#370: No-Regret Learning with High-Probability in Adversarial Markov Decision Processes

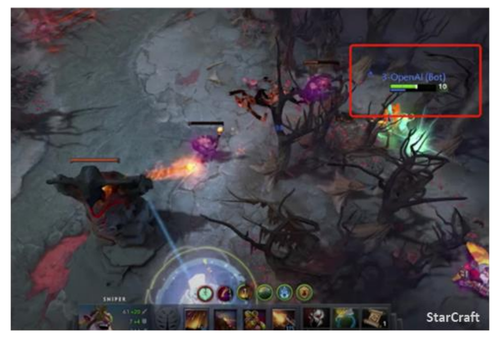
INSTITUTE FOR COMPUTATIONAL ENGINEERING & SCIENCES — ODEN

Mahsa Ghasemi*, Abolfazl Hashemi*, Haris Vikalo, Ufuk Topcu

The University of Texas at Austin
Electrical and Computer Engineering
Cockrell School of Engineering

SEQUENTIAL DECISION MAKING

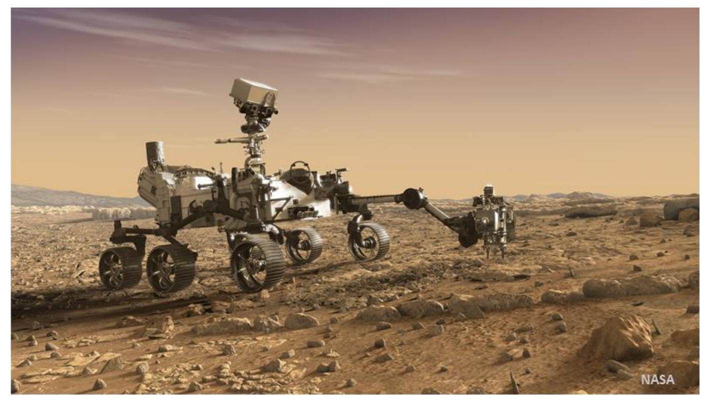
Sequential Interaction with the environment



Learning from a fixed reward

Offline: access to a lot of data

SEQUENTIAL DECISION MAKING WITH VARYING TASKS



Evolving environment and task

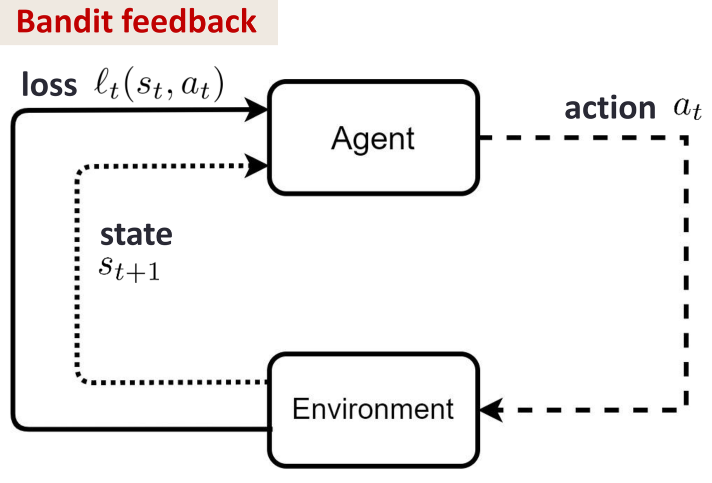
Safety-critical operation

Limited feedback from the environment

How can we design online algorithms with high probability guarantees for varying tasks?

ONLINE LEARNING FOR MDPS

Bandit feedback

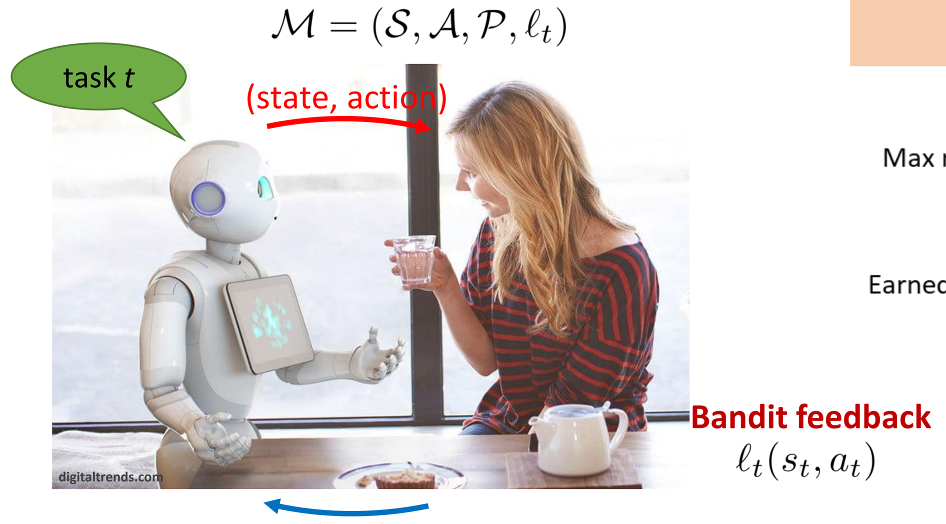


Uniform ergodicity:
For every policy over the MDP, the convergence rate of state distributions to a unique stationary distribution is exponentially fast.

$$\|\nu_1 \mathcal{P}^\pi - \nu_2 \mathcal{P}^\pi\|_1 \le e^{-\frac{1}{\tau}} \|\nu_1 - \nu_2\|_1$$

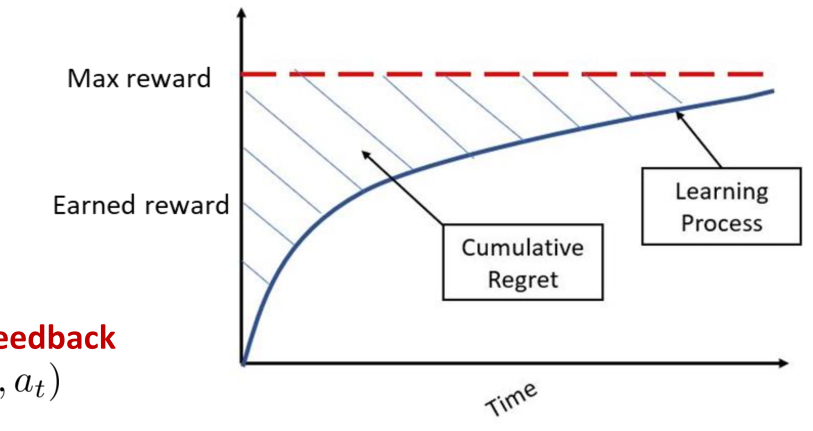
Unknown and time-varying loss function (A-MDP)

$$\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, \ell_t)$$



Learn a policy with sublinear regret:

$$\mathcal{R}_T := \max_\pi \mathcal{L}_T - \mathcal{L}_T(\pi)$$



LOSS ESTIMATION

Bandit feedback → Estimating the loss of all state-action pairs

Goal: Obtain a low-variance loss estimator

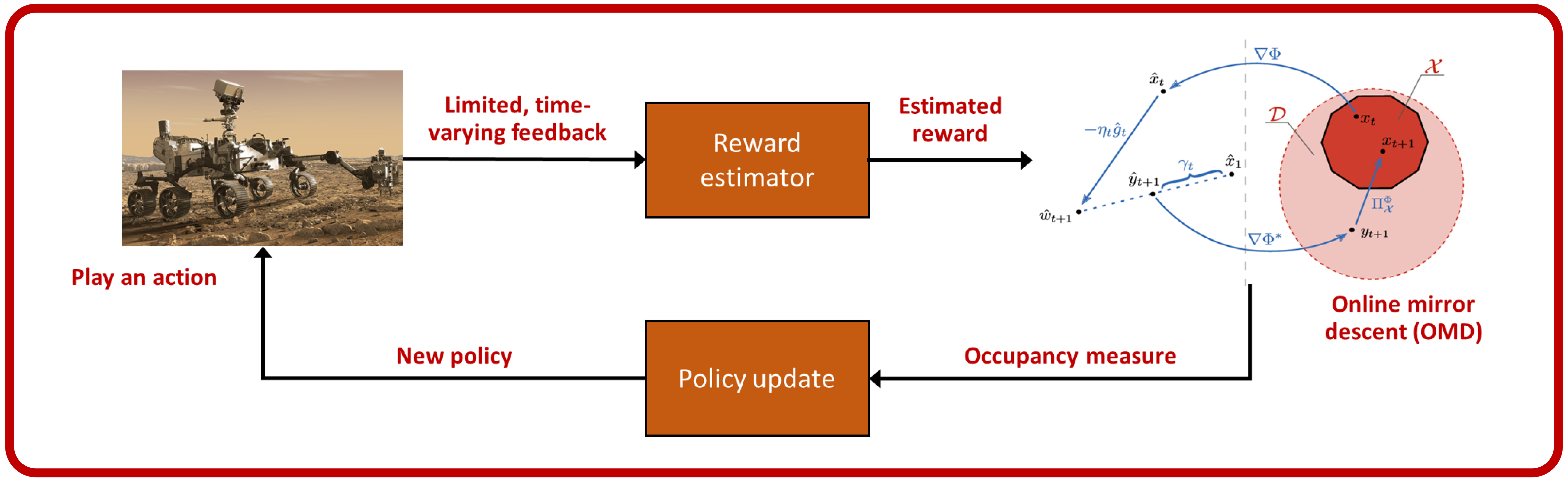
A novel optimistically biased estimator for the loss function:

$$\hat{\boldsymbol{\ell}}_t(s,a) := \frac{\ell_t(s,a)}{\underbrace{\boldsymbol{\nu}_{t|t-N}(s)}_{\text{moving-window estimate of state distribution}}\boldsymbol{\pi}_t(a|s) + \underbrace{\gamma}_{\text{exploration parameter}}} \mathbb{I}\{\boldsymbol{s}_t = s, \boldsymbol{a}_t = a\}$$

Optimistically biased → Implicit exploration

$$\mathbb{E}\left[\hat{\boldsymbol{\ell}}_t(s,a) | t-N\right] \le \ell_t(s,a)$$

Estimation-window parameter *N* delays the policy update which leads to lower variance of the random regret.



POLICY OPTIMIZATION VIA OMD

Goal: Compute a new policy from the estimated loss function

An OMD algorithm utilizing the proposed loss estimator:

$$\boldsymbol{\rho}_{t+1} = \arg\min_{\rho \in \Delta(\mathcal{M})} \left\{ \overset{\text{learning rate}}{\eta} \underbrace{\langle \rho, \hat{\boldsymbol{\ell}}_t \rangle}_{\text{loss}} + \underbrace{\overset{\text{unnormalized KL divergence}}{D}(\rho \| \boldsymbol{\rho}_t)}_{\text{policy change}} \right\}$$

Constrained optimization → Two-step procedure

$$\tilde{\boldsymbol{\rho}}_{t+1} = \arg\min_{\rho} \left\{ \eta \langle \rho, \hat{\boldsymbol{\ell}}_t \rangle + D(\rho \| \boldsymbol{\rho}_t) \right\}$$

$$\boldsymbol{\rho}_{t+1} = \arg\min_{\rho \in \Delta(\mathcal{M})} \{ D(\rho \| \tilde{\boldsymbol{\rho}}_{t+1}) \}$$

REGRET BOUND

Result: Establishing sublinear regret bounds both on expectation and with high-probability

Theorem: (high-probability regret bound for uniformly ergodic A-MDP)

Let $\delta \in (0,1)$. With probability at least $1-\delta$,

$$\text{regret} \le C \underbrace{T}_{\text{time horizon}}^{\frac{2}{3}} \underbrace{\tau}_{\text{mixing time}}^{\frac{1}{2}} |\mathcal{S}|^{\frac{2}{3}} |\mathcal{A}|^{\frac{2}{3}} \sqrt{\log(\underbrace{|\mathcal{S}|}_{\text{number of states}} \underbrace{|\mathcal{A}|}_{\text{number of actions}}) \log T \log \frac{1}{\delta}} + C' \tau \log T.$$

CONCLUSION

- Proposed an optimistic loss estimator for learning in episodic A-MDP under bandit feedback
- Developed an OMD policy optimization utilizing the proposed loss estimator
- Established a sublinear regret bound with high probability

Future Directions

- Parameter-free and anytime algorithms
- Unknown, time-varying dynamics and large-scale state spaces
- Structure-aware and game-theoretic online learning