Targeted Uncertainty Reduction in Robust MDPs

Uri Gadot Technion
Kaixin Wang Technion
Esther Derman Mila - Quebec AI Institute
Navdeep Kumar Technion
Kfir Levy Technion
Shie Mannor Technion

Abstract

Robust Markov decision processes (MDPs) provide a practical framework for generalizing trained agents to new environments. There, the objective is to maximize performance under the worst model of a given uncertainty set. By construction, this raises a performance-robustness dilemma: accounting for too large uncertainty yields guarantees against larger disturbances, whilst too small uncertainty may result in over-sensitivity to model misspecification. In this work, we introduce an online method that addresses the conservativeness of robust MDPs by strategically contracting the uncertainty set. First, we explicitly formulate the gradient of the robust return with respect to the uncertainty radius. This gradient derivation enables us to prioritize efforts in reducing uncertainty and leads us to interesting findings on the relation between the robust return and the uncertainty set. Second, we present a sampling-based algorithm aimed at enhancing our uncertainty estimation with respect to the robust return. Third, we illustrate the effectiveness of our algorithm within a tabular environment.

1 Introduction

The Markov Decision Process (MDP) model provides methodology to interact with a given environment and learn an optimal policy that maximizes the cumulative discounted reward [14]. In many cases, the environment is unknown or partially known which can cause the deterioration of the performance of the policy [11]. To alleviate this issue, robust MDP (RMDP) framework consider a set of possible environments (the uncertainty set) instead of a single environment, and maximize the robust performance [1, 4, 6]. The robust performance of a policy is it's worst performance over all of the environment within the uncertainty set. Robust MDPs are not only resilient to the perturbation of the environment but were also shown to have better generalization [18]. Unfortunately, robust MDPs are NP-Hard for general uncertainty sets, hence the uncertainty set is assumed to be `s`-rectangular and convex in order to make the problem tractable [17]. An uncertainty set is called `s`-rectangular (resp. `sa` rectangular) if it can be decomposed as the Cartesian product of states (resp. states and actions).

There is a plethora of work on the topic [12, 6, 17, 1, 10] including the most recent works [15, 8, 16, 7] that effectively exploits robustness-regularizer equivalence [3]. All these works consider a fixed uncertainty set, however, a large uncertainty set can result in extremely cautious solution leading to over-conservativeness. In many situations, such as designing simulators, approximating dynamics, designing robots etc, we do have some control on the precision of the model parameters. However, there might be a cost associated with that precision, and obtaining arbitrary precision in the model parameters might be prohibitively expensive and unnecessary. This necessitates the reduction in uncertainty in most sensitive areas. To the best of our knowledge, there does not exist any work that studies the strategic contraction of the uncertainty set in order to reduce the conservativeness while still enjoying the benefits of robust MDPs.

NeurIPS 2023 Workshop on Generalization in Planning (GenPlan 2023).

To illustrate the significance of addressing uncertainty reduction in RMDPs, imagine the following scenario in the context of autonomous robotics policy design, where different components of a robot are manufactured across multiple factories. Due to inherent imperfections in these component production processes, there exists a degree of uncertainty in the specifications of each part, influenced by the capabilities of the respective factories. Now, envision a scenario where you have the option to dispatch a single inspector to one of these factories, leading to improved precision in the delivered part parameters. Our central inquiry revolves around optimizing the worst-case return: which factory should you select as the destination for your inspector? Or in more general terms: where should we invest resources in reducing this uncertainty? That is, uncertainty in some states may be critical to the robust performance while in others it may be negligible. A natural approach is to look at the derivative of the robust return with respect to (w.r.t.) the uncertainty set, however, its computation may be non-trivial.

As is common in the robust MDPs literature, we consider $\mathtt{s}$ and $\mathtt{sa}$ rectangular uncertainty sets constrained by L_p norms as in [3, 8, 7, 5] and compute the gradient of robust return w.r.t. uncertainty radius. Our formulae shed light on the sensitivity of the robust return w.r.t. the uncertainty radius in different states. We also suggest a sampling algorithm to improve the robust return by reducing the uncertainty set in the parts that matter more.

2 Preliminaries

2.1 Markov Decision Processes

A Markov decision process (MDP) is a tuple $(\mathcal{S}, \mathcal{A}, \gamma, \mu, P, R)$ such that $\mathcal{S}$ and $\mathcal{A}$ are finite state and action spaces . $\gamma \in [0, 1)$ is a discount factor and $\mu \in \Delta_{\mathcal{S}}$ the initial state distribution. Denoting $\mathcal{X} := \mathcal{S} \times \mathcal{A}$, the couple (P, R) corresponds to the MDP model with $P : \mathcal{X} \to \Delta_{\mathcal{S}}$ being a transition kernel and $R : \mathcal{X} \to \mathbb{R}$ a reward function. A policy $\pi : \mathcal{S} \to \Delta_{\mathcal{A}}$ maps each state to a probability distribution over $\mathcal{A}$. For any policy π, $R^\pi \in \mathbb{R}^{\mathcal{S}}$ is the expected immediate reward defined as $R^\pi(s) := \langle \pi(\cdot|s), R(s, \cdot)\rangle_{\mathcal{A}}, \quad \forall s \in \mathcal{S}$. We similarly define the stochastic matrix induced by π as $P^\pi(s'|s) := \langle \pi(\cdot|s), P(s'|s, \cdot)\rangle_{\mathcal{A}}, \quad \forall s, s' \in \mathcal{S}$, and the policy induced occupation measure $d^\pi_P := \mu^\top (I_{\mathcal{S}} - \gamma P^\pi)^{-1}$. The performance measure we aim to maximize is the value function $v^\pi_{(P,R)} := (I_{\mathcal{S}} - \gamma P^\pi)^{-1} R^\pi$, or alternatively, the return $\rho^\pi_{(P,R)} := \langle \mu, v^\pi_{(P,R)}\rangle_{\mathcal{S}}$, which can be done using Bellman operators [13].

2.2 Robust Markov Decision Processes

In a robust MDP setting, we assume that $(P, R) \in \mathcal{U}$ and aim to maximize the return under the worst realization from the set. We denote the robust performance of a policy π by $\rho^\pi_{\mathcal{U}} := \min_{(P,R)\in\mathcal{U}} \rho^\pi_{(P,R)}$. It is optimal when it reaches $\rho^*_{\mathcal{U}} := \max_\pi \rho^\pi_{\mathcal{U}}$ at an optimal robust policy $\pi^*_{\mathcal{U}} \in \arg\max_\pi \rho^\pi_{\mathcal{U}}$. When considering the robust value function $v^\pi_{\mathcal{U}} := \min_{(P,R)\in\mathcal{U}} v^\pi_{(P,R)}$, we further need to assume that $\mathcal{U}$ is convex and rectangular so that an optimal robust policy realizing $v^*_{\mathcal{U}} := \max_\pi v^\pi_{\mathcal{U}}$ can be computed in polynomial time [17]. Specifically, we denote an $\mathtt{s}$ (resp. $\mathtt{sa}$)-rectangular uncertainty set by $\mathcal{U}^{\mathtt{s}} := \times_{s\in\mathcal{S}}(\mathcal{P}_s, \mathcal{R}_s)$ (resp. $\mathcal{U}^{\mathtt{sa}} := \times_{(s,a)\in\mathcal{X}}(\mathcal{P}_{(s,a)}, \mathcal{R}_{(s,a)})$). Similarly to non-robust MDPs, rectangular robust MDPs can be solved through Bellman recursion [6, 17].

The worst kernel $P^\pi_{\mathcal{U}}$ and worst reward function $R^\pi_{\mathcal{U}}$ under the policy π are defined as $(P^\pi_{\mathcal{U}}, R^\pi_{\mathcal{U}}) \in \arg\min_{(P,R)\in\mathcal{U}} \rho^\pi_{(P,R)}$ [12, 6]. The robust occupation measure can be defined w.r.t. worst parameters as $d^\pi_{\mathcal{U}} := d^\pi_{P^\pi_{\mathcal{U}}}$ [7].

2.2.1 Ball Constrained Uncertainty set

We consider uncertainty sets that are centered around a nominal model (P_0, R_0), i.e., of the form $\mathcal{U} = (P_0, R_0) + (\mathcal{P}, \mathcal{R})$ and constrained according to L_p-norm balls [3, 9, 5, 2]. In the (s, a)-rectangular case, the corresponding uncertainty set is denoted by $\mathcal{U}^{\mathtt{sa}}_p := \mathcal{R}^{\mathtt{sa}}_p \times \mathcal{P}^{\mathtt{sa}}_p = \times_{(s,a)\in\mathcal{X}}(\mathcal{P}_{(s,a)}, \mathcal{R}_{(s,a)})$ where for any $(s, a) \in \mathcal{X}$,

$$\mathcal{R}_{(s,a)} = \{r \in \mathbb{R} \mid |r| \le \alpha_{s,a}\}, \quad \text{and} \quad \mathcal{P}_{(s,a)} = \left\{p \in \mathbb{R}^{\mathcal{S}} \mid \langle p, \mathbf{1}\rangle_{\mathcal{S}} = 0, \|p\|_p \le \beta_{s,a}\right\}.$$

Similarly, an s-rectangular L_p uncertainty is denoted by $\mathcal{U}^{\mathtt{s}}_p := \times_{s\in\mathcal{S}}(\mathcal{P}_s, \mathcal{R}_s)$ where for any $s \in \mathcal{S}$,

$$\mathcal{R}_s = \{r \in \mathbb{R}^{\mathcal{A}} \mid \|r\|_p \le \alpha_s\}, \quad \text{and} \quad \mathcal{P}_s = \{p \in \mathbb{R}^{\mathcal{X}} \mid \langle p(\cdot, a), \mathbf{1}\rangle_{\mathcal{S}} = 0 \quad \forall a \in \mathcal{A}, \|p\|_p \le \beta_s\}.$$

3 The gradient of the robust return

In this section, we are interested in deriving the gradient of the robust return $\rho^\pi_\mathcal{U}$ with respect to the uncertainty level. To do so, we use the robust Bellman evaluation operators of [8][Thm. 1]:

$$\rho^\pi_\mathcal{U} = \sum_{s,a\in\mathcal{X}} d^\pi_{P_0}(s,a)\left(R_0(s,a) - (\alpha^\pi_{s,a} + \gamma\beta^\pi_{s,a}\kappa_q(v^\pi_\mathcal{U}))\right),$$

where $\kappa_q(v) := \min_{w\in\mathbb{R}}\|v - w\mathbf{1}\|_q, \quad \forall v \in \mathbb{R}^\mathcal{S}$, $d^\pi_P(s,a) = d^\pi_P(s)\pi(a|s)$, and q is the conjugate value of p. For s-rectangular L_p-constrained uncertainty set $\mathcal{U} = \mathcal{U}^\mathtt{s}_p$, we have:

$$\rho^\pi_\mathcal{U} = \sum_s d^\pi_{P_0}(s)\left[\, R^\pi_0(s) - \left(\,\alpha_s + \gamma\beta_s\kappa_q(v^\pi_\mathcal{U})\,\right)\|\pi_s\|_q\,\right]$$

from [8][Thm. 2]. When taking the derivative with respect to α and β on both sides, a key challenge is the dependence of the variance function κ on the robust value function, which in turn depends on α and β. The theorem below derives such gradient for sa-rectangular L_p constrained uncertainty set $\mathcal{U}^{sa}_p$.

Theorem 3.1. *For sa-rectangular L_p constrained uncertainty set $\mathcal{U} = \mathcal{U}^{sa}_p$, and any policy π, the gradient of the robust return $\rho^\pi_\mathcal{U}$ w.r.t. uncertainty radiuses α, β is given by*

$$\frac{\partial\rho^\pi_\mathcal{U}}{\partial\beta_{s,a}} = -\gamma\kappa_q(v^\pi_\mathcal{U})d^\pi_\mathcal{U}(s,a), \quad \textit{and} \quad \frac{\partial\rho^\pi_\mathcal{U}}{\partial\alpha_{s,a}} = -d^\pi_\mathcal{U}(s,a), \quad \forall s,a,$$

From the above theorem, we can make the following observations: (i) The gradient with respect to the reward radius differs from that of the kernel radius by a multiplicative constant $\gamma\kappa_q(v^\pi_\mathcal{U})$ only. Thus, more variance in the robust value function $\kappa_q(v^\pi_\mathcal{U})$ leads to a return that is more sensitive to the kernel than to the reward radius; (ii) The term $\frac{\partial\rho^\pi_\mathcal{U}}{\partial\beta}$ is proportional to the discount factor γ, so that longer horizon leads to a return that is more sensitive to kernel uncertainty than reward uncertainty. The extreme case of $\gamma = 0$ results in no sensitivity to kernel uncertainty.

The result below captures the overall sensitivity of the robust return instead of the local one.

Corollary 3.2. *The L_1 norm of the gradient robust return for sa-rectangular L_p constrained uncertainty set $\mathcal{U} = \mathcal{U}^{sa}_p$, and for any policy π, is given by*

$$(1-\gamma)\left\|\frac{\partial\rho^\pi_\mathcal{U}}{\partial\beta}\right\|_1 = \gamma\kappa_q(v^\pi_\mathcal{U}), \qquad \textit{and} \qquad (1-\gamma)\left\|\frac{\partial\rho^\pi_\mathcal{U}}{\partial\alpha}\right\|_1 = 1.$$

As the result above illustrates, the norm of the normalized robust return ($(1-\gamma)\rho^\pi_\mathcal{U}$) w.r.t. reward uncertainty is a scalar constant, and independent of the policy employed. In the case of uncertainty radius, it depends on the discount factor and the policy through the variance function.

We now generalize the result to the s-rectangular case.

Theorem 3.3. *For s-rectangular L_p constrained uncertainty set $\mathcal{U} = \mathcal{U}^s_p$, the gradient of the robust return $\rho^\pi(\mathcal{U})$ w.r.t. uncertainty radiuses α, β is given by*

$$\frac{\partial\rho^\pi(\mathcal{U})}{\partial\beta} = \gamma\kappa_q(v^\pi)\frac{\partial\rho^\pi(\mathcal{U})}{\partial\alpha} \quad \textit{and} \quad \frac{\partial\rho^\pi(\mathcal{U})}{\partial\alpha_s} = -d^\pi_\mathcal{U}(s) + \left[\,1 - \|\pi_s\|_q\,\right] d^\pi_{P_0}(s), \quad \forall s.$$

Observe the additional $\|\pi_s\|_q$ appearing here in the s-rectangular case, which establishes additional dependence of the robust return on the stochasticity of the policy. Stochastic policies lead to less variation in the return than deterministic policies. This is confirmed in the result below.

Corollary 3.4. *The L_1 norm of the gradient robust return for s-rectangular L_p constrained uncertainty set $\mathcal{U} = \mathcal{U}^s_p$, and for any policy π, is given by*

$$(1-\gamma)\left\|\frac{\partial\rho^\pi_\mathcal{U}}{\partial\beta}\right\|_1 = \gamma\kappa_q(v^\pi_\mathcal{U})\sum_s d^\pi_{P_0}(s)\|\pi_s\|_q, \quad \textit{and} \quad (1-\gamma)\left\|\frac{\partial\rho^\pi_\mathcal{U}}{\partial\alpha}\right\|_1 = \sum_s d^\pi_{P_0}(s)\|\pi_s\|_q.$$

The norm of the gradient of the robust return is proportional to $\sum_s d^\pi_{P_0}(s)\|\pi_s\|_q$, which indicates that the robust return of a stochastic policy is less sensitive to uncertainty than that of a deterministic policy.

4 Uncertainty set reduction

Given the ability to reduce your uncertainty set (e.g. using resources to improve accuracy or sampling a random variable more times) the question we aim to answer is: "Where should we put our effort?". Our previous results answer that question, provided that we seek to improve the robust return. Therefore, we establish the following example to further illustrate the benefits of using the uncertainty set radius gradients: In this problem, the true transition kernel is drawn from some distribution $\mathcal{D}$, and we try to learn the expected model $P_{goal} = \mathbb{E}_{p\sim\mathcal{D}}[p]$ but still being robust w.r.t our estimation error when we plan. We are given a finite number of samples T, and at each iteration $t \in [1:T]$, we can choose which part of the kernel we want to sample (e.g., $P(\cdot|s,a)$). By estimating P_{goal} and construct confidence intervals around it to be used as uncertainty sets, the difference between our robust return given that uncertainty set and the true return for P_{goal} is measured. Our goal is to reduce this difference as much as possible.

Algorithm 1 Sample w.r.t robust return gradient - `sa`-rectangular

Input: confidence δ
Initialize: $\forall s,a : \beta_{s,a} = \infty$
for t = $1, 2, \cdots T$ **do**
  Estimate current model: $\forall s,a : \hat{P}_t(\cdot|s,a) = \frac{1}{n_{s,a}} \sum_{i=1}^{n_{s,a}} P_i(\cdot|s,a)$
  Estimate uncertainty set (estimation accuracy): $\beta_{s,a}^t = \epsilon_{s,a}(\delta)$
  Calculate optimal robust policy $\pi^*_{\mathcal{U}_t}$ and optimal robust return $\rho^*_{\mathcal{U}_t}$ w.r.t to the uncertainty set.
  $\forall s,a$: Calculate the gradient w.r.t the uncertainty set radius $\frac{\partial \rho^*_{\mathcal{U}_t}}{\beta_{s,a}}$.
  Choose (s_t, a_t) pair for next sampling (e.g. Softmax, Hardmax of $\frac{\partial \rho^*_{\mathcal{U}_t}}{\beta_{s,a}}$).
  Sample another transition kernel $P(\cdot|s_t,a_t) \sim \mathcal{D}$
end for

4.1 Experimental results

In a tabular setting, each state-action transition kernel is sampled from a Dirichlet distribution with some unknown parameters vector $P_t(\cdot|s,a) \sim Dir(\theta_{sa}) \in \Delta_{\mathcal{S}}$. Utilizing Theorems 3.1 and 3.3 we suggest the sampling mechanism depicted in algorithm 1. To show our mechanism advantages we compare it with 2 naive sampling methods: Firstly, uniformly selecting the (s,a)-pair to sample from. Secondly, selecting the (s,a)-pair with the current biggest uncertainty set. The results can be seen in figure 1, it is clear that our method achieve a lower difference between the true optimal return and the robust return w.r.t the estimated uncertainty. In addition figure 2 shows the number of samples for each (s,a) - pair. This figure is an exemplar of the *"Not all parameters are born equal"* phenomenon - meaning that it is worth to put effort on reducing the uncertainty in some areas more than others. The same phenomenon can be seen in figure 3 where some (s,a)-pair uncertainty set does not change since it's not relevant for the robust return compared to other pairs.

5 Discussion

Our work is the first to analyze the sensitivity of the robust return with respect to the uncertainty set. It may help in designing robust planning and learning algorithms with a proper robustness level. While guided exploration is beyond the scope of this work, our study paves the path toward better exploration using our sensitivity-based formulae. One limitation of our method is that it does not directly facilitate policy optimization. Rather, policy optimization is implicitly done in most applications of interest. Interleaving policy improvement with uncertainty reduction may be very useful, as far as phasic policy iteration is concerned. Another limitation is our reliance on rectangularity. Although it is a common assumption in much of the RMDP literature, it may only serve as an approximation of reality and can by itself lead to overly conservative strategies. As such, we may consider extending our results to e.g., k-rectangular uncertainty sets [10].

Figure 1: Optimal robust return w.r.t uncertainty in each iteration. In red - the optimal robust return for the true expected model

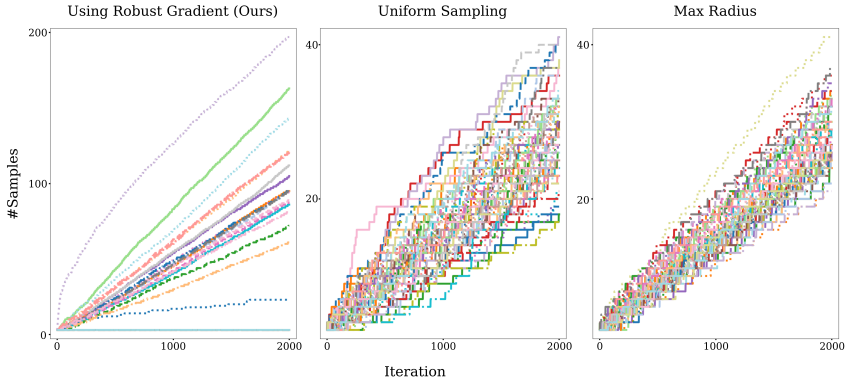

Figure 2: Number of sample at each time-step. Each line represent different (s, a)-pair

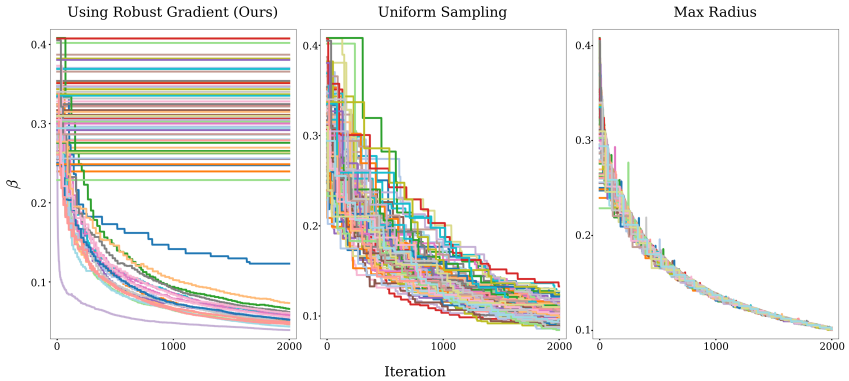

Figure 3: Size of uncertainty at each time-step. Each line represent different (s, a)-pair

Acknowledgments

This Project has received funding from the European Union's Horizon Europe Programme under grant agreement No.101070568.

References

[1] J. Andrew Bagnell, Andrew Y. Ng, and Jeff G. Schneider. Solving uncertain markov decision processes. Technical report, Carnegie Mellon University, 2001.

[2] Bahram Behzadian, Marek Petrik, and Chin Pang Ho. Fast algorithms for l_∞-constrained s-rectangular robust MDPs. *Advances in Neural Information Processing Systems*, 34:25982–25992, 2021.

[3] Esther Derman, Matthieu Geist, and Shie Mannor. Twice regularized mdps and the equivalence between robustness and regularization. *Advances in Neural Information Processing Systems*, 34:22274–22287, 2021.

[4] Esther Derman, Daniel J. Mankowitz, Timothy A. Mann, and Shie Mannor. Soft-robust actor-critic policy-gradient. *AUAI press for Association for Uncertainty in Artificial Intelligence*, pages 208–218, 2018.

[5] Chin Pang Ho, Marek Petrik, and Wolfram Wiesemann. Partial policy iteration for l_1-robust Markov decision processes. *J. Mach. Learn. Res.*, 22:275–1, 2021.

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

[7] Navdeep Kumar, Esther Derman, Matthieu Geist, Kfir Levy, and Shie Mannor. Policy gradient for s-rectangular robust markov decision processes, 2023.

[8] 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.

[9] Navdeep Kumar, Kfir Levy, Kaixin Wang, and Shie Mannor. Efficient policy iteration for robust markov decision processes via regularization, 2022.

[10] Shie Mannor, Ofir Mebel, and Huan Xu. Lightning does not strike twice: Robust MDPs with coupled uncertainty. *International Conference on Machine Learning*, 2012.

[11] Shie Mannor, Duncan Simester, Peng Sun, and John N. Tsitsiklis. Bias and variance in value function estimation. In *Proceedings of the Twenty-First International Conference on Machine Learning*, ICML '04, page 72, New York, NY, USA, 2004. Association for Computing Machinery.

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

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

[14] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, second edition, 2018.

[15] Yue Wang and Shaofeng Zou. Online robust reinforcement learning with model uncertainty, 2021.

[16] Yue Wang and Shaofeng Zou. Policy gradient method for robust reinforcement learning, 2022.

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

[18] Huan Xu and Shie Mannor. Robustness and generalization. *Machine learning*, 86:391–423, 2012.

A Helper results

We present here a simple mathematical result which is crucially used in deriving the main technical result of the paper, that is, gradient of robust return.

Proposition A.1. *For $u, d, k, f \in \mathbb{R}^{\mathcal{S}}, c \in \mathbb{R}$, the solution to*

$$u = -cd - \gamma\langle k, u\rangle f \tag{1}$$

is given by

$$u = -cd + \frac{c\gamma\langle k, d\rangle}{1 + \gamma\langle k, f\rangle} f \tag{2}$$

$$\tag{3}$$

Proof. We have

$$u = -cd - \gamma\langle k, u\rangle f. \tag{4}$$

Now taking both sides dot product with k, we get

$$\langle k, u\rangle = -c\langle k, d\rangle - \gamma\langle k, u\rangle\langle k, f\rangle \tag{5}$$

$$= -\frac{c\langle k, d\rangle}{1 + \gamma\langle k, f\rangle}. \tag{6}$$

$$\tag{7}$$

Putting it back in the original equation, we get

$$u = -cd - \gamma\langle k, u\rangle f \tag{8}$$

$$= -cd + \frac{c\gamma\langle k, d\rangle}{1 + \gamma\langle k, f\rangle} f \tag{9}$$

$$\tag{10}$$

□

B Gradient of robust return

Here we prove main results of the paper, that is, we derive the gradient of the robust return w.r.t. uncertainty radiuses. We begin with sa-rectangular case.

B.1 sa-rectangular Uncertainty

Proof. (**Theorem 3.1**) Recall that $v^{\pi}_{\mathcal{U}^{\texttt{sa}}_p}$ is the fixed point of the robust Bellman operator $\mathcal{T}^{\pi}_{\mathcal{U}^{\texttt{sa}}_p}$, so from proposition **??** (Theorem 3.1 of [9]), we have

$$v^{\pi}_{\mathcal{U}^{\texttt{sa}}_p}(s) = \sum_a \pi(a|s) \Big[-\alpha_{s,a} - \gamma\beta_{s,a}\kappa_q(v^{\pi}_{\mathcal{U}^{\texttt{sa}}_p}) + R_0(s,a) + \gamma \sum_{s'} P_0(s'|s,a) v^{\pi}_{\mathcal{U}^{\texttt{sa}}_p}(s') \Big] .$$

Using compact notations $\alpha^{\pi}(s) := \sum_a \pi(a|s)\alpha_{s,a}$, $\beta^{\pi}(s) := \sum_a \pi(a|s)\beta_{s,a}$, $P^{\pi}_0(s'|s) := \sum_a \pi(a|s)P_0(s'|s,a)$, we have

$$\begin{aligned}
& v^{\pi}_{\mathcal{U}^{\texttt{sa}}_p} = -\alpha^{\pi} - \gamma\beta^{\pi}\kappa_q(v^{\pi}_{\mathcal{U}^{\texttt{sa}}_p}) + R^{\pi}_0 + \gamma P^{\pi}_0 v^{\pi}_{\mathcal{U}^{\texttt{sa}}_p}, \\
\implies & (I - \gamma P^{\pi}_0) v^{\pi}_{\mathcal{U}^{\texttt{sa}}_p} = -\alpha^{\pi} - \gamma\beta^{\pi}\kappa_q(v^{\pi}_{\mathcal{U}^{\texttt{sa}}_p}) + R^{\pi}_0, \\
\implies & v^{\pi}_{\mathcal{U}^{\texttt{sa}}_p} = \Big(I - \gamma P^{\pi}_0 \Big)^{-1} \Big(-\alpha^{\pi} - \gamma\beta^{\pi}\kappa_q(v^{\pi}_{\mathcal{U}^{\texttt{sa}}_p}) + R^{\pi}_0 \Big), \\
\implies & v^{\pi}_{\mathcal{U}^{\texttt{sa}}_p}(\hat{s}) = \sum_{s,a} d^{\pi}_{P_0,\hat{s}}(s,a) \Big[R_0(s,a) - \alpha_{s,a} - \gamma\beta_{s,a}\kappa_q(v^{\pi}_{\mathcal{U}^{sa}_p}) \Big],
\end{aligned}$$

where $d^\pi(s,a)$ is shorthand for $d^\pi(s)\pi(a|s)$. We wish to take derivative of robust value function w.r.t. uncertainty radiuses, using the above expression. However, the main challenge is that the robust value function is present on the both sides of the equation, particularly inside variance term κ. We mitigate this issue in the case by case.

Reward derivative. Now taking the derivative both sides, we have

$$\frac{\partial v^\pi_{\mathcal{U}^{\text{sa}}_p}(\hat{s})}{\partial \alpha_{s,a}} = \frac{\partial}{\partial \alpha_{s,a}} \sum_{s',a'} d^\pi_{P_0,\hat{s}}(s',a') \Big[R_0(s',a') - \alpha_{s',a'} - \gamma \beta_{s',a'} \kappa_q(v^\pi_{\mathcal{U}^{sa}_p}) \Big], \tag{11}$$

$$\text{(The first term has no dependence on } \alpha_{s,a}) \tag{12}$$

$$= -\sum_{s',a'} d^\pi_{P_0,\hat{s}}(s',a') \frac{\partial}{\partial \alpha_{s,a}} \Big[\alpha_{s',a'} + \gamma \beta_{s',a'} \kappa_q(v^\pi_{\mathcal{U}^{sa}_p}) \Big], \tag{13}$$

$$= -d^\pi_{P_0,\hat{s}}(s,a) - \sum_{s',a'} \gamma d^\pi_{P_0,\hat{s}}(s',a') \beta_{s',a'} \frac{\partial \kappa_q(v^\pi_{\mathcal{U}^{sa}_p})}{\partial \alpha_{s,a}}, \tag{14}$$

$$= -d^\pi_{P_0,\hat{s}}(s,a) - \sum_{s',a'} \gamma d^\pi_{P_0,\hat{s}}(s',a') \beta_{s',a'} \sum_{s"} \frac{\partial \kappa_q(v^\pi_{\mathcal{U}^{sa}_p})}{\partial v^\pi_{\mathcal{U}}(s")} \frac{\partial v^\pi_{\mathcal{U}^{sa}_p}(s")}{\partial \alpha_{s,a}}. \tag{15}$$

$$\tag{16}$$

Kernel derivative. Similarly, we have

$$\frac{\partial v^\pi_{\mathcal{U}^{\text{sa}}_p}(\hat{s})}{\partial \beta_{s,a}} = \frac{\partial}{\partial \beta_{s,a}} \sum_{s,a} d^\pi_{P_0,\hat{s}}(s,a) \Big[R_0(s,a) - \alpha_{s,a} - \gamma \beta_{s,a} \kappa_q(v^\pi_{\mathcal{U}^{sa}_p}) \Big], \tag{17}$$

$$= -\frac{\partial}{\partial \beta_{s,a}} \Big[\gamma \sum_{s,a} d^\pi_{P_0,\hat{s}}(s,a) \beta_{s,a} \kappa_q(v^\pi_{\mathcal{U}^{sa}_p}) \Big], \tag{18}$$

$$= -\gamma d^\pi_{P_0,\hat{s}}(s,a) \kappa_q(v^\pi_{\mathcal{U}^{sa}_p}) - \Big(\sum_{s',a'} \gamma d^\pi_{P_0,\hat{s}}(s',a') \beta_{s',a'} \Big) \frac{\partial \kappa_q(v^\pi_{\mathcal{U}})}{\partial \beta_{s,a}}, \tag{19}$$

$$= -\gamma d^\pi_{P_0,\hat{s}}(s,a) \kappa_q(v^\pi_{\mathcal{U}^{sa}_p}) - \Big(\sum_{s',a'} \gamma d^\pi_{P_0,\hat{s}}(s',a') \beta_{s',a'} \Big) \sum_{s"} \frac{\partial \kappa_q(v^\pi_{\mathcal{U}})}{\partial v^\pi_{\mathcal{U}^{sa}_p}(s")} \frac{\partial v^\pi_{\mathcal{U}^{sa}_p}(s")}{\partial \beta_{s,a}}. \tag{20}$$

Note that we have derivative of value function w.r.t. uncertainty on the both sides of the equations. To summarize, we have

$$\frac{\partial v^\pi_{\mathcal{U}^{sa}_p}(\hat{s})}{\partial \alpha_{s,a}} = -d^\pi_{P_0,\hat{s}}(s,a) - \Big(\sum_{s',a'} \gamma d^\pi_{P_0,\hat{s}}(s',a') \beta_{s',a'} \Big) \Big(\sum_{s"} \frac{\partial \kappa_q(v^\pi_{\mathcal{U}^{sa}_p})}{\partial v^\pi_{\mathcal{U}}(s")} \frac{\partial v^\pi_{\mathcal{U}^{sa}_p}(s")}{\partial \alpha_{s,a}} \Big),$$

$$\frac{\partial v^\pi_{\mathcal{U}^{sa}_p}(\hat{s})}{\partial \beta_{s,a}} = -\gamma d^\pi_{P_0,\hat{s}}(s,a) \kappa_q(v^\pi_{\mathcal{U}^{sa}_p}) - \Big(\sum_{s',a'} \gamma d^\pi_{P_0,\hat{s}}(s',a') \beta_{s',a'} \Big) \Big(\sum_{s"} \frac{\partial \kappa_q(v^\pi_{\mathcal{U}})}{\partial v^\pi_{\mathcal{U}^{sa}_p}(s")} \frac{\partial v^\pi_{\mathcal{U}^{sa}_p}(s")}{\partial \beta_{s,a}} \Big).$$

For fixed s, a, the above equations are of the form

$$u = -cd - \gamma \langle k, u \rangle f. \tag{21}$$

To illustrate, in the last equation (gradient w.r.t. β), we can put $u(\hat{s}) = \frac{\partial v^\pi_{\mathcal{U}}(\hat{s})}{\partial \beta_{s,a}}$, $d(\hat{s}) := d^\pi_{\hat{s}}(s,a)$, $f(\hat{s}) := \sum_{s',a'} d_{\hat{s}}(s',a') \beta_{s',a'}$, $k(\hat{s}) = \frac{\partial \kappa_q(v^\pi_{\mathcal{U}})}{\partial v^\pi_{\mathcal{U}}(\hat{s})}$ and $c = \gamma \kappa_q(v^\pi_{\mathcal{U}^{sa}_p})$. We get the solution to the above equation, using proposition A.1 as

$$u = -cd + \frac{c\gamma \langle k, d \rangle}{1 + \gamma \langle k, f \rangle} f. \tag{22}$$

Using this result, we get

$$\frac{\partial v^\pi_{\mathcal{U}}(\hat{s})}{\partial \alpha_{s,a}} = -d^\pi_{\hat{s}}(s,a) + \gamma \Big(\frac{\sum_{s"} \frac{\partial \kappa_q(v^\pi_{\mathcal{U}})}{\partial v^\pi_{\mathcal{U}}(s")} d^\pi_{s"}(s,a)}{1 + \gamma \sum_{s"} \frac{\partial \kappa_q(v^\pi_{\mathcal{U}})}{\partial v^\pi_{\mathcal{U}}(s")} \sum_{s',a'} d^\pi_{s"}(s',a') \beta_{s',a'}} \Big) \Big(\sum_{s',a'} d^\pi_{\hat{s}}(s',a') \beta_{s',a'} \Big).$$

Taking dot product with initial distribution μ, we get

$$\frac{\partial \rho^\pi_{\mathcal{U}}}{\partial \alpha_{s,a}} = -d^\pi_\mu(s,a) + \gamma \frac{\sum_{s"} \frac{\partial \kappa_q(v^\pi_{\mathcal{U}})}{\partial v^\pi_{\mathcal{U}}(s")} d^\pi_{s"}(s,a)}{1+\gamma \sum_{s",s',a'} \frac{\partial \kappa_q(v^\pi_{\mathcal{U}})}{\partial v^\pi_{\mathcal{U}}(s")} d^\pi_{s"}(s',a')\beta_{s',a'}} \sum_{\hat{s},s',a'} \mu(\hat{s}) d^\pi_{\hat{s}}(s',a')\beta_{s',a'} \quad (23)$$

$$= -d^\pi_\mu(s,a) + \gamma \frac{\sum_{s"} \frac{\partial \kappa_q(v^\pi_{\mathcal{U}})}{\partial v^\pi_{\mathcal{U}}(s")} d^\pi_{s"}(s,a)}{1+\gamma \sum_{s",s',a'} \frac{\partial \kappa_q(v^\pi_{\mathcal{U}})}{\partial v^\pi_{\mathcal{U}}(s")} d^\pi_{s"}(s',a')\beta_{s',a'}} \langle d^\pi_\mu, \beta \rangle. \quad (24)$$

Similarly, we have

$$\frac{\partial \rho^\pi(\mathcal{U})}{\partial \beta_{s,a}} = \gamma \kappa_q(v^\pi_{\mathcal{U}}) \left[-d^\pi_\mu(s,a) + \gamma \frac{\sum_{s"} \frac{\partial \kappa_q(v^\pi_{\mathcal{U}})}{\partial v^\pi_{\mathcal{U}}(s")} d^\pi_{s"}(s,a)}{1+\gamma \sum_{s",s',a'} \frac{\partial \kappa_q(v^\pi_{\mathcal{U}})}{\partial v^\pi_{\mathcal{U}}(s")} d^\pi_{s"}(s',a')\beta_{s',a'}} \langle d^\pi_\mu, \beta \rangle \right]. \quad (25)$$

We get the desired results by using the fact: $u^\pi_{\mathcal{U}} = \nabla_v \kappa_q(v) \Big|_{v=v^\pi_{\mathcal{U}}}$ from [7] and expression of robust occupation measure [7] □

Special Cases. It is worth noting that for sa-rectangular L_1 constrained set $\mathcal{U}^{\mathtt{sa}}_1$, we have the following simplified expressions

$$\frac{\partial \rho^\pi(\mathcal{U}^{sa}_1)}{\partial \alpha_{s,a}} = -d^\pi_\mu(s,a) + \gamma \langle d^\pi_\mu, \beta \rangle \frac{d^\pi_{s_{\max}}(s,a) - d^\pi_{s_{\min}}(s,a)}{1+\gamma \langle d^\pi_{s_{\max}} - d^\pi_{s_{\min}}, \beta \rangle} \quad (26)$$

$$\frac{\partial \rho^\pi(\mathcal{U}^{sa}_1)}{\partial \beta_{s,a}} = \gamma \kappa_q(v^\pi_{\mathcal{U}^{sa}_1}) \left[-d^\pi_\mu(s,a) + \gamma \langle d^\pi_\mu, \beta \rangle \frac{d^\pi_{s_{\max}}(s,a) - d^\pi_{s_{\min}}(s,a)}{1+\gamma \langle d^\pi_{s_{\max}} - d^\pi_{s_{\min}}, \beta \rangle} \right]. \quad (27)$$

$$(28)$$

where $s_{max} \in arg\max_s v^\pi_{\mathcal{U}^{\mathtt{sa}}_1}(s)$, $s_{min} \in arg\min_s v^\pi_{\mathcal{U}^{\mathtt{sa}}_1}(s)$.

To get $\frac{\partial \rho^\pi(\mathcal{U})}{\partial \alpha_{s,a}}$, $\frac{\partial \rho^\pi(\mathcal{U})}{\partial \beta_{s,a}}$, we just need to replace π with optimal robust policy $\pi^*_{\mathcal{U}}$ in the above expressions.

B.2 s-rectangular Uncertainty

In this, we derive the gradient for s-rectangular case. We techniques used are very similar to sa-case. Nonetheless, we derive it in detail for sake of completeness.

Proof. (of Theorem 3.3) Precisely, we assume the uncertainty set $\mathcal{U} = \mathcal{U}^{\mathtt{sa}}_p$ is s-rectangular L_p constrained. From the above discussion, we have

$$v^\pi_{\mathcal{U}}(\hat{s}) = \sum_s d^\pi_{P_0,\hat{s}}(s) \left[R^\pi_0(s) - \left(\alpha_s + \gamma \beta_s \kappa_q(v^\pi_{\mathcal{U}}) \right) \|\pi_s\|_q \right]. \quad (29)$$

Now, taking the derivative $\rho^\pi(\mathcal{U})$ w.r.t. α_s, we get

$$\frac{\partial v^\pi_{\mathcal{U}}(\hat{s})}{\partial \alpha_s} = -\frac{\partial}{\partial \alpha_s} \sum_{s'} d^\pi_{\hat{s}}(s') \left(\alpha_{s'} + \gamma \beta_{s'} \kappa_q(v^\pi_{\mathcal{U}}) \right) \|\pi_{s'}\|_q \quad (30)$$

$$= -d^\pi_{\hat{s}}(s)\|\pi_s\|_q - \gamma \left[\sum_{s'} d^\pi_{\hat{s}}(s')\beta_{s'}\|\pi_{s'}\|_q \right] \frac{\partial}{\partial \alpha_s} \kappa_q(v^\pi_{\mathcal{U}}) \quad (31)$$

$$= -d^\pi_{\hat{s}}(s)\|\pi_s\|_q - \gamma \left[\sum_{s'} d^\pi_{\hat{s}}(s')\beta_{s'}\|\pi_{s'}\|_q \right] \sum_{s"} \frac{\partial \kappa_q(v^\pi_{\mathcal{U}})}{\partial v^\pi_{\mathcal{U}}(s")} \frac{\partial v^\pi_{\mathcal{U}}(s")}{\partial \alpha_s}. \quad (32)$$

Similarly, we have

$$\frac{\partial v_{\mathcal{U}}^{\pi}(\hat{s})}{\partial \beta_s} = -\gamma d_{\hat{s}}^{\pi}(s)\kappa_q(v_{\mathcal{U}}^{\pi})\|\pi_s\|_q - \Big[\sum_{s'} d_{\hat{s}}^{\pi}(s')\beta_{s'}\|\pi_{s'}\|_q\Big]\sum_{s"}\frac{\partial \kappa_q(v_{\mathcal{U}}^{\pi})}{\partial v_{\mathcal{U}}^{\pi}(s")}\frac{\partial v_{\mathcal{U}}^{\pi}(s")}{\partial \beta_s}. \tag{33}$$

Again observe that the above equations are of the following form,

$$u = ce + \gamma\langle k, u\rangle f \tag{34}$$

where $u = \frac{\partial v_{\mathcal{U}}^{\pi}}{\partial \cdot_s}, e(s') := d_s^{\pi}(s')\|\pi_{s'}\|_q, f(s') := \sum_{s"} d_{s'}(s)\beta_{s'}\|\pi_{s'}\|_q, k(s) = \frac{\partial \kappa_q(v_{\mathcal{U}}^{\pi})}{\partial v_{\mathcal{U}}^{\pi}(s)}$ and c is a constant. Using its solution, we get

$$\frac{\partial v_{\mathcal{U}}^{\pi}(\hat{s})}{\partial \alpha_s} = d_{\hat{s}}^{\pi}(s)\|\pi_s\|_q + \sum_{s'}\gamma d_{\hat{s}}^{\pi}(s')\beta_{s'}\|\pi_{s'}\|_q\sum_{s"}\frac{\partial \kappa_q(v_{\mathcal{U}}^{\pi})}{\partial v_{\mathcal{U}}^{\pi}(s")}\frac{\partial v_{\mathcal{U}}^{\pi}(s")}{\partial \alpha_s}, \tag{35}$$

$$= -d_{\hat{s}}^{\pi}(s)\|\pi_s\|_q - \gamma\frac{\sum_{s"}\frac{\partial \kappa_q(v_{\mathcal{U}}^{\pi})}{\partial v_{\mathcal{U}}^{\pi}(s")}d_{s"}^{\pi}(s)\|\pi_{s"}\|_q}{1-\gamma\sum_{s"}\frac{\partial \kappa_q(v_{\mathcal{U}}^{\pi})}{\partial v_{\mathcal{U}}^{\pi}(s")}\sum_{s'} d_{s"}^{\pi}(s')\beta_{s'}\|\pi_{s'}\|_q}\sum_{s'} d_{\hat{s}}^{\pi}(s')\beta_{s'}\|\pi_{s'}\|_q. \tag{36}$$

$$\tag{37}$$

Taking dot product with μ the both sides, we get

$$\frac{\partial \rho^{\pi}(\mathcal{U})}{\partial \alpha_s} = -d_{\mu}^{\pi}(s)\|\pi_s\|_q - \gamma\frac{\sum_{s"}\frac{\partial \kappa_q(v_{\mathcal{U}}^{\pi})}{\partial v_{\mathcal{U}}^{\pi}(s")}d_{s"}^{\pi}(s)\|\pi_{s"}\|_q}{1-\gamma\sum_{s"}\frac{\partial \kappa_q(v_{\mathcal{U}}^{\pi})}{\partial v_{\mathcal{U}}^{\pi}(s")}\sum_{s'} d_{s"}^{\pi}(s')\beta_{s'}\|\pi_{s'}\|_q}\sum_{s'} d_{\mu}^{\pi}(s')\beta_{s'}\|\pi_{s'}\|_q. \tag{38}$$

Similarly, we have

$$\frac{\partial \rho^{\pi}(\mathcal{U})}{\partial \beta_s} = -\gamma\kappa_q(v_{\mathcal{U}}^{\pi})\Big[\, d_{\mu}^{\pi}(s)\|\pi_s\|_q \tag{39}$$

$$+\gamma\frac{\sum_{s"}\frac{\partial \kappa_q(v_{\mathcal{U}}^{\pi})}{\partial v_{\mathcal{U}}^{\pi}(s")}d_{s"}^{\pi}(s)\|\pi_{s"}\|_q}{1-\gamma\sum_{s"}\frac{\partial \kappa_q(v_{\mathcal{U}}^{\pi})}{\partial v_{\mathcal{U}}^{\pi}(s")}\sum_{s'} d_{s"}^{\pi}(s')\beta_{s'}\|\pi_{s'}\|_q}\sum_{s'} d_{\mu}^{\pi}(s')\beta_{s'}\|\pi_{s'}\|_q\Big] \tag{40}$$

$$= \gamma\kappa_q(v_{\mathcal{U}}^{\pi})\frac{\partial \rho^{\pi}(\mathcal{U})}{\partial \alpha_s}. \tag{41}$$

Again, we get the desired results by using the fact: $u_{\mathcal{U}}^{\pi} = \nabla_v \kappa_q(v)\Big|_{v=v_{\mathcal{U}}^{\pi}}$ from [7] and expression of robust occupation measure [7].

Observe that here we have factor $\|\pi_s\|_q$ instead of $\pi(a|s)$ as compared to `sa`-case.

□