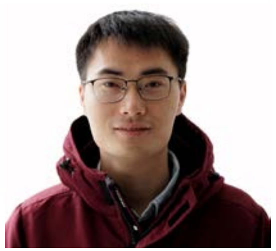
ICML | 2022
Thirty-ninth International Conference on Machine Learning

Dynamic Regret of Online Markov Decision Processes



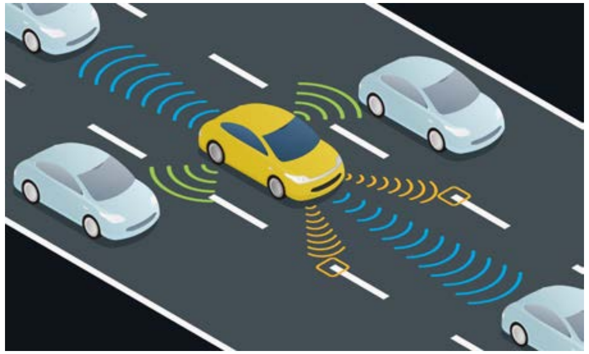
Peng Zhao Long-Fei Li Zhi-Hua Zhou

LAMDA Group
Nanjing University

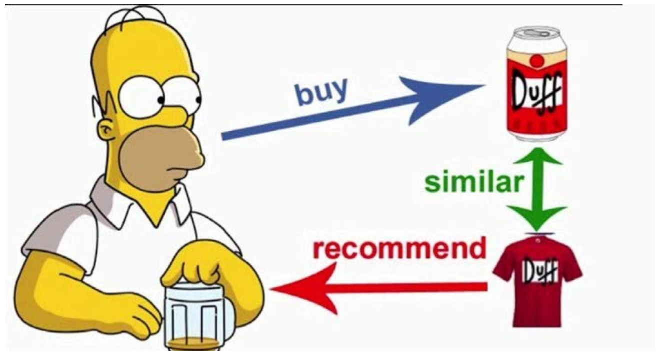
Introduction

- Learning adversarial MDPs with *static regret* is well studied.

  the single fixed strategy may perform poorly in the non-stationary or even adversarially changing environments.



autonomous driving



online recommendations

- A more strengthened performance measure: ***dynamic regret***.

  competes the performance against a sequence of ***changing*** policies

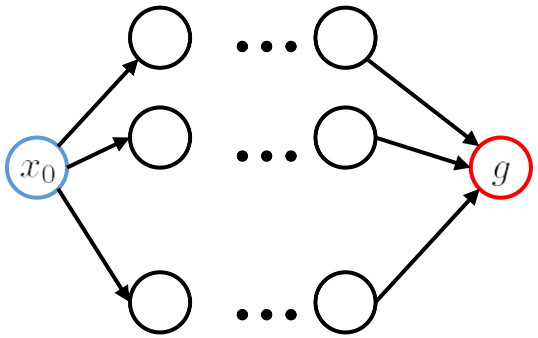
Online MDPs

Learning And Mining from DatA

- We consider the three foundational models of online MDPs:

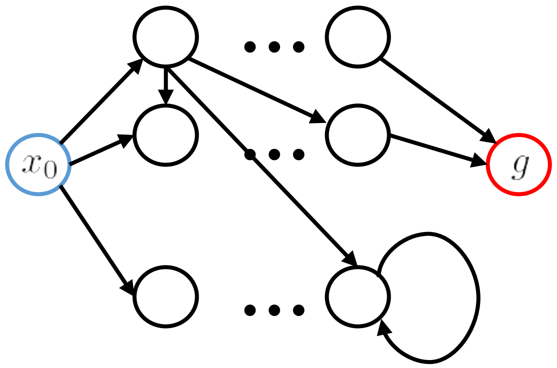
Episodic Setting

Loop-free



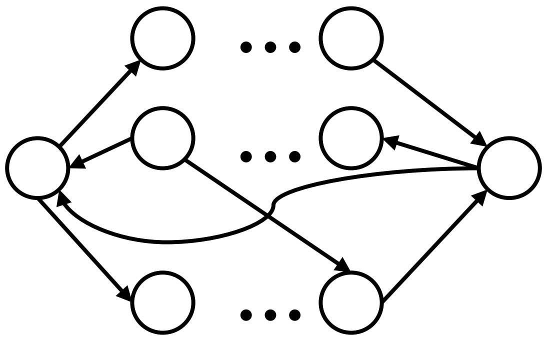
[Neu et al., COLT 2010;
Rosenberg et al., ICML 2019;
Jin et al., ICML 2020]

Non-loop-free



[Rosenberg et al., IJCAI 2021;
Chen et al., COLT 2021]

Infinite-horizon Setting



[Even-Dar et al.,, MathOR 2009;
Yu et al., MathOR 2009;
Neu et al., NeurIPS 2010]

- **Focus**: adversarial online MDPs with full info. and known trans.

Our Contributions

All three MDP models:

- propose ***parameter-free*** algorithms with ***dynamic regret*** guarantees that can recover best known static regret results
- establish relationship between variation of occupancy measures and policies

Episodic (loop-free) SSP:

- prove the obtained dynamic regret is minimax optimal

Infinite-horizon MDPs:

- present a reduction to the switching-cost expert problem

Adversarial Online MDPs

For each round $t = 1, \ldots, T$:

- learner observes current state x_t, decides a policy $\pi_t: X \times A \rightarrow [0,1]$, executes an action a_t sampled from $\pi_t(\cdot|x_t)$.
- environment chooses a loss function $\ell_t: X \times A \rightarrow [0,1]$ simultaneously.
- learner suffers loss $\ell_t(x_t, a_t)$ and observes loss function ℓ_t.

Dynamic regret:

$$\text{D-Regret}_T\left(\pi_{1:T}^c\right) = \sum_{t=1}^{T} \ell_t\big(x_t, \pi_t(x_t)\big) - \sum_{t=1}^{T} \ell_t\big(x_t, \pi_t^c(x_t)\big),$$

where $\pi_1^c, \ldots, \pi_T^c$ is any sequence of compared policies in the policy class Π.

$\Rightarrow$ recover the standard static regret when choosing a fixed compared policy

Our Result: Algorithm and Theory

- We propose parameter-free algorithms that can obtain the following dynamic regret guarantees for three MDP models.

MDP Model	Ours Result (dynamic regret)	Previous Work (static regret)
Episodic loop-free SSP (Section 2)	$\widetilde{\mathcal{O}}(H\sqrt{K(1+P_T)})$ [Theorem 1]	$\widetilde{\mathcal{O}}(H\sqrt{K})$ (Zimin & Neu, 2013)
Episodic SSP (Section 3)	$\widetilde{\mathcal{O}}(\sqrt{B_K(H_* + \bar{P}_K)} + \bar{P}_K)$ [Theorem 3]	$\widetilde{\mathcal{O}}(\sqrt{H^{\pi^*}DK})$ (Chen et al., 2021a)
Infinite-horizon MDPs (Section 4)	$\widetilde{\mathcal{O}}(\sqrt{\tau T(1+\tau P_T)} + \tau^2 P_T)$ [Theorem 6]	$\widetilde{\mathcal{O}}(\sqrt{\tau T})$ (Zimin & Neu, 2013)

- Our dynamic regret results can recover the best known static regret bounds for all three MDP models.
- The results for episodic (loop-free) SSP are minimax optimal in terms of time horizon and certain non-stationarity measures.

Summary

- An initial resolution for dynamic regret of online MDPs.
- Design *parameter-free algorithms* with dynamic regret bounds which can recover the best known static regret results for all three MDPs, and the results for episodic (loop-free) SSP are minimax optimal.
- Present a reduction to the switching-cost expert problem for the infinite-horizon MDPs, which is new to the best of our knowledge.
- The algorithm design is based on the *online ensemble* framework, and requires several new components (groupwise scheduling, correction terms, and weighted negative entropy regularizer, etc).

Thanks!