Improved Sample Complexity Bounds for Distributionally Robust Reinforcement Learning

TEXAS A&M UNIVERSITY
Department of Electrical & Computer Engineering

Zaiyan Xu[*,1] Kishan Panaganti[*,1] Dileep Kalathil[1]

[*]Equal Contribution, [1]Texas A&M University
Emails: {zxu43, kpb, dileep.kalathil}@tamu.edu

Motivation

- Standard reinforcement learning (RL) algorithms often fail to perform well when the training and testing environments are different (**sim-to-real gap**).
- The framework of robust Markov decision process (RMDP) (Iyengar, 2005) is one of the ways to address the issue. It characterizes an *uncertainty set* which is a collection of models, in contrast to just **one** model in non-robust MDP. That is, the goal is to find a **distributionally robust** solution against mismatches in distribution.
- The sample complexity of non-robust MDP is well-studied already. There are matching lower and upper bounds on sample complexity for learning an ϵ-optimal policy. However, it remains an open question for robust MDP.

Goal

How many samples from the nominal model are required to learn an ϵ-optimal robust policy with a high probability?

- Previous works all used uniform covering number argument: covering the generic value function class $\mathcal{V} = \{V \in \mathbb{R}^{|\mathcal{S}|} \mid 0 \preceq V \preceq H\}$.
- Key Idea: We develop uncertainty-set-specific covering number because the function class induced by dual reformulation of each uncertainty set turns out to be less complex than the generic function class.

Main Contributions

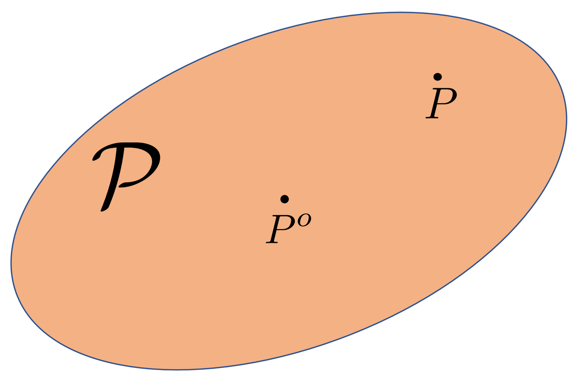
- We propose a new model-based DR-RL algorithm, RPVL, which takes advantage of the non-stationary dynamics in each *phase*.
- We provide the first-ever sample complexity result for the DR-RL problem with the Wasserstein uncertainty set.
- We demonstrate the performance of our RPVL algorithm on the Gambler's Problem for four different uncertainty sets.

Robust MDP Objective

Considering an RMDP $M = (\mathcal{S}, \mathcal{A}, \mathcal{P}, (r_h)_{h=1}^H, H)$, where the uncertainty set is defined as $\mathcal{P} = \bigotimes_{h,s,a \in [H]\times\mathcal{S}\times\mathcal{A}} \mathcal{P}_{h,s,a}$ such that $\mathcal{P}_{h,s,a} = \{P \in \Delta(\mathcal{S}) \colon D(P, P^o_{h,s,a}) \le \rho\}$. We seek to solve the following objective:

$$\sup_{\pi\in\Pi} \inf_{P\in\mathcal{P}} V_h^{\pi,P}, \quad \forall h \in [H],$$

where $V_h^{\pi,P}(s) := \mathbb{E}_{\pi,P}\left[\sum_{t=h}^H r_t(s_t, a_t) \mid s_h = s, \pi\right]$. Π is the policy class of all deterministic Markovian policies.



$\mathcal{P}$ is a collection of measures (models)! We want to find the **best** policy under the **worst** model. Note that we only have access to a generative model on the nominal model P^o.

Table: Comparison of Sample Complexity Results

Algorithm	Sample Complexity				
	TV	chi-square	Kullback-Leibler		Wasserstein
(Yang et al., 2021)	$\frac{\lvert\mathcal{S}\rvert^2\lvert\mathcal{A}\rvert H^5}{\rho^2\epsilon^2}$	$\frac{(1+\rho)^2\lvert\mathcal{S}\rvert^2\lvert\mathcal{A}\rvert H^5}{(\sqrt{1+\rho}-1)^2\epsilon^2}$	-	$\frac{\lvert\mathcal{S}\rvert^2\lvert\mathcal{A}\rvert H^5}{\rho^2\underline{p}^2\epsilon^2}$	-
(Zhou et al., 2021)	-	-	$\frac{\exp\mathcal{O}(H)\lvert\mathcal{S}\rvert^2\lvert\mathcal{A}\rvert H^5}{\rho^2\epsilon^2}$	-	-
(Panaganti and Kalathil, 2022)	$\frac{\lvert\mathcal{S}\rvert^2\lvert\mathcal{A}\rvert H^5}{\epsilon^2}$	$\frac{\rho\lvert\mathcal{S}\rvert^2\lvert\mathcal{A}\rvert H^5}{\epsilon^2}$	$\frac{\exp\mathcal{O}(H)\lvert\mathcal{S}\rvert^2\lvert\mathcal{A}\rvert H^5}{\rho^2\epsilon^2}$	-	-
This work	$\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert H^5}{\epsilon^2}$	$\frac{(1+\rho)^2\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert H^5}{(\sqrt{1+\rho}-1)^2\epsilon^2}$	$\frac{\exp\mathcal{O}(H)\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert H^5}{\rho^2\epsilon^2}$	$\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert H^5}{\rho^2\underline{p}^2\epsilon^2}$	$\frac{(B_p+\rho^p)^2\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert H^5}{\rho^{2p}\epsilon^2}$
(Non-robust) Lower bound (Li et al., 2020)	$\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert H^4/\epsilon^2$				

Theorem: Consider a finite-horizon RMDP. Let the uncertainty set be defined as one of the four distances considered in this work. Fix $\delta \in (0,1)$, $\rho > 0$, and $\epsilon \in (0, H)$. Consider the RPVL algorithm, with the total number of samples greater than or equal to the ones specified in the row of "This work" in the table above, then we have the PAC guarantee: $\|V^* - V^{\widehat{\pi}}\|_\infty \le \epsilon$ with probability at least $1-\delta$.

Algorithm: Robust Phased Value Learning (RPVL)

RPVL is a model-based algorithm. For each step (phase) $h \in [H]$, we use the generative model to generate N transitions for each state-action pairs $(s,a) \in \mathcal{S}\times\mathcal{A}$. Let $N_h(s,a,s')$ be the count of the state s' in the N total transitions from the state-action pair (s,a) in step $h \in [H]$. We then construct the maximum likelihood estimate of the nominal model as $\widehat{P}^o_{h,s,a}(s') = N_h(s,a,s')/N$.

Algorithm 1 Robust Phased Value Learning (RPVL)

1: **Input:** Uncertainty radius ρ
2: **Initialize:** $\widehat{V}_{H+1} = 0$
3: **for** $h = H, \ldots, 1$ **do**
4: Compute the empirical uncertainty set $\widehat{\mathcal{P}}_{h,s,a} = \{P \in \Delta(\mathcal{S}) \colon D(P, \widehat{P}^o_{h,s,a}) \le \rho\}$
5: $\widehat{V}_h(s) = \max_a (r(s,a) + L_{\widehat{\mathcal{P}}_{h,s,a}} \widehat{V}_{h+1}), \ \forall s \in \mathcal{S}$
6: $\widehat{\pi}_h(s) = \arg\max_a (r(s,a) + L_{\widehat{\mathcal{P}}_{h,s,a}} \widehat{V}_{h+1}), \ \forall s \in \mathcal{S}$
7: **end for**
8: **Output:** $\widehat{\pi} = (\widehat{\pi}_h)_{h=1}^H$

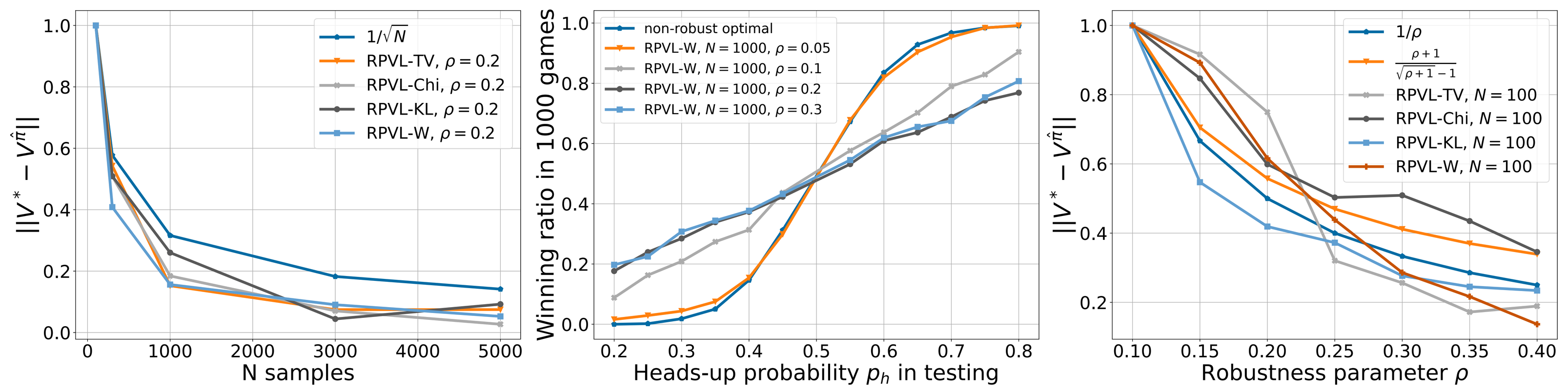
The Covering Trick - Total Variation Case

The operator $L_{\mathcal{P}^{\mathrm{TV}}_{h,s,a}} V = \inf\{PV \colon P \in \mathcal{P}^{\mathrm{TV}}_{h,s,a}\}$ is a difficult optimization problem. Using dual reformulation, we have the following equivalent form

$$L_{\mathcal{P}^{\mathrm{TV}}_{h,s,a}} V = -\inf_{\eta\in[0,2H/\rho]} \mathbb{E}_{s'\sim P^o_h(\cdot|s,a)}\left[(\eta - V(s'))_+\right] + \left(\eta - \inf_{s''\in\mathcal{S}} V(s'')\right)_+ \cdot \rho - \eta.$$

Note that in the dual reformulation, the expectation is only with respect to the nominal model P^o. With this, we discover that, in order to bound the error from using empirical uncertainty set $|L_{\mathcal{P}^{\mathrm{TV}}_{h,s,a}} V - L_{\widehat{\mathcal{P}}^{\mathrm{TV}}_{h,s,a}} V|$, we only need to cover the function class $\mathcal{U}_V = \{(\eta \cdot \mathbf{1} - V)_+ \colon \eta \in [0, H]\}$, rather than all possible value functions.

Simulations



The left plot shows the rate of convergence with respect to the number of sample N. The middle plot shows the level of robustness of Wasserstein robust policies in testing environments with perturbed model parameter p_h. The right plot shows how sub-optimality gap changes with respect to the robustness parameter ρ.

References

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

Li, G., Wei, Y., Chi, Y., Gu, Y., and Chen, Y. (2020). Breaking the sample size barrier in model-based reinforcement learning with a generative model. In *Advances in Neural Information Processing Systems*, volume 33, pages 12861–12872.

Panaganti, K. and Kalathil, D. (2022). Sample complexity of robust reinforcement learning with a generative model. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 9582–9602. PMLR.

Yang, W., Zhang, L., and Zhang, Z. (2021). Towards theoretical understandings of robust markov decision processes: Sample complexity and asymptotics. *arXiv preprint arXiv:2105.03863*.

Zhou, Z., Bai, Q., Zhou, Z., Qiu, L., Blanchet, J., and Glynn, P. (2021). Finite-sample regret bound for distributionally robust offline tabular reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pages 3331–3339.

The 26th International Conference on Artificial Intelligence and Statistics (AISTATS 2023)