Efficient RL with Impaired Observability: Learning to Act with Delayed and Missing State Observations

Minshuo Chen, Yu Bai, H. Vincent Poor, Mengdi Wang

November 12, 2023

Standard MDPs

Tabular MDP as a tuple $(\mathcal{S}, \mathcal{A}, P, R, H)$:

- $\mathcal{S}, \mathcal{A}$ are state and action spaces, respectively;
- P, R are the transition and reward, respectively;
- H is the horizon.

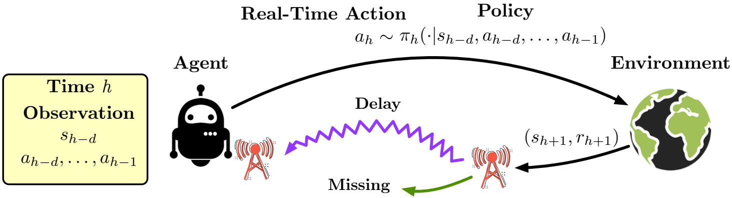
$\star$ Policy is a probability distribution conditioned on an *instantaneous state*, i.e.,

$$a_h \sim \pi_h(\cdot | s_h).$$

MDPs with Delayed and Missing Observations

Instantaneous state is **not available** due to

- Communication latency;
- Lossy channel.



Real-Time Policies

To take real-time decisions, the policy becomes

$$\Pi_{\text{exec}} = \{\pi_h(\cdot | s_{t_h}, \text{historical actions}) : h = 1, \ldots, H\},$$

where t_h is the nearest observed state index.

Major Challenges

- **Exponential complexity growth**: Π_{exec} can be exponentially large.
- **Information loss**: Missing observations are permanently lost.
- **POMDP formulation not working**: Complicated belief state propagation.

Contributions

Our contributions:

- Efficient policy learning algorithms with regret having optimal dependence on the size of the state-action space.
- Analysis of performance degradation caused impaired observability.

Delayed Observations

Stochastic Delay Distribution

Nonnegative Interarrival Time

The instantaneous state s_h is delayed for $d_h \in \{0, 1, \dots\}$ steps.
We denote the *interarrival time* as

$$\Delta_h = d_{h+1} - d_h,$$

taking value in $\{0, 1, \dots\}$.

Assumption

The distribution $\mathcal{D}_h(s_h, a_h)$ of Δ_h can depend on (s_h, a_h), but is conditionally independent of the MDP transition given (s_h, a_h).

Regret Analysis

Theorem

Let $\gamma \in (0,1)$ be any failure probability. With probabiblity $1-\gamma$, the regret of the proposed UCBVI-type learning algorithm satisfies

$$\texttt{Regret}(K) \leq c\left(H^4\sqrt{SAK\iota} + H^4 S^2 A \iota^2\right),$$

where K is the number of episodes, $\iota = \log \frac{SAHK}{\gamma}$ and c is a constant.

- Comparison to the best policy in Π_{exec};
- Sharp dependence on S and A;
- Sharp dependence on K.

Bounded Length of Delay

Assumption

The length of stochastic delay is bounded by a constant $D \leq H$.

Corollary

Running the proposed UCBVI-type algorithm leads to a regret of

$$\texttt{Regret}(K) = \widetilde{\mathcal{O}}\left(D^{5/2}\sqrt{SAKH^3\iota}\right).$$

Performance Degradation

We quantify the performance degradation by

$$\texttt{Gap}(s_1) = V^*_{1,\text{nodelay}}(s_1) - V^*_{1,\text{delay}}(s_1)$$

Proposition

Consider constant delays. Fix a positive integer $d < H$. Then there exists an MDP instance such that simultaneously,

- *When delay is d, it holds that*

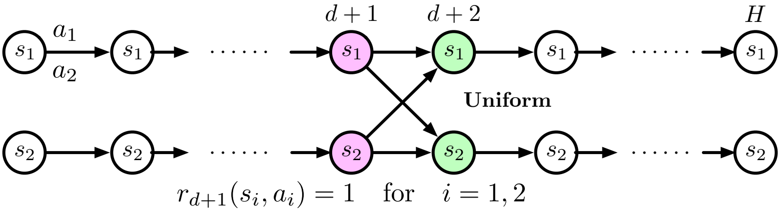
$$\frac{1}{K}\sum_{k=1}^{K} \texttt{Gap}(s_1^k) = 0.$$

- *When delay is $d+1$, it holds that*

$$\frac{1}{K}\sum_{k=1}^{K} \texttt{Gap}(s_1^k) \geq \frac{1}{2} - \sqrt{\frac{1}{2K}\log\frac{1}{\gamma}},$$

with probability $1-\gamma$.

Deterministic v.s. Stochastic Transition



- Deterministic transition is lossless with the presense of delay;
- Totally random transition incurs a large performance drop.

Missing Observations

Randomly Missing Observations

Assumption

Any pair of observation (state and reward) is independently observable. The observation rate is $\lambda_0 > 0$. (Equivalently, observation is missing at a rate of $1 - \lambda_0$.)

SA Regret When λ_0 Large

Theorem

Suppose $\lambda_0 \geq 1 - A^{-(1+v)}$ *for some positive constant* v. *Given a failure probability* γ, *with probability* $1-\gamma$, *running UCBVI-Type policy learning in* $\texttt{MDP}_{\mathrm{aug}}$, *the regret satisfies*

$$\texttt{Regret}(K) \leq c\left(H^4\sqrt{SAK\iota^3} + S^2\sqrt{H^9 K^{\frac{1}{(1+v)}}\iota^6}\right),$$

where $\iota = \log\frac{SAHK}{\gamma}$ *and* c *is some constant.*

- Sharp dependence on S and A;
- Sharp dependence on K;

Thank you!