LEARNING THE UNCERTAINTY SET IN ROBUST MARKOV DECISION PROCESS

Navdeep Kumar
Technion
navdeepkumar@alum.iisc.ac.in

Kaixin Wang
Technion
kaixin.wang@u.nus.edu

Uri Gadot
Technion
uri.gad@campus.technion.ac.il

Kfir Levy
Technion
kfirylevy@technion.ac.il

Shie Mannor
Technion
shie@ee.technion.ac.il

ABSTRACT

In robust Markov Decision Processes (MDPs), the uncertainty set is often assumed to be fixed and given. However, the size of the uncertainty set is crucial due to the inherent trade-off between robustness and conservatives: a larger uncertainty set fosters a more robust solution but tends towards increased conservativeness, while a smaller set may sacrifice robustness for higher performance. In this work, we introduce a novel method to learn the size of reward uncertainty set from data. Such a data-driven approach ensures that the learned uncertainty set is large enough to cover the underlying models implied by the data while being compact to minimize conservativeness.

1 INTRODUCTION

Robust reinforcement learning is a tool to tackle decision-making problems where the system parameters are uncertain or partially known (Nilim & El Ghaoui, 2005; Iyengar, 2005; Mannor et al., 2004). There are many works on solving the robust Markov Decision Processes (MDPs) for specified uncertainty set (Wolfram Wiesemann, 2012; Tamar et al., 2014; Ho et al., 2020; Derman et al., 2021; Abdullah et al., 2019; Wang & Zou, 2021; 2022; Kumar et al., 2022a; 2023a; Gadot et al., 2023; Kumar et al., 2023b; Wang et al., 2023; 2022). However, the performance-robustness trade-off receives less attention. That is, excessively big uncertainty leads to overly conservative solutions that have very sub-optimal performance. Conversely, overly restricted uncertainty sets can result in less robust solutions, vulnerable to changes in the environment. Thus, striking a careful balance when formulating assumptions about the uncertainty set becomes pivotal for achieving optimal performance.

In this work, we focus on learning reward uncertainty sets in a data-driven way, instead of manually specifying a fixed uncertainty set. More specifically, suppose we have a dataset of transitions sampled from several reward models within an unknown uncertainty set, we aim to learn a minimal radius of the uncertainty set that covers all these models. Such a minimal radius would give us a nice balance between robustness and conservativeness.

2 METHOD

A Markov decision process (MDP) is a tuple $(\mathcal{S}, \mathcal{A}, P, R, \mu, \gamma)$ such that $\mathcal{S}$, $\mathcal{A}, P : \mathcal{S} \times \mathcal{A} \to \Delta_{\mathcal{S}}, R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}, \mu \in \Delta_{\mathcal{S}}, \gamma \in [0, 1), \Delta_{\mathcal{X}}$ are state space, action space, transition kernel, reward function, an initial distribution over states, discount factor ensuring that the infinite-horizon return is well-defined, and probability simplex over the set $\mathcal{X}$ respectively (Sutton & Barto, 2018).

The policy $\pi \in (\Delta_A)^{\mathcal{S}}$ maps states to action, where $\pi(a|s)$ is the probability of playing action a in state s. The return ρ^π_R of a policy π and reward function R, is defined as $\rho^\pi_{(P,R)} = \langle R, d^\pi \rangle$ where $d^\pi(s,a) := \mathbb{E}[\sum_{n=0}^{\infty} \gamma^n \mathbf{1}(s_n = s, a_n = a) \mid s_0 \sim \mu, \pi, P]$ is the occupation measure associated with policy π (Puterman, 2014).

In most cases, the system parameters are not known exactly, but up to an uncertainty set, hence the robust return w.r.t. uncertainty set $\mathcal{R}$ under policy π is defined as $\rho^\pi_{\mathcal{R}} = \min_{R \in \mathcal{R}} \rho^\pi_R$ (Gadot et al., 2023). There exists a wide range of literature on solving robust MDPs given the uncertainty set (Gadot et al., 2023; Derman et al., 2021; Kumar et al., 2022a; 2023a), However, the uncertainty may not be available to us, making those approaches inapplicable.

We assume that there exists a true but unknown uncertainty set $\mathcal{R}$. We only have access to the trajectories $\{s_t, a_t, R(s_t, a_t) \mid P, \pi, \mu\}_{t=0}^{\infty}$ from different reward models $R \in \mathcal{R}$. The result below states that the uncertainty set $\mathcal{R}$ can be estimated from the occupation measure and returns from the different models. Note that the returns and the occupation measure are easily estimated from the trajectories.

Theorem 1. *(Radius of Ball uncertainty set) Let $B(R_0, \alpha) = \{R \mid \|R - R_0\|_p \leq \alpha\}$, then for every policy π, we have*

$$\alpha = \max_{R, R' \in B(R_0, \alpha)} \frac{|\rho^\pi_R - \rho^\pi_{R'}|}{2\|d^\pi\|_q}.$$

The occupation measure can be bootstrapped using the γ- contraction operator (in L_1 norm) $\mathcal{L}$ defined as (Kumar et al., 2022b)

$$(\mathcal{L}^\pi d)(s) = \mu(s) + \sum_{s'} P^\pi(s'|s) d(s),$$

with fixed point d^π. Here we present a sample-based method to learn the uncertainty radius.

Algorithm 1 Sample-based learning of uncertainty radius

Input: Sample trajectories $\{s^i_n, a^i_n, R_i(s^i_n, a^i_n) \mid \pi, P\}_{n=0}^{\infty}$ for different reward functions $R_i \in \mathcal{R}$. Learning rate η^i_n schedule.

while not converged **do**

  For all i, update the occupancy measure: $d(s^i_n) = d(s^i_n) + \eta^i_n \left[d_0(s^i_n) + \gamma d(s^i_{n+1}) - d(s^i_n)\right]$

  Keep track of highest and lowest return $\rho_i = \sum_{n=0}^{\infty} \gamma^n R^i(s^i_n, a^i_n)$.

  Compute $\alpha = \frac{\max_i \rho_i - \min_j \rho_j}{2\|d\|_q}$.

end while

Once the uncertainty set is learned (*i.e.*, the radius), we can employ the robust method to learn the robust optimal policy (Gadot et al., 2023). Moreover, it may also be possible to combine both *learning the uncertainty set* and *solving for robust optimal policy* together, in a single efficient algorithm.

3 Discussions

We believe learning the uncertainty set holds promise across various real-world applications. For instance, in robotics, learning the uncertainty set from past interaction data allows agents to discern areas within the environment that are more susceptible to disturbances. This knowledge enables training robust control policies with greater efficiency, mitigating the issues of overly conservative approaches.

Consider another example in autonomous driving. Each driver may possess distinct preferences regarding comfortable driving behaviors. By representing these differences as an uncertainty set and learning this uncertainty set from data, we can develop a driving policy that is not only safe and robust but also tailored to various driving preferences. This approach ensures an efficient yet secure driving experience for diverse drivers.

URM Statement

The authors acknowledge that at least one key author of this work meets the URM criteria of ICLR 2024 Tiny Papers Track.

References

Mohammed Amin Abdullah, Hang Ren, Haitham Bou Ammar, Vladimir Milenkovic, Rui Luo, Mingtian Zhang, and Jun Wang. Wasserstein robust reinforcement learning, 2019. URL https://arxiv.org/abs/1907.13196.

Esther Derman, Matthieu Geist, and Shie Mannor. Twice regularized mdps and the equivalence between robustness and regularization, 2021.

Uri Gadot, Esther Derman, Navdeep Kumar, Maxence Mohamed Elfatihi, Kfir Levy, and Shie Mannor. Solving non-rectangular reward-robust mdps via frequency regularization, 2023.

Chin Pang Ho, Marek Petrik, and Wolfram Wiesemann. Partial policy iteration for l1-robust markov decision processes, 2020. URL https://arxiv.org/abs/2006.09484.

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

Navdeep Kumar, Kfir Levy, Kaixin Wang, and Shie Mannor. Efficient policy iteration for robust markov decision processes via regularization, 2022a. URL https://arxiv.org/abs/2205.14327.

Navdeep Kumar, Kaixin Wang, Kfir Levy, and Shie Mannor. Policy gradient for reinforcement learning with general utilities, 2022b. URL https://arxiv.org/abs/2210.00991.

Navdeep Kumar, Esther Derman, Matthieu Geist, Kfir Levy, and Shie Mannor. Policy gradient for s-rectangular robust markov decision processes, 2023a. URL https://arxiv.org/abs/2301.13589.

Navdeep Kumar, Ilnura Usmanova, Kfir Yehuda Levy, and Shie Mannor. Towards faster global convergence of robust policy gradient methods. In *Sixteenth European Workshop on Reinforcement Learning*, 2023b. URL https://openreview.net/forum?id=cWrwdbEBx5.

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, pp. 72, New York, NY, USA, 2004. Association for Computing Machinery. ISBN 1581138385. doi: 10.1145/1015330.1015402. URL https://doi.org/10.1145/1015330.1015402.

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

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

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, second edition, 2018. URL http://incompleteideas.net/book/the-book-2nd.html.

Aviv Tamar, Shie Mannor, and Huan Xu. Scaling up robust mdps using function approximation. In Eric P. Xing and Tony Jebara (eds.), *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pp. 181–189, Bejing, China, 22–24 Jun 2014. PMLR. URL https://proceedings.mlr.press/v32/tamar14.html.

Kaixin Wang, Navdeep Kumar, Kuangqi Zhou, Bryan Hooi, Jiashi Feng, and Shie Mannor. The geometry of robust value functions. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 22727–22751. PMLR, 17–23 Jul 2022. URL `https://proceedings.mlr.press/v162/wang22k.html`.

Qiuhao Wang, Chin Pang Ho, and Marek Petrik. Policy gradient in robust mdps with global convergence guarantee. *Proceedings of the 40th International Conference on Machine Learning, PMLR 202:35763-35797*, 2023.

Yue Wang and Shaofeng Zou. Online robust reinforcement learning with model uncertainty, 2021. URL `https://arxiv.org/abs/2109.14523`.

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

Berç Rustem Wolfram Wiesemann, Daniel Kuhn. Robust markov decision processes. *Mathematics of Operations Research 38(1):153-183*, 2012.

A Appendix

A.1 Reward Uncertainty

Lemma. *(Radius of Ball uncertainty set) Let $B(R_0, \alpha) = \{R \mid \|R - R_0\|_p \leq \alpha\}$, then for every policy π, we have*

$$\alpha = \max_{R,R' \in B(R_0,\alpha)} \frac{|\rho^\pi_R - \rho^\pi_{R'}|}{2\|d^\pi\|_q}.$$

Proof. Then from the previous result, we have

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

Similarly, it is easy to see,

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

Adding both, we get

$$\max_{R\in\mathcal{R}} \rho^\pi_R - \min_{R\in\mathcal{R}} \rho^\pi_R = 2\alpha\|d^\pi\|_q.$$

This implies

$$\max_{R,R'\in\mathcal{R}} |\rho^\pi_R - \rho^\pi_{R'}| = \max_{R\in\mathcal{R}} \rho^\pi_R - \min_{R\in\mathcal{R}} \rho^\pi_R = 2\alpha\|d^\pi\|_q, \qquad \forall R, R' \in \mathcal{R}.$$

This proves the desired claim.

□

Lemma 1. *(Uncertainty Radius Lower Bound) The L_p radius τ_p is lower bounded as*

$$\tau_p(\mathcal{R}) \geq \max_\pi \max_{R,R'\in\mathcal{R}} \frac{|\rho^\pi_R - \rho^\pi_{R'}|}{2\|d^\pi\|_q}.$$

Let L_p radius $\tau_p(\mathcal{R})$ of the uncertainty set $\mathcal{R}$, be defined as

$$\tau_p(\mathcal{R}) := \frac{1}{2} \max_{R,R'\in\mathcal{R}} \|R - R'\|_p.$$

Proof. From the above lemma, we have

$$\tau_p(\mathcal{R}) = \max_{R,R'\in B(R_0,\tau_p(\mathcal{R}))} \frac{|\rho^\pi_R - \rho^\pi_{R'}|}{2\|d^\pi\|_q}, \qquad \forall\pi, \qquad (R_0 \text{ is such that } \mathcal{R} \subset B(R_0, \tau_p(\mathcal{R}))), \quad (1)$$

$$\geq \max_{R,R'\in\mathcal{R}} \frac{|\rho^\pi_R - \rho^\pi_{R'}|}{2\|d^\pi\|_q}, \qquad \forall\pi \qquad (\text{as } \mathcal{R} \subset B(R_0, \tau_p(\mathcal{R}))). \quad (2)$$

This implies the desired result. □

Theorem 2. *The radius of the smallest L_p ball that contains $\mathcal{R}$, is given by*

$$\alpha = \max_{\pi} \max_{R,R' \in \mathcal{R}} \frac{|\rho_R^\pi - \rho_{R'}^\pi|}{2\|d^\pi\|_q}.$$

Proof. Note that we have

$$\|R - R'\|_p = \frac{|\rho_R^\pi - \rho_{R'}^\pi|}{2\|d^\pi\|_q}$$

□