Solving Non-rectangular Reward-Robust MDPs via Frequency Regularization

Uri Gadot[1], **Esther Derman**[2], **Navdeep Kumar**[1], **Maxence Elfatihi**[4], **Kfir Levy**[1], **Shie Mannor**[1,3]

[1]Technion - Israel Institute of Technology
[2]MILA, Université de Montréal
[3]NVIDIA Research
[4]IMT Atlantique
ugdaot@gmail.com, esther.derman@mila.quebec, navdeepkumar@campus.technion.ac.il

Abstract

In robust Markov decision processes (RMDPs), it is assumed that the reward and the transition dynamics lie in a given uncertainty set. By targeting maximal return under the most adversarial model from that set, RMDPs address performance sensitivity to misspecified environments. Yet, to preserve computational tractability, the uncertainty set is traditionally independently structured for each state. This so-called rectangularity condition is solely motivated by computational concerns. As a result, it lacks a practical incentive and may lead to overly conservative behavior. In this work, we study coupled reward RMDPs where the transition kernel is fixed, but the reward function lies within an α-radius from a nominal one. We draw a direct connection between this type of non-rectangular reward-RMDPs and applying policy visitation frequency regularization. We introduce a policy-gradient method and prove its convergence. Numerical experiments illustrate the learned policy's robustness and its less conservative behavior when compared to rectangular uncertainty.

Introduction

The Markov decision process (MDP) framework formalizes sequential decision-making problems where the goal is to find a policy that maximizes the agent's performance in a particular environment (Sutton and Barto 2018; Puterman 2014). In most scenarios, the environment's dynamics and/or the reward function are partially known, perturbed by noise, or attacked in an adversarial way. For example, considering a self-driving car simulator, the discrepancy between the idealized virtual environment and unexpectedly varying weather, traffic, and road conditions raises significant challenges during training. Ignoring such model uncertainty can have detrimental effects on the agent's performance, potentially leading to catastrophic failure (Mannor et al. 2004).

On the other hand, solving RMDPs with general uncertainty sets is known to be NP-hard (Wiesemann, Kuhn, and Rustem 2013). To address this issue, previous studies have focused on identifying sub-classes of coupled RMDPs that are still solvable in polynomial time (Mannor, Mebel, and Xu 2016; Goyal and Grand-Clement 2023). Yet, the above studies have mostly focused on RMDPs with a known reward model but uncertain dynamics. Hence, little attention has been given to RMDPs with coupled reward uncertainty and known transition.

Even when the model is comprehensively understood, the challenge of obtaining a precise reward function persists in many practical applications. This predicament can arise when employing a reward model trained on a subset of labeled data or when learning relies on human feedback or preferences. Additionally, although allowing ambiguity on the reward only can seem restrictive, it models a large class of sequential decision-making problems, including MDPs with deterministic transitions such as path planning. Consider again our self-driving car example and assume that its policy is deployed on real road conditions to drive towards a destination point. In this setting, not accounting for reward uncertainty during training could lead the car toward a different destination. On the other hand, a robust policy under rectangular reward uncertainty could yield overly conservative behavior and prevent the car from approaching its goal. The rationale behind this is visually depicted in Figure 1, showcasing why opting for a rectangular uncertainty set might lead to excessive conservatism. This phenomenon is further elaborated upon in Section within the context of a tabular model-based setting.

In this work, we study a subclass of RMDPs where the transition model is known and the reward is uncertain but coupled. We first characterize the nice properties induced by this type of RMDP, as well as the challenges raised by reward coupling. Specifically, we show that without rectangularity, resorting to the common robust Bellman recursion leads to an incorrect and overly conservative value function. Then, under a (general) convex and compact reward uncertainty set, we establish the sufficiency of stationary policies to reach optimal robust return and prove strong duality. For reward uncertainty sets that are further specified as a norm ball centered around a nominal, we explicitly formulate the worst-case reward. The norm of interest being over the whole state-action space, the resulting set is non-rectangular. In this setting, the robust return comes out to be a regularized version of the non-robust return, where the regularization function involves the visitation frequency. This finding also enables us to: *(i)* devise an

Copyright © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

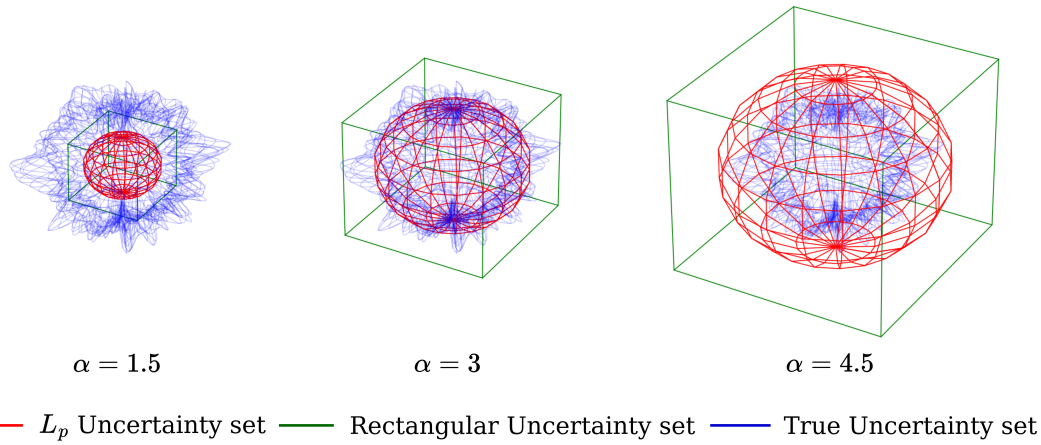

Figure 1: An illustrative example of conservatism in a lower-dimensional context: When faced with an unfamiliar coupled uncertainty set (depicted in blue, see appendix for more info on this particular coupled set), we explore two potential modeling approaches. One involves an s-rectangular uncertainty set with a constant radius parameter α for each state independently (displayed in green). The other chooses a coupled uncertainty set (in red) with the same radius. By increasing α we are increasing conservativness. The rectangular set encompasses the actual uncertainty more swiftly. Nevertheless, this approach results in a rapid expansion of the uncertainty set to a considerable size. Conversely, the coupled set representation covers the genuine uncertainty set at a later point, yet it exhibits a lower degree of conservatism.

efficient policy evaluation algorithm for coupled reward RMDPs; *(ii)* introduce a robust policy-gradient method that trains a reward robust policy with convergence guarantees. Numerical experiments show the advantage of coupling the reward uncertainty set and illustrate the applicability of our method to high-dimensional environments. Moreover, our approach is agnostic to the reinforcement learning (RL) method being used, so it can be added on top of any learning algorithm.

Contributions. To summarize, we make the following contributions: (1) We explicitly formulate the worst-case reward when the reward uncertainty set is a norm ball centered around a nominal, and show that it induces a regularized return whose regularizer is given by state visitation frequency; (2) We provide tractable solutions to this type of reward RMDPs and numerically test their robust behavior against relevant baselines. The proofs of all our theoretical statements can be found in the appendix at (Gadot et al. 2023).

Related Work

Since the work of Wiesemann, Kuhn, and Rustem (2013), uncertainty sets in RMDPs are commonly assumed to be s-rectangular, besides being convex and compact (Ho, Petrik, and Wiesemann 2018, 2021; Derman, Geist, and Mannor 2021). In fact, except for those considered in (Mannor, Mebel, and Xu 2016; Goyal and Grand-Clement 2023) which are locally coupled, s-rectangular uncertainty sets represent the largest class of tractable RMDPs. On the other hand, if not the studies (Xu and Mannor 2010; Mannor, Mebel, and Xu 2016; Derman, Geist, and Mannor 2021; Kumar et al. 2023) that treat both reward and transition uncertainty, RMDP literature has mostly focused just on transition uncertainty. We believe this is due to the greater challenge it represents, as the repercussions of transition ambiguity are epistemic and can lead to a butterfly effect: a small kernel deviation at some state can have an unpredictable effect on another state so we are no longer able to track how local kernel uncertainty propagates across the state space.

Recent works have established a formal connection between reward robustness and policy regularization (Husain, Ciosek, and Tomioka 2021; Brekelmans et al. 2022; Eysenbach and Levine 2022), while others have generalized the robustness-regularization equivalence to general RMDPs to facilitate robust RL (Derman, Geist, and Mannor 2021; Kumar et al. 2022). All these studies focused on a rectangular uncertainty set, whereas we tackle the robust problem induced by coupled reward uncertainty. This coupling leads us to derive a regularization function involving the visitation frequency, which we leverage in our policy gradient method.

In that respect, the robust policy gradient methods recently introduced in (Wang and Zou 2022; Kumar et al. 2023; Li, Zhao, and Lan 2022) assume the uncertainty set to be rectangular. Although Wang and Zou (2022) did prove convergence in the non-rectangular case, their analysis exclusively focused on transition uncertainty while they assumed oracle access to the policy gradient. To the best of our knowledge, our work is the first to propose a provably converging policy gradient method for general reward RMDPs.

A different line of works addresses the problem of corrupted reward signals (Everitt et al. 2017; Wang, Liu, and Li 2020; Rakhsha et al. 2020; Huang and Zhu 2020, 2022; Nika, Singla, and Radanovic 2023). There, the question is how to modify the reward so that the agent is misled to a prescribed policy, but does not detect the attacking signal. In

(Rakhsha et al. 2020), the latter criterion is thought of as a budget constraint, which is formalized as the same coupled norm bound as ours. Although related to the robust setting, the two problems are complementary: a robust agent asks how to cope with an adversary while knowing its deviation level, whereas an attacker asks how to deviate the least from the observed reward so the agent is fooled and chooses a prescribed policy. Moreover, besides tackling the problem from the attacker's viewpoint, this type of study generally focuses on stealthy attacks, i.e., the attacking reward value stays the same across multiple visits of the same state-action pair (Everitt et al. 2017; Huang and Zhu 2020). In the robust setup, the agent can deal with arbitrary time-varying rewards within the uncertainty set.

Preliminaries

Notations

For a set $\mathcal{S}$, $|\mathcal{S}|$ denotes its cardinal. $\langle u, v\rangle := \sum_{s\in\mathcal{S}} u_s v_s$ denotes the dot product between functions $u, v : \mathcal{S} \to \mathbb{R}$ while $\|v\|_p^q := (\sum_s |v(s)|^p)^{\frac{1}{p}}$ is the L_p norm of function v. For $p \in [1,\infty]$, its Hölder conjugate $q \in [1,\infty]$ is the (extended) real number such that $\frac{1}{p} + \frac{1}{q} = 1$. Finally, we denote the probability simplex over $\mathcal{S}$ by $\Delta_{\mathcal{S}} := \{a : \mathcal{S} \to \mathbb{R} | \sum_{s\in\mathcal{S}} a_s = 1, a_s \geq 0 \quad \forall s\}$.

Markov Decision Processes

A Markov decision process (MDP) is a tuple $(\mathcal{S}, \mathcal{A}, P, R, \gamma, \mu)$ such that $\mathcal{S}$, $\mathcal{A}$ are state and action spaces respectively, $P : \mathcal{S}\times\mathcal{A} \to \Delta_{\mathcal{S}}$ is a transition kernel, $R : \mathcal{S}\times\mathcal{A} \to \mathbb{R}$ a reward function, $0 < \mu \in \Delta_{\mathcal{S}}$ an initial distribution over states and $\gamma \in [0, 1)$ a discount factor ensuring that the infinite-horizon return is well-defined. At step t, the agent is in some state $s_t \in \mathcal{S}$, executes an action a_t according to a decision rule π_t that maps past information to a probability distribution over the action space, receives a reward $R(s_t, a_t)$, and transits to another state $s_{t+1} \sim P(\cdot|s_t, a_t)$.

A decision rule can be history-dependent or Markovian, and randomized or deterministic. A policy $\pi = (\pi_t)_{t\geq 0}$ is a sequence of decision rules whose type determines that of the policy. If the decision rules are constant over time, i.e., $\pi_t = \pi_{t+1}$ for all $t \geq 0$, then the corresponding policy is said to be stationary, and we shall define it as $\pi : \mathcal{S} \to \Delta_{\mathcal{A}}$ with a slight abuse of notation. We further denote by $\Pi := \Delta_{\mathcal{A}}^{\mathcal{S}}$ the set of all stationary policies.

Let $R^\pi(s) := \sum_{a\in\mathcal{A}} \pi_s(a)R(s,a)$ and $P^\pi(s'|s) := \sum_{a\in\mathcal{A}} \pi_s(a)P(s'|s,a), \forall s, s' \in \mathcal{S}$, the expected reward and transition, respectively, where $\pi_s := \pi(\cdot|s)$ is a shorthand notation for policy π at state s. The overall goal is to maximize the following return over the policy space:

$$\rho_R^\pi := \langle R, d^\pi\rangle = \langle \mu, v_R^\pi\rangle,$$

where $d^\pi := \mu^\top(\mathbf{I}_{\mathcal{S}} - \gamma P^\pi)^{-1}$ is the occupation measure associated with policy π and $v_R^\pi := (\mathbf{I}_{\mathcal{S}} - \gamma P^\pi)^{-1}R^\pi$ the value function under policy π and model parameters (P, R). In this setting, it is known that there exists a stationary policy achieving maximal return (Puterman 2014). We thus denote the optimum by $\rho_R^{\pi^*}$. In practice, the problem can be solved through Bellman operators, respectively given by $\mathcal{T}_R^\pi v := R^\pi + \gamma P^\pi v$ and $\mathcal{T}_R^* v := \max_{\pi\in\Pi} \mathcal{T}_R^\pi v, \quad \forall v \in \mathbb{R}^{\mathcal{S}}$. The subscript R in the operator notation indicates the dependence on the reward function R, which will be useful in the reward-robust setting we introduce next.

Reward-Robust MDPs

In a reward-robust MDP (reward RMDP), the reward function R is unknown but lies in a given uncertainty set $\mathcal{R}$. This set is commonly assumed to be s-rectangular, i.e., it can be decomposed over states as $\mathcal{R} = \times_{s\in\mathcal{S}}\mathcal{R}_s$, in which case we denote it by $\mathcal{R}^{\mathrm{s}}$. If it can further be decomposed across states and actions, i.e., if $\mathcal{R} = \times_{s\in\mathcal{S},a\in\mathcal{A}}\mathcal{R}_{(s,a)}$, we will denote it by $\mathcal{R}^{\mathrm{sa}}$.

The objective is to maximize the robust performance $\rho_{\mathcal{R}}^\pi := \min_{R\in\mathcal{R}} \rho_R^\pi$ over Π. For any policy $\pi \in \Pi$, the reward model realizing the worst return is denoted by $R_{\mathcal{R}}^\pi \in \arg\min_{R\in\mathcal{R}} \rho_R^\pi$. Its corresponding robust value and robust Q-value functions are respectively defined as:

$$v_{\mathcal{R}}^\pi := v_{R_{\mathcal{R}}^\pi}^\pi, \quad Q_{\mathcal{R}}^\pi := Q_{R_{\mathcal{R}}^\pi}^\pi. \tag{1}$$

Based on non-robust definitions, they are related through:

$$v_{\mathcal{R}}^\pi(s) = \langle \pi_s, Q_{\mathcal{R}}^\pi(s,\cdot)\rangle, \quad \forall s \in \mathcal{S}\,.$$

When the uncertainty set is s-rectangular, the above value function coincides with the worst value, that is: $v_{\mathcal{R}^{\mathrm{s}}}^\pi = \min_{R\in\mathcal{R}^{\mathrm{s}}} v_R^\pi$. On the other hand, one needs (s,a)-rectangularity for the same to hold for Q-values, i.e., $Q_{\mathcal{R}^{\mathrm{sa}}}^\pi = \min_{R\in\mathcal{R}^{\mathrm{sa}}} Q_R^\pi$ (Nilim and El Ghaoui 2005; Iyengar 2005; Wiesemann, Kuhn, and Rustem 2013; Kumar et al. 2023).

The optimal robust return is defined as

$$\rho_{\mathcal{R}}^* := \max_{\pi\in\Pi} \rho_{\mathcal{R}}^\pi.$$

A standard way to solve RMDPs is through Bellman recursion. The robust Bellman evaluation operator is

$$\mathcal{T}_{\mathcal{R}}^\pi v = \min_{R\in\mathcal{R}} \mathcal{T}_R^\pi v, \quad \forall v \in \mathbb{R}^{\mathcal{S}}\,.$$

Although non-linear, it is still a γ-contraction for any uncertainty set $\mathcal{R}$ (Wiesemann, Kuhn, and Rustem 2013). The same applies to the robust Bellman optimal operator defined as

$$\mathcal{T}_{\mathcal{R}}^* v := \max_{\pi\in\Pi} \mathcal{T}_{\mathcal{R}}^\pi v, \quad \forall v \in \mathbb{R}^{\mathcal{S}}\,.$$

In the s-rectangular case, the robust value function $v_{\mathcal{R}^{\mathrm{s}}}^\pi$ (respectively, the optimal robust value function $v_{\mathcal{R}^{\mathrm{s}}}^*$) is the fixed point of the robust Bellman evaluation operator (resp., of the robust Bellman optimal operator) (Wiesemann, Kuhn, and Rustem 2013). Thus, these RMDPs can be solved using policy iteration (Wiesemann, Kuhn, and Rustem 2013; Ho, Petrik, and Wiesemann 2021; Derman, Geist, and Mannor 2021; Kumar et al. 2022).

Analyzing Reward-Robust MDPs

In this section, we show that the above robust operators can no longer be used for general (non-rectangular) uncertainty sets $\mathcal{R}$. Indeed, as stated in Prop. 1, the robust Bellman evaluation operator (resp., the robust Bellman optimal operator) does not admit the robust value function $v^\pi_\mathcal{R}$ (resp., the optimal robust value function $v^*_{\mathcal{R}^s}$) as a fixed point.

Proposition 1. *For non-rectangular uncertainty set $\mathcal{R}$, the robust Bellman operator $\mathcal{T}^\pi_\mathcal{R}$ (resp., $\mathcal{T}^*_\mathcal{R}$) has $v^\pi_{C(\mathcal{R})}$ (resp., $v^*_{C(\mathcal{R})}$) as its fixed point, where $C(\mathcal{R})$ is the smallest s-rectangular uncertainty set containing $\mathcal{R}$, that is*

$$C(\mathcal{R}) = \cap_{\mathcal{R}\subseteq\mathcal{R}^s}\mathcal{R}^s.$$

Hence, robust value iteration on a general (non-rectangular) uncertainty set can lead to an overly conservative solution, as $C(\mathcal{R})$ can be much larger than $\mathcal{R}$ in large state spaces. This is illustrated in Fig. 1. Therefore, other methods need to be used to solve coupled reward RMDPs. Before introducing our solution, we begin by presenting key overarching findings that apply to any convex and compact reward uncertainty set.

Lemma 2 (Stationary policies are enough)**.** *Assume that $\mathcal{R}$ is a compact and convex set. Then, there exists a stationary policy $\pi \in \Pi$ that achieves maximal robust return:*

$$\min_{R\in\mathcal{R}} \mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t R(s_t,a_t)\,\middle|s_0\sim\mu, a_t\sim\pi_t(\cdot|s_t), \right.$$
$$\left. s_{t+1}\sim P(\cdot|s_t,a_t), \forall t\geq 0\right].$$

The aforementioned result establishes that even though the optimal policy may be non-Markovian for general RMDPs (Wiesemann, Kuhn, and Rustem 2013), in our setting, we can focus on the set of stationary policies Π, similar to non-robust MDPs (Puterman 2014). Moreover, strong duality holds, as stated below.

Lemma 3 (Duality)**.** *For all convex uncertainty sets $\mathcal{R}$, the order of optimization can be interchanged, that is*

$$\max_{\pi\in\Pi}\min_{R\in\mathcal{R}}\rho^\pi_\mathcal{R} = \min_{R\in\mathcal{R}}\max_{\pi\in\Pi}\rho^\pi_\mathcal{R}.$$

In our framework, we examine particular constraints on reward perturbations within the aforementioned setting. Given a nominal reward denoted by $R_0 \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ and a positive radius $\alpha > 0$, the uncertainty set we focus on is an L_p-ball centered around this nominal:

$$\mathcal{R}_p := \{R\in\mathbb{R}^{\mathcal{S}\times\mathcal{A}} \mid \|R-R_0\|_p \leq \alpha\}.$$

We note that although this constraint is restricted to L_p norm balls, it is non-rectangular and still enjoys the benefit of generality. L_p norm balls encompass a wide range of uncertainty patterns such as worst-case and probabilistic uncertainties, by selecting appropriate values of p (Mannor, Mebel, and Xu 2012; Delage and Mannor 2010).

Remark 4. *For the sake of simplicity and to enhance the clarity of our expression, we limit our study to L_p-ball constrained uncertainty sets. Nonetheless, our approach readily holds for weighted L_p-norms. Further elaboration on this extension can be found in the appendix.*

Worst Reward Function

The ball structure enables us to derive the worst reward function in closed form and illuminates its effect on the occupation measure. This worst-case reward expression is formalized below and in fact, represents a key component of the robust learning methods introduced later on.

Theorem 5 (Worst-case reward)**.** *For any policy $\pi\in\Pi$ and state-action pair $(s,a)\in\mathcal{S}\times\mathcal{A}$, the worst-case reward at (s,a) is given by:*

$$R^\pi_{\mathcal{R}_p}(s,a) = R_0(s,a) - \alpha\left(\frac{d^\pi(s,a)}{\|d^\pi\|_q}\right)^{q-1}.$$

For simplicity, we will write $R^\pi_p := R^\pi_{\mathcal{R}_p}$.

Thm. 5 highlights the adversarial strategy reward-robust MDPs model. The occupation measure in the numerator shows a diminution of the reward in states that the agent frequently visits. As for the denominator, it can be thought of as the entropy of the occupancy measure: evenly distributed occupancy leads to a lower norm and a weaker adversary, whereas concentrated occupancy leads to a higher norm and a stronger adversary. Please refer to Tab. 1, for an example of the worst reward penalties for different values of p.

p	$R^\pi_p(s,a) - R_0(s,a)$	Type of penalty
p	$\alpha\left(\frac{d^\pi(s,a)}{\|d^\pi\|_q}\right)^{q-1}$	General norm penalty
∞	α	Uniform penalty
2	$\alpha\frac{d^\pi(s,a)}{\|d^\pi\|_2}$	α-normed frequency
1	$\frac{\alpha}{\lvert\mathcal{X}^*\rvert}\mathbb{1}\{(s,a)\in\mathcal{X}^*\}$	One-hot penalty

Table 1: Reward penalty induced by different coupled-reward uncertainty sets. For $p=1$, $\mathcal{X}^* := \arg\max_{(s,a)\in\mathcal{S}\times\mathcal{A}} d^\pi(s,a)$.

Furthermore, Thm. 5 gives us one of our main findings:

Corollary 6 (Reward robust return)**.** *For a general L_p norm uncertainty set, the robust return is given by:*

$$\rho^\pi_{\mathcal{R}_p} = \rho^\pi_{R_0} - \alpha\|d^\pi\|_q.$$

The factor $-\alpha\|d^\pi\|_q$ behaves like an entropy. Indeed, it increases as the occupation measure is more distributed and vice versa. The preceding results unveil an intriguing connection between reward-robust MDPs and regularized MDPs that employ a variant of 'frequency' regularization. This correlation mirrors earlier research efforts that explored the relationship between policy regularization and

robust RL, as demonstrated in prior studies (Eysenbach and Levine 2022; Derman, Geist, and Mannor 2021; Brekelmans et al. 2022). In the context of general L_p norm uncertainty sets, we establish an explicit formulation for this regularizer and ascertain its reliance on the occupancy measure. Consequently, the resolution of general reward RMDPs becomes achievable by effectively addressing regularized MDPs (Geist, Scherrer, and Pietquin 2019) that encompass the aspect of 'frequency' regularization. Tab. 2 in the appendix provides a comprehensive overview of the regularization function for different L_p-norm ball uncertainty sets. It is evident that assuming different levels of rectangularity can be likened to imposing distinct budget constraints on an adversarial entity, or 'world'. In the case of (s, a)-rectangularity, the optimal strategy is to account for the most adverse penalty associated with each (s, a) pair. On the other hand, adopting s-rectangularity permits the adversary to manipulate the reward function independently for each state within certain limits, thereby prompting the robust policy to distribute its visitation more evenly across actions, yet independently for each state. This requires the potential employment of entropy-based regularization techniques. By relinquishing the constraints of rectangularity and considering a more general adversarial 'budget', a robust policy would strive to distribute its visitation frequency across the entire $\mathcal{S} \times \mathcal{A}$ space, which may involve implementing a form of 'frequency' regularization.

Policy Evaluation

As outlined in Prop. 1, utilizing the robust Bellman operator in the non-rectangular setting might not yield the robust value function. Nevertheless, Thm. 5 yields the formulation of the 'worst reward' Bellman operator, as articulated below.

Theorem 7. *Let an uncertainty set of the form $\mathcal{R} := \mathcal{R}_p$. Then, for any policy $\pi \in \Pi$, the robust value iteration*

$$\begin{aligned} v_{n+1}(s) &= T^\pi_{R_0} v_n(s) - \alpha \frac{\sum_a \pi_s(a) d^\pi(s,a)^{q-1}}{\|d^\pi\|_q^{q-1}} \\ &=: [\mathcal{T}^{\pi,\mathrm{REG}}_{\mathcal{R}_p} v_n](s), \quad \forall s \in \mathcal{S}, \end{aligned}$$

converges linearly to the robust value function $v^\pi_{\mathcal{R}_p}$.

This is nothing more than the non-robust Bellman operator for the MDP with the worst reward function. The new operator $\mathcal{T}^{\pi,\mathrm{REG}}_{\mathcal{R}_p}$ preserves the γ-contracting property of the non-robust Bellman operator. Thus, the sequence given by $v_{n+1} := \mathcal{T}^{\pi,\mathrm{REG}}_{\mathcal{R}_p} v_n$ converges to $v^\pi_{\mathcal{R}_p}$ (as defined in Eq. (1)).

A remaining question is how the robust Q-value $Q^\pi_{\mathcal{R}_p}$ relates to the robust value function $v^\pi_{\mathcal{R}_p}$, namely, to the fixed point of the Bellman operator introduced before. The theorem below establishes the connection between these measures. It ties the robust Q-function to the robust value function by the nominal non-robust Bellman operator and the 'frequency' regularization term. Different expressions of this regularizer are displayed in Tab. 1.

Corollary 8. *For the uncertainty set $\mathcal{R}_\alpha$, the robust Q-value can be obtained from the robust value function via*

$$Q^\pi_{\mathcal{R}_p}(s,a) = T^\pi_{R_0} v^\pi_{\mathcal{R}_p}(s) - \alpha \left(\frac{d^\pi(s,a)}{\|d^\pi\|_q} \right)^{q-1}.$$

Complexity Analysis We note that the complexity of computing an occupation measure of a given policy is $O(S^2 A \log(\frac{1}{\epsilon}))$. This implies that the complexity of policy evaluation in our algorithm is also $O(S^2 A \log(\frac{1}{\epsilon}))$ for reward robust MDPs, similarly to non-robust MDPs (Sutton et al. 1999), (s, a), and s-rectangular robust MDPs (Derman, Geist, and Mannor 2021; Wang and Zou 2022). A detailed analysis can be found in the appendix. Notably, the tractability of robust policy gradient estimation for non-rectangular convex kernel uncertainty sets is still an open question.

Reward-Robust Policy Gradient

As mentioned in Prop. 1, employing the optimal robust Bellman operator within the non-rectangular setting may not necessarily yield the optimal robust value function. Furthermore, transforming the robust operator introduced in Thm. 7 into an optimal robust operator is not straightforward. Indeed, a greedy update in π also impacts the 'frequency' regularization component. Hence, we cannot utilize a value iteration method to achieve an optimal robust policy. Alternatively, we introduce a policy gradient method for this type of RMDPs and provide convergence guarantees.

As a main prerequisite, we first establish a policy-gradient theorem for general reward RMDPs.

Theorem 9. *The reward robust policy-gradient is given by:*

$$\frac{\partial \rho^\pi_{\mathcal{R}_p}}{\partial \pi} = \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} d^\pi(s) Q^\pi_{\mathcal{R}_p}(s,a) \nabla \pi_s(a),$$

where $Q^\pi_{\mathcal{R}_p}$ is simply the non-robust Q-value under the worst reward, i.e., $Q^\pi_{\mathcal{R}_p} := Q^\pi_{R^\pi_p}$ obtained using Cor. 8.

Global Convergence

We use the gradient derived in Thm. 9 to define our projected policy gradient ascent rule as

$$\pi_{k+1} := \mathbf{proj}_\Pi \left[\pi_k + \eta_k \frac{\partial \rho^{\pi_k}_{\mathcal{R}_p}}{\partial \pi} \right].$$

The robust return can be non-differentiable for general uncertainty sets (Wang and Zou 2022; Wang, Ho, and Petrik 2023). However, the result below establishes the differentiability of the robust return when it is constrained by an L_p-ball.

Lemma 10 (Smoothness)**.** *For all $p \in (1, \infty)$, the robust return $\rho^\pi_{\mathcal{R}_p}$ is β-smooth in π, where β is a constant that depends on the problem parameters and is described in the appendix.*

Taking step size $\eta_k = \frac{1}{\beta}$, we have the following convergence result.

Theorem 11 (Convergence). *The suboptimality gap at the k^{th} iteration decays as*

$$\rho^*_{\mathcal{R}_p} - \rho^{\pi_k}_{\mathcal{R}_p} \leq c|\mathcal{S}|\beta\frac{\rho^*_{\mathcal{R}_p} - \rho^{\pi_0}_{\mathcal{R}_p}}{k},$$

where c is a constant that depends on the discount factor γ and on a mismatch coefficient described in the appendix.

The outcome presented here establishes the global convergence of the reward-robust policy gradient for the first time. This convergence holds with an iteration complexity of $O(\frac{1}{\epsilon})$ to attain an ϵ-optimal policy, similarly to non-robust MDPs (Agarwal et al. 2021; Xiao 2022). It is worth noting that the aforementioned convergence result also holds for (s, a) and s-rectangular L_p constrained reward robust MDPs, maintaining the same convergence rates. Differently, under kernel uncertainty, robust policy gradient exhibits an iteration complexity of $O(\frac{1}{\epsilon^4})$ (Wang, Ho, and Petrik 2023). Further comparison is detailed in the appendix.

Scaling Reward-Robust Policy-Gradient

We now propose an online actor-critic algorithm that employs our method but is adaptable to high-dimensional settings (see Alg. 1). To achieve this, we utilize Thm. 7 to approximate the robust value function, a key component of Thm. 9. Estimating the occupancy measure is also imperative for applying the 'frequency' regularizer. In this regard, we introduce the following result:

Proposition 12. *(Lemma 1 of (Kumar et al. 2023)) For all policies π and kernels P, the iterative sequence given by*

$$d_{n+1} := \mu + \gamma P^\pi d_n, \quad \forall n \in \mathbb{N},$$

converges linearly to d^π.

Experiments

This section is dedicated to two categories of experiments. In Sec. , we illustrate the conservative nature of rectangularity assumptions. In Sec. , we assess the efficacy of our proposed algorithm in a high-dimensional setting. For reproducibility, we have provided a link to our source code in the appendix, along with comprehensive experiment details and supplementary results.

Conservative Rectangularity

To further illustrate the conservatism inherent to the rectangularity assumption, we explore a tabular problem. Imagine a model-based scenario where we possess knowledge of P, μ, and $R_0 \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$. However, during testing, the reward function is drawn from a multivariate Gaussian distribution as follows: $R \sim \mathcal{N}(R_0, \Sigma)$, where the covariance matrix Σ is a non-diagonal positive semi-definite matrix. It is crucial to note that the agent remains unaware of this perturbation. To derive a robust policy, we tackle this scenario using two types of uncertainty sets.

The first approach consists of treating this as an s-rectangular reward-RMDP with an L_2-norm uncertainty set, where the radius around each state remains constant, that is $\alpha_s \equiv \alpha$. The second approach adopts a coupled reward-RMDP framework with an L_2 norm uncertainty set, where the radius pertains to the entire reward function, labeled as α. For both models, a soft-max parameterization is applied, and a model-based policy gradient (PG) is employed.

For the s-rectangular RMDP, we utilize the method described in (Kumar et al. 2022). In the case of the general RMDP, we employ Alg. 1 in its simplified model-based version. Subsequently, we train the robust policy, subject it to testing over 1000 samples drawn from the unknown distribution, and measure the Conditional Value-at-Risk (CVaR) for the worst-performing 5%. This process is repeated across various α values. The results depicted in Figure 2 underscore that the general model attains superior 'worst' performance and exhibits greater stability against radius estimation errors. This highlights that opting for a rectangular uncertainty set can significantly reduce the worst-case performance within the true uncertainty framework. For more findings from this experiment, please refer to the appendix.

Algorithm 1: Actor-Critic for General Reward RMDPs

Input: Differentiable policy $\pi_\theta(a|s)$; Q-value $Q_\omega(s,a)$; Frequency $d^\pi_\zeta(a|s)$, Step-sizes $\eta_\theta, \eta_\omega, \eta_\zeta$; Batch size N; Robustness radius α

1: **for** $t = 0, 1, 2, \cdots$ **do**
2: Using current policy π_{θ_t}, collect current batch $\{(s_i, a_i, r_i, s'_i)\}_{i=1}^N$.
3: Update policy parameters
$\theta_{t+1} = \theta_t + \eta_\theta \frac{1}{N}\sum_{i=1}^N (Q_\omega(s_i, a_i)\nabla_\theta \pi(a_i|s_i))$
4: Update robust Q function parameters
$\omega_{t+1} = \omega_t + \eta_\omega \delta_t \nabla_\omega Q$, where $\delta_t =$ robust TD-error
5: Update occupancy measure parameters
$\zeta_{t+1} = \zeta_t + \eta_\zeta \Delta_t \nabla_\zeta d_\zeta$,
where Δ_t is the visitation frequency error (Prop. 12)
6: **end for**

Output: Robust value Q_ω; Robust policy π_θ

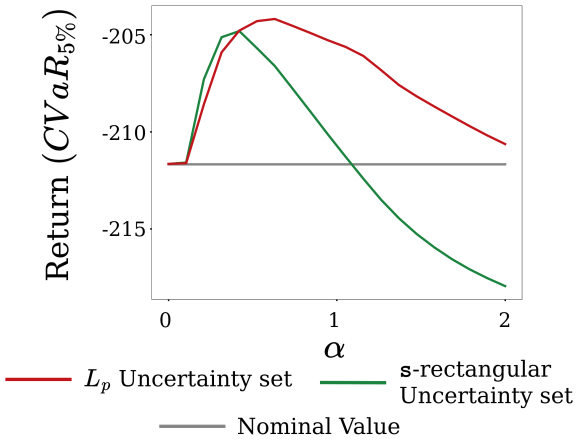

Figure 2: $CVaR_{5\%}$ results for different α

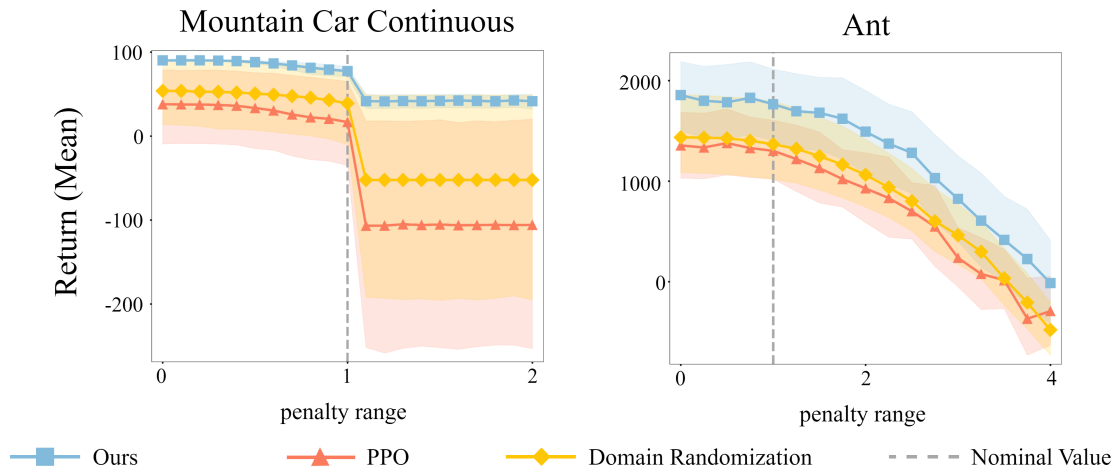

Figure 3: Evaluation results on both environments for different reward perturbations.

PG For High-Dimensional Setting

We now undertake experiments within the online robust RL framework and evaluate the effectiveness of Alg. 1 for learning robust policies. This involves training the agent using the nominal reward function and evaluating its performance under perturbed reward functions. We examine two continuous control tasks of high dimension from OpenAI's Gym (Brockman et al. 2016): 'Mountain Car Continuous' (Moore 1990) and Mujoco's (Todorov, Erez, and Tassa 2012) 'Ant-v3' environment. As the baseline RL algorithm, we opt for PPO (Schulman et al. 2017). In addition, to compare our method with other robust methods we consider another commonly-used robust RL approach: domain randomization. Domain randomization trains the agent across a range of scenarios by introducing variations in the reward function during training. This equips the trained agent with robustness against analogous perturbations during testing. Notably, it is important to acknowledge that domain randomization holds an advantage over our proposed algorithm in that it can utilize multiple perturbed reward functions during training. In contrast, our algorithm remains entirely agnostic to such parameters and solely necessitates samples from the nominal reward function. To obtain stable results, we run each experiment with 10 random seeds, and report the mean and 95% stratified bootstrap confidence intervals (CIs) (Efron 1992).

In both environments, we introduce reward perturbation by incorporating a 'penalized' segment marked by a single range parameter. Whenever the agent is within this range, it incurs a penalty proportional to its location in the area. While the range remains consistent during training, it varies during testing. The agent's performance under diverse perturbed rewards is illustrated in Figure 3. Both results demonstrate that when the range significantly deviates from the nominal reward, our method outperforms the baseline PPO and the domain randomization method. While it may appear surprising that our approach exhibited superior performance compared to the non-robust algorithm under the nominal reward function, it is worth noting that previous works have shown that applying regularization may also enhance average performance (Liu et al. 2019).

Conclusion And Discussion

In this paper, we explore the often-overlooked realm of coupled RMDPs. Our attention is directed toward the context of reward uncertainty, wherein we demonstrate that the challenges posed might be less formidable than previously thought. Our study establishes that achieving tractability does not necessitate adhering to rectangularity assumptions. By drawing a direct connection between coupled L_p reward RMDPs and regularized MDPs with a policy visitation frequency regularizer, we can prove the convergence of reward robust policy gradient. We present an online-based scalable algorithm for learning a robust policy within this framework and empirically substantiate our algorithm's capability to learn a robust policy.

Furthermore, we provide a rationale for employing a coupled uncertainty set. In the case where the uncertainty set is unknown but needs to be learned from samples (Lim, Xu, and Mannor 2016), our coupled approach greatly facilitates learning, as it reduces the uncertainty set parameter to only one radius size. It is also more interpretable in safe RL since it can be thought of as the attacker's budget. As such, one interesting direction would be to extend our setting to the case where the uncertainty radius is unknown but needs to be inferred from trajectories.

One limitation of our work is that it is relevant for L_p-norm balls for $p > 1$. Engaging future avenues of research could involve extending the framework to accommodate an adaptive adversary or pursuing analogous outcomes within the realm of coupled kernel uncertainty RMDPs, a domain that currently remains largely unexplored.

References

Agarwal, A.; Kakade, S. M.; Lee, J. D.; and Mahajan, G. 2021. On the theory of policy gradient methods: Optimality,

approximation, and distribution shift. *The Journal of Machine Learning Research*, 22(1): 4431–4506.

Brekelmans, R.; Genewein, T.; Grau-Moya, J.; Delétang, G.; Kunesch, M.; Legg, S.; and Ortega, P. 2022. Your Policy Regularizer is Secretly an Adversary. *Transactions on Machine Learning Research (TMLR)*.

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

Delage, E.; and Mannor, S. 2010. Percentile optimization for Markov decision processes with parameter uncertainty. *Operations research*, 58(1): 203–213.

Derman, E.; Geist, M.; and Mannor, S. 2021. Twice regularized MDPs and the equivalence between robustness and regularization. *Advances in Neural Information Processing Systems*, 34: 22274–22287.

Efron, B. 1992. Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics: Methodology and distribution*, 569–593. Springer.

Everitt, T.; Krakovna, V.; Orseau, L.; Hutter, M.; and Legg, S. 2017. Reinforcement learning with a corrupted reward channel. *arXiv preprint arXiv:1705.08417*.

Eysenbach, B.; and Levine, S. 2022. Maximum Entropy RL (Provably) Solves Some Robust RL Problems. *International Conference on Learning Representations*.

Gadot, U.; Derman, E.; Kumar, N.; Elfatihi, M. M.; Levy, K.; and Mannor, S. 2023. Solving Non-Rectangular Reward-Robust MDPs via Frequency Regularization. *arXiv preprint arXiv:2309.01107*.

Geist, M.; Scherrer, B.; and Pietquin, O. 2019. A theory of regularized Markov decision processes. In *International Conference on Machine Learning*, 2160–2169. PMLR.

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

Ho, C. P.; Petrik, M.; and Wiesemann, W. 2018. Fast Bellman updates for robust MDPs. In *International Conference on Machine Learning*, 1979–1988. PMLR.

Ho, C. P.; Petrik, M.; and Wiesemann, W. 2021. Partial Policy Iteration for l_1-Robust Markov Decision Processes. *J. Mach. Learn. Res.*, 22: 275–1.

Huang, Y.; and Zhu, Q. 2020. Manipulating reinforcement learning: Poisoning attacks on cost signals. *arXiv preprint arXiv:2002.03827*.

Huang, Y.; and Zhu, Q. 2022. Reinforcement learning for linear quadratic control is vulnerable under cost manipulation. *arXiv preprint arXiv:2203.05774*.

Husain, H.; Ciosek, K.; and Tomioka, R. 2021. Regularized Policies are Reward Robust. In *International Conference on Artificial Intelligence and Statistics*, 64–72. PMLR.

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

Kumar, N.; Derman, E.; Geist, M.; Levy, K.; and Mannor, S. 2023. Policy gradient for s-rectangular robust markov decision processes. *arXiv preprint arXiv:2301.13589*.

Kumar, N.; Levy, K.; Wang, K.; and Mannor, S. 2022. Efficient Policy Iteration for Robust Markov Decision Processes via Regularization. *arXiv preprint arXiv:2205.14327*.

Li, Y.; Zhao, T.; and Lan, G. 2022. First-order Policy Optimization for Robust Markov Decision Process. *arXiv preprint arXiv:2209.10579*.

Lim, S. H.; Xu, H.; and Mannor, S. 2016. Reinforcement Learning in Robust Markov Decision Processes. *Mathematics of Operations Research*, 41(4): 1325–1353.

Liu, Z.; Li, X.; Kang, B.; and Darrell, T. 2019. Regularization matters in policy optimization. *arXiv preprint arXiv:1910.09191*.

Mannor, S.; Mebel, O.; and Xu, H. 2012. Lightning does not strike twice: Robust MDPs with coupled uncertainty. *arXiv preprint arXiv:1206.4643*.

Mannor, S.; Mebel, O.; and Xu, H. 2016. Robust MDPs with K-Rectangular Uncertainty. *Math. Oper. Res.*, 41(4): 1484–1509.

Mannor, S.; Simester, D.; Sun, P.; and Tsitsiklis, J. N. 2004. Bias and Variance in Value Function Estimation. In *Proceedings of the Twenty-First International Conference on Machine Learning*, ICML '04, 72. New York, NY, USA: Association for Computing Machinery. ISBN 1581138385.

Moore, A. W. 1990. Efficient Memory-based Learning for Robot Control. Technical report, University of Cambridge.

Nika, A.; Singla, A.; and Radanovic, G. 2023. Online Defense Strategies for Reinforcement Learning Against Adaptive Reward Poisoning. In *26th International Conference on Artificial Intelligence and Statistics*. PMRL.

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

Puterman, M. L. 2014. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons.

Rakhsha, A.; Radanovic, G.; Devidze, R.; Zhu, X.; and Singla, A. 2020. Policy teaching via environment poisoning: Training-time adversarial attacks against reinforcement learning. In *International Conference on Machine Learning*, 7974–7984. PMLR.

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Sutton, R. S.; and Barto, A. G. 2018. *Reinforcement Learning: An Introduction*. The MIT Press, second edition.

Sutton, R. S.; McAllester, D. A.; Singh, S. P.; Mansour, Y.; et al. 1999. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems*, volume 99, 1057–1063. Citeseer.

Todorov, E.; Erez, T.; and Tassa, Y. 2012. MuJoCo: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 5026–5033. IEEE.

Wang, J.; Liu, Y.; and Li, B. 2020. Reinforcement learning with perturbed rewards. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 6202–6209.

Wang, Q.; Ho, C. P.; and Petrik, M. 2023. Policy Gradient in Robust MDPs with Global Convergence Guarantee. *Proceedings of the 40th International Conference on Machine Learning, PMLR 202:35763-35797*.

Wang, Y.; and Zou, S. 2022. Policy Gradient Method For Robust Reinforcement Learning. *International Conference on Machine Learning*, 162: 23484–23526.

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

Xiao, L. 2022. On the convergence rates of policy gradient methods. *The Journal of Machine Learning Research*, 23(1): 12887–12922.

Xu, H.; and Mannor, S. 2010. Distributionally robust Markov decision processes. *Advances in Neural Information Processing Systems*, 23.