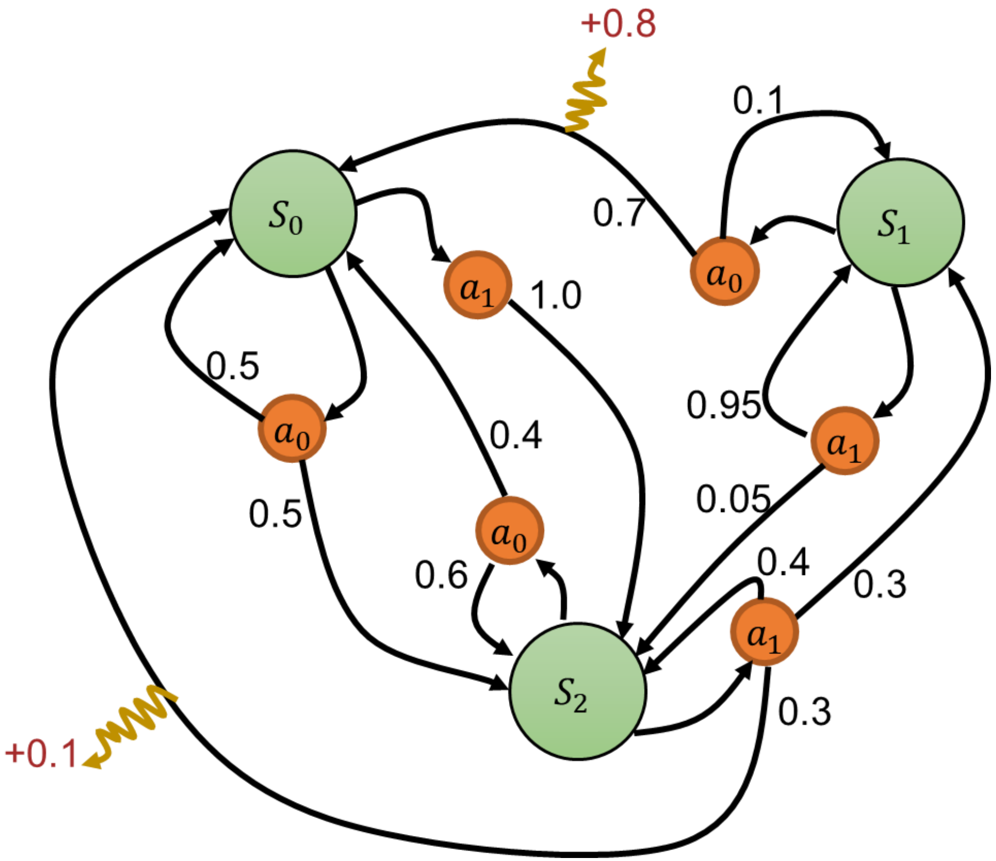
Horizon-Free and Variance-Dependent Reinforcement Learning for Latent Markov Decision Processes

Runlong Zhou

Joint work with Ruosong Wang and Simon Du

Introduction

- Markov decision processes (MDPs)
  - **State** space $\mathcal{S}$, with size S
  - **Action** space $\mathcal{A}$, with size A
  - **Reward** function $R(s, a)$
  - **Initial state** distribution $\nu(s)$
  - **Transition** probability $P(s'|s, a)$
  - **Maximum transition degree** $\Gamma = \max_{s,a} \left|\left|P(\cdot\,|s, a)\right|\right|_0$



Introduction

- Episodic RL for MDPs
  - Planning **horizon** H
  - Play a policy π^k in episode k
  - **Regret**
    - Play for K episodes
    - Regret = Optimal cumulative reward – sum of expected cumulative rewards of each π^k
    - **Minimax** regret = worst case regret

Motivation of LMDPs

- (Generic) MDPs have been solved
  - Minimax regret matches lower-bound
- Partially Observable MDPs are quite hard
  - Instead of observing s, the agent observes $o \sim O(o|s)$
  - Sample complexity lower-bound is **exponential**

Motivation of LMDPs

- Something in the middle – **Latent MDPs (LMDPs)**
  - Hidden state is ”decomposable”, and the unobserved part is static
    - $[s = (m, o), a] \rightarrow s' = (m, o')$
  - **Strictly harder** than MDPs!
  - Under some assumptions, LMDP is **strictly easier** than POMDPs!
    - **Context in hindsight**
- LMDPs are useful
  - Modeling combinatorial optimization problems
  - Multi-task learning

Motivation of **variance-dependent** regrets

- Some LMDPs are **easier** than the others
  - Applying an algorithm on them yields regrets better than **worst-case**
  - Example: **Deterministic MDP** as a special case of LMDP
    - Regret lower bound $\Omega(SA)$, no dependency on K
    - **Specially designed** algorithms can have regret upper bound $O(SA)$

Formulation

- LMDPs [Kwon et al.,2021]
  - A distribution of $\boldsymbol{M}$ MDPs $\mathcal{M} = \{\mathcal{M}_1, \ldots, \mathcal{M}_M\}$
  - Each with **weight (probability)** $\boldsymbol{w_1}, \ldots, \boldsymbol{w_M}$
  - All MDPs share the same **states, actions and horizon**
  - But have their own **reward function** $\boldsymbol{R_m}$**, transition** $\boldsymbol{P_m}$ **and initial state distribution** $\boldsymbol{\nu_m}$
  - **Core difficulty** is learning P, so assume w, ν and R are **known**

Planning in LMDPs

- Alpha vectors
  - History dependent π, let h be any history, a be any action

$$\alpha_m^\pi(h) := \mathbb{E}\left[\sum_{t'=t}^{H} R_m(s_{t'}, a_{t'}) \,\middle|\, \mathcal{M}_m, \pi, h_t = h \right],$$

$$\alpha_m^\pi(h, a) := \mathbb{E}\left[\sum_{t'=t}^{H} R_m(s_{t'}, a_{t'}) \,\middle|\, \mathcal{M}_m, \pi, (h_t, a_t) = (h, a) \right].$$

  - Generalized value function and Q-function of each MDP

Problem setup

- Generally, follow the episodic RL for MDPs setting
- At the beginning of each episode, draw an MDP $\mathcal{M}_m$ according to the probability
- When the episode **ends** (completes H steps), **tell the agent $\boldsymbol{m}$**
  - **Context in hindsight** [Kwon et al.,2021]
  - Drastically reduce the sample complexity lower-bound to polynomial.

Related works

- Minimax regret with context in hindsight

Theorem 3.3 *The regret of the Algorithm 1 is bounded by:*

$$Regret(K) \leq \sum_{k=1}^{K} (V^{\pi_k}_{\widetilde{\mathcal{M}}_k} - V^{\pi_k}_{\mathcal{M}^*}) \lesssim HS\sqrt{MAN},$$

where $N = HK$, i.e., total number of taken actions.

- Often assume reward of each episode **bounded by 1 almost surely**, so regret is $\tilde{O}\left(\sqrt{MS^2AHK}\right)$
- [Kwon et al.,2021]

$\tilde{O}$ hides poly log terms

Contribution

- Define variance to **characterize the difficulty** of each LMDP
- A regret better than $\tilde{O}\left(\sqrt{MS^2AHK}\right)$ **with** context in hindsight
- A lower bound for any class of variance-bounded LMDPs

Maximum policy-value variance

- For any policy π:

$$\mathsf{Var}^{\pi} := \mathbb{V}(w \circ \nu, \alpha^{\pi}_{\cdot}(\cdot)) + \mathbb{E}_{\pi}\left[\sum_{t=1}^{H} \mathbb{V}(P_m(\cdot|s_t, a_t), \alpha^{\pi}_{m}(h_t a_t r_t \cdot))\right].$$

- Further define $\mathrm{Var}^{\pi} \coloneqq \max_{s} \mathrm{Var}_1^{\pi}(s)$.
- Maximum policy-value variance is defined as $\mathrm{Var}^{\star} \coloneqq \max_{\pi} \mathrm{Var}^{\pi}$.

Regret upper bound

Theorem 2. *For both the Bernstein confidence set for LMDP (Algorithm 1 combined with Algorithm 2) and the Monotonic Value Propagation for LMDP (Algorithm 1 combined with Algorithm 3), with probability at least* $1-\delta$*, we have that*

$$\mathsf{Regret}(K) \leqslant \tilde{O}(\sqrt{\mathsf{Var}^{\star} M\Gamma SAK} + MS^2A).$$

- For deterministic MDPs:
  - $\mathrm{Var}^{\star} = 0$, $M = 1$ and $\Gamma = 1$
  - $\mathrm{Regret}(K) \leq \tilde{O}(S^2 A)$
    - Only depend on $\mathrm{poly}(\log(K))$ instead of $\mathrm{poly}(K)$!
    - Have a gap of S compared to SA

Regret lower bound

Theorem 3. *Assume that* $S \geqslant 6$, $A \geqslant 2$, $M \leqslant \lfloor \frac{S}{2} \rfloor !$ *and* $0 < \mathcal{V} \leqslant O(1)$. *For any algorithm* $\boldsymbol{\pi}$, *there exists an LMDP* $\mathcal{M}_{\boldsymbol{\pi}}$ *such that:*

- $\mathsf{Var}^\star = \Theta(\mathcal{V})$;
- *For* $K \geqslant \widetilde{\Omega}(M^2 + MSA)$, *its expected regret in* $\mathcal{M}_{\boldsymbol{\pi}}$ *after* K *episodes satisfies*

$$\mathbb{E}\left[\sum_{k=1}^{K} (V^\star - V^k) \,\middle|\, \mathcal{M}_{\boldsymbol{\pi}}, \boldsymbol{\pi} \right] \geqslant \Omega(\sqrt{\mathcal{V} MSAK}).$$

References

- Jeongyeol Kwon, Yonathan Efroni, Constantine Caramanis, and Shie Mannor. Rl for latent mdps:Regret guarantees and a lower bound, 2021.
- Zihan Zhang, Xiangyang Ji, and Simon Shaolei Du. Is reinforcement learning more difficult than bandits?
a near-optimal algorithm escaping the curse of horizon. In COLT, 2021.
- Cohen, A., Kaplan, H., Mansour, Y., & Rosenberg, A. (2020). Near-optimal Regret Bounds for Stochastic Shortest Path. *ICML*.
- Santosh S Vempala, Ruosong Wang, and David P Woodruff. The communication complexity of optimization. In Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms, pages 1733–1752. SIAM, 2020.