INSTITUTE FOR COMPUTATIONAL ENGINEERING & SCIENCES ODEN

autonomous SYSTEMS GROUP

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

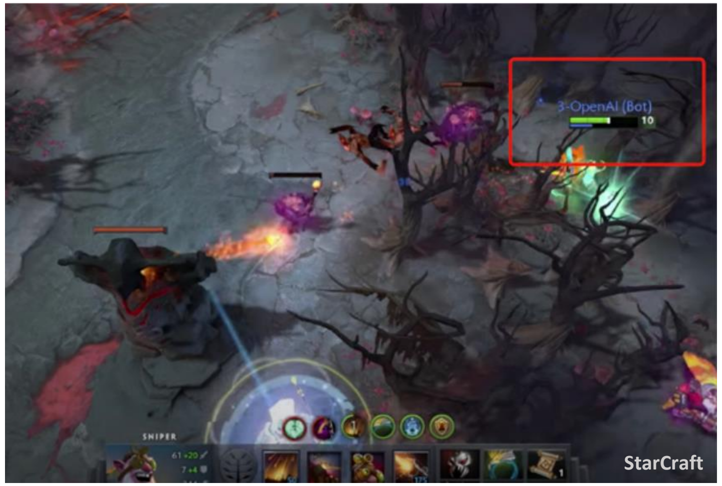
No-Regret Learning with High-Probability in Adversarial Markov Decision Processes

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

Uncertainty in Artificial Intelligence (UAI)
July 27th - July 29th, 2021

Sequential Decision Making

Sequential Interaction with the environment

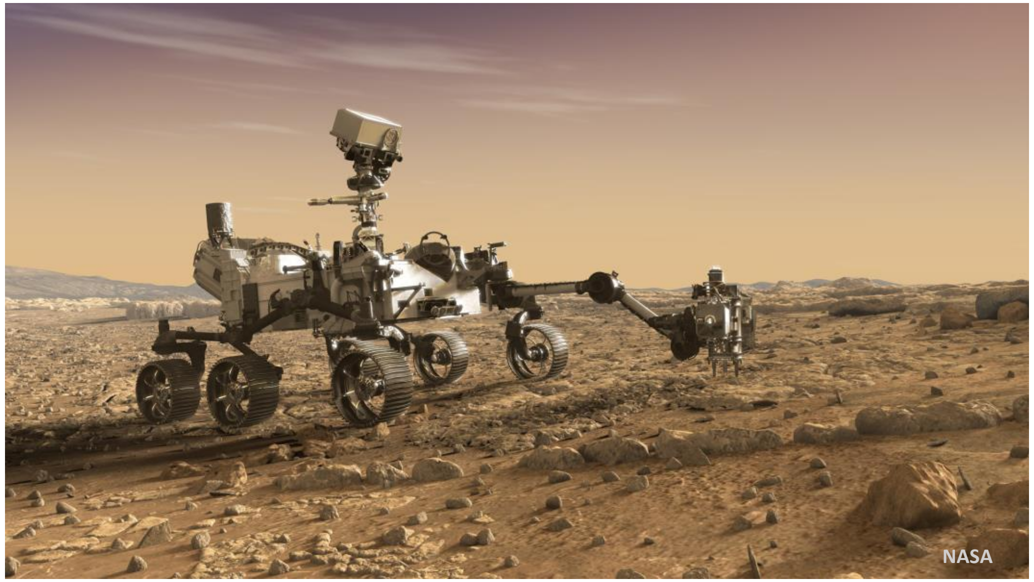


Learning from a fixed reward

Offline: access to a lot of data

KUKA

Sequential Decision Making with Varying Tasks



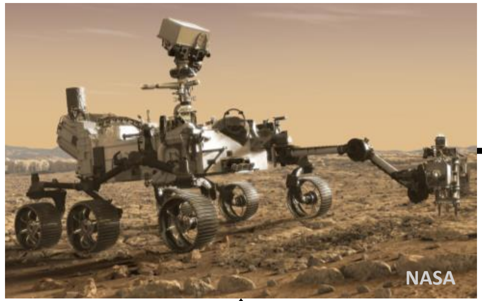
Evolving environment and task

Safety-critical operation

Limited feedback from the environment

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

Online Policy Learning with Implicit Exploration

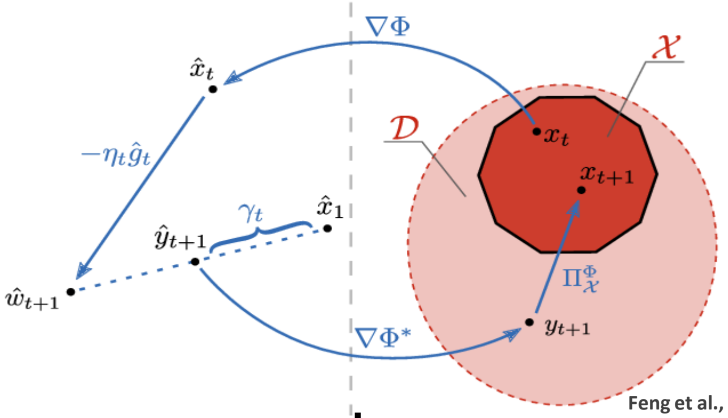


Play an action

Limited, time-varying feedback

Reward estimator

Estimated reward



Feng et al., 2020

Online mirror descent (OMD)

Occupancy measure

Policy update

New policy

Contributions:

- A novel optimistically-biased reward estimator for implicit exploration
- Policy search using online mirror descent (OMD)
- Sublinear regret bound with high probability

Adversarial Markov Decision Process (A-MDP)

finite state space

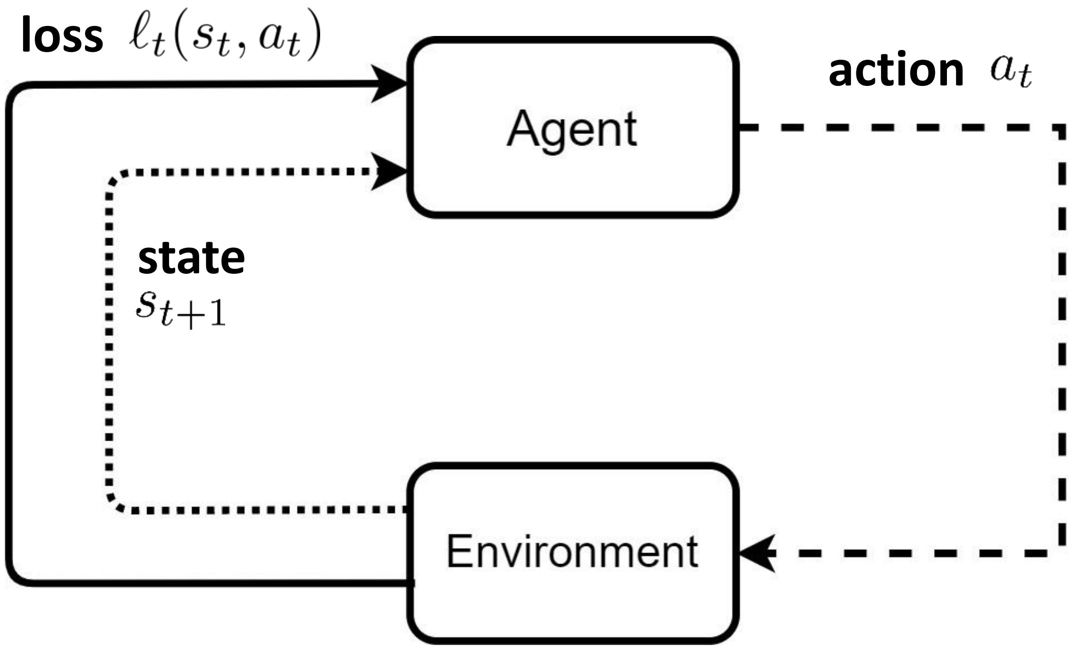
finite action space

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

probabilistic transition function

time-varying loss function

Bandit feedback



Uniform ergodicity:

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

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

Agent's Policy Representation via Occupancy Measure

Looking for a time-varying stochastic policy $\pi_t : \mathcal{S} \times \mathcal{A} \to [0, 1]$

Occupancy measure: the probability induced over state-action pairs by executing a policy, asymptotically.

$$\rho^{\pi}(s, a) = \lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} \Pr(\mathbf{s}_t = s, \mathbf{a}_t = a | \pi)$$

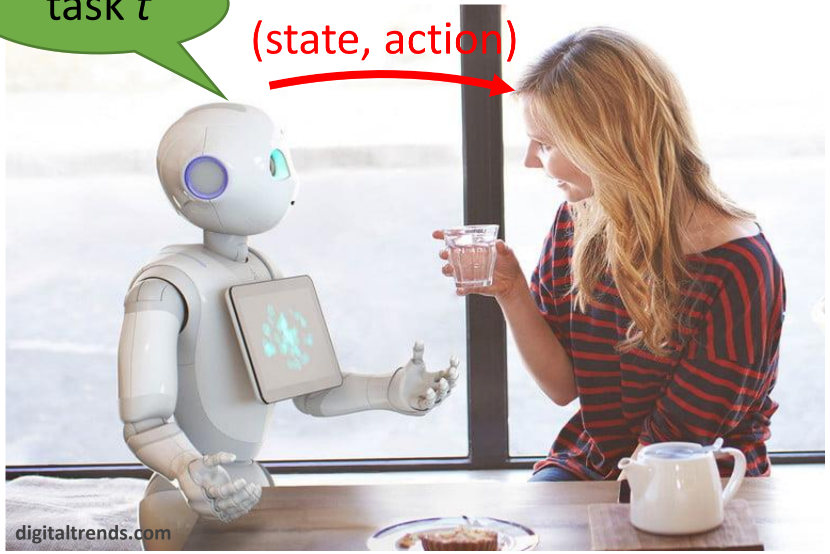
Stochastic stationary policy given an occupancy measure

$$\pi^{\rho}(a|s) = \frac{\rho(s, a)}{\sum_{a' \in \mathcal{A}} \rho(s, a')}, \quad \forall (s, a) \in \mathcal{S} \times \mathcal{A}$$

Regret Minimization

task *t*

(state, action)



(loss)

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

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

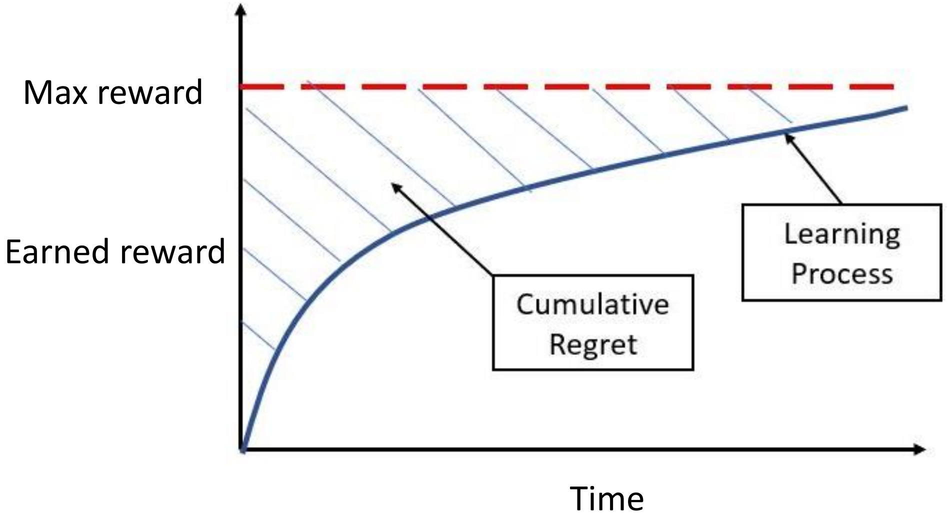
Bandit feedback

$$\ell_t(s_t, a_t)$$

Learn a policy with sublinear regret:

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

best fixed policy in hindsight



Question: Can we obtain low regret with high probability?

Optimistic Loss Estimator

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:

$$\hat{\boldsymbol{\ell}}_t(s,a) := \frac{\ell_t(s,a)}{\boldsymbol{\nu}_{t|t-N}(s)\boldsymbol{\pi}_t(a|s)+\gamma}\mathbb{I}\{\boldsymbol{s}_t = s, \boldsymbol{a}_t = a\}$$

moving-window estimate of state distribution

exploration parameter

Optimistically biased

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

→ Implicit exploration

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

Policy Optimization via Online Mirror Descent

Goal: Compute a new policy from the estimated loss function

An OMD algorithm utilizing the proposed loss estimator:

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

Constrained optimization $\longrightarrow$ Two-step procedure

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

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

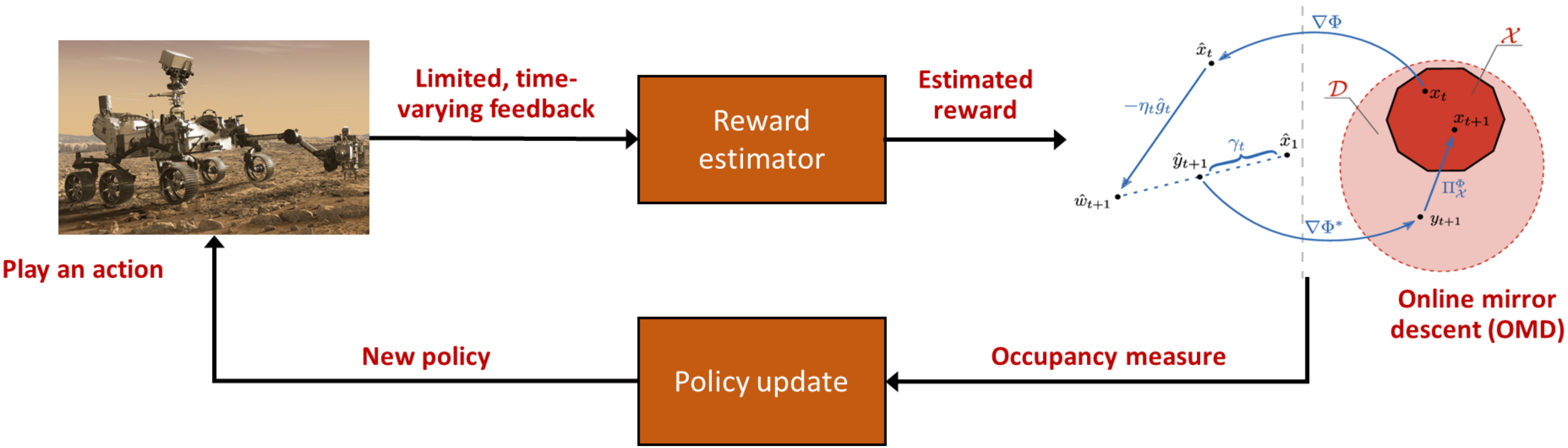
No-Regret Learning with High-Probability

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 $1-\delta$,

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



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

Mahsa Ghasemi, Abolfazl Hashemi, Haris Vikalo, Ufuk Topcu

supported in part by NSF ECCS grant 1809327, DARPA grant D19AP00004, and AFRL grant FA9550-19-1-0169