as u_{π^*}, then by taking $\alpha \downarrow 0$ in (2.14), we obtain

$$\lim_{\alpha \downarrow 0} \langle \mathcal{F}_s(\pi_\alpha, \pi^*), \pi - \pi^* \rangle \geq 0 \Rightarrow \langle \mathcal{G}_s(\pi^*), \pi - \pi^* \rangle \geq 0, \tag{2.15}$$

$$\mathcal{G}_s(\pi^*)[s', a] = \tfrac{1}{1-\gamma} d_s^{\pi^*,u_{\pi^*}}(s') Q_r^{\pi^*}(s', a), \quad \forall (s', a) \in \mathcal{S} \times \mathcal{A}. \tag{2.16}$$

In view of the discussions in the last two paragraphs, (2.15) suggests us to find the optimal robust policy via solving the following variational inequality (VI),

$$\langle \mathbb{E}_{s \sim \rho} \left[\mathcal{G}_s(\pi^*) \right], \pi - \pi^* \rangle \geq 0, \quad \forall \pi \in \Pi. \tag{2.17}$$

Interested readers might find that VI (2.17) has a close analogy for solving non-robust MDPs, constructed in [20]. Specifically, for a non-robust, discounted finite MDP, the optimal policy satisfies the following VI,

$$\begin{aligned} &\langle \mathbb{E}_{s \sim \nu^*} \left[\mathcal{G}_s^N(\pi^*) \right], \pi - \pi^* \rangle \geq 0, \\ &\mathcal{G}_s^N(\pi^*)[s', a] = \tfrac{1}{1-\gamma} d_s^{\pi^*}(s') Q^{\pi^*}(s', a), \quad \forall (s', a) \in \mathcal{S} \times \mathcal{A}, \end{aligned} \tag{2.18}$$

where ν^* denotes the stationary state distribution induced by the optimal policy π^*. One can clearly see that the only difference between the non-robust VI (2.18) and the robust VI (2.17) is the additional role of worst-case environment for the latter. To solve the non-robust version of VI constructed therein, the author exploits the key fact that the constructed VI satisfies the so-called generalized monotonicity:

$$\langle \mathbb{E}_{s \sim \nu^*} \left[\mathcal{G}_s^N(\pi) \right], \pi - \pi^* \rangle = \mathbb{E}_{s \sim \nu^*} \left[V^\pi(s) - V^{\pi^*}(s) \right] = f_{\nu^*}(\pi) - f_{\nu^*}(\pi^*) > 0, \tag{2.19}$$

for any policy $\pi \notin \Pi^*$.

However, we next show that unlike the performance difference lemma for standard MDPs [12, 20], for which (2.13) holds with equality, the inequality in Lemma 2.6 seems unavoidable in the robust setting. Consequently, the VI for the robust MDP 2.17 no longer satisfies the generalized monotonicity as defined in (2.19).

Proposition 2.3. There exists a robust MDP instance $\mathcal{M}_{\mathcal{U}}$ such that

$$V_r^{\pi^*}(s) - V_r^{\pi}(s) < 0, \quad \tfrac{1}{1-\gamma}\mathbb{E}_{s'\sim d_s^{\pi^*,u_{\pi^*}}}\langle Q_r^{\pi}, \pi^* - \pi\rangle_{s'} = 0 \tag{2.20}$$

holds for at least one state $s \in \mathcal{S}$. In particular, π^* corresponds to a solution of (1.4), and π is a strictly suboptimal policy. Moreover, denoting $\nu^* := \nu^{\pi^*,u_{\pi^*}}$ as the stationary state distribution of the optimal policy π^* within its worst-case MDP $\mathcal{M}_{u_{\pi^*}}$, we have

$$0 = \langle \mathbb{E}_{s\sim\nu^*}\left[\mathcal{G}_s(\pi)\right], \pi - \pi^*\rangle < f_{\nu^*}(\pi) - f_{\nu^*}(\pi^*). \tag{2.21}$$

Proof. Consider the MDP $\mathcal{M}$ that contains three states $\mathcal{S} = \{S_a, S_b, S_c\}$, and each state is associated with two actions $\mathcal{A} = \{L, R\}$. The transition of $\mathcal{M}$, denoted by $\mathbb{P}$, is fully deterministic, and is illustrated in Figure 2. Each edge starts from the current state, and ends at the next state, with its arc (a, c) consisting of the action a, and the cost c associated with the action a. We assume such a cost occurs immediately after the action is made, and is independent of the next state. In addition, we assume $C > 1$.

Let us consider the scenario where the uncertain environment has maximum manipulation strength at state S_c except there is no returning to itself, and has no manipulation strength at other states. That is, for any $a \in \mathcal{A}$, we have $\mathcal{P}_{S_c,a} = \Delta_{\{S_a,S_b\}}$. On the other hand, for any $s \neq S_c$, and for any $a \in \mathcal{A}$, we have $\mathcal{U}_{s,a} = \mathbf{0}$.

We consider a pair of policies π, π^*, defined by

$$\pi(L|S_a) = 1,\ \pi^*(R|S_a) = 1,\ \pi(\cdot|S_b) = \pi(\cdot|S_c) = \mathrm{Unif}(\mathcal{A}),\ \pi^*(\cdot|S_b) = \pi^*(\cdot|S_c) = \mathrm{Unif}(\mathcal{A}).$$

Since $C > 1$, and the fact that $\pi(L|S_a) = 1$, it should be clear that for policy π, the worst case transition $\mathbb{P}_{u_\pi}$ should satisfy

$$\mathbb{P}_{u_\pi}(S_a|S_c, L) = 1, \mathbb{P}_{u_\pi}(S_a|S_c, R) = 1.$$

Consequently, the robust value of π is given by

$$V_r^{\pi}(S_c) = \tfrac{\gamma C}{1-\gamma^2},\ V_r^{\pi}(S_b) = 1 + \tfrac{\gamma^2 C}{1-\gamma^2},\ V_r^{\pi}(S_a) = \tfrac{C}{1-\gamma^2}.$$

On the other hand, for policy π^*, since $\pi^*(R|S_a) = 1$ and $1 > 0$, the worst case transition $\mathbb{P}_{u_{\pi^*}}$ should satisfy

$$\mathbb{P}_{u_\pi}(S_b|S_c, L) = 1, \mathbb{P}_{u_\pi}(S_b|S_c, R) = 1.$$

It should also be clear that π^* is an optimal robust policy. From the previous observations, simple calculation yields

$$\sum_{s'\in\mathcal{S}} \mathbb{P}_{u_{\pi^*}}(s'|S_c, L)V_r^{\pi}(s') = V_r^{\pi}(S_b) < V_r^{\pi}(S_a) = \max_{u\in\mathcal{U}} \sum_{s'\in\mathcal{S}} \mathbb{P}_u(s'|S_c, L)V_r^{\pi}(s'), \tag{2.22}$$

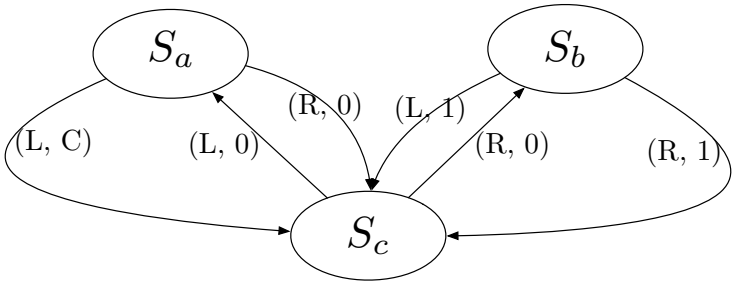


Figure 2: Example of a robust MDP where Lemma 2.6 holds with strict inequality.

$$\sum_{s'\in\mathcal{S}} \mathbb{P}_{u_{\pi^*}}(s'|S_c,R)V_r^{\pi}(s') = V_r^{\pi}(S_b) < V_r^{\pi}(S_a) = \max_{u\in\mathcal{U}} \sum_{s'\in\mathcal{S}} \mathbb{P}_{u}(s'|S_c,R)V_r^{\pi}(s'). \tag{2.23}$$

Now let $\xi_u^{\pi^*}(S_c)$ denote the trajectory generated by π^* within $\mathcal{M}_{\pi^*}$, starting from state S_c. Then we have

$$\begin{aligned}
\mathbb{E}_{\xi_u^{\pi^*}(S_c)}\left[c(s_0,a_0)+\gamma V_r^{\pi}(s_1)\right] &= \mathbb{E}_{\xi_u^{\pi^*}(S_c)}\left[c(s_0,a_0)+\gamma\sum_{s'\in\mathcal{S}}\mathbb{P}_{u_{\pi^*}}(s'|s_0,a_0)V_r^{\pi}(s')\right] \\
&= \left[c(S_c,L)+\gamma\sum_{s'\in\mathcal{S}}\mathbb{P}_{u_{\pi^*}}(s'|S_c,L)V_r^{\pi}(s')\right]\pi^*(L|S_c) \\
&\quad + \left[c(S_c,R)+\gamma\sum_{s'\in\mathcal{S}}\mathbb{P}_{u_{\pi^*}}(s'|S_c,R)V_r^{\pi}(s')\right]\pi^*(R|S_c) \\
&< \left[c(S_c,L)+\gamma\max_{u\in\mathcal{U}}\sum_{s'\in\mathcal{S}}\mathbb{P}_{u}(s'|S_c,L)V_r^{\pi}(s')\right]\pi^*(L|S_c) \\
&\quad + \left[c(S_c,R)+\gamma\max_{u\in\mathcal{U}}\sum_{s'\in\mathcal{S}}\mathbb{P}_{u}(s'|S_c,R)V_r^{\pi}(s')\right]\pi^*(R|S_c) \\
&= \mathbb{E}_{\xi_u^{\pi^*}(S_c)}\left[c(s_0,a_0)+\gamma\max_{u}\sum_{s'\in\mathcal{S}}\mathbb{P}_{u}(s'|s_0,a_0)V_r^{\pi}(s')\right] \\
&= \mathbb{E}_{\xi_u^{\pi^*}(S_c)}\left[Q_r^{\pi}(s_0,a_0)\right],
\end{aligned}$$

where the strict inequality follows from observation (2.22) and (2.23). Thus by repeating the same argument in the proof of Lemma 2.6, we know that inequality (d) therein holds with strict inequality for $s = S_c$. Since (d) is the only inequality in the proof of Lemma 2.6, we conclude that

$$V_r^{\pi^*}(S_c) - V_r^{\pi}(S_c) < \tfrac{1}{1-\gamma}\mathbb{E}_{s'\sim d_{S_c}^{\pi^*,u_{\pi^*}}}\langle Q_r^{\pi}, \pi^*-\pi\rangle_{s'}. \tag{2.24}$$

We proceed to show that $\mathbb{E}_{s'\sim d_{S_c}^{\pi^*,u_{\pi^*}}}\langle Q_r^{\pi}, \pi^*-\pi\rangle_{s'} = 0$ and hence establishing (2.20). To see this, note that $Q_r^{\pi}(S_c,\cdot) = \gamma V_r^{\pi}(S_a)$, and hence $\langle Q_r^{\pi}, \pi^*-\pi\rangle_{S_c} = 0$. In addition, we also have $Q_r^{\pi}(S_b,\cdot) = 1+\gamma V_r^{\pi}(S_c)$, and hence $\langle Q_r^{\pi}, \pi^*-\pi\rangle_{S_b} = 0$. Finally, note that $d_{S_c}^{\pi^*,u_{\pi^*}}(S_a) = 0$, and hence $\mathbb{E}_{s'\sim d_{S_c}^{\pi^*,u_{\pi^*}}}\langle Q_r^{\pi}, \pi^*-\pi\rangle_{s'} = 0$.

To show (2.21), it suffices to note that $\nu^* = \mathrm{Unif}(\{S_b, S_c\})$ is the stationary state distribution of π^* within MDP $\mathcal{M}_{u_{\pi^*}}$. With the same arguments as in the last paragraph we can show that $\mathbb{E}_{s'\sim d_{S_b}^{\pi^*,u_{\pi^*}}}\langle Q_r^{\pi}, \pi^*-\pi\rangle_{s'} = 0$. Hence we obtain

$$0 = \mathbb{E}_{s\sim\nu^*}\mathbb{E}_{s'\sim d_s^{\pi^*,u_{\pi^*}}}\langle Q_r^{\pi}, \pi^*-\pi\rangle_{s'} < (1-\gamma)\left(f_{\nu^*}(\pi) - f_{\nu^*}(\pi^*)\right),$$

where the strict inequality follows from (2.24). The claim then follows from the definition of $\mathcal{G}_s$ in (2.16). □

The construction of strict inequality in Proposition 2.3 suggests that (2.13) in Lemma 2.6 fails to characterize even the simplest change of robust values when switching the policy. In particular, as illustrated in (2.20), improvement of value when switching from π to the optimal π^* seems not being captured by the aggregated inner product defined in the second term of (2.20) at all. A closer look into the constructed example shows that the state S_a is the only state where the inner product is nonzero (positive), but when changing to the optimal policy, the state S_a is never visited by the optimal policy via $d_{\widetilde{s}}^{\pi^*,u_{\pi^*}}$, for any $\widetilde{s} \neq S_a$. It seems to suggest that the fundamental difficulty causing the insufficiency of Lemma 2.6 is due to the reason that the very state (i.e., state S_a) that leads to the improvement of policy is *never visited by the improved policy* π^* *within its worst-case environment* $\mathcal{M}_{u^{\pi^*}}$.

The following lemma aims to address the previously observed difficulty.

Lemma 2.7. For any policy π, let u_π denote its worst-case uncertainty defined in (2.2), then we have

$$\mathbb{E}_{s'\sim d_s^{\pi^*,u_\pi}}\left[\langle Q_r^\pi, \pi - \pi^*\rangle_{s'}\right] \geq (1-\gamma)\left(V_r^\pi(s) - V_r^{\pi^*}(s)\right). \tag{2.25}$$

Proof. We have

$$\begin{aligned}
\mathbb{E}_{s'\sim d_s^{\pi^*,u_\pi}}\left[\langle Q_r^\pi, \pi - \pi^*\rangle_{s'}\right] &= -\mathbb{E}_{s'\sim d_s^{\pi^*,u_\pi}}\left[\langle Q_r^\pi, \pi^* - \pi\rangle_{s'}\right] \\
&\overset{(a)}{=} -\mathbb{E}_{s'\sim d_s^{\pi^*,u_\pi}}\left[\langle Q_{u_\pi}^\pi, \pi^* - \pi\rangle_{s'}\right] \\
&\overset{(b)}{=} (1-\gamma)\left(V_{u_\pi}^\pi(s) - V_{u^\pi}^{\pi^*}(s)\right) \\
&\overset{(c)}{=} (1-\gamma)\left(V_r^\pi(s) - V_{u^\pi}^{\pi^*}(s)\right) \\
&\overset{(d)}{\geq} (1-\gamma)\left(V_r^\pi(s) - V_r^{\pi^*}(s)\right),
\end{aligned}$$

where (a) follows from (2.7), (c) follows from the (2.3), (d) follows from the definition of $V_r^{\pi^*} = \max_{u\in\mathcal{U}} V_u^{\pi^*}$. We now proceed to establish (b).

The proof of (b) follows similar lines as in the proof of Lemma 2.6. Let $\xi_{u_\pi}^{\pi^*}(s) \coloneqq \{(s_t, a_t)\}_{t\geq 0}$ denote the trajectories generated by policy π^*, within MDP $\mathcal{M}_{u_\pi}$, with initial state set as s. That is, $a_t \sim \pi^*(\cdot|s_t)$, $s_{t+1} \sim \mathbb{P}_{u_{\pi^*}}(\cdot|s_t, a_t)$ for all $t \geq 0$ and $s_0 = s$. We then have

$$\begin{aligned}
V_{u_\pi}^{\pi^*}(s) - V_{u_\pi}^\pi(s) &= \mathbb{E}_{\xi_{u_\pi}^{\pi^*}(s)}\left[\sum_{t=0}^\infty \gamma^t c(s_t, a_t)\right] - V_{u_\pi}^\pi(s) \\
&= \mathbb{E}_{\xi_{u_\pi}^{\pi^*}(s)}\left[\sum_{t=0}^\infty \gamma^t \left[c(s_t, a_t) + V_{u_\pi}^\pi(s_t) - V_{u_\pi}^\pi(s_t)\right]\right] - V_{u_\pi}^\pi(s) \\
&= \mathbb{E}_{\xi_{u_\pi}^{\pi^*}(s)}\left[\sum_{t=0}^\infty \gamma^t \left[c(s_t, a_t) + \gamma V_{u_\pi}^\pi(s_{t+1}) - V_{u_\pi}^\pi(s_t)\right]\right] + \mathbb{E}_{\xi_{u_\pi}^{\pi^*}(s)}\left[V_{u_\pi}^\pi(s_0)\right] - V_{u_\pi}^\pi(s) \\
&= \mathbb{E}_{\xi_{u_\pi}^{\pi^*}(s)}\left[\sum_{t=0}^\infty \gamma^t \left[c(s_t, a_t) + \gamma V_{u_\pi}^\pi(s_{t+1}) - V_{u_\pi}^\pi(s_t)\right]\right] \\
&= \mathbb{E}_{\xi_{u_\pi}^{\pi^*}(s)}\left[\sum_{t=0}^\infty \gamma^t \left[Q_{u_\pi}^\pi(s_t, a_t) - V_{u_\pi}^\pi(s_t)\right]\right] \\
&\overset{(e)}{=} \tfrac{1}{1-\gamma}\sum_{s'\in\mathcal{S}} d_s^{\pi^*,u_\pi}(s') \sum_{a'\in\mathcal{A}} \pi^*(a'|s') Q_{u_\pi}^\pi(s', a') - \tfrac{1}{1-\gamma}\sum_{s'\in\mathcal{S}} d_s^{\pi^*,u_\pi}(s') \sum_{a'\in\mathcal{A}} \pi(a'|s') Q_{u_\pi}^\pi(s', a') \\
&= \tfrac{1}{1-\gamma}\mathbb{E}_{s'\sim d_s^{\pi^*,u_\pi}}\langle Q_{u_\pi}^\pi, \pi^* - \pi\rangle_{s'},
\end{aligned}$$

where (e) follows from (1.6). □

By comparing Proposition 2.3 and Lemma 2.7, it should be clear that the key to inequality (2.25) is the combination of policy and the environment when choosing the state-visitation measure. Instead of choosing the worst-case environment u_{π^*} for π^* as in Proposition 2.3 (Lemma 2.6), we choose the worst-case environment u_π of the current policy π. In particular, returning back to the constructed example in Proposition 2.3, we see that the state (i.e., state S_a) that leads to the improvement of policy can be now visited by the improved policy π^*, if the environment is still fixed as the worst-case environment $\mathcal{M}_{u^\pi}$ for the original policy π.

Before we end our discussion in this section, it is worth mentioning pointing out some similarities that robust state-action value function Q_r^π shares with the (sub)gradient in the convex optimization literature,

in the sense that with proper aggregation over states: (1) it provides an upper bound on the local value changes, as shown in Lemma 2.6, much similar to using gradient-based local quadratic approximation for smooth objectives [19] (with the quadratic term being identically 0); (2) its inner product with the direction to the optimal policy is lower bounded by the optimality gap, as shown in Lemma 2.7, similar to the (sub)-gradient for convex objectives [19, 25]. These observations thus suggest using Q_r^π as the first-order information to update the policy. Nevertheless, it should be noted that the objective (1.5) is neither convex nor smooth [1, 42]. In the following section, we develop the robust policy mirror descent method that formalizes this intuition, and establish its computational efficiency in finding an optimal policy.

3 Robust Policy Mirror Descent

In this section, we introduce the deterministic robust policy mirror descent method (RPMD) for solving (1.5). RPMD assumes an access to an oracle that outputs the robust Q-function Q_r^π of a given policy π. At each iteration, the RPMD method (Algorithm 1) updates the policy according to

$$\pi_{k+1}(\cdot|s) = \operatorname*{argmin}_{p(\cdot|s)\in\Delta_{|\mathcal{A}|}} \eta_k \langle Q_r^{\pi_k}(s,\cdot), p(\cdot|s)\rangle + D^p_{\pi_k}(s),\ \forall s\in\mathcal{S}. \tag{3.1}$$

Here $D^\pi_{\pi'}(s)$ denotes the Bregman divergence between policy $\pi(\cdot|s)$ and $\pi'(\cdot|s)$, defined as

$$D^\pi_{\pi'}(s) = w(\pi(\cdot|s)) - w(\pi'(\cdot|s)) - \langle \partial w(\pi'(\cdot|s)), \pi(\cdot|s) - \pi'(\cdot|s)\rangle, \tag{3.2}$$

where $w : \mathbb{R}^{|\mathcal{A}|} \to \mathbb{R}$ is a strictly convex function, also known as the distance-generating function, and $\partial w(p)$ denotes a subgradient of w at $p \in \mathbb{R}^{|\mathcal{A}|}$. Common distance-generating functions include $w(p) = \|p\|_2^2$, which induces $D^\pi_{\pi'}(s) = \|\pi(\cdot|s) - \pi'(\cdot|s)\|_2^2$; and $w(p) = \sum_{a\in\mathcal{A}} p_a \log p_a \coloneqq -\mathrm{Ent}(p)$, which induces

$$D^\pi_{\pi'}(s) = \sum_{a\in\mathcal{A}} \pi(a|s) \log\big(\pi(a|s)/\pi'(a|s)\big) \coloneqq \mathrm{KL}(\pi(\cdot|s)\|\pi'(\cdot|s)).$$

We will also denote $D_w = \max_{\pi\in\Pi}\max_{s\in\mathcal{S}} D^\pi_{\pi_0}(s)$, which is finite for many practical divergences when π_0 corresponds to the uniform policy, including the previously introduced KL-divergence and the squared ℓ_2 distance.

It is also worth mentioning that RPMD with KL-divergence yields an equivalent policy update for the natural policy gradient method [13] applied to the softmax parameterization for the robust MDP objective (1.5), by directly applying Lemma 2.5. The same equivalence has also been observed for solving the non-robust MDP problem in the literature.

We proceed to establish some general convergence properties of RPMD. To begin with, the following lemma characterizes each policy update of RPMD.

Lemma 3.1. For any $p \in \Delta_{|\mathcal{A}|}$ and any $s \in \mathcal{S}$, we have

$$\eta_k \langle Q_r^{\pi_k}(s,\cdot), \pi_{k+1}(\cdot|s) - p\rangle + D^{\pi_{k+1}}_{\pi_k}(s) \le D^p_{\pi_k}(s) - D^p_{\pi_{k+1}}(s). \tag{3.3}$$

Algorithm 1 The robust policy mirror descent (RPMD) method

Input: Initial policy π_0 and stepsizes $\{\eta_k\}_{k\ge 0}$.
for $k = 0, 1, \ldots$ **do**
Update policy:

$$\pi_{k+1}(\cdot|s) = \operatorname*{argmin}_{p(\cdot|s)\in\Delta_{|\mathcal{A}|}} \eta_k \langle Q_r^{\pi_k}(s,\cdot), p(\cdot|s)\rangle + D^p_{\pi_k}(s),\ \forall s\in\mathcal{S}.$$

end for

Proof. From the optimality condition of the RPMD update (3.1), we have for any $p \in \Delta_{|\mathcal{A}|}$,

$$\eta_k \langle Q_r^{\pi_k}(s,\cdot), p - \pi_{k+1}(\cdot|s)\rangle + \langle \nabla D_{\pi_k}^{\pi_{k+1}}(s), p - \pi_{k+1}(\cdot|s)\rangle \geq 0, \tag{3.4}$$

In addition, given the definition of Bregman divergence, we have the following identity

$$\langle \nabla D_{\pi_k}^{\pi_{k+1}}(s), p - \pi_{k+1}(\cdot|s)\rangle = D_{\pi_k}^{p}(s) - D_{\pi_k}^{\pi_{k+1}}(s) - D_{\pi_{k+1}}^{p}(s).$$

Combining the previous observation with (3.4), we immediately obtain the result. □

The next lemma then establishes the basic convergence properties of RPMD.

Lemma 3.2. At each iteration of RPMD, we have

$$f_\rho(\pi_{k+1}) - f_\rho(\pi^*) \leq \left(1 - \tfrac{1-\gamma}{M}\right)\left(f_\rho(\pi_k) - f_\rho(\pi^*)\right) + \tfrac{1}{M\eta_k}\mathbb{E}_{s\sim d_\rho^{\pi^*,u_k}} D_{\pi_k}^{\pi^*}(s) - \tfrac{1}{M\eta_k}\mathbb{E}_{s\sim d_\rho^{\pi^*,u_k}} D_{\pi_{k+1}}^{\pi^*}(s), \tag{3.5}$$

where $M \coloneqq \sup_{u\in\mathcal{U}} \left\| d_\rho^{\pi^*,u}/\rho \right\|_\infty < \infty$ for every ρ with $\mathrm{supp}(\rho) = \mathcal{S}$.

Proof. By plugging in $p = \pi_k$ in (3.3), we obtain

$$\eta_k \langle Q_r^{\pi_k}(s,\cdot), \pi_{k+1}(\cdot|s) - \pi_k(\cdot|s)\rangle \leq -D_{\pi_{k+1}}^{\pi_k}(s) - D_{\pi_k}^{\pi_{k+1}}(s) \leq 0, \ \forall s \in \mathcal{S}. \tag{3.6}$$

On the other hand, plugging in $p = \pi^*$ in (3.3), we obtain

$$\underbrace{\eta_k \langle Q_r^{\pi_k}(s,\cdot), \pi_{k+1}(\cdot|s) - \pi_k(\cdot|s)\rangle}_{(A)} + \underbrace{\eta_k \langle Q_r^{\pi_k}(s,\cdot), \pi_k(\cdot|s) - \pi^*(\cdot|s)\rangle}_{(B)} + D_{\pi_k}^{\pi_{k+1}}(s) \leq D_{\pi_k}^{\pi^*}(s) - D_{\pi_{k+1}}^{\pi^*}(s). \tag{3.7}$$

We let $u_k = u_{\pi_k}$ denote the worst-case uncertainty of policy π_k for any $k \geq 0$.

To handle term (A), note that

$$\begin{aligned} V_r^{\pi_{k+1}}(s) - V_r^{\pi_k}(s) &\overset{(a)}{\leq} \tfrac{1}{1-\gamma}\mathbb{E}_{s'\sim d_s^{\pi_{k+1},u_{k+1}}} \langle Q_r^{\pi_k}, \pi_{k+1} - \pi_k\rangle_{s'} \\ &\overset{(b)}{\leq} \tfrac{d_s^{\pi_{k+1},u_{k+1}}(s)}{1-\gamma} \langle Q_r^{\pi_k}(s,\cdot), \pi_{k+1}(\cdot|s) - \pi_k(\cdot|s)\rangle \\ &\overset{(c)}{\leq} \langle Q_r^{\pi_k}(s,\cdot), \pi_{k+1}(\cdot|s) - \pi_k(\cdot|s)\rangle \leq 0, \end{aligned} \tag{3.8}$$

where (a) is due to Lemma 2.6; (b) is due to (3.6); (c) is due to (3.6), and the observation that $d_s^{\pi_{k+1},u_{k+1}}(s) \geq (1-\gamma)$ for all $s \in \mathcal{S}$.

To handle term (B), we make use of Lemma 2.7. Specifically, we obtain from (2.25) that

$$\mathbb{E}_{s'\sim d_s^{\pi^*,u_k}} \left[\langle Q_r^{\pi_k}, \pi_k - \pi^*\rangle_{s'}\right] \geq (1-\gamma)\left(V_r^{\pi}(s) - V_r^{\pi^*}(s)\right) \geq 0. \tag{3.9}$$

Hence by aggregating (3.7) across different states with weights set as $d_s^{\pi^*,u_k}$, and making use of (3.8) and (3.9), we obtain

$$\begin{aligned} &\mathbb{E}_{s'\sim d_s^{\pi^*,u_k}} \left[\eta_k \left(V_r^{\pi_{k+1}}(s') - V_r^{\pi_k}(s')\right)\right] + (1-\gamma)\eta_k \left(V_r^{\pi_k}(s) - V_r^{\pi^*}(s)\right) \\ \leq &\mathbb{E}_{s'\sim d_s^{\pi^*,u_k}} D_{\pi_k}^{\pi^*}(s') - \mathbb{E}_{s'\sim d_s^{\pi^*,u_k}} D_{\pi_{k+1}}^{\pi^*}(s'). \end{aligned}$$

By further taking expectation with respect to $s \sim \rho$, and making use of $V_r^{\pi_{k+1}}(s) \leq V_r^{\pi_k}(s)$ given (3.8), and the definition of $M \coloneqq \sup_{u\in\mathcal{U}} \left\| d_\rho^{\pi^*,u}/\rho \right\|_\infty < \infty$,

$$M\left[f_\rho(\pi_{k+1}) - f_\rho(\pi_k)\right] + (1-\gamma)\left[f_\rho(\pi_k) - f_\rho(x^*)\right] \leq \tfrac{1}{\eta_k}\mathbb{E}_{s\sim d_\rho^{\pi^*,u_k}} D_{\pi_k}^{\pi^*}(s) - \tfrac{1}{\eta_k}\mathbb{E}_{s\sim d_\rho^{\pi^*,u_k}} D_{\pi_{k+1}}^{\pi^*}(s),$$

which after simple arrangement, gives

$$f_\rho(\pi_{k+1}) - f_\rho(\pi^*) \leq \left(1 - \tfrac{1-\gamma}{M}\right)\left(f_\rho(\pi_k) - f_\rho(\pi^*)\right) + \tfrac{1}{M\eta_k}\mathbb{E}_{s\sim d_\rho^{\pi^*,u_k}} D_{\pi_k}^{\pi^*}(s) - \tfrac{1}{M\eta_k}\mathbb{E}_{s\sim d_\rho^{\pi^*,u_k}} D_{\pi_{k+1}}^{\pi^*}(s).$$

□

3.1 Convergence with Increasing Stepsizes

We start by showing that by applying exponentially increasing stepsizes, RPMD achieves linear convergence in the optimality gap.

Theorem 3.1. *Suppose the stepsizes* $\{\eta_k\}$ *satisfy*

$$\eta_k \geq \eta_{k-1}\left(1-\tfrac{1-\gamma}{M}\right)^{-1} M', \quad \forall k \geq 1, \tag{3.10}$$

where $M' = \sup_{u,u' \in \mathcal{U}} \left\| d_\rho^{\pi^*,u}/d_\rho^{\pi^*,u'} \right\|_\infty < \infty$ *for every* ρ *with* $\operatorname{supp}(\rho) = \mathcal{S}$. *Then for any iteration* k, *RPMD produces policy* π_k *satisfying*

$$f_\rho(\pi_k) - f_\rho(\pi^*) \leq \left(1-\tfrac{1-\gamma}{M}\right)^k (f_\rho(\pi_0) - f_\rho(\pi^*)) + \left(1-\tfrac{1-\gamma}{M}\right)^{k-1} \tfrac{D_w}{M\eta_0}.$$

Proof. By recursively applying inequality (3.5) from $t=0$ to $k-1$, we obtain

$$\begin{aligned} f_\rho(\pi_k) - f_\rho(\pi^*) \leq{}& \left(1-\tfrac{1-\gamma}{M}\right)^k (f_\rho(\pi_0) - f_\rho(\pi^*)) + \left(1-\tfrac{1-\gamma}{M}\right)^{k-1} \tfrac{1}{M\eta_0} \mathbb{E}_{s\sim d_\rho^{\pi^*,u_0}} D_{\pi_0}^{\pi^*}(s) \\ &+ \underbrace{\tfrac{1}{M}\sum_{t=1}^{k-1} \left[\left(1-\tfrac{1-\gamma}{M}\right)^{k-t-1} \tfrac{1}{\eta_t} \mathbb{E}_{s\sim d_\rho^{\pi^*,u_t}} D_{\pi_t}^{\pi^*}(s) - \left(1-\tfrac{1-\gamma}{M}\right)^{k-t} \tfrac{1}{\eta_{t-1}} \mathbb{E}_{s\sim d_\rho^{\pi^*,u_{t-1}}} D_{\pi_t}^{\pi^*}(s) \right]}_{(C)}. \end{aligned}$$

Now define $M' = \sup_{u,u' \in \mathcal{U}} \left\| d_\rho^{\pi^*,u}/d_\rho^{\pi^*,u'} \right\|_\infty < \infty$, and let stepsizes $\{\eta_k\}$ satisfy

$$\eta_k \geq \eta_{k-1}\left(1-\tfrac{1-\gamma}{M}\right)^{-1} M',$$

then it should be clear that term $(C) \leq 0$. In conclusion, we obtain that whenever (3.10) holds, then

$$\begin{aligned} f_\rho(\pi_k) - f_\rho(\pi^*) &\leq \left(1-\tfrac{1-\gamma}{M}\right)^k (f_\rho(\pi_0) - f_\rho(\pi^*)) + \left(1-\tfrac{1-\gamma}{M}\right)^{k-1} \tfrac{1}{M\eta_0} \mathbb{E}_{s\sim d_\rho^{\pi^*,u_0}} D_{\pi_0}^{\pi^*}(s) \\ &\leq \left(1-\tfrac{1-\gamma}{M}\right)^k (f_\rho(\pi_0) - f_\rho(\pi^*)) + \left(1-\tfrac{1-\gamma}{M}\right)^{k-1} \tfrac{D_w}{M\eta_0}. \end{aligned}$$

□

It should be noted that whenever $\mathcal{U} = \{\mathbf{0}\}$, solving the robust MDP $\mathcal{M}_U$ is equivalent to solving the nominal MDP $\mathcal{M}$. Then we have $M' = 1$, and $M = \left\| d_\rho^{\pi^*,\mathbf{0}}/\rho \right\|_\infty$ reduces to the mismatch coefficient defined in the analysis of policy gradient methods for solving nominal MDP [1, 12, 44]. In this case, RPMD admits a simple learning rate scaling rule given by $\eta_t \geq \gamma^{-1}\eta_{t-1}$, and the obtained convergence rate for RPMD matches exactly the fastest existing rate of convergence of first-order methods for solving the nominal MDP [20, 23, 44].

By applying a less aggressive stepsize scheme, we can also obtain the following sublinear convergence of RPMD.

Theorem 3.2. *Suppose the stepsizes* $\{\eta_k\}$ *satisfy* $\eta_k \geq \eta_{k-1}M'$ *for all* $k \geq 1$, *where* M' *is defined as in Theorem 3.1. Then for every* ρ *with* $\operatorname{supp}(\rho) = \mathcal{S}$, *at any iteration* k, *RPMD produces policy* π_k *satisfying*

$$f_\rho(\pi_k) - f(\pi^*) \leq \tfrac{M}{(1-\gamma)k}\left(f_\rho(\pi_0) - f_\rho(\pi^*) + \tfrac{D_w}{M\eta_0}\right),$$

where M *is defined as in Lemma 3.2.*

Proof. By summing up inequality (3.5) from $t = 0$ to $k-1$, we obtain

$$\begin{aligned}
&f_\rho(\pi_k) - f_\rho(\pi^*) + \tfrac{1-\gamma}{M}\sum_{t=1}^{k-1}\left(f_\rho(\pi_t) - f_\rho(\pi^*)\right) \\
\leq & f_\rho(\pi_0) - f_\rho(\pi^*) + \tfrac{1}{M\eta_0}\mathbb{E}_{s\sim d_\rho^{\pi^*,u_0}} D_{\pi_0}^{\pi^*}(s) + \underbrace{\sum_{t=1}^{k-1}\left(\tfrac{1}{M\eta_t}\mathbb{E}_{s\sim d_\rho^{\pi^*,u_t}} D_{\pi_t}^{\pi^*}(s) - \tfrac{1}{M\eta_{t-1}}\mathbb{E}_{s\sim d_\rho^{\pi^*,u_{t-1}}} D_{\pi_t}^{\pi^*}(s)\right)}_{(A)}.
\end{aligned}$$

Now suppose stepsizes $\{\eta_k\}$ satisfy $\eta_k \geq \eta_{k-1}M'$, we then obtain term $(A) \leq 0$, and hence

$$f_\rho(\pi_k) - f_\rho(\pi^*) + \tfrac{1-\gamma}{M}\sum_{t=1}^{k-1}\left(f_\rho(\pi_t) - f_\rho(\pi^*)\right) \leq f_\rho(\pi_0) - f_\rho(\pi^*) + \tfrac{1}{M\eta_0}\mathbb{E}_{s\sim d_\rho^{\pi^*,u_0}} D_{\pi_0}^{\pi^*}(s).$$

In addition, given (3.8), we know that $f_\rho(\pi_{t+1}) \leq f_\rho(\pi_t)$ for all $t \geq 0$. Thus we further obtain

$$\tfrac{(1-\gamma)k}{M}\left(f_\rho(\pi_k) - f_\rho(\pi^*)\right) \leq f_\rho(\pi_0) - f_\rho(\pi^*) + \tfrac{1}{M\eta_0}\mathbb{E}_{s\sim d_\rho^{\pi^*,u_0}} D_{\pi_0}^{\pi^*}(s),$$

which is equivalent to

$$\begin{aligned}
f_\rho(\pi_k) - f(\pi^*) &\leq \tfrac{M}{(1-\gamma)k}\left(f_\rho(\pi_0) - f_\rho(\pi^*) + \tfrac{1}{M\eta_0}\mathbb{E}_{s\sim d_\rho^{\pi^*,u_0}} D_{\pi_0}^{\pi^*}(s)\right) \\
&\leq \tfrac{M}{(1-\gamma)k}\left(f_\rho(\pi_0) - f_\rho(\pi^*) + \tfrac{D_w}{M\eta_0}\right).
\end{aligned}$$

□

We add a few remarks regarding the convergence characterizations developed in this subsection. Firstly, both the linear convergence and the $\mathcal{O}(1/k)$ sublinear convergence rates seem to be new in the existing literature of first-order policy-based methods applied to solving the robust MDPs. Secondly, although the constant M' is in general unknown, the stepsize requirements in both Theorem 3.1 and 3.2 only needs $\eta_k \gtrapprox M'\eta_{k-1}$, hence one can simply provide a pessimistic upper bound of M', and still retain the established convergence rates. In particular, for $\rho = \text{Unif}(\mathcal{S})$, one can simply take $M' = |\mathcal{S}|/(1-\gamma)$. Finally, the increasing-stepsize schemes are not the only ones that can certify the convergence of RPMD. In the next subsection, we will establish the convergence of RPMD with a constant-stepsize scheme, which allows one to completely avoid the estimation of M'.

3.2 Convergence with Constant Stepsizes

In this subsection, we establish the convergence of RPMD with constant stepsize. For RPMD with euclidean Bregman divergence ($w(\cdot) = \|\cdot\|_2^2$) and a constant stepsize of η, we establish an $\mathcal{O}(\max\{1/k, 1/\sqrt{\eta k}\})$ rate of convergence for arbitrary rectangular uncertainty set. For RPMD with a general class of Bregman divergences, we establish a similar convergence rate for rectangular uncertainty set that satisfies relative strong convexity.

3.2.1 Euclidean Divergence

We proceed to show that when $w(\cdot) = \|\cdot\|_2^2$, i.e., the Bregman divergence $D_x^{x'} = \|x - x'\|_2^2$ reduces to the standard euclidean distance, then RPMD achieves sublinear convergence of the last-iterate with constant stepsizes. We first establish the following simple fact on the asymptotic stationarity of the RPMD iterates.

Lemma 3.3. For any $k \geq 1$, the iterates in RPMD with constant stepsizes $\eta_k = \eta > 0$ satisfy

$$\tfrac{1}{\eta}\sum_{t=0}^{k-1}\left(D_{\pi_{t+1}}^{\pi_t}(s) + D_{\pi_t}^{\pi_{t+1}}(s)\right) \leq V_r^{\pi_0}(s) - V_r^{\pi^*}(s). \tag{3.11}$$

Proof. Given (3.8) and (3.6), we obtain that

$$V_r^{\pi_{k+1}}(s) - V_r^{\pi_k}(s) \leq -\tfrac{1}{\eta} D_{\pi_{k+1}}^{\pi_k}(s) - \tfrac{1}{\eta} D_{\pi_k}^{\pi_{k+1}}(s) \leq 0, \ \forall s \in \mathcal{S}.$$

Summing up the prior relation from $t = 0$ to $k-1$, we obtain

$$\tfrac{1}{\eta} \sum_{t=0}^{k-1} \left(D_{\pi_{t+1}}^{\pi_t}(s) + D_{\pi_t}^{\pi_{t+1}}(s) \right) \leq V_r^{\pi_0}(s) - V_r^{\pi_k}(s) \leq V_r^{\pi_0}(s) - V_r^{\pi^*}(s).$$

□

Combining Lemma 3.3 and Lemma 3.2, we are able to establish the following convergence characterization for RPMD with euclidean Bregman divergence, when adopting any constant-stepsize scheme.

Theorem 3.3. *Let $w(\cdot) = \|\cdot\|_2^2$ be the distance-generating function, and $\eta_t = \eta$ for all $t \geq 0$ and any $\eta > 0$, then at each iteration, RPMD outputs policy π_k satisfying*

$$f_\rho(\pi_k) - f_\rho(\pi^*) \leq \tfrac{M}{(1-\gamma)k} \left(f_\rho(\pi_0) - f_\rho(\pi^*) \right) + \sqrt{\tfrac{18|\mathcal{S}|^2}{\eta k (1-\gamma)^3}}, \tag{3.12}$$

where M is defined as in Lemma 3.2.

Proof. By summing up inequality (3.5) from $t = 0$ to $k-1$, we obtain

$$\begin{aligned} & f_\rho(\pi_k) - f_\rho(\pi^*) + \tfrac{1-\gamma}{M} \sum_{t=1}^{k-1} \left(f_\rho(\pi_t) - f_\rho(\pi^*) \right) \\ \leq & f_\rho(\pi_0) - f_\rho(\pi^*) + \sum_{t=0}^{k-1} \left(\tfrac{1}{M\eta} \mathbb{E}_{s \sim d_\rho^{\pi^*, u_t}} D_{\pi_t}^{\pi^*}(s) - \tfrac{1}{M\eta} \mathbb{E}_{s \sim d_\rho^{\pi^*, u_t}} D_{\pi_{t+1}}^{\pi^*}(s) \right) \end{aligned}$$

Note that $D_\pi^{\pi'}(s) = \|\pi'(\cdot|s) - \pi(\cdot|s)\|_2^2$, and hence

$$D_{\pi_t}^{\pi^*}(s) - D_{\pi_{t+1}}^{\pi^*}(s) \leq \sqrt{18 D_{\pi_{t+1}}^{\pi_t}(s)},$$

where we use the fact that for any $a, b, c \in \Delta_{\mathcal{A}}$,

$$\begin{aligned} \sum_i (a_i - b_i)^2 - \sum_i (a_i - c_i)^2 &\leq \|b + c - 2a\|_2 \|b - c\|_2 \\ &\leq \left(3 \left(\|b\|_2^2 + \|c\|_2^2 + 4 \|a\|_2^2 \right) \right)^{1/2} \|b - c\|_2 \leq \sqrt{18} \|b - c\|_2 . \end{aligned}$$

Thus we obtain

$$f_\rho(\pi_k) - f_\rho(\pi^*) + \tfrac{1-\gamma}{M} \sum_{t=1}^{k-1} \left(f_\rho(\pi_t) - f_\rho(\pi^*) \right) \leq f_\rho(\pi_0) - f_\rho(\pi^*) + \tfrac{\sqrt{18}}{M\eta} \sum_{t=0}^{k-1} \mathbb{E}_{s \sim d_\rho^{\pi^*, u_t}} \sqrt{D_{\pi_{t+1}}^{\pi_t}(s)}.$$

To proceed, we will make use of Lemma 3.3, which gives

$$\left(\sum_{t=0}^{k-1} \sqrt{D_{\pi_{t+1}}^{\pi_t}(s)} \right)^2 \leq k \sum_{t=0}^{k-1} D_{\pi_{t+1}}^{\pi_t}(s) \leq k\eta \left(V_r^{\pi_0}(s) - V_r^{\pi^*}(s) \right) \leq \tfrac{k\eta}{1-\gamma}, \quad \forall s \in \mathcal{S},$$

which combined with the previous inequality, gives

$$f_\rho(\pi_k) - f_\rho(\pi^*) + \tfrac{1-\gamma}{M} \sum_{t=1}^{k-1} \left(f_\rho(\pi_t) - f_\rho(\pi^*) \right) \leq f_\rho(\pi_0) - f_\rho(\pi^*) + \tfrac{|\mathcal{S}|}{M} \sqrt{\tfrac{18k}{\eta(1-\gamma)}}.$$

By (3.8), we know that $f_\rho(\pi_{t+1}) \le f_\rho(\pi_t)$ for all $t \ge 0$. Hence we can conclude that

$$f_\rho(\pi_k) - f_\rho(\pi^*) \le \frac{M}{(1-\gamma)k}\left(f_\rho(\pi_0) - f_\rho(\pi^*)\right) + \sqrt{\frac{18|\mathcal{S}|^2}{\eta k(1-\gamma)^3}}.$$

□

It should be noted that the dependence on the size of the state space in the second term of (3.12) can be simply eliminated by using a large stepsize. Theorem 3.3 states that to find an ϵ-optimal policy using constant stepsizes RPMD with euclidean divergence, the iteration complexity is bounded by $\mathcal{O}\left(\max\left\{1/\epsilon, 1/(\eta\epsilon^2)\right\}\right)$, which reduces to $\mathcal{O}(1/\epsilon)$ when choosing $\eta = \Theta(1/\epsilon)$. To the best of our knowledge, the fastest existing rate of convergence of euclidean-divergence based first-order methods for solving robust MDP is $\mathcal{O}(1/\epsilon^3)$ with $\eta = \Theta(\epsilon)$ [42], which studies a more restrictive subclass of polyhedral rectangular uncertainty set, and further requires a delicate smoothing technique and selecting the best policy among historical policy iterates. In contrast, the RPMD comes with a much stronger convergence guarantee for general rectangular uncertainty sets, while at the same time enjoying greater algorithmic simplicity. In particular, RPMD finds an ϵ-optimal policy within $\mathcal{O}(1/\epsilon)$ iterations for large stepsizes, and one can simply choose the last-iterate as the output policy.

3.2.2 A General Class of Bregman Divergences

In this subsection, we show that for a general class of Bregman divergences, whenever the set of uncertain transitions $\mathcal{P}_{s,a}$ form a relatively strongly convex set in $\mathbb{R}^{|\mathcal{S}|}$, then RPMD converges with any constant stepsize for a general class of divergences. To proceed, we first recall the following definition of strongly convex sets.

Definition 3.1 (Strongly Convex Set)**.** A set $\mathcal{C}$ in $\mathbb{R}^d$ is called strongly convex with respect to $R > 0$, if $\mathcal{C}$ is bounded, and for any $x, y \in \mathcal{C}$, any $\lambda \in [0,1]$, we have $\mathcal{B}(\lambda x + (1-\lambda)y), \delta) \subset \mathcal{C}$, where $\delta = \frac{\lambda(1-\lambda)}{2R}\|x-y\|_2^2$. Here $\mathcal{B}(z,r) = \left\{z' \in \mathbb{R}^d : \|z'-z\|_2 \le r\right\}$ denotes the ball centered around z with radius r.

It is well known that any level set of a strongly convex function is a strongly convex set [40]. In particular, any full-dimensional ellipsoid is a strongly convex set. A strongly convex convex set also satisfies the following useful property.

Lemma 3.4 (Theorem 1, [40])**.** For any set $\mathcal{C} \subset \mathbb{R}^d$ that is strongly convex with respect to $R > 0$, let $\mathrm{Bd}(C)$ denote its boundary, and $\mathcal{N}_\mathcal{C}(x)$ denote its normal cone at $x \in X$. Then for any pair of vectors (x_i, p_i), $i = 1, 2$, where $x_i \in \mathrm{Bd}(\mathcal{C})$ and $p_i \in \mathcal{N}_\mathcal{C}(x_i)$ with $\|p_i\|_2 = 1$, we have

$$\|x_1 - x_2\|_2 \le R\,\|p_1 - p_2\|\,. \tag{3.13}$$

Note that by definition, a strongly convex set must have full dimension, which can not be satisfied by $\mathcal{P}_{s,a}$ as $\mathcal{P}_{s,a} \subset \Delta_{|\mathcal{A}|}$ lies in a lower dimensional hyperplane. Given this observation, we define the following strong convexity of a set $\mathcal{S}$ restricted to its affine span.

Definition 3.2 (Relatively Strongly Convex Set)**.** A set $\mathcal{C}$ in $\mathbb{R}^d$ is called relatively strongly convex with respect to $R > 0$, if $\mathcal{C}$ is bounded, and for any $x, y \in \mathcal{C}$, any $\lambda \in [0,1]$, we have $\mathcal{B}(\lambda x + (1-\lambda)y), \delta) \cap \mathcal{H}_\mathcal{C} \subset \mathcal{C}$, where $\delta = \frac{\lambda(1-\lambda)}{2R}\|x-y\|_2^2$. Here $\mathcal{B}(z,r) = \left\{z' \in \mathbb{R}^d : \|z'-z\|_2 \le r\right\}$ denotes the ball centered around z with radius r, and $\mathcal{H}_\mathcal{C} = \mathrm{Aff}(\mathcal{C})$ denotes the affine subspace spanned by $\mathcal{C}$.

Similar to Lemma 3.4, the relatively strongly convex set also satisfies an analogy to (3.13).

Lemma 3.5. For any set $\mathcal{C} \subset \mathbb{R}^d$ that is relatively strongly convex with respect to $R > 0$, let $\mathrm{ReBd}(C)$ denote its relative boundary, and $\mathcal{N}_\mathcal{C}^r(x)$ denote its normal cone at $x \in X$ restricted to $\mathcal{H}_\mathcal{C}$, i.e.,

$$\mathcal{N}_\mathcal{C}^r(x) = \left\{z \in \mathrm{Lin}(\mathcal{C}) : (y-x)^\top z \le 0, \forall y \in \mathcal{C}\right\},$$

where $\mathrm{Lin}(\mathcal{C})$ denotes the unique linear subspace defining $\mathcal{H}_{\mathcal{C}}$, which satisfies $\mathcal{H}_{\mathcal{C}} = x + \mathrm{Lin}(C)$ for any $x \in \mathcal{C}$. Then for any pair of vectors (x_i, p_i), $i = 1, 2$, where $x_i \in \mathrm{ReBd}(\mathcal{C})$ and $p_i \in \mathcal{N}_{\mathcal{C}}^r(x_i)$ with $\|p_i\|_2 = 1$, we have

$$\|x_1 - x_2\|_2 \le R \|p_1 - p_2\| . \tag{3.14}$$

Proof. Note that $\mathcal{N}_{\mathcal{C}}^r(x)$ is invariant to translation of set C. Moreover, the both sides of statement (3.14) are also invariant to translation of set C, and hence we can without loss of generality assume $\mathbf{0} \in \mathcal{C}$, and hence $\mathcal{H}_{\mathcal{C}} = \mathrm{Lin}(\mathcal{C})$. The rest of the claim follows directly from a change of coordinates argument , and working in the lower-dimensional space $\mathrm{Lin}(\mathcal{C})$, in which we can apply Lemma 3.4. □

We then show that if set of uncertain transitions $\mathcal{P}_{s,a}$ is a relatively strongly convex set in $\mathbb{R}^{|\mathcal{S}|}$ with dimension $|\mathcal{S}| - 1$, then the worst-case transition kernel $\mathbb{P}_{u_\pi}$ is uniquely defined, and is also Lipschitz continuous with respect to policy π. En route, we will make use of the following simple fact.

Fact 3.4. For $x, y \in \mathbb{R}^d$ and $x, y \neq \mathbf{0}$, we have

$$\left\| \tfrac{x}{\|x\|_2} - \tfrac{y}{\|y\|_2} \right\|_2 \le \|x - y\|_2 / \min\{\|x\|_2, \|y\|_2\} . \tag{3.15}$$

Proof. Since both side of claim (3.15) are symmetric with respect to (x, y), we can without loss of generality assume $\|y\|_2 \ge \|x\|_2$. Note that $\left\| \tfrac{x}{\|x\|_2} - \tfrac{y}{\|y\|_2} \right\|_2 = \tfrac{1}{\|x\|_2} \left\| x - \tfrac{y}{\|y\|_2} \|x\|_2 \right\|_2$. We make the following observations.

Denote $\mathrm{Proj}_y : \mathbb{R}^d \to \mathbb{R}^d$ as the projection operator onto $\mathrm{span}(\{y\})$. We know that $\mathrm{Proj}_y(x) = \delta_x y$ for $\delta_x = \langle x, y \rangle / \|y\|_2^2$ and $\mathrm{Proj}_y(x) \le \|x\|_2$. On the other hand, we have $x' \coloneqq \tfrac{\|x\|_2}{\|y\|_2} y = \delta'_x y$ with $\|x'\|_2 = \|x\|_2$, hence $\delta'_x \ge |\delta_x|$. Since $x' \in \mathrm{span}(\{y\})$, we have

$$\|x - x'\|_2^2 = \|x - \mathrm{Proj}_y(x)\|_2^2 + \|\mathrm{Proj}_y(x) - x'\|_2^2 = \|x - \mathrm{Proj}_y(x)\|_2^2 + (\delta'_x - \delta_x)^2 \|y\|_2^2 .$$

On the other hand, since $\delta'_x \le 1$, we also have

$$\|x - \mathrm{Proj}_y(x)\|_2^2 + (\delta'_x - \delta_x)^2 \|y\|_2^2 \le \|x - \mathrm{Proj}_y(x)\|_2^2 + (1 - \delta_x)^2 \|y\|_2^2 = \|x - y\|_2^2 .$$

Hence, we obtain $\|x - x'\|_2^2 \le \|x - y\|_2^2$. In conclusion, since we have assumed $\|x\|_2 \le \|y\|_2$, we obtain

$$\left\| \tfrac{x}{\|x\|_2} - \tfrac{y}{\|y\|_2} \right\|_2 \le \|x - y\|_2 / \min\{\|x\|_2, \|y\|_2\} .$$

□

Lemma 3.6. Suppose $\mathcal{P}_{s,a}$ is relatively strongly convex with respect to $R_{s,a} > 0$ for all $(s, a) \in \mathcal{S} \times \mathcal{A}$, and $\dim(\mathcal{P}_{s,a}) = |\mathcal{S}| - 1$. Let $r_* = \min_{\pi \in \Pi} \|\mathrm{Proj}_{\mathcal{H}}(V_r^\pi)\|_2$, where $\mathcal{H} = \mathrm{Lin}(\mathcal{P}_{s,a}) = \{x \in \mathbb{R}^{|\mathcal{S}|} : \mathbf{1}^\top x = 0\}$. Suppose $r_* > 0$, then for any policy $\pi \in \Pi$, the worst-case environment $\mathbb{P}_{u_\pi}$ defined in (2.2) is unique, and

$$\left\| \mathbb{P}_{u_\pi}(\cdot|s,a) - \mathbb{P}_{u_{\pi'}}(\cdot|s,a) \right\|_2 \le \tfrac{R_{s,a}}{r_*} \left\| V_r^\pi - V_r^{\pi'} \right\|_2 , \quad \forall (s,a) \in \mathcal{S} \times \mathcal{A}. \tag{3.16}$$

Proof. We first recall from (2.2) that for any policy $\pi \in \Pi$, the worse case transition $\mathbb{P}_{u_\pi}$ is given by

$$\mathbb{P}_{u_\pi}(\cdot|s,a) \in \operatorname*{argmax}_{u(\cdot|s,a) \in \mathcal{U}_{s,a}} \sum_{s' \in \mathcal{S}} \mathbb{P}_u(s'|s,a) V_r^\pi(s') = \operatorname*{argmax}_{p \in \mathcal{P}_{s,a}} p^\top V_r^\pi, \quad \forall (s,a) \in \mathcal{S} \times \mathcal{A}. \tag{3.17}$$

It is worth mentioning that the solution to (3.17) remains unchanged when we shift V_r^π by $\delta \mathbf{1}$ for any $\delta \in \mathbb{R}$, where $\mathbf{1} \in \mathbb{R}^{|\mathcal{S}|}$ denotes the all-one vector. To see this, note that

$$p^\top (V_r^\pi + \delta \mathbf{1}) = p^\top V_r^\pi + \delta \tag{3.18}$$

due to $p \in \mathcal{P}_{s,a}$, and hence shifting V_r^π only changes the objective by a constant δ for every feasible p. Let $\text{Proj}_{\mathcal{H}} : \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}^{|\mathcal{S}|}$ denote the projection operator onto $\mathcal{H}$, then given observation (3.18), we must have

$$\mathbb{P}_{u_\pi}(\cdot|s,a) \in \operatorname*{argmax}_{p \in \mathcal{P}_{s,a}} p^\top \text{Proj}_{\mathcal{H}}(V_r^\pi), \quad \forall (s,a) \in \mathcal{S} \times \mathcal{A}, \ \forall \pi \in \Pi. \tag{3.19}$$

Let $\text{ReBd}(\mathcal{P}_{s,a})$ denote the relative boundary of $\mathcal{P}_{s,a}$. We claim that for any p^* that is a solution of (3.17) (or equivalently (3.19)), we must have $p \in \text{ReBd}(\mathcal{P}_{s,a})$. Suppose not, and p^* is in the relative interior of $\mathcal{P}_{s,a}$, then there exists $\delta > 0$ such that $\mathcal{B}(p^*, \delta) \cap \mathcal{H} \subset \mathcal{P}_{s,a}$. Now it is immediate to see that $p^* + \delta \text{Proj}_{\mathcal{H}}(V_r^\pi) / \left\|\text{Proj}_{\mathcal{H}}(V_r^\pi)\right\|_2$ is a strictly better solution than p^* unless $\text{Proj}_{\mathcal{H}}(V_r^\pi) = \mathbf{0}$, which can not happen since $r_* > 0$.

We proceed to show that p^* is indeed unique. Suppose not, and $p_1 \neq p_2 \in \mathcal{P}_{s,a}$ are two solutions of (3.17). From the relative strong convexity of set $\mathcal{P}_{s,a}$, we know that for any $\lambda \in (0,1)$, we have $p_\lambda = \lambda p_1 + (1-\lambda) p_2$ in the relative interior of $\mathcal{P}_{s,a}$, which contradicts with the previously established fact that any solution of (3.17) is in $\text{ReBd}(\mathcal{P}_{s,a})$.

By reading the optimality condition of the (3.19), we have

$$(p - \mathbb{P}_{u_\pi}(\cdot|s,a))^\top \text{Proj}_{\mathcal{H}}(V_r^\pi) \leq 0, \quad \forall p \in \mathcal{P}_{s,a}. \tag{3.20}$$

Hence we immediately see that $\mathbb{P}_{u_\pi}(\cdot|s,a) \in \text{ReBd}(\mathcal{P}_{s,a})$, and $\text{Proj}_{\mathcal{H}}(V_r^\pi) \in \mathcal{N}_{\mathcal{P}_{s,a}}(\mathbb{P}_{u_\pi}(\cdot|s,a))$. Then we can apply Lemma 3.4 and obtain

$$\left\|\mathbb{P}_{u_\pi}(\cdot|s,a) - \mathbb{P}_{u_{\pi'}}(\cdot|s,a)\right\|_2 \leq R_{s,a} \left\|\overline{V}_r^\pi - \overline{V}_r^{\pi'}\right\|_2, \quad \forall (s,a) \in \mathcal{S} \times \mathcal{A},$$

where $\overline{V}_r^\pi = \text{Proj}_{\mathcal{H}}(V_r^\pi) / \left\|\text{Proj}_{\mathcal{H}}(V_r^\pi)\right\|_2$, $\overline{V}_r^{\pi'} = \text{Proj}_{\mathcal{H}}(V_r^{\pi'}) / \left\|\text{Proj}_{\mathcal{H}}(V_r^{\pi'})\right\|_2$. Now applying Fact 3.4, we obtain

$$\begin{aligned}\left\|\mathbb{P}_{u_\pi}(\cdot|s,a) - \mathbb{P}_{u_{\pi'}}(\cdot|s,a)\right\|_2 &\leq R_{s,a} \left\|\text{Proj}_{\mathcal{H}}(V_r^\pi) - \text{Proj}_{\mathcal{H}}(V_r^{\pi'})\right\|_2 / r_* \\ &\leq \tfrac{R_{s,a}}{r_*} \left\|V_r^\pi - V_r^{\pi'}\right\|_2,\end{aligned}$$

where $r_* = \min_{\pi \in \Pi} \left\|\text{Proj}_{\mathcal{H}}(V_r^\pi)\right\|_2$, and the last inequality uses the non-expansiveness of the projection operator. □

It should be noted that the conditions on the relative strong convexity of $\mathcal{P}_{s,a}$ and $\dim(\mathcal{P}_{s,a}) = |\mathcal{S}| - 1$ are readily satisfied by many common uncertainty sets. For instance, one can take $\mathcal{P}_{s,a}$ as the intersection of a full-dimensional ellipsoid with $\Delta_{\mathcal{S}}$ (i.e., ellipsoidal uncertainty [32]). On the other hand, Lemma 3.6 relies on the key condition that $r_* = \min_{\pi \in \Pi} \left\|\text{Proj}_{\mathcal{H}}(V_r^\pi)\right\|_2 > 0$. We next provide a simple necessary and sufficient condition that certifies $r_* > 0$.

Lemma 3.7. Suppose the cost function c satisfies $\cap_{s \in \mathcal{S}} \left[\min_{a \in \mathcal{A}} c(s,a), \max_{a \in \mathcal{A}} c(s,a)\right] = \emptyset$, then we must have $r_* = \min_{\pi \in \Pi} \left\|\text{Proj}_{\mathcal{H}}(V_r^\pi)\right\|_2 > 0$, where $\mathcal{H} = \left\{x \in \mathbb{R}^{|\mathcal{S}|} : \mathbf{1}^\top x = 0\right\}$. The converse of the statement is also true.

Proof. "⇒": Suppose the cost function satisfies c satisfies $\cap_{s \in \mathcal{S}} \left[\min_{a \in \mathcal{A}} c(s,a), \max_{a \in \mathcal{A}} c(s,a)\right] = \emptyset$, but $r_* = 0$. Then we have $V_r^\pi = \lambda \mathbf{1}$ for some $\lambda > 0$ (since $V^{\pi_r} > 0$ must hold) and $\pi \in \Pi$. Let us define $\mathbb{P}_{u_\pi}^\pi : \mathcal{S} \times \mathcal{S} \to [0,1]$ by $\mathbb{P}_{u_\pi}^\pi(s,s') = \sum_{a \in \mathcal{A}} \mathbb{P}_{u_\pi}(s'|s,a)\pi(a|s)$, and $c^\pi : \mathcal{S} \to \mathbb{R}$ by $c^\pi(s) = \sum_{a \in \mathcal{A}} \pi(a|s) c(a|s)$, then given observation (2.3) and the standard Bellman condition of $V_{u_\pi}^\pi$, we know that $V_r^\pi = \left(I - \gamma \mathbb{P}_{u_\pi}^\pi\right)^{-1} c^\pi$, which gives

$$c^\pi = \left(I - \gamma \mathbb{P}_{u_\pi}^\pi\right) V_r^\pi = V_r^\pi - \gamma \mathbb{P}_{u_\pi}^\pi V_r^\pi \overset{(a)}{=} \lambda \mathbf{1} - \gamma \lambda \mathbf{1} = (1-\gamma)\lambda \mathbf{1}, \tag{3.21}$$

where in term (a) we uses the fact that $\mathbb{P}^{\pi}_{u_\pi}\mathbf{1}=\mathbf{1}$. However, (3.21) must contradict with the condition that $\cap_{s\in\mathcal{S}}\left[\min_{a\in\mathcal{A}}c(s,a),\max_{a\in\mathcal{A}}c(s,a)\right]=\emptyset$, since for each $s\in\mathcal{S}$, we have $c^{\pi}(s)\in\left[\min_{a\in\mathcal{A}}c(s,a),\max_{a\in\mathcal{A}}c(s,a)\right]$.

"$\Leftarrow$": On the other hand, suppose $r_*>0$ but $\cap_{s\in\mathcal{S}}\left[\min_{a\in\mathcal{A}}c(s,a),\max_{a\in\mathcal{A}}c(s,a)\right]\neq\emptyset$, then one can readily find a policy π such that $c^{\pi}=\lambda\mathbf{1}$ for some $\lambda\in\mathbb{R}$. Thus we have

$$V_r^{\pi}=\left(I-\gamma\mathbb{P}^{\pi}_{u_\pi}\right)^{-1}c^{\pi}\overset{(b)}{=}\sum_{t=0}^{\infty}\gamma^t\left(\mathbb{P}^{\pi}_{u_\pi}\right)^t c^{\pi}=\lambda\sum_{t=0}^{\infty}\gamma^t\mathbf{1}=\tfrac{\lambda}{1-\gamma}\mathbf{1},$$

where in term (b) we use the fact that $\rho(\gamma P^{\pi}_{u_\pi})\leq\gamma$, where $\rho(A)$ denotes the spectral radius of matrix A. Note that the previous observation then implies $r_*\leq\left\|\mathrm{Proj}_{\mathcal{H}}(V_r^{\pi})\right\|_2=0$. □

We proceed to show that $V_r^{\pi}(s)$ is Lipschitz continuous in π.

Lemma 3.8. For any policies $\pi,\pi'\in\Pi$, we have

$$\left\|V_r^{\pi}-V_r^{\pi'}\right\|_{\infty}\leq\tfrac{1}{1-\gamma}\left\|\pi-\pi'\right\|_{\infty},\tag{3.22}$$

where we define $\left\|\pi-\pi'\right\|_{\infty}=\sup_{s\in\mathcal{S}}\left\|\pi(\cdot|s)-\pi'(\cdot|s)\right\|_1$, i.e., the matrix ℓ_∞-norm when viewing π as a matrix in $\mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$.

Proof. Note that for any $u\in\mathcal{U}$, from the performance difference lemma of standard MDPs [12, 20], we have

$$|V_u^{\pi}(s)-V_u^{\pi}(s)|=\left|\mathbb{E}_{s'\sim d_s^{\pi,u}}\left[\langle Q_u^{\pi},\pi-\pi'\rangle_{s'}\right]\right|\leq\tfrac{1}{1-\gamma}\left\|\pi-\pi'\right\|_{\infty},$$

where in the last inequality we use the Cauchy-Schwarz inequality, and the fact that $\left\|Q_u^{\pi}\right\|_{\infty}\leq\frac{1}{1-\gamma}$. Hence we obtain

$$|V_r^{\pi}(s)-V_r^{\pi'}(s)|=\left|\sup_{u\in\mathcal{U}}V_u^{\pi}(s)-\sup_{u\in\mathcal{U}}V_u^{\pi}(s)\right|\leq\sup_{u\in\mathcal{U}}|V_u^{\pi}(s)-V_u^{\pi}(s)|\leq\tfrac{1}{1-\gamma}\left\|\pi-\pi'\right\|_{\infty}.$$

□

By combining Lemma 3.8 and Lemma 3.6, we are ready to establish the Lipschitz continuity of d^{π^*,u_π} with respect to policy π.

Lemma 3.9. Suppose $\mathcal{P}_{s,a}$ is relatively strongly convex with respect to $R_{s,a}>0$ for all $(s,a)\in\mathcal{S}\times\mathcal{A}$ with $R=\max_{s\in\mathcal{S},a\in\mathcal{A}}R_{s,a}$, and $\dim(\mathcal{P}_{s,a})=|\mathcal{S}|-1$. Let $r_*=\min_{\pi\in\Pi}\left\|\mathrm{Proj}_{\mathcal{H}}(V_r^{\pi})\right\|_2$, where $\mathcal{H}=\left\{x\in\mathbb{R}^{|\mathcal{S}|}:\mathbf{1}^{\top}x=0\right\}$. Suppose $r_*>0$, then we have

$$\left\|d_\rho^{\pi^*,u_\pi}-d_\rho^{\pi^*,u_{\pi'}}\right\|_1\leq\tfrac{\gamma|\mathcal{A}||\mathcal{S}|R}{(1-\gamma)^2r_*}\left\|\pi-\pi'\right\|_{\infty}.\tag{3.23}$$

Proof. Let us define $\mathbb{P}_u^{\pi}:\mathcal{S}\times\mathcal{S}\to[0,1]$ by $\mathbb{P}_u^{\pi}(s',s)=\sum_{a\in\mathcal{A}}\mathbb{P}_u(s'|s,a)\pi(a|s)$, then for any $\pi\in\Pi$, and $\rho\in\Delta_{\mathcal{S}}$, we obtain

$$d_\rho^{\pi,u}=(1-\gamma)\sum_{t=0}^{\infty}\gamma^t(\mathbb{P}_u^{\pi})^t\rho=(1-\gamma)\left(1-\gamma\mathbb{P}_u^{\pi}\right)^{-1}\rho,\tag{3.24}$$

where in the last inequality we use the fact that $\rho(\gamma\mathbb{P}_u^{\pi})\leq\gamma<1$. Hence we have

$$\begin{aligned}d_\rho^{\pi^*,u_\pi}-d_\rho^{\pi^*,u_{\pi'}}&=(1-\gamma)\left(\left(I-\gamma\mathbb{P}^{\pi^*}_{u_\pi}\right)^{-1}-\left(I-\gamma\mathbb{P}^{\pi^*}_{u'_\pi}\right)^{-1}\right)\rho\\&=(1-\gamma)\gamma\left(I-\gamma\mathbb{P}^{\pi^*}_{u_\pi}\right)^{-1}\underbrace{\left(\mathbb{P}^{\pi^*}_{u_\pi}-\mathbb{P}^{\pi^*}_{u'_\pi}\right)}_{\Delta^{\pi}_{\pi'}}\left(I-\gamma\mathbb{P}^{\pi^*}_{u'_\pi}\right)^{-1}\rho,\end{aligned}$$

where the last equality uses the matrix identity $A^{-1} - B^{-1} = A^{-1}(B - A)B^{-1}$ for any invertible matrix A, B. Note that $\left\|\left(I - \gamma \mathbb{P}_{u_\pi}^{\pi^*}\right)^{-1}\right\|_1 \leq (1-\gamma)^{-1}$ as we have shown in (2.10), it suffices to control the ℓ_1-norm of $\Delta_{\pi'}^{\pi} = \mathbb{P}_{u_\pi}^{\pi^*} - \mathbb{P}_{u'_\pi}^{\pi^*}$. We now make the following observations

$$\begin{aligned}
\sum_{s' \in \mathcal{S}} |\mathbb{P}_{u_\pi}^{\pi^*}(s', s) - \mathbb{P}_{u'_\pi}^{\pi^*}(s', s)| &= \sum_{s' \in \mathcal{S}} |\sum_{a \in \mathcal{A}} \left(\mathbb{P}_{u_\pi}(s'|s, a) - \mathbb{P}_{u_{\pi'}}(s'|s, a)\right) \pi^*(a|s)| \\
&\leq \sum_{a \in \mathcal{A}} \sum_{s' \in \mathcal{S}} |\mathbb{P}_{u_\pi}(s'|s, a) - \mathbb{P}_{u_{\pi'}}(s'|s, a)| \pi^*(a|s) \\
&\leq \sup_{a \in \mathcal{A}} \left\|\mathbb{P}_{u_\pi}(\cdot|s, a) - \mathbb{P}_{u_{\pi'}}(\cdot|s, a)\right\|_1 \\
&\overset{(a)}{\leq} \frac{R\sqrt{|\mathcal{S}|}}{r_*} \left\|V_r^{\pi} - V_r^{\pi'}\right\|_2 \\
&\overset{(b)}{\leq} \frac{R|\mathcal{S}|}{r_*(1-\gamma)} \left\|\pi - \pi'\right\|_\infty
\end{aligned}$$

where (a) uses Lemma 3.6 and the definition of R, (b) uses Lemma 3.8, From the prior inequality, we obtain $\left\|\Delta_{\pi'}^{\pi}\right\|_1 \leq \frac{R|\mathcal{S}|}{(1-\gamma)r_*} \left\|\pi - \pi'\right\|_\infty$. Putting everything together, we conclude that

$$\begin{aligned}
\left\|d_\rho^{\pi^*, u_\pi} - d_\rho^{\pi^*, u_{\pi'}}\right\|_1 &\leq (1-\gamma)\gamma \left\|\left(I - \gamma \mathbb{P}_{u_\pi}^{\pi^*}\right)^{-1}\right\|_1 \left\|\Delta_{\pi'}^{\pi}\right\|_1 \left\|\left(I - \gamma \mathbb{P}_{u'_\pi}^{\pi^*}\right)^{-1}\right\|_1 \|\rho\|_1 \\
&\leq \frac{\gamma|\mathcal{S}|R}{(1-\gamma)^2 r_*} \left\|\pi - \pi'\right\|_\infty .
\end{aligned}$$

□

Combining elements established above, we are now ready to establish the main result of this subsection.

Theorem 3.5. *Suppose $\mathcal{P}_{s,a}$ is relatively strongly convex with respect to $R_{s,a} > 0$ for all $(s, a) \in \mathcal{S} \times \mathcal{A}$ with $R = \max_{s \in \mathcal{S}, a \in \mathcal{A}} R_{s,a}$, and $\dim(\mathcal{P}_{s,a}) = |\mathcal{S}| - 1$. Let $r_* = \min_{\pi \in \Pi} \|\mathrm{Proj}_{\mathcal{H}}(V_r^{\pi})\|_2$, where $\mathcal{H} = \left\{x \in \mathbb{R}^{|\mathcal{S}|} : \mathbf{1}^\top x = 0\right\}$, and suppose $r_* > 0$.*

For any $\eta > 0$, let $\eta_k = \eta > 0$ for all $k \geq 0$. If (1) Distance-generating function w is μ-strongly convex with respect to $\|\cdot\|_1$-norm; (2) $D_w^* = \sup_{\pi \in \Pi} D_\pi^{\pi^*}(s) < \infty$. Then RPMD outputs policy π_k satisfying

$$f_\rho(\pi_k) - f_\rho(\pi^*) \leq \frac{M}{(1-\gamma)k} \left(f_\rho(\pi_0) - f_\rho(\pi^*)\right) + \frac{D_w}{(1-\gamma)\eta k} + \frac{2\gamma|\mathcal{S}|^2 R D_w^*}{(1-\gamma)^{7/2} r_* \sqrt{\eta \mu k}}, \tag{3.25}$$

where M is defined as in Lemma 3.2.

Proof. Recall that from the proof of Theorem 3.2, we have

$$\begin{aligned}
&f_\rho(\pi_k) - f_\rho(\pi^*) + \tfrac{1-\gamma}{M} \sum_{t=1}^{k-1} \left(f_\rho(\pi_t) - f_\rho(\pi^*)\right) \\
\leq &f_\rho(\pi_0) - f_\rho(\pi^*) + \tfrac{1}{M\eta} \mathbb{E}_{s \sim d_\rho^{\pi^*, u_0}} D_{\pi_0}^{\pi^*}(s) + \underbrace{\sum_{t=1}^{k-1} \left(\tfrac{1}{M\eta} \mathbb{E}_{s \sim d_\rho^{\pi^*, u_t}} D_{\pi_t}^{\pi^*}(s) - \tfrac{1}{M\eta} \mathbb{E}_{s \sim d_\rho^{\pi^*, u_{t-1}}} D_{\pi_t}^{\pi^*}(s)\right)}_{(A)}
\end{aligned} \tag{3.26}$$

We will make use of Lemma 3.9 to bound term (A). Specifically, we have

$$\begin{aligned}
|\mathbb{E}_{s \sim d_\rho^{\pi^*, u_t}} D_{\pi_t}^{\pi^*}(s) - \mathbb{E}_{s \sim d_\rho^{\pi^*, u_{t-1}}} D_{\pi_t}^{\pi^*}(s)| &\overset{(a)}{\leq} \left\|d_\rho^{\pi^*, u_t} - d_\rho^{\pi^*, u_{t-1}}\right\|_1 D_w^* \\
&\overset{(b)}{\leq} \frac{\gamma|\mathcal{S}|RD_w^*}{(1-\gamma)^2 r_*} \left\|\pi_t - \pi_{t-1}\right\|_\infty ,
\end{aligned} \tag{3.27}$$

where (a) uses Holder's inequality, and (b) uses Lemma 3.9, and the definition that $u_t = u_{\pi_t}$ for all $t \geq 0$. Since $w(\cdot)$ is μ-strongly convex with respect to $\|\cdot\|_1$-norm, we have

$$\|\pi_t - \pi_{t-1}\|_\infty = \sup_{s\in\mathcal{S}} \|\pi_t(\cdot|s) - \pi_{t-1}(\cdot|s)\|_1 \leq \sup_{s\in\mathcal{S}} \sqrt{2D^{\pi_t}_{\pi_{t-1}}(s)/\mu},$$

which combined with (3.26) and (3.27), gives

$$\begin{aligned}
&f_\rho(\pi_k) - f_\rho(\pi^*) + \tfrac{1-\gamma}{M}\sum_{t=1}^{k-1}\left(f_\rho(\pi_t) - f_\rho(\pi^*)\right)\\
\leq& f_\rho(\pi_0) - f_\rho(\pi^*) + \tfrac{1}{M\eta}\mathbb{E}_{s\sim d_\rho^{\pi^*,u_0}} D^{\pi^*}_{\pi_0}(s) + \tfrac{\gamma|\mathcal{S}|RD^*_w}{(1-\gamma)^2 r_* M\eta}\sum_{t=1}^{k-1}\sum_{s\in\mathcal{S}}\sqrt{\tfrac{2D^{\pi_t}_{\pi_{t-1}}(s)}{\mu}}.
\end{aligned}$$

We then make use of Lemma 3.3, which gives

$$\left(\sum_{t=1}^{k-1}\sqrt{D^{\pi_t}_{\pi_{t-1}}(s)}\right)^2 \leq k\sum_{t=1}^{k-1} D^{\pi_t}_{\pi_{t-1}}(s) \leq k\eta\left(V_r^{\pi_0}(s) - V_r^{\pi^*}(s)\right) \leq \tfrac{k\eta}{1-\gamma}, \quad \forall s\in\mathcal{S}.$$

By combining two prior relations, we obtain

$$\begin{aligned}
&f_\rho(\pi_k) - f_\rho(\pi^*) + \tfrac{1-\gamma}{M}\sum_{t=1}^{k-1}\left(f_\rho(\pi_t) - f_\rho(\pi^*)\right)\\
\leq& f_\rho(\pi_0) - f_\rho(\pi^*) + \tfrac{1}{M\eta}\mathbb{E}_{s\sim d_\rho^{\pi^*,u_0}} D^{\pi^*}_{\pi_0}(s) + \tfrac{2\gamma|\mathcal{S}|^2RD^*_w\sqrt{k}}{(1-\gamma)^{5/2}r_* M\sqrt{\eta\mu}}.
\end{aligned}$$

By (3.8), we know that $f_\rho(\pi_{t+1}) \leq f_\rho(\pi_t)$ for all $t \geq 0$. Hence we can conclude that

$$f_\rho(\pi_k) - f_\rho(\pi^*) \leq \tfrac{M}{(1-\gamma)k}\left(f_\rho(\pi_0) - f_\rho(\pi^*)\right) + \tfrac{D_w}{(1-\gamma)\eta k} + \tfrac{2\gamma|\mathcal{S}|^2RD^*_w}{(1-\gamma)^{7/2}r_*\sqrt{\eta\mu k}}.$$

□

In view of Theorem 3.5, the RPMD method with constant stepsize η outputs an ϵ-optimal policy within $\mathcal{O}(\max\{1/\epsilon, 1/(\eta\epsilon^2)\})$ iterations, for a general class of Bregman divergences. Moreover, conditions (1) and (2) in Theorem 3.5 are both satisfied by some common distance-generating functions, including the squared ℓ_2-norm $w(p) = \|p\|_2^2$, and the negative Tsallis entropy $w(p) = \frac{k}{q-1}\left(\sum_i p_i^q - 1\right)$ for any entropic-index $q \in (1, 2]$. To the best of our knowledge, this result seem to be the first convergence result of policy-based first-order method with non-euclidean divergences.

In addition, we remark that the second term (3.25) also establishes a link on geometry of the set of uncertain probabilities $\mathcal{P}_{s,a}$ to the convergence rate. In addition, the dependence of convergence on the size of the state space is largely due to Lipschitz constant of $d_\rho^{\pi^*,u_\pi}$ with respect to policy π. The current characterization of such Lipschitz constant seems improvable with additional efforts. Nevertheless, it is worth mentioning that one can simply get rid of such a dependence by using a large stepsize.

4 Stochastic Robust Policy Mirror Descent

In this section, we extend the deterministic RPMD method to the stochastic settings, where the exact information of the robust state-action value function Q_r^π is not available. The stochastic robust policy mirror descent (SRPMD) instead uses the stochastic estimate $Q_r^{\pi,\xi}$ to update the policy, where ξ denotes the samples used for the construction of the stochastic estimate.

At iteration k, given a stochastic estimate $Q_r^{\pi_k,\xi_k}$, the SRPMD method (Algorithm 2) updates the policy according to

$$\pi_{k+1}(\cdot|s) = \operatorname*{argmin}_{p(\cdot|s)\in\Delta_{|\mathcal{A}|}} \eta_k \langle Q_r^{\pi_k,\xi_k}(s,\cdot), p(\cdot|s)\rangle + D_{\pi_k}^p(s),\ \forall s\in\mathcal{S}. \tag{4.1}$$

The convergence of SRPMD assumes the following noise condition on the noisy estimate $\left\{Q^{\pi_k,\xi_k}\right\}$,

$$\mathbb{E}_{\xi_k}\left\|Q_r^{\pi_k,\xi_k} - Q^{\pi_k}\right\|_\infty \le e_k. \tag{4.2}$$

We will also define $\delta_k = Q_r^{\pi_k,\xi_k} - Q_r^{\pi_k}$ for all $k\ge 0$.

Similar to Lemma 3.2, we can establish the following generic convergence property of the SRPMD method.

Lemma 4.1. At each iteration of SRPMD, we have

$$\begin{aligned} f_\rho(\pi_{k+1}) - f_\rho(\pi^*) \le \left(1-\tfrac{1-\gamma}{M}\right)\left(f_\rho(\pi_k) - f_\rho(\pi^*)\right) + \tfrac{1}{M\eta_k}\mathbb{E}_{s\sim d_\rho^{\pi^*,u_k}} D_{\pi_k}^{\pi^*}(s) \\ -\tfrac{1}{M\eta_k}\mathbb{E}_{s\sim d_\rho^{\pi^*,u_k}} D_{\pi_{k+1}}^{\pi^*}(s) + \tfrac{4}{1-\gamma}\left\|\delta_k\right\|_\infty, \end{aligned} \tag{4.3}$$

where M is defined as in Lemma 3.2.

Proof. First, following the same lines as in the proof of Lemma 3.1, we have

$$\eta_k\langle Q_r^{\pi_k,\xi_k}(s,\cdot), \pi_{k+1}(\cdot|s) - p\rangle + D_{\pi_k}^{\pi_{k+1}}(s) \le D_{\pi_k}^p(s) - D_{\pi_{k+1}}^p(s). \tag{4.4}$$

Thus by letting $p=\pi_k$ in (4.4), we obtain

$$\eta_k\langle Q_r^{\pi_k,\xi_k}(s,\cdot), \pi_{k+1}(\cdot|s) - \pi_k(\cdot|s)\rangle \le -D_{\pi_{k+1}}^{\pi_k}(s) - D_{\pi_k}^{\pi_{k+1}}(s) \le 0, \tag{4.5}$$

On the other hand, by plugging in $p=\pi^*$ in the above relation, we obtain

$$\underbrace{\eta_k\langle Q_r^{\pi_k,\xi_k}(s,\cdot), \pi_{k+1}(\cdot|s) - \pi_k(\cdot|s)\rangle}_{(A)} + \underbrace{\eta_k\langle Q_r^{\pi_k,\xi_k}(s,\cdot), \pi_k(\cdot|s) - \pi^*(\cdot|s)\rangle}_{(B)} + D_{\pi_k}^{\pi_{k+1}}(s) \le D_{\pi_k}^{\pi^*}(s) - D_{\pi_{k+1}}^{\pi^*}(s). \tag{4.6}$$

We let $u_k = u_{\pi_k}$ denote the worst-case uncertainty of policy π_k for any $k\ge 0$. To handle term (A), note that

$$\begin{aligned} V_r^{\pi_{k+1}}(s) - V_r^{\pi_k}(s) &\overset{(a)}{\le} \tfrac{1}{1-\gamma}\mathbb{E}_{s'\sim d_s^{\pi_{k+1},u_{k+1}}}\langle Q_r^{\pi_k}, \pi_{k+1}-\pi_k\rangle_{s'} \\ &= \sum_{s'\in\mathcal{S}} \tfrac{d_s^{\pi_{k+1},u_{k+1}}(s')}{1-\gamma}\left[\langle Q_r^{\pi_k,\xi_k}(s',\cdot), \pi_{k+1}(\cdot|s') - \pi_k(\cdot|s')\rangle + \langle \delta_k(s',\cdot), \pi_{k+1}(\cdot|s') - \pi_k(\cdot|s')\rangle\right] \end{aligned}$$

Algorithm 2 The stochastic robust policy mirror descent (SRPMD) method

Input: Initial policy π_0 and stepsizes $\{\eta_k\}_{k\ge 0}$.
for $k=0,1,\ldots$ **do**
Update policy:

$$\pi_{k+1}(\cdot|s) = \operatorname*{argmin}_{p(\cdot|s)\in\Delta_{|\mathcal{A}|}} \eta_k \langle Q_r^{\pi_k,\xi_k}(s,\cdot), p(\cdot|s)\rangle + D_{\pi_k}^p(s),\ \forall s\in\mathcal{S}.$$

end for

$$
\begin{aligned}
&\leq \tfrac{1}{1-\gamma} \sum_{s' \in \mathcal{S}} d_s^{\pi_{k+1}, u_{k+1}}(s') \left[\langle Q_r^{\pi_k, \xi_k}(s', \cdot), \pi_{k+1}(\cdot|s') - \pi_k(\cdot|s') \rangle + 2 \left\| \delta_k \right\|_\infty \right] \\
&\overset{(b)}{\leq} \tfrac{d_s^{\pi_{k+1}, u_{k+1}}(s)}{1-\gamma} \langle Q_r^{\pi_k, \xi_k}(s, \cdot), \pi_{k+1}(\cdot|s) - \pi_k(\cdot|s) \rangle + \tfrac{2}{1-\gamma} \left\| \delta_k \right\|_\infty \\
&\overset{(c)}{\leq} \langle Q_r^{\pi_k, \xi_k}(s, \cdot), \pi_{k+1}(\cdot|s) - \pi_k(\cdot|s) \rangle + \tfrac{2}{1-\gamma} \left\| \delta_k \right\|_\infty \\
&\leq \tfrac{2}{1-\gamma} \left\| \delta_k \right\|_\infty , \qquad (4.7)
\end{aligned}
$$

where (a) uses Lemma 2.6, (b) uses (4.5), and (c) uses again (4.5) and the fact that $d_s^{\pi_{k+1}, u_{k+1}}(s) \geq 1 - \gamma$. Hence we obtain from inequality (c) in the last relation that

$$
(A) \geq \eta_k \left(V_r^{\pi_{k+1}}(s) - V_r^{\pi_k}(s) \right) - \tfrac{2\eta_k}{1-\gamma} \left\| \delta_k \right\|_\infty . \qquad (4.8)
$$

For term (B), we have

$$
\begin{aligned}
\mathbb{E}_{s' \sim d_s^{\pi^*, u_k}} \left[\langle Q_r^{\pi_k, \xi_k}, \pi_k - \pi^* \rangle_{s'} \right] &= \mathbb{E}_{s' \sim d_s^{\pi^*, u_k}} \left[\langle Q_r^{\pi_k}, \pi_k - \pi^* \rangle_{s'} + \langle \delta_k, \pi_k - \pi^* \rangle_{s'} \right] \\
&\geq (1 - \gamma) \left(V_r^{\pi_k}(s) - V_r^{\pi^*}(s) \right) - 2 \left\| \delta_k \right\|_\infty , \qquad (4.9)
\end{aligned}
$$

where the inequality follows from (3.9). Hence by combining (4.6), (4.8) and (4.9), we obtain

$$
\begin{aligned}
&\mathbb{E}_{s' \sim d_s^{\pi^*, u_k}} \left(V_r^{\pi_{k+1}}(s') - V_r^{\pi_k}(s') - \tfrac{2\|\delta_k\|_\infty}{1-\gamma} \right) + (1 - \gamma) \left(V_r^{\pi}(s) - V_r^{\pi^*}(s) \right) + \tfrac{1}{\eta_k} \mathbb{E}_{s' \sim d_s^{\pi^*, u_k}} D_{\pi_k}^{\pi_{k+1}}(s') \\
&\qquad \leq \tfrac{1}{\eta_k} \mathbb{E}_{s' \sim d_s^{\pi^*, u_k}} D_{\pi_k}^{\pi^*}(s') - \tfrac{1}{\eta_k} \mathbb{E}_{s' \sim d_s^{\pi^*, u_k}} D_{\pi_{k+1}}^{\pi^*}(s') + 2 \left\| \delta_k \right\|_\infty .
\end{aligned}
$$

By using (4.7), the definition of M, and further taking $s \sim \rho$ in the above relation, we conclude that

$$
\begin{aligned}
&M \mathbb{E}_{s \sim \rho} \left[V_r^{\pi_{k+1}}(s) - V_r^{\pi_k}(s) \right] + (1 - \gamma) \mathbb{E}_{s \sim \rho} \left(V_r^{\pi}(s) - V_r^{\pi^*}(s) \right) + \mathbb{E}_{s \sim d_\rho^{\pi^*, u_k}} D_{\pi_k}^{\pi_{k+1}}(s) \\
&\qquad \leq \mathbb{E}_{s \sim d_\rho^{\pi^*, u_k}} D_{\pi_k}^{\pi^*}(s) - \mathbb{E}_{s \sim d_\rho^{\pi^*, u_k}} D_{\pi_{k+1}}^{\pi^*}(s) + \tfrac{2(1-\gamma)+2M}{1-\gamma} \left\| \delta_k \right\|_\infty .
\end{aligned}
$$

The claim follows immediately after simple rearrangement to the above inequality. □

By specializing Lemma 4.1 with exponentially increasing stepsizes, we obtain the following linear convergence of SRPMD up to a noise level determined by the noise in the stochastic estimation Q^{π_k, ξ_k}.

Theorem 4.1. *Suppose the stepsizes $\{\eta_k\}$ in SRPMD satisfy*

$$
\eta_k \geq \eta_{k-1} \left(1 - \tfrac{1-\gamma}{M} \right)^{-1} M', \quad \forall k \geq 1, \qquad (4.10)
$$

where M' is defined as in Theorem 3.1. Then for any iteration k, SRPMD produces policy π_k satisfying

$$
\begin{aligned}
\mathbb{E} \left[f_\rho(\pi_k) - f_\rho(\pi^*) \right] \leq \left(1 - \tfrac{1-\gamma}{M} \right)^k \left(f_\rho(\pi_0) - f_\rho(\pi^*) \right) + \left(1 - \tfrac{1-\gamma}{M} \right)^{k-1} \tfrac{D_w}{M \eta_0} \\
+ \tfrac{4}{1-\gamma} \sum_{t=0}^{k-1} \left(1 - \tfrac{1-\gamma}{M} \right)^{k-t-1} e_t ,
\end{aligned}
$$

where M is defined as in Lemma 3.2. In particular, if we have $\mathbb{E}_{\xi_k} \left\| Q_r^{\pi_k, \xi_k} - Q^{\pi_k} \right\|_\infty \leq e$ for all $k \geq 0$, then

$$
\mathbb{E} \left[f_\rho(\pi_k) - f_\rho(\pi^*) \right] \leq \left(1 - \tfrac{1-\gamma}{M} \right)^k \left(f_\rho(\pi_0) - f_\rho(\pi^*) \right) + \left(1 - \tfrac{1-\gamma}{M} \right)^{k-1} \tfrac{D_w}{M \eta_0} + \tfrac{4Me}{(1-\gamma)^2} . \qquad (4.11)
$$

Proof. The proof follows from similar lines as the proof of Theorem 3.1, except we will make use of Lemma 4.1 instead of Lemma 3.2, and taking expectation with respect to $\{\xi_t\}$ in the end. □

Given (4.11), Theorem 4.1 states that the last-iterate of SRPMD converges linearly up to a noise-level of $\mathcal{O}(Me/(1-\gamma)^2)$, where e characterizes the quality of the estimation of the robust state-action value function.

Similarly to Theorem 3.2 for the deterministic RPMD method, we can also show that with a less aggressive choice of stepsize schedule, SRPMD attains sublinear convergence.

Theorem 4.2. *Suppose the stepsizes $\{\eta_k\}$ satisfy $\eta_k \geq \eta_{k-1}M'$ for all $k \geq 1$, where M' is defined as in Theorem 3.1. Then for every ρ with $\mathrm{supp}(\rho) = \mathcal{S}$, at any iteration $k \geq 1$, RPMD produces policy π_R satisfying*

$$\mathbb{E}\left[f_\rho(\pi_R) - f(\pi^*)\right] \leq \tfrac{M}{(1-\gamma)k}\left(f_\rho(\pi_0) - f_\rho(\pi^*) + \tfrac{D_w}{M\eta_0} + \tfrac{4}{1-\gamma}\sum_{t=0}^{k-1} e_t\right),$$

where M is defined as in Lemma 3.2, and R is a random integer uniformly sampled from $\{1\dots k\}$. In particular, if we have $\mathbb{E}_{\xi_k}\left\|Q_r^{\pi_k,\xi_k} - Q^{\pi_k}\right\|_\infty \leq e$ for all $k \geq 0$, then

$$\mathbb{E}\left[f_\rho(\pi_R) - f(\pi^*)\right] \leq \tfrac{M}{(1-\gamma)k}\left(f_\rho(\pi_0) - f_\rho(\pi^*) + \tfrac{D_w}{M\eta_0}\right) + \tfrac{4Me}{(1-\gamma)^2}. \tag{4.12}$$

Proof. By summing up inequality (4.3) from $t = 0$ to $k-1$, and taking expectation with respect to $\{\xi_t\}$, we obtain

$$\begin{aligned}\mathbb{E}\left[f_\rho(\pi_k) - f_\rho(\pi^*)\right] + \tfrac{1-\gamma}{M}\sum_{t=1}^{k-1}\mathbb{E}\left[f_\rho(\pi_t) - f_\rho(\pi^*)\right] &\leq f_\rho(\pi_0) - f_\rho(\pi^*) + \tfrac{1}{M\eta_0}\mathbb{E}_{s\sim d_\rho^{\pi^*,u_0}} D_{\pi_0}^{\pi^*}(s) \\ &+ \underbrace{\sum_{t=1}^{k-1}\mathbb{E}\left(\tfrac{1}{M\eta_t}\mathbb{E}_{s\sim d_\rho^{\pi^*,u_t}} D_{\pi_t}^{\pi^*}(s) - \tfrac{1}{M\eta_{t-1}}\mathbb{E}_{s\sim d_\rho^{\pi^*,u_{t-1}}} D_{\pi_t}^{\pi^*}(s)\right)}_{(A)} + \tfrac{4}{1-\gamma}\sum_{t=0}^{k-1} e_t.\end{aligned}$$

Now suppose stepsizes $\{\eta_k\}$ satisfy $\eta_k \geq \eta_{k-1}M'$, we then obtain term $(A) \leq 0$, and hence

$$\mathbb{E}\left[f_\rho(\pi_k) - f_\rho(\pi^*)\right] + \tfrac{1-\gamma}{M}\sum_{t=1}^{k-1}\mathbb{E}\left[f_\rho(\pi_t) - f_\rho(\pi^*)\right] \leq f_\rho(\pi_0) - f_\rho(\pi^*) + \tfrac{1}{M\eta_0}\mathbb{E}_{s\sim d_\rho^{\pi^*,u_0}} D_{\pi_0}^{\pi^*}(s) + \tfrac{4}{1-\gamma}\sum_{t=0}^{k-1} e_t.$$

Combining the above relation with $R \sim \mathrm{Unif}(\{1\dots k\}$, we further obtain

$$\tfrac{(1-\gamma)k}{M}\mathbb{E}\left[f_\rho(\pi_R) - f_\rho(\pi^*)\right] \leq f_\rho(\pi_0) - f_\rho(\pi^*) + \tfrac{1}{M\eta_0}\mathbb{E}_{s\sim d_\rho^{\pi^*,u_0}} D_{\pi_0}^{\pi^*}(s) + \tfrac{4}{1-\gamma}\sum_{t=0}^{k-1} e_t,$$

the claim then follows immediately. □

Given (4.12), Theorem 4.2 then states that SRPMD converges at the rate of $\mathcal{O}(1/k)$, up to a noise level of $\mathcal{O}(Me/(1-\gamma)^2)$. In addition, compared to Theorem 4.1, Theorem 4.2 requires randomly sampling a historical policy π_R as an output, and converges slower up to a similar noise level. This comparison thus suggests using the more aggressive stepsize scheme among the two increasing-stepsize schemes in most practical scenarios.

We then proceed to establish the convergence of SRPMD with a constant stepsize, by focusing on the euclidean divergence considered in Section 3.2.1. Similar to Lemma 3.3, we first make the following simple observations regarding the policies generated by SRPMD.

Lemma 4.2. For any $k \geq 1$, the iterates in SRPMD with constant stepsizes $\eta_k = \eta > 0$ satisfy

$$\tfrac{1}{\eta}\sum_{t=0}^{k-1}\left(D_{\pi_{t+1}}^{\pi_t}(s) + D_{\pi_t}^{\pi_{t+1}}(s)\right) \leq V_r^{\pi_0}(s) - V_r^{\pi^*}(s) + \tfrac{2}{1-\gamma}\sum_{t=0}^{k-1}\|\delta_k\|_\infty . \tag{4.13}$$

Proof. Given (4.5) and inequality (c) in (4.7), we obtain that

$$\tfrac{1}{\eta}\left(D_{\pi_{k+1}}^{\pi_k}(s) + D_{\pi_k}^{\pi_{k+1}}(s)\right) \leq V_r^{\pi_k}(s) - V_r^{\pi_{k+1}}(s) + \tfrac{2}{1-\gamma}\|\delta_k\|_\infty , \ \forall s \in \mathcal{S}.$$

Summing up the prior relation from $t = 0$ to $k-1$, we obtain the desired result. □

Combining Lemma 4.2 and Lemma 4.1, we are able to establish the following convergence characterization for SRPMD with euclidean Bregman divergence, when adopting any constant-stepsize scheme.

Theorem 4.3. *Let $w(\cdot) = \|\cdot\|_2^2$ be the distance-generating function, and $\eta_k = \eta$ for some fixed $\eta > 0$ and $k \geq 0$, then at any iteration $k \geq 1$, RPMD produces policy π_R satisfying*

$$\mathbb{E}\left[f_\rho(\pi_R) - f_\rho(\pi^*)\right] \leq \tfrac{M}{(1-\gamma)k}\left(f_\rho(\pi_0) - f_\rho(\pi^*)\right) + \tfrac{4M}{(1-\gamma)^2 k}\sum_{t=0}^{k-1} e_t + \sqrt{\tfrac{18|\mathcal{S}|^2}{k\eta(1-\gamma)^3}}\sqrt{1 + 2\sum_{t=0}^{k-1} e_t}, \tag{4.14}$$

where M is defined as in Lemma 3.2, and and R is a random integer uniformly sampled from $\{1 \dots k\}$. In particular, if we have $\mathbb{E}_{\xi_k}\left\|Q_r^{\pi_k,\xi_k} - Q^{\pi_k}\right\|_\infty \leq e$ for all $k \geq 0$, then

$$\mathbb{E}\left[f_\rho(\pi_R) - f_\rho(\pi^*)\right] \leq \tfrac{M}{(1-\gamma)k}\left(f_\rho(\pi_0) - f_\rho(\pi^*)\right) + \tfrac{4Me}{(1-\gamma)^2} + \sqrt{\tfrac{18|\mathcal{S}|^2}{\eta k(1-\gamma)^3}} + \sqrt{\tfrac{36|\mathcal{S}|^2 e}{\eta(1-\gamma)^3}}. \tag{4.15}$$

Proof. By summing up inequality (4.3) from $t = 0$ to $k-1$, we obtain

$$\begin{aligned}
&f_\rho(\pi_k) - f_\rho(\pi^*) + \tfrac{1-\gamma}{M}\sum_{t=1}^{k-1}\left(f_\rho(\pi_t) - f_\rho(\pi^*)\right) \\
\leq &f_\rho(\pi_0) - f_\rho(\pi^*) + \sum_{t=0}^{k-1}\left(\tfrac{1}{M\eta}\mathbb{E}_{s\sim d_\rho^{\pi^*,u_t}} D_{\pi_t}^{\pi^*}(s) - \tfrac{1}{M\eta}\mathbb{E}_{s\sim d_\rho^{\pi^*,u_t}} D_{\pi_{t+1}}^{\pi^*}(s)\right) + \tfrac{4}{1-\gamma}\sum_{t=0}^{k-1}\|\delta_t\|_\infty .
\end{aligned}$$

Following the same lines as in the proof of Theorem 3.3, we can obtain from the above relation that

$$\begin{aligned}
f_\rho(\pi_k) - f_\rho(\pi^*) + \tfrac{1-\gamma}{M}\sum_{t=1}^{k-1}\left(f_\rho(\pi_t) - f_\rho(\pi^*)\right) \leq f_\rho(\pi_0) - f_\rho(\pi^*) + \tfrac{4}{1-\gamma}\sum_{t=0}^{k-1}\|\delta_t\|_\infty \\
+ \tfrac{\sqrt{18}}{M\eta}\sum_{t=0}^{k-1}\mathbb{E}_{s\sim d_\rho^{\pi^*,u_t}}\sqrt{D_{\pi_{t+1}}^{\pi_t}(s)}.
\end{aligned} \tag{4.16}$$

We then make use of Lemma 4.2, which gives

$$\begin{aligned}
\left(\sum_{t=0}^{k-1}\sqrt{D_{\pi_{t+1}}^{\pi_t}(s)}\right)^2 \leq k\sum_{t=0}^{k-1} D_{\pi_{t+1}}^{\pi_t}(s) &\leq k\eta\left(V_r^{\pi_0}(s) - V_r^{\pi^*}(s) + \tfrac{2}{1-\gamma}\sum_{t=0}^{k-1}\|\delta_t\|_\infty\right) \\
&\leq \tfrac{k\eta}{1-\gamma}\left(1 + 2\sum_{t=0}^{k-1}\|\delta_t\|_\infty\right), \quad \forall s \in \mathcal{S},
\end{aligned}$$

which combined with (4.16), gives

$$\begin{aligned}
&f_\rho(\pi_k) - f_\rho(\pi^*) + \tfrac{1-\gamma}{M}\sum_{t=1}^{k-1}\left(f_\rho(\pi_t) - f_\rho(\pi^*)\right) \\
\leq & f_\rho(\pi_0) - f_\rho(\pi^*) + \tfrac{4}{1-\gamma}\sum_{t=0}^{k-1}\|\delta_t\|_\infty + \tfrac{|\mathcal{S}|}{M}\sqrt{\tfrac{18k}{\eta(1-\gamma)}}\sqrt{1+2\sum_{t=0}^{k-1}\|\delta_t\|_\infty}.
\end{aligned}$$

Finally, given the definition of $R \sim \text{Unif}\{1\dots k\}$, we take expectation with respect to $\{\xi_t\}$ and R, and conclude that

$$\tfrac{(1-\gamma)k}{M}\mathbb{E}\left[f_\rho(\pi_R) - f_\rho(\pi^*)\right] \leq f_\rho(\pi_0) - f_\rho(\pi^*) + \tfrac{4}{1-\gamma}\sum_{t=0}^{k-1} e_t + \tfrac{|\mathcal{S}|}{M}\sqrt{\tfrac{18k}{\eta(1-\gamma)}}\sqrt{1+2\sum_{t=0}^{k-1} e_t},$$

where in the last inequality we also uses that fact that $\sqrt{1+x}$ is concave and hence $\mathbb{E}\sqrt{1+2\sum_{t=0}^{k-1}\|\delta_t\|_\infty} \leq \sqrt{1+2\sum_{t=0}^{k-1} e_t}$. Hence the desired claim (4.14) follows immediately after simple rearrangement. In addition, (4.15) follows from the fact that $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$. □

Given (4.15), Theorem 4.3 states that for large enough constant stepsize, SRPMD with euclidean divergence converges at the rate of $\mathcal{O}(1/k)$, until a noise level of $\mathcal{O}(Me/(1-\gamma)^2)$ is reached, where e quantifies the noise in the estimation of the robust state-action value function.

5 Sample Complexity of Stochastic Robust Policy Mirror Descent

In this section, we discuss an online method of estimating the robust state-action value function Q_r^π for a given policy π, by using samples ξ collected during the interaction with the nominal environment $\mathcal{M}_{\text{N}}$. By incorporating this online estimation method into the previously discussed SRPMD methods, we are able to learn a robust policy without the need of training policy within its worst-case environment. Consequently, we will also establish the sample complexity of the SRPMD methods with different stepsize schemes discussed in Section 4.

To facilitate our presentation, let us define operator $F : \mathbb{R}^{|\mathcal{S}||\mathcal{A}|} \to \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ by

$$F(x) = \text{diag}(\nu^\pi)\left(\mathcal{T}^\pi(x) - x\right) + x, \tag{5.1}$$

where ν^π denotes the stationary state-action pair distribution induced by policy π, and operator $\mathcal{T}^\pi : \mathbb{R}^{|\mathcal{S}||\mathcal{A}|} \to \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ is defined by

$$\begin{aligned}
\left[\mathcal{T}^\pi(x)\right](s,a) &= c(s,a) + \gamma \max_{u\in\mathcal{U}} \sum_{s'\in\mathcal{S}}\sum_{a'\in\mathcal{A}} \mathbb{P}_u(s'|s,a)\pi(a'|s')x(s',a') \\
&= c(s,a) + \gamma \max_{u\in\mathcal{U}} \sum_{s',a'} \mathbb{P}_u^\pi(s',a'|s,a)x(s',a'),
\end{aligned}$$

where in the second equality we denote $\mathbb{P}_u^\pi(s',a'|s,a) = \mathbb{P}_u(s'|s,a)\pi(a'|s')$. We will also write the previous definition in matrix form as

$$\mathcal{T}^\pi(x) = c + \gamma \max_{u\in\mathcal{U}} \mathbb{P}_u^\pi x.$$

Clearly, given the rectangularity of uncertainty set in (1.2), we have the following equivalent definition of $\mathcal{T}^\pi$,

$$\left[\mathcal{T}^\pi(x)\right](s,a) = c(s,a) + \gamma \sum_{s'\in\mathcal{S}}\sum_{a'\in\mathcal{A}} \mathbb{P}_{\text{N}}(s'|s,a)\pi(a'|s')x(s',a') + \gamma \max_{u\in\mathcal{U}} \sum_{s'\in\mathcal{S}} u(s'|s,a) \sum_{a'\in\mathcal{A}} \pi(a'|s')x(s',a')$$

Algorithm 3 The robust temporal difference learning method

Input: Policy π to be evaluated. Initial iterate $\theta_0 \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ and stepsizes $\{\alpha_t\}_{t\geq 0}$, initial state $s_0 \in \mathcal{S}$, initial action $a_0 \sim \pi(\cdot|s_0)$.

for $t = 0, 1, \ldots$ **do**

Collect $s_{t+1} \sim \mathbb{P}_{\mathrm{N}}(\cdot|s_t, a_t)$, and make action $a_{t+1} \sim \pi(\cdot|s_{t+1})$.

Update the iterate:

$$\theta_{t+1} = \theta_t + \alpha_t \left(c(s_t, a_t) + \gamma\theta_t(s_{t+1}, a_{t+1}) + \sigma_{\mathcal{U}_{s_t,a_t}}(M(\pi, \theta_t)) - \theta_t(s_t, a_t)\right) e(s_t, a_t).$$

end for

$$= c(s, a) + \gamma \sum_{s'\in\mathcal{S}} \sum_{a'\in\mathcal{A}} \mathbb{P}_{\mathrm{N}}(s'|s, a)\pi(a'|s')x(s', a') + \gamma\sigma_{\mathcal{U}_{s,a}}(M(\pi, x)), \tag{5.2}$$

where $M(\pi, x) \in \mathbb{R}^{|\mathcal{S}|}$ is defined as $[M(\pi, x)](s) = \sum_{a\in\mathcal{A}} \pi(a|s)x(s, a)$, and σ_X denotes the support function of set X.

Given (2.4) in Proposition 2.2, it should be clear that the robust state-action value function Q_r^π is a fixed point of operator F. On the other hand, since

$$\|\mathcal{T}^\pi(x) - \mathcal{T}^\pi(y)\|_\infty = \gamma \left\| \max_{u\in\mathcal{U}} \mathbb{P}_u^\pi x - \max_{u\in\mathcal{U}} \mathbb{P}_u^\pi y \right\|_\infty \leq \gamma \max_{u\in\mathcal{U}} \|\mathbb{P}_u^\pi x - \mathbb{P}_u^\pi y\|_\infty \leq \gamma \|x - y\|_\infty, \tag{5.3}$$

$\mathcal{T}^\pi$ is a γ-contraction in $\|\cdot\|_\infty$-norm. Thus whenever $\min(\nu^\pi) > 0$, Q_r^π is the unique fixed-point of operator F.

Based on the prior observation, we propose the robust temporal difference (RTD) method (Algorithm 3) and establish its sample complexity for finding a stochastic estimate of Q_r^π. The RTD method assumes the access to an stochastic operator $f(x; \zeta) : \mathbb{R}^{|\mathcal{S}||\mathcal{A}|} \to \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, where $\zeta = (s, a, s', a')$ denotes a random quadruple sampled from a Markov chain defined over $\mathcal{Z} = (\mathcal{S} \times \mathcal{A})^2$. Specifically, the operator takes the form of

$$f(x; \zeta) = \left(c(s, a) + \gamma x(s', a') + \sigma_{\mathcal{U}_{s,a}}(M(\pi, x)) - x(s, a)\right) e(s, a) + x. \tag{5.4}$$

For a given policy, an initial state $s_0 \in \mathcal{S}$, and initial action $a_0 \sim \pi(\cdot|s_0)$, the RTD method, at any iteration $t \geq 0$, (1) Given (s_t, a_t), collects $s_{t+1} \sim \mathbb{P}_{\mathrm{N}}(\cdot|s_t, a_t)$, and make actions $a_{t+1} \sim \pi(\cdot|s_{t+1})$; (2) Forms $\zeta_t = (s_t, a_t, s_{t+1}, a_{t+1})$, and performs the following update

$$\theta_{t+1} = \theta_t + \alpha_t \left(f(\theta_t; \zeta_t) - \theta_t\right).$$

It should be clear that by construction, $\{\zeta_t\}$ indeed forms a Markov chain over $\mathcal{Z}$. Furthermore, given (5.1) and (5.2), by letting $\overline{\nu}^\pi$ denotes the stationary distribution of $\{\zeta_t\}$, we have $\mathbb{E}_{\zeta\sim\overline{\nu}^\pi} f(x; \zeta) = F(x)$.

Through out the rest of our discussions, we make the following assumption on the to-be-evaluated policy and the nominal environment $\mathcal{M}_{\mathrm{N}}$, which is commonly assumed in the literature of reinforcement learning.

Assumption 1. *The policy π satisfies $\min_{s\in\mathcal{S},a\in\mathcal{A}} \pi(a|s) > 0$, and the Markov chain $\{s_t\}$ induced by π within the nominal MDP $\mathcal{M}_{\mathrm{N}}$ is aperiodic and irreducible.*

Combining Assumption 1 and the finiteness of the state space $\mathcal{S}$, the Markov chain $\{s_t\}$ satisfies geometric-mixing property [21]. In addition, the stationary distribution of $\{s_t\}$, denoted by μ^π, satisfies $\mu^\pi(s) > 0$ for all $s \in \mathcal{S}$. Consequently, we also have $\nu_{\min} = \min_{s\in\mathcal{S},a\in\mathcal{A}} \nu^\pi(s, a) = \min_{s\in\mathcal{S},a\in\mathcal{A}} \mu^\pi(s)\pi(a|s) > 0$ for all $(s, a) \in \mathcal{S} \times \mathcal{A}$.

The following lemma characterizes the sample complexity of the RTD method for obtaining a stochastic estimation of Q_r^π, which utilizes the machinery of stochastic approximation applied to contraction operators developed in [4].

Lemma 5.1. Under Assumption 1, let $\alpha_t = \alpha$, with α satisfying $\alpha T_\alpha(\{\zeta_t\}) \le C \min(\nu^\pi)^2 / \log(|\mathcal{S}||\mathcal{A}|)$[1] for all $t \ge 0$, Then for any $\epsilon > 0$, the RTD method needs at most

$$T = \widetilde{\mathcal{O}}\left(\frac{\log^2(1/\epsilon)}{(1-\gamma)^5 \nu_{\min}^3 \epsilon^2}\right) \tag{5.5}$$

iterations to find an estimate θ_T satisfying $\mathbb{E}_\xi \|\theta_T - Q^\pi\|_\infty \le \epsilon$, where $\xi = \{\zeta_t\}_{t=0}^T$ denotes the trajectory collected by the RTD method, and $\widetilde{\mathcal{O}}(\cdot)$ ignores all polylogarithmic terms.

Proof. We begin by establishing several properties of the operator F defined in (5.1), the stochastic operator f defined in (5.4), and the Markov chain $\{\zeta_t\}$. For operator F, note that

$$\begin{aligned}\|F(x) - F(y)\|_\infty &= \|\mathrm{diag}(\nu^\pi)\left(\mathcal{T}^\pi(x) - \mathcal{T}^\pi(y)\right) + \left(I - \mathrm{diag}(\nu^\pi)\right)(x-y)\|_\infty \\ &\le \left(1 - \min(\nu^\pi)(1-\gamma)\right)\|x-y\|_\infty,\end{aligned} \tag{5.6}$$

where the inequality uses (5.3). Hence F is a $(1-\min(\nu^\pi))$-contraction in $\|\cdot\|_\infty$ norm. Consequently, Q_r^π is the unique fixed point of F.

For operator f, we have for any ζ,

$$\begin{aligned}&\|f(x,\zeta) - f(y,\zeta)\|_\infty \\ =&\left\|\left[\gamma\left(x(s',a') - y(s',a')\right) - \left(x(s,a) - y(s,a)\right) + \sigma_{\mathcal{U}_{s,a}}(M(\pi,x)) - \sigma_{\mathcal{U}_{s,a}}(M(\pi,y))\right] e(s,a) + x - y\right\|_\infty \\ \le& 3\|x-y\|_\infty + |\sigma_{\mathcal{U}_{s,a}}(M(\pi,x)) - \sigma_{\mathcal{U}_{s,a}}(M(\pi,y))|.\end{aligned}$$

Now by defining $z_x = M(\pi,x)$ for any x, we know $|z_x(s) - z_y(s)| = |\sum_{a\in\mathcal{A}} \pi(a|s)\left(x(s,a) - y(s,a)\right)| \le \|x-y\|_\infty$ holds for any $s \in \mathcal{S}$. Hence we have

$$\begin{aligned}|\sigma_{\mathcal{U}_{s,a}}(M(\pi,x)) - \sigma_{\mathcal{U}_{s,a}}(M(\pi,y))| &= |\sigma_{\mathcal{U}_{s,a}}(z_x) - \sigma_{\mathcal{U}_{s,a}}(z_y)| \\ &\le \max_{u(\cdot|s,a)\in\mathcal{U}_{s,a}} |\langle u(\cdot|s,a), z_x - z_y\rangle| \\ &\le \|u(\cdot|s,a)\|_1 \|z_x - z_y\|_\infty \\ &\overset{(a)}{\le} 2\|z_x - z_y\|_\infty \le 2\|x-y\|_\infty,\end{aligned}$$

where inequality (a) uses the fact that $\|u(\cdot|s,a)\|_1 = \|\mathbb{P}_u(\cdot|s,a) - \mathbb{P}_{\mathrm{N}}(\cdot|s,a)\|_1 \le 2$. Thus we obtain

$$\|f(x,\zeta) - f(y,\zeta)\|_\infty \le 5\|x-y\|_\infty. \tag{5.7}$$

Additionally, one can readily verify that

$$\|f(\mathbf{0},\zeta)\|_\infty \le 1, \quad \forall \zeta \in \mathcal{Z}. \tag{5.8}$$

Lastly, for the Markov chain $\{\zeta_t\}$, we proceed to establish its fast-mixing property under Assumption 1. Note that the stationary distribution of $\{\zeta_t\}$, denoted by $\overline{\nu}^\pi$, is given by $\overline{\nu}^\pi(s,a,s',a') = \mu^\pi(s)\pi(a|s)\mathbb{P}_{\mathrm{N}}(s'|s,a)\pi(a'|s')$. Let us denote the transition kernel of $\{\zeta_t\}$ by $\mathbb{P}_\zeta$, and accordingly denote the transition kernel of $\{s_t\}$ by $\mathbb{P}_{\mathrm{S}}$. Then for any $\zeta \in \mathcal{Z}$,

$$\begin{aligned}\left\|\mathbb{P}_\zeta^{k+1}(\zeta,\cdot) - \overline{\nu}^\pi(\cdot)\right\|_{\mathrm{TV}} &= \tfrac{1}{2}\sum_{\widetilde{\zeta}} |\mathbb{P}^{k+1}(\zeta,\widetilde{\zeta}) - \overline{\nu}^\pi(\widetilde{\zeta})| \\ &\overset{(a)}{=} \tfrac{1}{2}\sum_{\widetilde{\zeta}} |\mathbb{P}_{\mathrm{S}}^k(s',\widetilde{s})\pi(\widetilde{a}|\widetilde{s})\mathbb{P}_{\mathrm{N}}(\widetilde{s}'|\widetilde{s},\widetilde{a})\pi(\widetilde{a}'|\widetilde{s}') - \mu^\pi(\widetilde{s})\pi(\widetilde{a}|\widetilde{s})\mathbb{P}_{\mathrm{N}}(\widetilde{s}'|\widetilde{s},\widetilde{a})\pi(\widetilde{a}'|\widetilde{s}')|\end{aligned}$$

[1]Here $T_\alpha(\{\zeta_t\}) \coloneqq \min\left\{k : k \ge 0, \max_{\zeta\in\mathcal{Z}} \left\|\mathbb{P}_\zeta^k(\zeta,\cdot) - \overline{\nu}^\pi(\cdot)\right\|_{\mathrm{TV}} \le \alpha\right\}$, where $\mathbb{P}_\zeta$ denotes the transitional kernel of Markov chain $\{\zeta_t\}$. Note $T_\alpha(\{\zeta_t\})$ is well defined, see (5.9) in the proof of Lemma 5.1.

$$\leq \tfrac{1}{2}\sum_{\widetilde{s}} |\mathbb{P}_{\mathrm{S}}^{k}(s',\widetilde{s}) - \mu^{\pi}(\widetilde{s})| \leq C\alpha^{k}$$

for some $\alpha \in (0,1)$ and $C > 0$, where the last inequality follows from the geometric-mixing property of $\{s_t\}$ given Assumption 1, and equality (a) follows from the Markov property. Thus we obtain that

$$\max_{\zeta} \left\| \mathbb{P}_{\zeta}^{k+1}(\zeta,\cdot) - \overline{\nu}^{\pi}(\cdot) \right\|_{\mathrm{TV}} \leq C\alpha^{k}. \tag{5.9}$$

From (5.6), (5.7), (5.8), and (5.9), it should be clear that the operators F, f and the stochastic process $\{\zeta_t\}$ satisfy Assumption 2.1 - 2.3 in [4], and the rest of the proof follows the same lines as in Corollary 3.1.1 therein. □

Given Lemma 5.1, we proceed to establish the sample complexity of the SRPMD method with different stepsize schemes discussed in Section 4. We begin with the stepsize scheme (4.10) considered in Theorem 4.1, which demonstrates linear convergence up to policy evaluation error.

Proposition 5.1. Suppose the stepsizes $\{\eta_k\}$ in SRPMD satisfy $\eta_k \geq \eta_{k-1}\left(1 - \frac{1-\gamma}{M}\right)^{-1} M'$ for all $k \geq 1$, M' is defined as in Theorem 3.1. Furthermore, for any $\epsilon > 0$, suppose $\mathbb{E}_{\xi_k}\left\| Q_r^{\pi_k,\xi_k} - Q^{\pi_k} \right\|_{\infty} \leq e$ with $4Me/(1-\gamma)^2 \leq \epsilon/2$. Then SRPMD outputs a policy π_k with $\mathbb{E}\left[f_\rho(\pi_k) - f_\rho(\pi^*)\right] \leq \epsilon$ in

$$k = \mathcal{O}\left(\tfrac{M}{1-\gamma}\left[\log\left(\tfrac{\Delta_0}{\epsilon}\right) + \log\left(\tfrac{D_w}{M\eta_0\epsilon}\right)\right]\right)$$

iterations, where M is defined as in Lemma 3.2, and $\Delta_0 = f_\rho(\pi_0) - f_\rho(\pi^*)$. In addition, the total number of samples required by SRPMD can be bounded by

$$\widetilde{\mathcal{O}}\left(\tfrac{M^3\log^2\left(4M/(\epsilon(1-\gamma)^2)\right)}{(1-\gamma)^{10}\nu_{\min}^3\epsilon^2}\left[\log\left(\tfrac{\Delta_0}{\epsilon}\right) + \log\left(\tfrac{D_w}{M\eta_0\epsilon}\right)\right]\right).$$

Proof. The bound on the total number of iterations k can be readily obtained from (4.11) in Theorem 4.1, if $4Me/(1-\gamma)^2 \leq \epsilon/2$. To satisfy this condition, suppose one needs to run RTD for T iterations when evaluating for each $Q_r^{\pi_k}$, then from (5.5) in Lemma 5.1, one can bound T by

$$T = \widetilde{\mathcal{O}}\left(\tfrac{M^2\log^2(4M/(\epsilon(1-\gamma)^2))}{(1-\gamma)^9\nu_{\min}^3\epsilon^2}\right).$$

The bound on the total number of samples follows immediately by combining the previous two observations. □

Given Proposition 5.1, we remark that the sample complexity of applying SRPMD to solving the robust MDP with $(\mathbf{s},\mathbf{a})$-rectangular uncertainty sets is comparable to that of solving standard MDPs with linearly converging policy mirror descent methods in terms of its dependence on the optimality gap [20], and is slightly worse in terms of its dependence on the effective horizon $(1-\gamma)^{-1}$. To the best of our knowledge, this is the first sample complexity result for first-order policy-based method that is optimal in terms of the dependence on the optimality gap. A closer look to the analysis shows that this worse dependence on the effective horizon comes from the current convergence characterization of the RTD method, which exhibits a worse dependence on the effective horizon compared to the CTD method considered in [20] for evaluating policy in standard MDPs. See Section 6 for more detailed discussions.

We then proceed to establish the sample complexity for SRPMD with another increasing-stepsize scheme specified in Theorem 4.2, which demonstrates sublinear convergence up to the policy evaluation error.

Proposition 5.2. Suppose the stepsizes $\{\eta_k\}$ in SRPMD satisfy $\eta_k \geq \eta_{k-1} M'$ for all $k \geq 1$, M' is defined as in Theorem 3.1. Furthermore, for any $\epsilon > 0$, suppose $\mathbb{E}_{\xi_k} \left\| Q_r^{\pi_k, \xi_k} - Q^{\pi_k} \right\|_\infty \leq e$ with $4Me/(1-\gamma)^2 \leq \epsilon/2$. Then SRPMD outputs a policy π_R with $\mathbb{E}\left[f_\rho(\pi_R) - f_\rho(\pi^*)\right] \leq \epsilon$, where $R \sim \text{Unif}(\{1 \dots k\})$, in

$$k = \mathcal{O}\left(\tfrac{M\Delta_0}{(1-\gamma)\epsilon} + \tfrac{D_w}{(1-\gamma)\eta_0 \epsilon}\right)$$

iterations, where M is defined as in Lemma 3.2, and $\Delta_0 = f_\rho(\pi_0) - f_\rho(\pi^*)$. In addition, the total number of samples required by SRPMD can be bounded by

$$\widetilde{\mathcal{O}}\left(\tfrac{M^3 \log^2(4M/(\epsilon(1-\gamma)))}{(1-\gamma)^{10}\nu_{\min}^3 \epsilon^3}\left(\Delta_0 + \tfrac{D_w}{\eta_0 M}\right)\right).$$

Proof. The bound on the total number of iterations k can be readily obtained from (4.12) in Theorem 4.2, if $4Me/(1-\gamma)^2 \leq \epsilon/2$. To satisfy this condition, suppose one needs to run RTD for T iterations when evaluating for each $Q_r^{\pi_k}$, then from (5.5) in Lemma 5.1, one can bound T by

$$T = \widetilde{\mathcal{O}}\left(\tfrac{M^2 \log^2(4M/(\epsilon(1-\gamma)))}{(1-\gamma)^9 \nu_{\min}^3 \epsilon^3}\right).$$

The bound on the total number of samples follows immediately by combining the previous two observations. □

Finally, we establish the sample complexity of SRPMD when using $w(\cdot) = \|\cdot\|_2^2$ as the distance-generating function, which allows a constant-stepsize scheme and attains sublinear convergence.

Proposition 5.3. Let $w(\cdot) = \|\cdot\|_2^2$ be the distance-generating function, and $\eta_k = \eta$ for all $k \geq 0$. Furthermore, for any $\epsilon > 0$, suppose $\mathbb{E}_{\xi_k} \left\| Q_r^{\pi_k, \xi_k} - Q^{\pi_k} \right\|_\infty \leq e$ with $4Me/(1-\gamma)^2 \leq \epsilon/2$. Then by taking $\eta = 72|\mathcal{S}|^2/\left((1-\gamma)^2 M\epsilon\right)$, SRPMD outputs a policy π_R with $\mathbb{E}\left[f_\rho(\pi_R) - f_\rho(\pi^*)\right] \leq \epsilon$, where $R \sim \text{Unif}(\{1 \dots k\})$, in

$$k = \mathcal{O}\left(\tfrac{M\Delta_0}{(1-\gamma)\epsilon}\right)$$

iterations, where M is defined as in Lemma 3.2, and $\Delta_0 = f_\rho(\pi_0) - f_\rho(\pi^*)$. In addition, the total number of samples required by SRPMD can be bounded by

$$\widetilde{\mathcal{O}}\left(\tfrac{M^3 \log^2\left(4M/(\epsilon(1-\gamma)^2\right)}{(1-\gamma)^{10}\nu_{\min}^3 \epsilon^3}\right).$$

Proof. The bound on the total number of iterations k can be readily obtained from (4.15) in Theorem 4.3, provided

$$\tfrac{4Me}{(1-\gamma)^2} \leq \tfrac{\epsilon}{4}, \quad \sqrt{\tfrac{18|\mathcal{S}|^2}{\eta k (1-\gamma)^3}} \leq \tfrac{\epsilon}{4}, \quad \sqrt{\tfrac{36|\mathcal{S}|^2 e}{\eta(1-\gamma)^3}} \leq \tfrac{\epsilon}{4}.$$

To satisfy the first condition above, suppose one needs to run RTD for T iterations when evaluating for each $Q_r^{\pi_k}$, then from (5.5) in Lemma 5.1, one can bound T by

$$T = \widetilde{\mathcal{O}}\left(\tfrac{M^2 \log^2(4M/(\epsilon(1-\gamma)^2))}{(1-\gamma)^9 \nu_{\min}^3 \epsilon^3}\right).$$

To satisfies the second and third conditions, it suffices to have $\eta \geq \max\left\{\tfrac{72|\mathcal{S}|^2}{k(1-\gamma)^3\epsilon^2}, \tfrac{36|\mathcal{S}|^2 e}{(1-\gamma)^3\epsilon^2}\right\}$, which can be readily satisfied by $\eta = 72|\mathcal{S}|^2/\left((1-\gamma)^2 M\epsilon\right)$ given the bound on k and e. The bound on the total number of samples Tk then follows immediately by combining the previous observations. □

To the best of our knowledge, all the obtained sample complexities in this section appears to be new in the literature of first-order methods applied to the robust MDP problem. The best sample complexity for PGM applied to this problem in the existing literature is at the order of $\mathcal{O}(1/\epsilon^7)$ [42], which focuses on the euclidean Bregman divergence and a special subclass of polyhedral uncertainty sets. In comparison, as shown in Proposition 5.3, SRPMD with the same divergence improves this sample complexity by orders of magnitude, and applies to a much more general class of uncertainty sets.

6 Concluding Remarks

In this manuscript, we develop the robust policy mirror descent method and its stochastic variants for controlling Markov decision process with uncertain transition kernels. Our established iteration and sample complexity seems to be new in the literature of policy-space first-order methods applied to this problem class. We highlight a few future directions worthy of continuing explorations from our perspective.

First, the analysis of constant stepsize RPMD yields an additional dependence on the size of the state space. Though this dependence can be bypassed with a large stepsize, removing this dependence completely remains not only as a theoretical interest, but can also potentially help improving the sample complexity of the SRPMD methods.

Second, the current analysis of SRPMD uses only a single characterization on the noise of the stochastic estimate (see (4.2)), which contrasts with more delicate approach of separating bias and variance for solving standard MDPs [20]. As a result, it is unclear whether the dependence of obtained sample complexities on the effective horizon is optimal. The reason for our simplified treatment is due to the fact that the robust TD method in Section 5 does not have a separate characterizations for the bias and variance in the obtained stochastic estimate given the nonlinearity of operator F in (5.1). This hinder applying similar treatments of bias and variance in [20] for standard MDPs, where the author heavily exploits the fact that bias converges much faster than the variance. It would be highly interesting to develop a robust TD method that can have separate convergence characterizations for the bias and the variance in the resulting stochastic estimate. Another question related to the robust TD method is to relax Assumption 1, which requires handling the rarely visited state-action pair in evaluating the robust state-action value function.

Lastly, it would also be rewarding to develop RPMD variants for solving robust MDP beyond the $(\mathbf{s}, \mathbf{a})$-rectangular uncertainty sets considered in this manuscript.

References

[1] Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Journal of Machine Learning Research*, 22(98):1–76, 2021.

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

[3] Shicong Cen, Chen Cheng, Yuxin Chen, Yuting Wei, and Yuejie Chi. Fast global convergence of natural policy gradient methods with entropy regularization. *Operations Research*, 2021.

[4] Zaiwei Chen, Siva Theja Maguluri, Sanjay Shakkottai, and Karthikeyan Shanmugam. A lyapunov theory for finite-sample guarantees of asynchronous q-learning and td-learning variants. *arXiv preprint arXiv:2102.01567*, 2021.

[5] John M. Danskin. The theory of max-min and its application to weapons allocation problems. 1967.

[6] Esther Derman, Matthieu Geist, and Shie Mannor. Twice regularized mdps and the equivalence between robustness and regularization. *Advances in Neural Information Processing Systems*, 34, 2021.

[7] Vineet Goyal and Julien Grand-Clement. A first-order approach to accelerated value iteration. *Operations Research*, 2022.

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

[9] Julien Grand-Clément and Christian Kroer. Scalable first-order methods for robust mdps. *arXiv preprint arXiv:2005.05434*, 2020.

[10] Chin Pang Ho, Marek Petrik, and Wolfram Wiesemann. Partial policy iteration for l1-robust markov decision processes. *Journal of Machine Learning Research*, 22(275):1–46, 2021.

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

[12] Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *In Proc. 19th International Conference on Machine Learning*. Citeseer, 2002.

[13] Sham M Kakade. A natural policy gradient. *Advances in neural information processing systems*, 14, 2001.

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

[15] Sajad Khodadadian, Prakirt Raj Jhunjhunwala, Sushil Mahavir Varma, and Siva Theja Maguluri. On the linear convergence of natural policy gradient algorithm. *arXiv preprint arXiv:2105.01424*, 2021.

[16] Umit Kose and Andrzej Ruszczynski. Risk-averse learning by temporal difference methods. *arXiv preprint arXiv:2003.00780*, 2020.

[17] A Ya Kruger. On fréchet subdifferentials. *Journal of Mathematical Sciences*, 116(3):3325–3358, 2003.

[18] Navdeep Kumar, Kfir Levy, Kaixin Wang, and Shie Mannor. Efficient policy iteration for robust markov decision processes via regularization. *arXiv preprint arXiv:2205.14327*, 2022.

[19] Guanghui Lan. *First-order and stochastic optimization methods for machine learning*. Springer, 2020.

[20] Guanghui Lan. Policy mirror descent for reinforcement learning: Linear convergence, new sampling complexity, and generalized problem classes. *arXiv preprint arXiv:2102.00135*, 2021.

[21] David A Levin and Yuval Peres. *Markov chains and mixing times*, volume 107. American Mathematical Soc., 2017.

[22] Yan Li, Ethan X.Fang, Huan Xu, and Tuo Zhao. Implicit bias of gradient descent based adversarial training on separable data. In *International Conference on Learning Representations*, 2020.

[23] Yan Li, Tuo Zhao, and Guanghui Lan. Homotopic policy mirror descent: Policy convergence, implicit regularization, and improved sample complexity. *arXiv preprint arXiv:2201.09457*, 2022.

[24] Ashvin Nair, Bob McGrew, Marcin Andrychowicz, Wojciech Zaremba, and Pieter Abbeel. Overcoming exploration in reinforcement learning with demonstrations. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 6292–6299. IEEE, 2018.

[25] Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2003.

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

[27] 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. PMLR, 2022.

[28] Deepak Pathak, Pulkit Agrawal, Alexei A Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *International conference on machine learning*, pages 2778–2787. PMLR, 2017.

[29] Warren B Powell. *Approximate Dynamic Programming: Solving the curses of dimensionality*, volume 703. John Wiley & Sons, 2007.

[30] Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

[31] R Tyrrell Rockafellar. *Convex analysis*, volume 18. Princeton university press, 1970.

[32] Aurko Roy, Huan Xu, and Sebastian Pokutta. Reinforcement learning under model mismatch. *Advances in neural information processing systems*, 30, 2017.

[33] Andrzej Ruszczyński. Risk-averse dynamic programming for markov decision processes. *Mathematical programming*, 125(2):235–261, 2010.

[34] Lior Shani, Yonathan Efroni, and Shie Mannor. Adaptive trust region policy optimization: Global convergence and faster rates for regularized mdps. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 5668–5675, 2020.

[35] Qianli Shen, Yan Li, Haoming Jiang, Zhaoran Wang, and Tuo Zhao. Deep reinforcement learning with robust and smooth policy. In *International Conference on Machine Learning*, pages 8707–8718. PMLR, 2020.

[36] Leon Simon. *Lectures on Geometric Measure Theory*, volume 3. The Australian National University, Mathematical Sciences Institute, Centre for Mathematics and its Applications, 1 1983.

[37] Leon Simon. Introduction to geometric measure theory. *Tsinghua Lectures*, 2(2):3–1, 2014.

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

[39] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. *arXiv preprint arXiv:1805.12152*, 2018.

[40] Jean-Philippe Vial. Strong convexity of sets and functions. *Journal of Mathematical Economics*, 9(1-2):187–205, 1982.

[41] Yue Wang and Shaofeng Zou. Online robust reinforcement learning with model uncertainty. *Advances in Neural Information Processing Systems*, 34, 2021.

[42] Yue Wang and Shaofeng Zou. Policy gradient method for robust reinforcement learning. *arXiv preprint arXiv:2205.07344*, 2022.

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

[44] Lin Xiao. On the convergence rates of policy gradient methods. *arXiv preprint arXiv:2201.07443*, 2022.

[45] Huan Xu, Constantine Caramanis, and Shie Mannor. Robust regression and lasso. *Advances in neural information processing systems*, 21, 2008.

[46] Huan Xu, Constantine Caramanis, and Shie Mannor. Robustness and regularization of support vector machines. *Journal of machine learning research*, 10(7), 2009.

[47] Tengyu Xu, Zhe Wang, and Yingbin Liang. Improving sample complexity bounds for actor-critic algorithms. *arXiv preprint arXiv:2004.12956*, 2020.

[48] Rui Yang, Chenjia Bai, Xiaoteng Ma, Zhaoran Wang, Chongjie Zhang, and Lei Han. Rorl: Robust offline reinforcement learning via conservative smoothing. *arXiv preprint arXiv:2206.02829*, 2022.

A Supplementary Proofs

Proof of Proposition 2.1. Fix the policy $\pi \in \Pi$, for any $u \in \mathcal{U}$, define operator $\mathcal{T}_u^\pi : \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}^{|\mathcal{S}|} : \mathcal{T}_u^\pi(V) = r^\pi + \gamma \mathbb{P}_u^\pi V$, where $r^\pi(s) = \sum_{a\in\mathcal{A}} \pi(a|s) r(s,a)$, and $\mathbb{P}_u^\pi(s,s') = \sum_{a\in\mathcal{A}} \mathbb{P}_u(s'|s,a)\pi(a|s)$. It is well known that the value V_u^π is the solution of the following linear program [30]:

$$\max_{v\in\mathbb{R}^{|\mathcal{S}|}} \mathbf{1}^\top v \;\; \text{s.t.} \;\; v \le \mathcal{T}_u^\pi v. \tag{A.1}$$

Moreover, we have active constraints at V_u^π : $V_u^\pi = \mathcal{T}_u^\pi V_u^\pi$. It is also useful to make note of the following properties of $\mathcal{T}_u^\pi$: (1) $\mathcal{T}_u^\pi$ is monotone, in the sense that $v \le v' \Rightarrow \mathcal{T}_u^\pi v \le \mathcal{T}_u^\pi v'$; (2) $\mathcal{T}_u^\pi$ is a γ-contraction in $\|\cdot\|_\infty$-norm, with the unique fixed-point being V_u^π. Since both are trivial to verify, we omit their proofs here.

By varying the uncertainty $u \in \mathcal{U}$, the robust value function V_r^π is the solution of the following program

$$\max_{v\in\mathbb{R}^{|\mathcal{S}|}, u\in\mathcal{U}} \mathbf{1}^\top v \text{ s.t. } v \le \mathcal{T}_u^\pi v. \tag{A.2}$$

We proceed to show that formulation (A.2) is equivalent to the following

$$\max_{v\in\mathbb{R}^{|\mathcal{S}|}} \mathbf{1}^\top v \text{ s.t. } v \le \sup_{u\in\mathcal{U}} \{\mathcal{T}_u^\pi v\} = \max_{u\in\mathcal{U}} \{\mathcal{T}_u^\pi v\}, \tag{A.3}$$

where operation $\sup_{u\in\mathcal{U}} \{\mathcal{T}_u^\pi v\}$ is element-wise supremum, which is is well-defined due to the rectangular nature of $\mathcal{U}$. The equality holds since $\mathcal{U}$ is compact and $\mathcal{T}_u^\pi v$ is continuous in u.

To establish equivalence between (A.2) and (A.3). Note that for any feasible solution (v,u) to (A.2), we know that v must also be feasible to (A.3). Hence we obtain Opt(A.3) $\ge$ Opt(A.2). On the other hand, suppose v^* is a solution of (A.3), we know that there exists $u^* \in \mathcal{U}$, such that $v \le \mathcal{T}_{u^*}^\pi v = \max_{u\in\mathcal{U}} \{\mathcal{T}_u^\pi v\}$. Thus (v^*, u^*) is a feasible solution to (A.2), and hence Opt(A.3) $\ge$ Opt(A.2), which further implies Opt(A.3) $=$ Opt(A.2). Moreover, we also know that (v^*, u^*) is an optimal solution of (A.2). The equivalence is then established.

Now define operator $\mathcal{T}^\pi : \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}^{|\mathcal{S}|}$ as $\mathcal{T}^\pi v = \max_{u\in\mathcal{U}} \mathcal{T}_u^\pi v$, then we know that $\mathcal{T}^\pi$ is monotone since $\mathcal{T}_u^\pi$ is monotone for every $u \in \mathcal{U}$. We proceed to show that $\mathcal{T}^\pi$ is also a γ-contraction in $\|\cdot\|_\infty$-norm.

$$\left\|\mathcal{T}^\pi v - \mathcal{T}^\pi v'\right\|_\infty = \left\|\sup_{u\in\mathcal{U}} \mathcal{T}_u^\pi v - \sup_{u\in\mathcal{U}} \mathcal{T}_u^\pi v'\right\|_\infty \le \sup_{u\in\mathcal{U}} \left\|\mathcal{T}_u^\pi v - \mathcal{T}_u^\pi v'\right\|_\infty \le \gamma \left\|v - v'\right\|_\infty,$$

where the first inequality uses the fact that $\|\sup_{u\in\mathcal{U}} f(u,v) - \sup_{u\in\mathcal{U}} f(u,v')\|_\infty \le \sup_{u\in\mathcal{U}} \|f(u,v) - f(u,v')\|_\infty$ for any vector-valued function f, and the second inequality uses the contraction property of operator $\mathcal{T}_u^\pi$ for any $u \in \mathcal{U}$.

Now given a solution v^* to (A.3), we claim that v^* must satisfies

$$v^* = \mathcal{T}^\pi v^* = \sup_{u\in\mathcal{U}} \mathcal{T}_u^\pi v^*,$$

and hence (2.1) is proved. Consequently, any u^* with $v^* = \mathcal{T}^\pi v^* = \mathcal{T}_{u^*}^\pi v^*$ will yield a pair (v^*, u^*) that is feasible to (A.2), and hence an optimal solution to (A.2), with $v^* = \mathcal{T}_{u^*}^\pi v^*$, and hence (2.2) holds. To show the claim, let v' denote the unique fixed point of $\mathcal{T}^\pi$, then clearly v' is a feasible solution to (A.3). Since $\mathcal{T}^\pi$ is monotone, we also have $v^* \le \lim_{t\to\infty} (\mathcal{T}^\pi)^{(t)} v^* = v'$. Thus in order for v^* to be the optimal solution, we must have $v^* = v'$ being the fixed point of $\mathcal{T}^\pi$, hence the proof is completed. □

Proof of Lemma 2.3. It suffices to consider the differentiability of f_ρ inside $\mathrm{ReInt}(\Pi)$, as its relative boundary is a zero-measure set when taking the $(|\mathcal{A}|-1)|\mathcal{S}|$-dimensional Hausdorff measure. We begin by noting that there exists $\{e_i\}_{i=0}^{|\mathcal{A}|-1} \subset \mathbb{R}^{|\mathcal{A}|}$, such that for any $\pi \in \Pi$, we have uniquely defined $\{a_i^\pi(s)\}_{i=1,\ldots,|\mathcal{A}|-1, s\in\mathcal{S}}$ satisfying

$$\pi(\cdot|s) = \sum_{i=1}^{|\mathcal{A}|-1} a_i^\pi(s)\cdot e_i + e_0 \coloneqq E\cdot a^\pi(s) + e_0, \quad \forall s\in\mathcal{S}, \tag{A.4}$$

where we denote $a^\pi(s) = (a_1^\pi(s), \ldots, a_{|\mathcal{A}|-1}^\pi(s))$, and $E = [e_1, \ldots, e_{|\mathcal{A}|-1}]$ has independent columns. We also write the above relation in short as $\pi = \mathcal{M}(a^\pi)$. It is clear that $\mathcal{M}$ is a Lipschitz continuous mapping, and we denote its Lipschitz constant by $L_{\mathcal{M}}$. Alternatively, since E has independent columns, we also have

$$a^\pi(s) = E^\dagger(\pi(\cdot|s) - e_0), \quad \forall s\in\mathcal{S}, \tag{A.5}$$

where $E^\dagger$ denotes the Moore–Penrose inverse of E. We will write the above relation in short as $a^\pi = \mathcal{M}^{-1}(\pi)$. In addition, we can write the objective (1.5) equivalently as

$$f_\rho(\pi) = g(a^\pi; \mathcal{E}), \tag{A.6}$$

where $a^\pi = (a^\pi(s))_{s\in\mathcal{S}} = \mathcal{M}^{-1}(\pi) \subset \mathbb{R}^{(|\mathcal{A}|-1)|\mathcal{S}|}$ is defined as in (A.5), and $\mathcal{E} = (E; e_0)$. Now consider the set

$$\mathcal{A} = \left\{(a(1), \ldots, a(|\mathcal{S}|)) :\ E\cdot a(s) + e_0 \in \mathrm{ReInt}(\Delta_{|\mathcal{A}|}),\ \forall s\in\mathcal{S}\right\} \subset \mathbb{R}^{(|\mathcal{A}|-1)|\mathcal{S}|}.$$

It is clearly that $\mathcal{A}$ is an open set in $\mathbb{R}^{(|\mathcal{A}|-1)|\mathcal{S}|}$. In addition, for any $a, a' \in \mathcal{A}$, by letting $\pi = \mathcal{M}(a)$, $\pi' = \mathcal{M}(a')$, we have

$$\begin{aligned}
|g(a;\mathcal{E}) - g(a';\mathcal{E})| &= |f_\rho(\pi) - f_\rho(\pi')| \\
&\overset{(a)}{\le} \tfrac{1}{1-\gamma} \sup_{s\in\mathcal{S}} \left\|\pi(\cdot|s) - \pi'(\cdot|s)\right\|_1 \\
&\overset{(b)}{\le} \tfrac{\sqrt{|\mathcal{A}|}}{1-\gamma} \sup_{s\in\mathcal{S}} \|E\|_2 \left\|a(s) - a'(s)\right\|_2 \\
&\le \tfrac{\sqrt{|\mathcal{A}|}}{1-\gamma} \|E\|_2 \left\|a - a'\right\|_2,
\end{aligned}$$

where (a) follows from Lemma 3.8, and (b) follows from the definition (A.4). From the prior relation, we know that $g(\cdot;\mathcal{E}) : \mathcal{A} \to \mathbb{R}$ is a Lipschitz continuous mapping. Combined with the fact that $\mathcal{A}$ is open, we conclude from the Rademacher's theorem [36] that $g(\cdot;\mathcal{E})$ is almost everywhere differentiable in $\mathcal{A}$, when the measure is taken to be the $\mathbb{R}^{(|\mathcal{A}|-1)|\mathcal{S}|}$-dimensional Lebesgue measure. Let us define $\mathcal{A}_z \subset \mathcal{A} \subset \mathbb{R}^{(|\mathcal{A}|-1)|\mathcal{S}|}$ as the set of non-differentiable points of $g(\cdot;\mathcal{E})$. Accordingly, we define $\Pi_z =$

$\{\pi \in \Pi : \pi = \mathcal{M}(a), a \in \mathcal{A}_z\} \subset \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$. We proceed to show that Π_z is a zero-measure set when taking the measure to be the $(|\mathcal{A}|-1)|\mathcal{S}|$-dimensional Hausdorff measure.

Recall that the m-dimensional Hausdorff measure of any set A is defined as (see [37])

$$\mathcal{H}^m(A) = \lim_{\delta \to 0, \delta > 0} \mathcal{H}^m_\delta(A) = \sup_{\delta > 0} \mathcal{H}^m_\delta(A), \tag{A.7}$$

$$\mathcal{H}^m_\delta(A) = w_m \inf_{\{C_j\}_{j=1}^\infty} \left\{ \sum_{j=1}^\infty \left(\frac{\mathrm{diam}(C_j)}{2} \right)^m : \ \mathrm{diam}(C_j) < \delta, \ A \subset \cup_{j=1}^\infty C_j \right\}, \tag{A.8}$$

where $w_m = \pi^{m/2}/\Gamma(\frac{m}{2}+1)$. In addition, by letting $\mathcal{L}^m$ denote the m-dimensional Lebesgure measure in $\mathbb{R}^m$, we have the following relation [37],

$$\mathcal{L}^m(A) = \mathcal{H}^m(A) = \mathcal{H}^m_\delta(A), \ \forall \delta > 0, \ \forall A \subset \mathbb{R}^m. \tag{A.9}$$

Now fix $\delta > 0$, for any collection of subset $\{C_j\}_{j=1}^m \subset \mathbb{R}^{(|\mathcal{A}|-1)|\mathcal{S}|}$ with $\mathrm{diam}(C_j) < \delta$, and $\mathcal{A}_z \subset \cup_{j=1}^\infty C_j$, we know that $\Pi_z \subset \cup_{j=1}^\infty \mathcal{M}(C_j)$, and $\mathrm{diam}(\mathcal{M}(C_j)) \le L_{\mathcal{M}} \mathrm{diam}(C_j)$. Thus,

$$\mathcal{H}^{(|\mathcal{A}|-1)|\mathcal{S}|}_{(L_{\mathcal{M}}\delta)}(\Pi_z) \le \sum_{j=1}^\infty \left(\frac{\mathrm{diam}(\mathcal{M}(C_j))}{2} \right)^{(|\mathcal{A}|-1)|\mathcal{S}|} \le \sum_{j=1}^\infty \left(\frac{\mathrm{diam}(C_j) L_{\mathcal{M}}}{2} \right)^{(|\mathcal{A}|-1)|\mathcal{S}|}.$$

Now by taking infimum over $\{C_j\}_{j=1}^m \subset \mathbb{R}^{(|\mathcal{A}|-1)|\mathcal{S}|}$ of the right hand side, we obtain

$$\begin{aligned}
\mathcal{H}^{(|\mathcal{A}|-1)|\mathcal{S}|}_{(L_{\mathcal{M}}\delta)}(\Pi_z) &\le L_{\mathcal{M}}^{(|\mathcal{A}|-1)|\mathcal{S}|} \inf_{\{C_j\}_{j=1}^\infty} \sum_{j=1}^\infty \left(\frac{\mathrm{diam}(C_j)}{2} \right)^{(|\mathcal{A}|-1)|\mathcal{S}|} \\
&\overset{(a)}{=} L_{\mathcal{M}}^{(|\mathcal{A}|-1)|\mathcal{S}|} \cdot \mathcal{H}^{(|\mathcal{A}|-1)|\mathcal{S}|}_\delta(\mathcal{A}_z) \\
&\overset{(b)}{=} L_{\mathcal{M}}^{(|\mathcal{A}|-1)|\mathcal{S}|} \cdot \mathcal{L}^{(|\mathcal{A}|-1)|\mathcal{S}|}(\mathcal{A}_z) \\
&\overset{(c)}{=} 0,
\end{aligned}$$

where (a) follows from the definition in (A.8), (b) follows from equivalence of $\mathcal{L}^{(|\mathcal{A}|-1)|\mathcal{S}|}$ and $\mathcal{H}^{(|\mathcal{A}|-1)|\mathcal{S}|}_\delta$ for any $\delta > 0$ given (A.9), and the fact that $\mathcal{A}_z \subset \mathbb{R}^{(|\mathcal{A}|-1)|\mathcal{S}|}$. Finally, (c) follows from the fact that $\mathcal{L}^{(|\mathcal{A}|-1)|\mathcal{S}|}(\mathcal{A}_z) = 0$. Thus, by letting $\delta \to 0$ on the left hand side, and making use of the definition of Hausdorff measure (A.7), we obtain $\mathcal{H}^{(|\mathcal{A}|-1)|\mathcal{S}|}(\Pi_z) = 0$.

We then proceed to show that f_ρ is differentiable within $\Pi_d = \mathrm{ReInt}(\Pi) \setminus \Pi_z = \mathcal{M}(\mathcal{A} \setminus \mathcal{A}_z)$, where the differentiability is defined in the sense of Definition 2.1. To see this, we first note that from the differentiability of $g(\cdot; \mathcal{E})$, for any $a' \in \mathcal{A} \setminus \mathcal{A}_z$,

$$g(a; \mathcal{E}) - g(a'; \mathcal{E}) - \langle \nabla g(a'; \mathcal{E}), a - a' \rangle = \mathcal{O}(\|a - a'\|), \ \forall a. \tag{A.10}$$

Let us denote $\nabla g(a; \mathcal{E})[s]$ as the partial derivative of $g(a; \mathcal{E})$ with respect to $a(s)$. Now given any $\pi' \in \Pi_d$, consider any policy π, we know that there exists $a = \mathcal{M}^{-1}(\pi)$, $a' = \mathcal{M}^{-1}(\pi')$, with $a' \in \mathcal{A} \setminus \mathcal{A}_z$. Hence we obtain from the differentiability of $g(\cdot; \mathcal{E})$ at a' that

$$\begin{aligned}
& f_\rho(\pi) - f_\rho(\pi') - \sum_{s \in \mathcal{S}} \langle \nabla g(a'; \mathcal{E})[s], E^\dagger(\pi(\cdot|s) - \pi'(\cdot|s)) \rangle \\
\overset{(a)}{=} & g(a; \mathcal{E}) - g(a'; \mathcal{E}) - \sum_{s \in \mathcal{S}} \langle \nabla g(a'; \mathcal{E})[s], a(s) - a'(s) \rangle \\
\overset{(b)}{=} & \mathcal{O}(\|a - a'\|)
\end{aligned}$$

$$\overset{(c)}{=}\mathcal{O}(\|\pi-\pi'\|),$$

where (a) follows from (A.5) and (A.6), (b) follows from (A.10) and the differentiability of $g(\cdot;\mathcal{E})$ at a', and (c) follows again from (A.5). Thus from the above relation and Definition 2.1, we know that f_ρ is differentiable at any $\pi\in\Pi^d$, whose (s,a)-entry is given by

$$\nabla f_\rho(\pi)[s,a]=\left[(E^\dagger)^\top\nabla g(\mathcal{M}^{-1}(\pi);\mathcal{E})[s]\right][a],\quad\forall(s,a)\in\mathcal{S}\times\mathcal{A}.$$

□

Proof of Lemma 2.5. The essential arguments can be understood as an application of Danskin's Theorem [5], but with additional care to handle the low-dimensional nature of the domain Π. This is due to the reason that Danskin's Theorem requires the domain to be an open set in the euclidean space, which is full-dimensional, and hence can not be directly applied in our setup, see Theorem II of Chapter 3 in [5].

We will inherit the same notations and definitions as in the proof of Lemma 2.3. Let $\Pi_{\mathcal{M}}=\{a^\pi:\pi\in\Pi\}$. Note that the mapping $\mathcal{M}:\Pi_{\mathcal{M}}\to\Pi$ is one-to-one and onto, and $\mathrm{Int}(\Pi_{\mathcal{M}})$ is an open set in $\mathbb{R}^{(|\mathcal{A}|-1)|\mathcal{S}|}$. To proceed, we first show that $g(a;\mathcal{E})$ is differentiable inside $\mathrm{Int}(\Pi_{\mathcal{M}})$. To this end, note that for any $\pi_a=\mathcal{M}(a)$,

$$g(a;\mathcal{E})=f_\rho(\pi_a)=\sum_{s\in\mathcal{S}}\max_{u\in\mathcal{U}}V_u^{\pi_a}(s)\rho(s).$$

To apply Danskin's Theorem, it suffices to show that

(O) $V_u^\pi(s)$ is continuous in (π,u).

(A) For any $a\in\mathrm{Int}(\Pi_{\mathcal{M}})$ and $u\in\mathcal{U}$, $V_u^{\pi_a}(s)$ is differentiable in a, and the partial gradient is continuous in (a,u).

(B) For any $a\in\mathrm{Int}(\Pi_{\mathcal{M}})$, the worst-case uncertainty $\{u\in\mathcal{U}:V_u^{\pi_a}(s)=\max_{u\in\mathcal{U}}V_u^{\pi_a}(s)\}$ is a singleton, denoted by u_{π_α}.

Note that condition (O) is trivial to verify, and condition (B) is readily implied by the precondition of the lemma. We then turn to show (A). For any $a',a\in\mathrm{Int}(\Pi_{\mathcal{M}})$, by letting $\delta=\pi'_a-\pi_a$,

$$\begin{aligned}V_u^{\pi_{a'}}(s)-V_u^{\pi_a}(s)&\overset{(a)}{=}\tfrac{1}{1-\gamma}\sum_{s'\in\mathcal{S}}\sum_{\widetilde{a}\in\mathcal{A}}d_s^{\pi_a,u}(s')Q_u^{\pi_a}(s',\widetilde{a})\delta(\widetilde{a}|s')+\mathcal{O}(\|\delta\|_2),\\&\overset{(b)}{=}\mathcal{L}_u^{\pi_a}(\delta)+\mathcal{O}(\|a'-a\|_2)\\&\overset{(c)}{=}\mathcal{L}_u^{\pi_a}(\mathcal{M}(a'-a))+\mathcal{O}(\|a'-a\|_2),\end{aligned}\tag{A.11}$$

where (a) is due to Lemma 2.1, and (b) and (c) follow from the definition in (A.4), and $\mathcal{L}_u^\pi$ denotes a linear operator of δ mapping to $\mathbb{R}$ and implicitly defined via the first term in equality (a). Hence given (c), since $\mathcal{L}_u^a\circ M$ is again a linear operator, we know that $V_u^{\pi_a}(s)$ is differentiable at any point $a\in\mathrm{Int}(\Pi_{\mathcal{M}})$. To show the continuity of the gradient, it suffices to show that the operator $\mathcal{L}_u^\pi$ is continuous, which simply follows from the fact that Q_u^π and $d_s^{\pi,u}$ is continuous in (π,u) (following from similar arguments of (2.11) and 3.22). Thus term (A) is proved.

Applying the Danskin's Theorem, we obtain that the robust value $V_r^{\pi_a}(s)=\max_{u\in\mathcal{U}}V_u^{\pi_a}(s)$ is also differentiable in a, for any $a\in\mathrm{Int}(\Pi_{\mathcal{M}})$. Specifically, we have

$$V_r^{\pi_{a'}}(s)-V_r^{\pi_a}(s)=\mathcal{L}_{u_{\pi_\alpha}}^{\pi_a}(\mathcal{M}(a'-a))+\mathcal{O}(\|a'-a\|_2),\tag{A.12}$$

for any $a, a' \in \mathrm{Int}(\Pi_{\mathcal{M}})$. Now given any $\pi, \pi' \in \mathrm{ReInt}(\Pi)$, it is clear that $a = a^{\pi}, a' = a^{\pi'}$ both belong to $\mathrm{Int}(\Pi_{\mathcal{M}})$, hence

$$\begin{aligned} V_r^{\pi'}(s) - V_r^{\pi}(s) = V_r^{\pi_{a'}}(s) - V_r^{\pi_a}(s) &\overset{(a)}{=} \mathcal{L}_{u_{\pi_\alpha}}^{\pi_a}(\mathcal{M}(a' - a)) + \mathcal{O}(\left\|a' - a\right\|_2) \\ &\overset{(b)}{=} \mathcal{L}_{u_\pi}^{\pi}(\pi' - \pi) + \mathcal{O}(\left\|\pi' - \pi\right\|_2), \end{aligned}$$

where (a) follows from (A.12), and (b) follows from the definition of a, a' and (A.4), (A.5). Since $\mathcal{L}_{u_\pi}^{\pi}$ is a linear operator, we obtain that $V_r^{\pi}(s)$ is differentiable at $\pi \in \mathrm{ReInt}(\Pi)$ in the sense of Definition 2.1. The concrete form of gradient can be simplify read from the definition of operator $\mathcal{L}_u^{\pi}$ in (A.11), the proof is then completed.

□