Sample Complexity of Distributionally Robust Off-Dynamics Reinforcement Learning with Online Interaction

Yiting He*, Zhishuai Liu*, Weixin Wang, Pan Xu
(*Equal contribution)

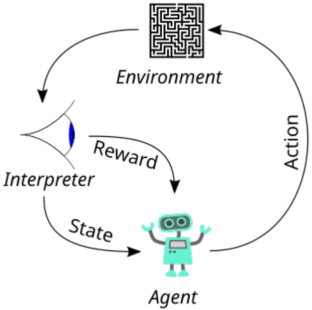
June 11th, 2025

Outline

Contents

Reinforcement Learning (RL)

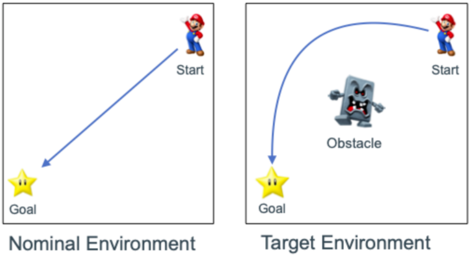
Goal: learning the optimal policy under an unknown environment
Examples: AlphaGo, ChatGPT, etc.



Off-Dynamics RL

Off-Dynamics RL considers the **distribution shift**

Goal: learning a robust policy to optimize the worst-case scenario



Constrained Robust Markov Decision Process (CRMDP)

CRMDPs seek the best policy under the worst-case transition within a predefined uncertainty set.

$$\text{CRMDP}(\mathcal{S}, \mathcal{A}, P^o, r, \mathcal{U}^\rho(P^o), H)$$

The optimal robust policy:

$$\pi^* = \arg\max_\pi V^{\pi,\rho}(s),$$

$$V^{\pi,\rho}(s) = \inf_{P \in \mathcal{U}^\rho(P^o)} \mathbb{E}_{\pi,P}\left[\sum_{t=1}^{H} r_t(s_t, a_t) \,\middle|\, s_1 = s\right].$$

Regularized Robust Markov Decision Process (RRMDP)

RRMDPs replaces the hard constraint on uncertainty sets with a regularization term.

$$\text{RRMDP}(\mathcal{S}, \mathcal{A}, P^o, r, \beta, \mathrm{D}, H)$$

The optimal robust policy:

$$\pi^* = \arg\max_{\pi} V^{\pi,\rho}(s),$$

$$V_h^{\pi,\beta}(s) = \inf_{P \in \Delta(S)} \mathbb{E}_{\pi,P}\Bigg[\sum_{t=1}^{H} r_t(s_t, a_t) + \beta \cdot \mathrm{D}(P_t(\cdot|s_t, a_t), P_t^o(\cdot|s_t, a_t)) \,\Bigg|\, s_1 = s\Bigg].$$

Learning Goal

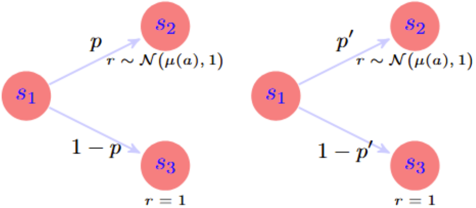
Minimize the average sub-optimality through online interaction:

$$\mathrm{Regret}(K) = \sum_{k=1}^{K} \left[V_1^{*,\rho}(s_1^k) - V_1^{\pi^k,\rho}(s_1^k) \right] \text{ for CRMDPs,}$$

$$\mathrm{Regret}(K) = \sum_{k=1}^{K} \left[V_1^{*,\beta}(s_1^k) - V_1^{\pi^k,\beta}(s_1^k) \right] \text{ for RRMDPs.}$$

Contents

Hard Instances



(a) The nominal MDP. (b) The perturbed MDP.

$\boxed{\text{Learning CRMDPs can be exponentially hard without a proper assumption}}$

Proposed Assumption

$P_h^{w,\pi}(\cdot|s,a)$: worst-case transition w.r.t. V_{h+1}^{π}

- $\mathrm{q}_h^{\pi}(\cdot)$: visitation measure induced by policy π under $P^{w,\pi}$,
- $\mathrm{d}_h^{\pi}(\cdot)$: visitation measure induced by policy π under P^{o}.

Bounded visitation measure ratio:

$$C_{vr} := \sup_{\pi,h,s} \frac{\mathrm{q}_h^{\pi}(s)}{\mathrm{d}_h^{\pi}(s)}$$

We assume that C_{vr} is polynomial in H, S and A.

Algorithm Design

Online Robust Bellman Iteration (ORBIT)

At each episode, ORBIT consists of two phases:

1. *Phase 1*: Robust Bellman Iteration with Optimistic Estimation
2. *Phase 2*: Explore the Nominal Environment & Collect Data

Robust Q-function Estimation

$$\widehat{Q}_h^k(s,a) = \underbrace{\mathrm{RB}_h^k(s,a)}_{\text{robust Bellman estimator}} + \underbrace{b_h^k(s,a)}_{\text{bonus term}}$$

CRMDP results

upper bounds:

$$\text{Regret}(K) = \begin{cases} \widetilde{\mathcal{O}}\left(C_{vr} S^2 A H^2 + C_{vr}^{\frac{1}{2}} S^{\frac{3}{2}} A^{\frac{1}{2}} H^2 \sqrt{K}\right) & \text{(TV)} \\ \widetilde{\mathcal{O}}\left(\left(1 + \frac{H\sqrt{S}}{\rho C_{MP}}\right)\left(C_{vr} SAH + C_{vr}^{\frac{1}{2}} S^{\frac{1}{2}} A^{\frac{1}{2}} H \sqrt{K}\right)\right) & \text{(KL)} \\ \widetilde{\mathcal{O}}\left(C_{vr} S^2 A H^2 + C_{vr}^{\frac{1}{2}} S^{\frac{3}{2}} A^{\frac{1}{2}} H^2 \sqrt{K}\right) & (\chi^2) \end{cases}.$$

lower bounds: $\forall$ learning algorithm ξ, $\exists$ CRMDP $\mathcal{M}$, s.t.

$$\mathbb{E}\left[\text{Regret}_{\mathcal{M}}(\xi, K)\right] = \Omega\left(C_{vr}^{\frac{1}{2}} \sqrt{K}\right)$$

$\boxed{\text{Upper bounds matches lower bounds in } C_{vr} \text{ and } K}$

RRMDP results

upper bounds:

$$\mathrm{Regret}(K) = \begin{cases} \widetilde{\mathcal{O}}\big(C_{vr}S^{\frac{3}{2}}AH^2 + C_{vr}^{\frac{1}{2}}SA^{\frac{1}{2}}H^2\sqrt{K}\big) & \text{(TV)} \\ \widetilde{\mathcal{O}}_{((1+\beta e^{\beta^{-1}H}\sqrt{S})}(C_{vr}SAH + C_{vr}^{\frac{1}{2}}S^{\frac{1}{2}}A^{\frac{1}{2}}H\sqrt{K})) & \text{(KL)} \\ \widetilde{\mathcal{O}}\big(C_{vr}S^2AH^3 + C_{vr}^{\frac{1}{2}}S^{\frac{3}{2}}A^{\frac{1}{2}}H^3\sqrt{K}\big) & (\chi^2) \end{cases}.$$

lower bounds: $\forall$ learning algorithm ξ, $\exists$ RRMDP $\mathcal{M}$, s.t.

$$\mathbb{E}\big[\mathrm{Regret}_{\mathcal{M}}(\xi, K)\big] = \Omega\big(C_{vr}^{\frac{1}{2}}\sqrt{K}\big)$$

Upper bounds matches lower bounds in C_{vr} and K

Comparison with Related Works

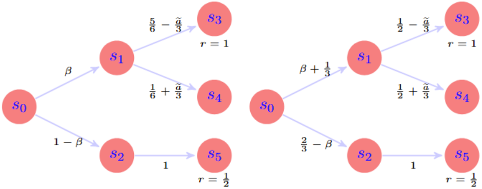
	Assumption	CRMDPs	RRMDPs
Liu and Xu (2024a)[1]	Fail-states	TV	$\times$
Lu et al. (2024)[2]	Vanishing minimal value	TV	$\times$
Our work	Bounded visitation measure ratio	$\checkmark$	$\checkmark$

[1] Liu, Zhishuai, and Pan Xu. "Distributionally robust off-dynamics reinforcement learning: Provable efficiency with linear function approximation." International Conference on Artificial Intelligence and Statistics. PMLR, 2024.

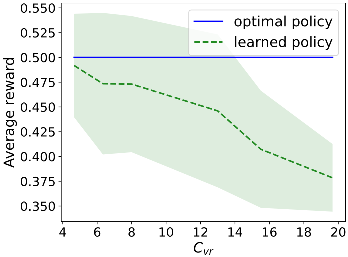
[2] Lu, Miao, et al. "Distributionally robust reinforcement learning with interactive data collection: Fundamental hardness and near-optimal algorithm." arXiv preprint arXiv:2404.03578 (2024).

Contents

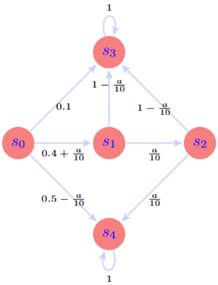
Effect of C_{vr}



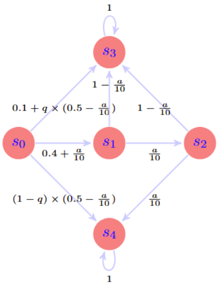
(a) The nominal MDP environment. (b) The perturbed MDP environment.



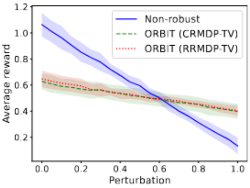
Learning on Simulated RMDPs



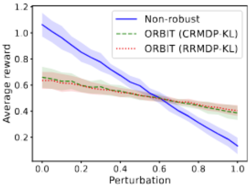
(a) The source RMDP environment.



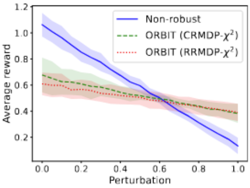
(b) The target RMDP environment.



TV settings

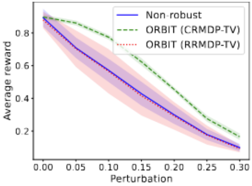


KL settings

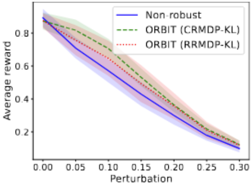


χ^2 settings

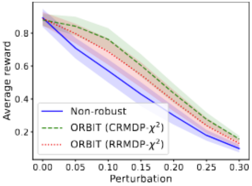
Learning the Frozen Lake Problem



TV settings



KL settings



χ^2 settings