Sample Complexity of Distributionally Robust Average-Reward Reinforcement Learning

Zijun Chen[1] Shengbo Wang[2] Nian Si[3]

[1] Department of Computer Science and Engineering, HKUST
[2] Daniel J. Epstein Department of Industrial and Systems Engineering, USC
[3] Department of Industrial Engineering and Decision Analytics, HKUST

November 5, 2025

Distributionally Robust Reinforcement Learning (DR-RL)

Given a Markov Decision Process (MDP) with nominal kernel P, the agent in DR-RL aims to maximize the **long-term** reward in uncertainty set $\mathcal{P}$

$$
\begin{aligned}
g_{\mathcal{P}}^{\pi}(s) &:= \inf_{Q \in \mathcal{P}} \limsup_{T \to \infty} \frac{1}{T} E_Q^{\pi}[\sum_{t=0}^{T} r(S_t, A_t) | S_0 = s] \\
g_{\mathcal{P}}^{*}(s) &:= \max_{\pi \in \Pi} g_{\mathcal{P}}^{\pi}(s)
\end{aligned}
$$

Distributionally Robust Reinforcement Learning (DR-RL)

Given a Markov Decision Process (MDP) with nominal kernel P, the agent in DR-RL aims to maximize the **long-term** reward in uncertainty set $\mathcal{P}$

$$g_{\mathcal{P}}^{\pi}(s) := \inf_{Q\in\mathcal{P}} \limsup_{T\to\infty} \frac{1}{T} E_Q^{\pi}[\sum_{t=0}^{T} r(S_t, A_t)|S_0 = s]$$
$$g_{\mathcal{P}}^{*}(s) := \max_{\pi\in\Pi} g_{\mathcal{P}}^{\pi}(s)$$

where $\mathcal{P}$ is the *uncertainty set* of candidate transition kernel for MDPs, constructed as a δ-ball centered as P

$$\mathcal{P}_{s,a}(D,\delta) = \{p : D(p||p_{s,a}) \leq \delta\} \quad \mathcal{P} = \times_{(s,a)}\mathcal{P}_{s,a}$$

Distributionally Robust Reinforcement Learning (DR-RL)

Given a Markov Decision Process (MDP) with nominal kernel P, the agent in DR-RL aims to maximize the **long-term** reward in uncertainty set $\mathcal{P}$

$$g_{\mathcal{P}}^{\pi}(s) := \inf_{Q \in \mathcal{P}} \limsup_{T \to \infty} \frac{1}{T} E_{Q}^{\pi}[\sum_{t=0}^{T} r(S_t, A_t) | S_0 = s]$$
$$g_{\mathcal{P}}^{*}(s) := \max_{\pi \in \Pi} g_{\mathcal{P}}^{\pi}(s)$$

where $\mathcal{P}$ is the *uncertainty set* of candidate transition kernel for MDPs, constructed as a δ-ball centered as P

$$\mathcal{P}_{s,a}(D, \delta) = \{p : D(p||p_{s,a}) \leq \delta\} \quad \mathcal{P} = \times_{(s,a)} \mathcal{P}_{s,a}$$

Objective: find optimal policy π^* and average-reward function $g_{\mathcal{P}}^{*}$ such that $g_{\mathcal{P}}^{\pi^*}(s) = g_{\mathcal{P}}^{*}$ for all $s \in \mathbf{S}$

Bellman optimality

It is shown that optimal average-reward function $g^*_{\mathcal{P}}$ satisfies the Robust Bellman equation:

Theorem 1 (Wang et al., 2023)

If $\mathcal{P}$ is uniformly ergodic with a uniformly bounded minorization time, then $g^*_{\mathcal{P}}(s) \equiv g^*_{\mathcal{P}}$ is a constant, and there exists a solution (g, v) of

$$v(s) = \max_{a \in A}\{r(s,a) + \inf_{q \in \mathcal{P}_{s,a}} q[v]\} - g$$

such solution satisfies $g(s) = g^*_{\mathcal{P}}$. And policy

$$\pi(s) \in \arg\max_{a \in A}\{r(s,a) + \inf_{q \in \mathcal{P}_{s,a}} q[v]\}$$

is optimal.

DR-AMDP Algorithms

- Step 1: Draw n samples for each (s, a) pair to compute empirical kernel $\widehat{P}$
- Step 2: Compute empirical uncertainty set $\widehat{\mathcal{P}}$ centered at $\widehat{P}$

Reduction to DMDP & Anchored AMDP

- Step 1: Draw n samples for each (s, a) pair to compute empirical kernel $\widehat{P}$
- Step 2: Compute empirical uncertainty set $\widehat{\mathcal{P}}$ centered at $\widehat{P}$
- Step 3:
  - **Reduction to DMDP**: Solve *Robust Discounted Bellman equation*

$$V^*_{\widehat{\mathcal{P}}}(s) = \max_{a \in A}\{r(s,a) + \gamma \inf_{q \in \widehat{\mathcal{P}}_{s,a}} q[V^*_{\widehat{\mathcal{P}}}]\}$$

    with $\gamma = 1 - 1/\sqrt{n}$
  - **Anchored AMDP**: Solve *Robust Average Bellman equation*

$$v^*_{\underline{\widehat{\mathcal{P}}}}(s) = \max_{a \in A}\{r(s,a) + \inf_{q \in \underline{\widehat{\mathcal{P}}}_{s,a}} q[v^*_{\underline{\widehat{\mathcal{P}}}}]\} - g^*_{\underline{\widehat{\mathcal{P}}}}$$

    where

$$\underline{\widehat{\mathcal{P}}}_{s,a} = \{(1 - \frac{1}{\sqrt{n}})q + \frac{1}{\sqrt{n}}\mathbf{1}e_{s_0}^\top, q \in \widehat{\mathcal{P}}\}$$

- Step 4: Extract $\widehat{\pi}^*$, $V^*_{\widehat{\mathcal{P}}}/\sqrt{n}$, $g^*_{\underline{\widehat{\mathcal{P}}}}$

Main Contribution: DR-DMDP

DR-DMDP

The solution $V^*_{\widehat{\mathcal{P}}}$ to empirical Bellman equation

$$V^*_{\widehat{\mathcal{P}}}(s) = \max_{a \in A}\{r(s,a) + \gamma \inf_{q \in \mathcal{P}_{s,a}} q[V^*_{\widehat{\mathcal{P}}}]\}$$

satisfies:

$$\|V^*_{\widehat{\mathcal{P}}} - V^*_{\mathcal{P}}\|_\infty = \widetilde{O}\left(\frac{t_{min}}{(1-\gamma)\sqrt{np_\wedge}}\right)$$

This improves the horizon dependence from $(1-\gamma)^{-2}$ (Shi and Chi, 2024; Wang et al., 2024) to $(1-\gamma)^{-1}$

$p_\wedge$: minimal support probability
t_{min}: minorization time of nominal kernel

Main Contribution: DR-AMDP

DR-AMDP

Reduction to DMDP & Anchored AMDP are **priori knowledge-free** algorithm that learn approximate optimal policy and average reward with error

$$0 \le \underbrace{g_{\mathcal{P}}^{\widehat{\pi}^*} - g_{\mathcal{P}}^*}_{\text{policy error}}, \overbrace{\underbrace{\left\| \frac{V_{\widehat{\mathcal{P}}}^*}{\sqrt{n}} - g_{\mathcal{P}}^* \right\|_\infty}_{\text{Reduction to DMDP}}, \underbrace{\left\| g_{\underline{\widehat{\mathcal{P}}}}^* - g_{\mathcal{P}}^* \right\|_\infty}_{\text{Anchored AMDP}}}^{\text{value error}} = O\left(\frac{t_{min}}{\sqrt{np_\wedge}} \sqrt{\log(\frac{|S|^2|A|}{\beta})} \right)$$

i.e., DR-AMDP achieves ε-optimality with $\widetilde{O}(\frac{t_{min}^2}{p_\wedge \varepsilon^2})$ samples.

Main Technique used

- Challenge 1: Minorization time could be unbounded over $\mathcal{P}$-issued by making constraints on uncertainty size δ
  - When $\mathcal{P} = \mathcal{P}(D_{KL}, \delta)$, $\delta \leq \frac{1}{8m_\vee^2} p_\wedge$
  - When $\mathcal{P} = \mathcal{P}(D_{f_k}, \delta)$, $\delta \leq \frac{1}{\max\{8,4k\} m_\vee^2} p_\wedge$

  Then

$$\sup_{Q \in \mathcal{P}, \pi \in \Pi} t_{min}(Q_\pi) = O(t_{min}) \quad t_{min} := \max_{\pi \in \Pi} t_{min}(P_\pi)$$

Main Technique used

- Challenge 2: Sub-optimal rate for DR-DMDP-issued by dual form

$$\inf_{q\in\mathcal{P}_{s,a}(D_{KL},\delta)} q[V] = \sup_{\alpha\geq}\{-\alpha\delta - \alpha\log p_{s,a}[e^{-V/\alpha}]\}$$

$$\inf_{q\in\mathcal{P}_{s,a}(D_{f_k},\delta)} q[V] = \sup_{\alpha\in\mathbb{R}}\{\alpha - c_k(\delta)p_{s,a}[(\alpha - V)_+^{k^*}]^{1/k^*}\}$$

The empirical error:

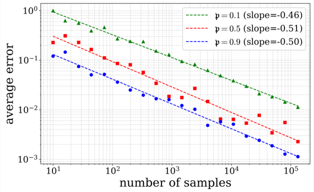
$$\|\inf_{q\in\widehat{\mathcal{P}}_{s,a}} q[V] - \inf_{q\in\mathcal{P}_{s,a}} q[V]\|_\infty = \widetilde{O}\left(\frac{Span(V)}{\sqrt{np_\wedge}}\right) = \widetilde{O}\left(\frac{t_{min}}{\sqrt{np_\wedge}}\right)$$

With this result, we show:

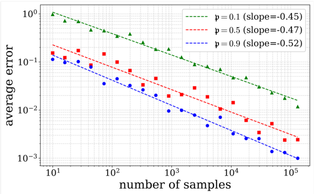
$$\|V^*_{\widehat{\mathcal{P}}} - V^*_{\mathcal{P}}\|_\infty = \widetilde{O}\left(\frac{t_{min}}{(1-\gamma)\sqrt{np_\wedge}}\right)$$

improves the horizon dependence from $(1-\gamma)^{-2}$ (Shi and Chi, 2024; Wang et al., 2024) to $(1-\gamma)^{-1}$

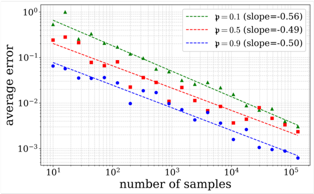
Experiment



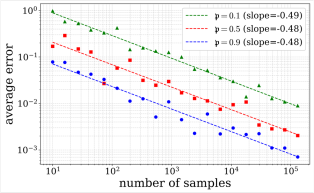
KL-divergence case for Reduction to DMDP



χ^2-divergence case for Reduction to DMDP



KL-divergence case for Anchored AMDP



χ^2-divergence case for Anchored AMDP

Figure: Reduction to DMDP & Anchored AMDP performance on hard MDP instance under KL-divergence and χ^2-divergence.

Thank You!

References

Shi, L. and Chi, Y. (2024). Distributionally robust model-based offline reinforcement learning with near-optimal sample complexity. *Journal of Machine Learning Research*, 25(200):1–91.

Wang, S., Si, N., Blanchet, J., and Zhou, Z. (2024). Sample complexity of variance-reduced distributionally robust q-learning. *Journal of Machine Learning Research*, 25(341):1–77.

Wang, Y., Velasquez, A., Atia, G., Prater-Bennette, A., and Zou, S. (2023). Robust average-reward markov decision processes. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(12):15215–15223.