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

Runlong Zhou[1] Ruosong Wang[1] Simon S. Du[1]

[1]University of Washington

Preliminaries

Markov Decision Processes

- S states
- A actions
- Planning horizon H
- Reward function $R(s,a)$, such that for any possible trajectory, its total reward in a single episode is upper-bounded by 1
- Transition probability $P(s'|s,a)$
- Initial state distribution $\nu(s)$
- Maximum transition support $\Gamma = \max_{s,a} \|P(\cdot|s,a)\|_0$

Latent MDPs

- A distribution of M MDPs $\mathcal{M} = \{\mathcal{M}_1, \ldots, \mathcal{M}_M\}$
- Each with weight (probability) $w_1, \ldots, w_M$
- All MDPs share the same states, actions and horizon
- But have their own reward function R_m, transition P_m and initial state distribution ν_m

Policy and alpha vectors

- $\pi : \mathcal{H} \to \mathcal{A}$, where $\mathcal{H}$ is the set of all histories
- Alpha vectors (generalization of value and Q-functions)

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

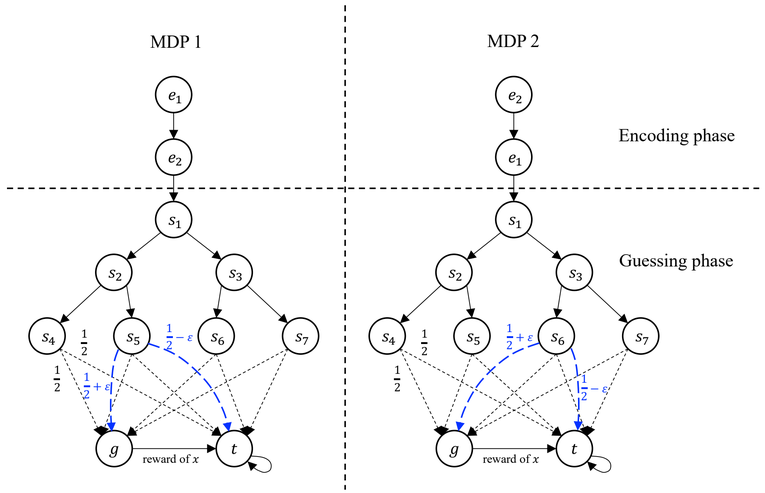
$$\alpha_m^\pi(h,a) := \mathbb{E}_{\pi,\mathcal{M}_m}\left[\sum_{t'=t}^{H} R_m(s_{t'}, a_{t'}) \,\middle|\, (h_t, a_t) = (h,a)\right].$$

- Value function $V^\pi = \sum_{m,s} w_m \nu_m(s) \alpha_m^\pi(s)$

Episodic reinforcement learning with context in hindsight

- K episodes, with policies $\pi^1, \pi^2, \ldots, \pi^k$
- **Context in hindsight:** At the beginning, sample $m \sim \{w\}$, but only tell the agent m when the episode ends
- Performance measure

$$\mathsf{Regret}(K) := \sum_{k=1}^{K} (V^\star - V^{\pi^k}).$$



Motivation

A problem harder than MDPs while easier than POMDPs

- LMDPs are collections of MDPs with a hidden context
- LMDPs are POMDPs with invariant unobservables throughout an episode: $s = (m, o) \to s' = (m, o')$

RL environments have different randomness:

We desire a regret that:

- Better than a previous work of $\widetilde{O}(\sqrt{MHS^2AK})$
- Reduce to better guarantee if the LMDP is special, e.g., a deterministic MDP

Maximum Policy-Value Variance

For deterministic policy π, define

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

and $\mathsf{Var}^\star := \max_{\pi\in\Pi} \mathsf{Var}^\pi$.

Discussion on Variances

- Under **Condition 1**, we have $\mathsf{Var}_\tau^\Sigma \leqslant \widetilde{O}(1)$ with high probability if τ is sampled by any policy and $\mathsf{Var}^\star \leqslant 1$. Without **Condition 1**, $\mathsf{Var}_\tau^\Sigma \leqslant \widetilde{O}(H^2)$ and $\mathsf{Var}^\star \leqslant H^2$.
- For deterministic MDPs, $\mathsf{Var}_\tau^\Sigma = \mathsf{Var}^\star = 0$.
- $\mathsf{Var}^\star = 0 \implies \mathsf{Var}_\tau^\Sigma = 0$, while reverse is not true.

Regret Upper Bound

We propose an algorithmic framework for solving LMDPs, which takes planning oracles as plug-in solvers. With two oracles we design (Bernstein confidence set and Monotonic Value Propagation for LMDP), we can guarantee a regret upper bound of

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

where $\widetilde{O}$ hides polylog factors.

Remark:

- $\mathsf{Var}^\star \leqslant 1$ and $\Gamma \leqslant S$, so the worst case regret is $\widetilde{O}(\sqrt{MS^2AK} + MS^2A)$, which is the first horizon-free bound for LMDPs
- If we view deterministic MDP as a special LMDP, we have $\mathsf{Var}^\star = 0$ and $M = 1$, regret is $\widetilde{O}(S^2A)$, which is a constant

Regret Lower Bound

For any variance level $0 < \mathcal{V} \leqslant O(1)$ and 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

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

High level idea: See the illustration in the top figure. We transform context in hindsight into context being told beforehand, while not affecting the optimal value function. Use a small portion of states to encode the context, then the optimal policy can extract information from them.

vectorzh@cs.washington.edu