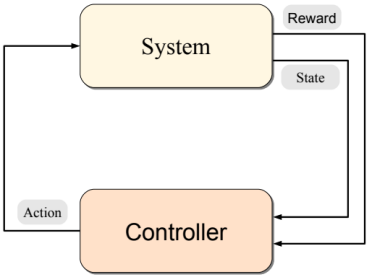
Reinforcement Learning with Action-Triggered Observations

Alexander Ryabchenko, Wenlong Mou

University of Toronto and Vector Institute

Standard Markov Decision Process



Basic RL loop [Sze10]

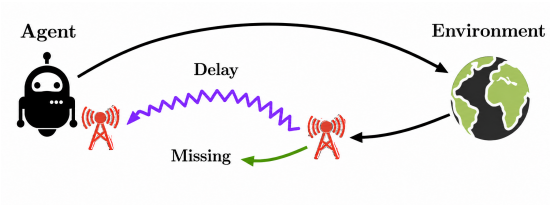
$$\textbf{MDP} = (\mathcal{S}, \mathcal{A}, \mathbb{P}, r)$$

- From state $s \in \mathcal{S}$, agent plays action $a \in \mathcal{A}$.
- Environment draws new state $s' \sim \mathbb{P}(\cdot \mid s, a)$.
- Agent observes s' and reward $r(s, a) \in [0, 1]$.

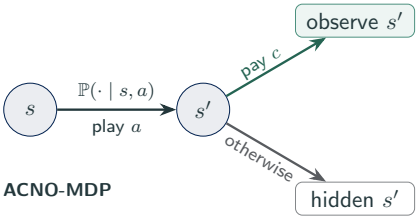
implicit assumption: state is observed after every action

Sporadic State Observations

impaired observability [Che+23]



paid observations [NFB21]



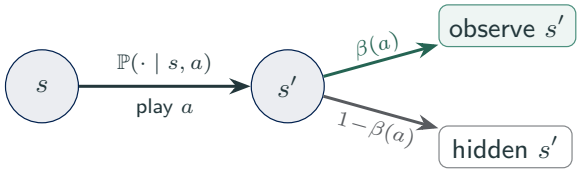
ACNO-MDP

real world: state observations may be sporadic and triggered by actions

Introducing Action-Triggered Sporadically Traceable MDP

ATST-MDP $= (\mathcal{S}, \mathcal{A}, \mathbb{P}, r, \beta)$

Extend an MDP with action-dependent state observation probabilities $\beta : \mathcal{A} \to [0, 1]$.



Intermittent feedback.
Fixed observation probability:

$$\beta(a) = \beta_0.$$

Paid observations (ACNO-MDPs).
Each action $a \in \mathcal{A}$ comes in two forms a^0, a^1 with

$$\beta(a^i) = i \quad \text{with} \quad r(s, a^i) = r(s, a) - i\,c.$$

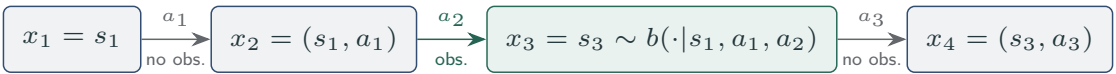
ATST-MDPs $\subset$ POMDPs.
ATST-MDPs are structured POMDPs: full state observations reset uncertainty.

ATST-MDP: Augmented States and Value Function

Augmented states. The last observed state s and the actions $a_1, \dots, a_m$ played since comprise

$$x = (s, a_1, \dots, a_m) \quad \text{from the } \textbf{augmented state space} \quad \mathcal{X} := \mathcal{S} \times \mathcal{A}^{<\mathbb{N}}.$$

Each augmented state x corresponds to a **belief** $b(\cdot|x) \in \Delta(\mathcal{S})$ over the latent current state.



Value function. For an augmented policy $\pi : \mathcal{X} \to \mathcal{A}$ and fixed discount factor $\gamma \in (0, 1)$,

$$V^\pi(x) \;=\; \mathbb{E}_{s_1 \sim b(.|x)}\Big[\textstyle\sum_{h=1}^{\infty} \gamma^{h-1}\, r(s_h, a_h) \,\Big|\, x_1 = x,\, a_h = \pi(x_h) \Big].$$

ATST-MDP: Bellman Equation and Optimal Policy

Tailored Bellman equation. For all $\pi : \mathcal{X} \to \mathcal{A}$ and $x \in \mathcal{X}$,

$$V^\pi(x) \;=\; \underbrace{\mathbb{E}_{s\sim b(\cdot|x)}[r(s,\pi(x))]}_{\text{immediate reward}} + \gamma\beta(\pi(x)) \cdot \underbrace{\mathbb{E}_{s'\sim b(\cdot|x\oplus\pi(x))}[V^\pi(s')]}_{\text{future reward with observation}} + \gamma(1-\beta(\pi(x))) \cdot \underbrace{V^\pi(x\oplus\pi(x))}_{\text{w/o observation}}.$$

The resulting Bellman operator $\mathbb{T} : \mathcal{V} \to \mathcal{V}$ for $\mathcal{V} := \{V : \mathcal{X} \to [0, 1/(1-\gamma)]\}$:

$$(\mathbb{T}V)(x) = \max_{a\in\mathcal{A}} \left\{\mathbb{E}_{s\sim b(\cdot|x)}[r(s,a)] + \gamma\beta(a)\,\mathbb{E}_{s'\sim b(\cdot|x\oplus a)}[V(s')] + \gamma(1-\beta(a))V(x\oplus a)\right\}$$

is a γ-contraction under the sup norm (i.e., $\|\mathbb{T}V_1 - \mathbb{T}V_2\|_\infty \le \gamma\|V_1 - V_2\|_\infty$).

Optimal augmented policy. We prove the existence of $\pi^* : \mathcal{X} \to \mathcal{A}$ such that

$$V^{\pi^*}(x) = V^*(x) := \sup_{\pi\in\mathcal{A}^\mathcal{X}} V^\pi(x) \qquad \text{for all } x \in \mathcal{X}.$$

ATST-MDP: Action-Sequence Paradigm and Linearity

Observation-to-observation (O2O) rollouts. After observing $s \in \mathcal{S}$, commit to an action sequence $\boldsymbol{a} = (a_1, a_2, \ldots) \in \mathcal{A}^{\mathbb{N}}$ until the next observation.

Value of committing to $\boldsymbol{a}$ until the next observation time $\mathcal{T}$ and following $\pi : \mathcal{X} \to \mathcal{A}$ thereafter:

$$K^{\pi}(s, \boldsymbol{a}) = \mathbb{E}\left[\sum_{h=1}^{\mathcal{T}} \gamma^{h-1} r(s_h, a_h) + \gamma^{\mathcal{T}} V^{\pi}(s_{\mathcal{T}+1})\right].$$

Linear MDP. Assume $\mathbb{P}(\cdot|s,a) = \langle \boldsymbol{\phi}(s,a), \boldsymbol{\mu}(\cdot) \rangle$ and $r(s,a) = \langle \boldsymbol{\phi}(s,a), \boldsymbol{\theta} \rangle$ for some $\boldsymbol{\phi} : \mathcal{S} \times \mathcal{A} \to \mathbb{B}^d$.

Theorem (Linearity of K^{π}, informal)

There exists $\psi : \mathcal{S} \times \mathcal{A}^{\mathbb{N}} \to \mathbb{B}^{2d}$ such that, for every $\pi : \mathcal{X} \to \mathcal{A}$, there exists $\mathbf{v}^{\pi} \in \mathbb{R}^{2d}$ satisfying

$$K^{\pi}(s, \boldsymbol{a}) = \langle \psi(s, \boldsymbol{a}), \mathbf{v}^{\pi} \rangle.$$

$K^{*}(s, \boldsymbol{a}) = \langle \psi(s, \boldsymbol{a}), \mathbf{v}^{*} \rangle$ **can be learned using regression-based methods**

Episodic Learning

Episodic setting. The agent interacts with the ATST-MDP for K episodes.

- Episode k starts from observed state s_1^k and lasts for $H^k \sim \mathrm{Geom}(1-\gamma)$ steps.
- Regret is measured against the best augmented policy:

$$\mathcal{R}_K = \sum_{k=1}^{K} \left(V^*(s_1^k) - V^k(s_1^k) \right),$$

where $V^k(s_1^k)$ is the expected return obtained by the learning algorithm in episode k.

Reduction. Each O2O rollout is one transition in an induced MDP on $(\mathcal{S}, \mathcal{A}^{\mathbb{N}})$:

state $s \in \mathcal{S}$, meta-action $\boldsymbol{a} \in \mathcal{A}^{\mathbb{N}}$, $(s, \boldsymbol{a}) \rightsquigarrow (s_N, R)$, $(s, \boldsymbol{a}, s_N, R)$ collected.

This induced MDP is linear for the **action-sequence feature map** $\psi : \mathcal{S} \times \mathcal{A}^{\mathbb{N}} \to \mathbb{B}^{2d}$.

Episodic Learning: Algorithm and Guarantee

We adapt LSVI-UCB of [Jin+20] to the induced Linear MDP $(\mathcal{S}, \mathcal{A}^{\mathbb{N}})$ with feature map $\boldsymbol{\psi}(s, \boldsymbol{a})$.

Planning Phase (before episode k)

Construct optimistic estimates $(K_u^k)_{u=1}^{\infty}$ for K^*.
For $u > H$, set $K_u^k = \frac{1}{1-\gamma}$.
For $u \leq H$, fit K_u^k on O2O tuples $(s, \boldsymbol{a}, s_N, R)$:

$$K_u^k(s, \boldsymbol{a}) \approx R + \max_{\boldsymbol{a}_N \in \mathcal{A}^{\mathbb{N}}} K_{u+1}^k(s_N, \boldsymbol{a}_N).$$

Execution Phase (during episode k)

Upon the u^{th} state observation s_u^k, commit to

$$\boldsymbol{a}_u^k \in \operatorname*{argmax}_{\boldsymbol{a} \in \mathcal{A}^{\mathbb{N}}} K_u^k(s_u^k, \boldsymbol{a})$$

until the next observation or termination.

Regret guarantee. Our algorithm guarantees

$$\mathcal{R}_K = \widetilde{O}\Big(\sqrt{K d^3 (1-\gamma)^{-3}}\Big).$$

Conclusion and Future Directions

Future work.

1. Delayed state observations.
2. State-dependent observation mechanisms.
3. Budget-constrained sensing.

Thank you!

References

[Che+23] M. Chen, Y. Bai, H. V. Poor, and M. Wang. **"Efficient RL with Impaired Observability: Learning to Act with Delayed and Missing State Observations".** In: *Advances in Neural Information Processing Systems*. Vol. 36. 2023.

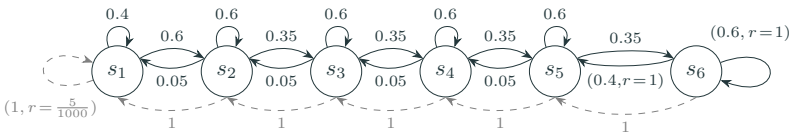
[Jin+20] C. Jin, Z. Yang, Z. Wang, and M. I. Jordan. **"Provably Efficient Reinforcement Learning with Linear Function Approximation".** In: *Proceedings of the 33rd Conference on Learning Theory (COLT)*. Vol. 125. Proceedings of Machine Learning Research. PMLR, 2020, pp. 2137–2143.

[NFB21] H. A. Nam, S. L. Fleming, and E. Brunskill. **"Reinforcement learning with state observation costs in action-contingent noiselessly observable Markov decision processes".** In: *Advances in Neural Information Processing Systems 34 (NeurIPS)*. 2021.

[Sze10] C. Szepesvári. **Algorithms for Reinforcement Learning.** Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool, 2010.

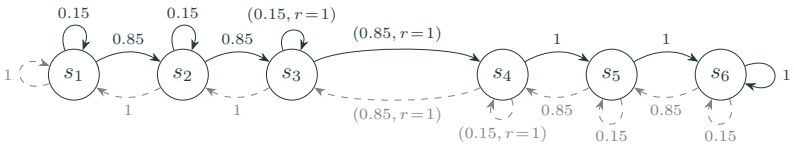
Appendix: Simulation Environments

RiverSwim



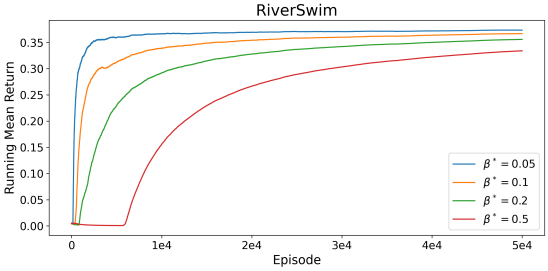
Hard exploration: large reward at s_6, small reward at s_1, current pushes back.

RiverBalance

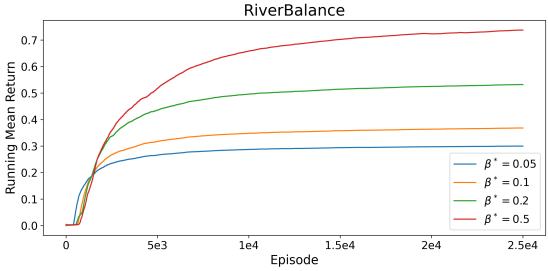


Stay near the middle: current pulls the agent away from rewarding s_3, s_4.

Appendix: Simulation Results



Sparse observations encourage longer open-loop commitments and faster exploration.



Frequent observations help when high reward requires state-dependent corrections.

Takeaway. The value of observing often is task-dependent.