INSTITUTE FOR COMPUTATIONAL ENGINEERING & SCIENCES
ODEN

autonomous
SYSTEMS GROUP

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

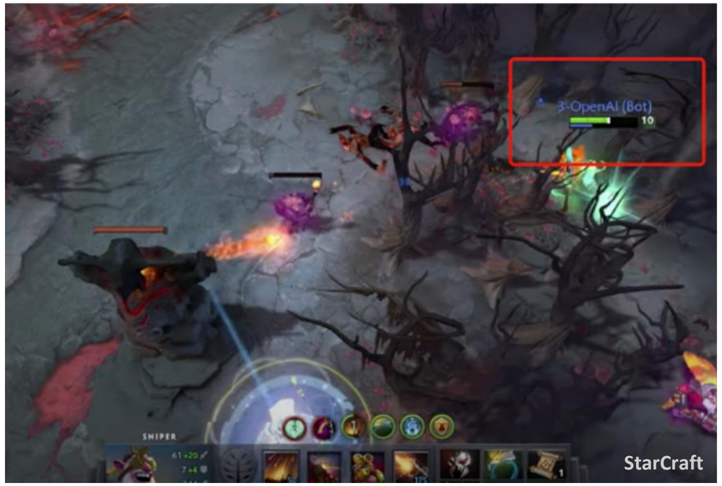
Online Policy Learning for Unknown and Varying Tasks in Adversarial Environments

Mahsa Ghasemi, Abolfazl Hashemi, Haris Vikalo, Ufuk Topcu

CoE Review

April 30th, 2021

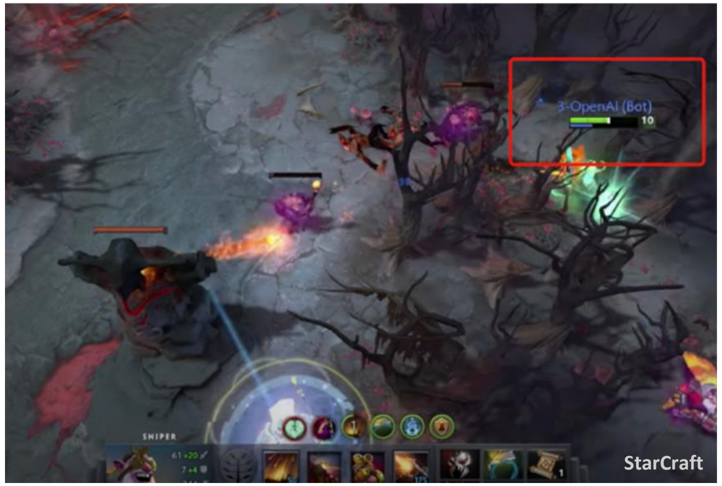
Sequential Decision Making



KUKA

Sequential Decision Making

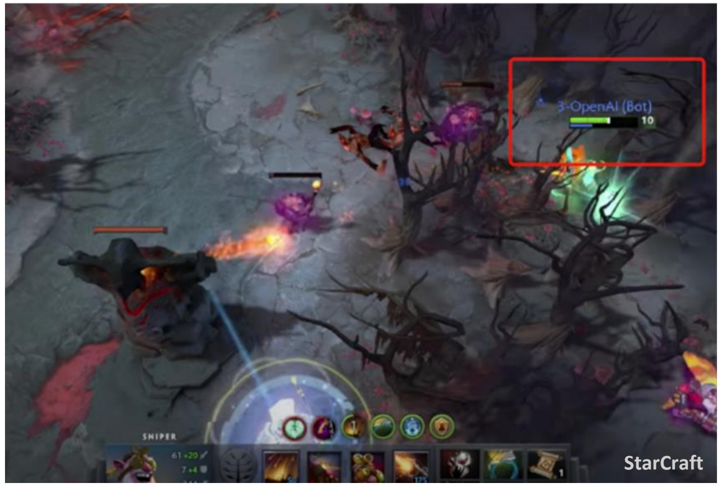
Sequential Interaction with the environment



KUKA

Sequential Decision Making

Sequential Interaction with the environment



Learning from a fixed reward

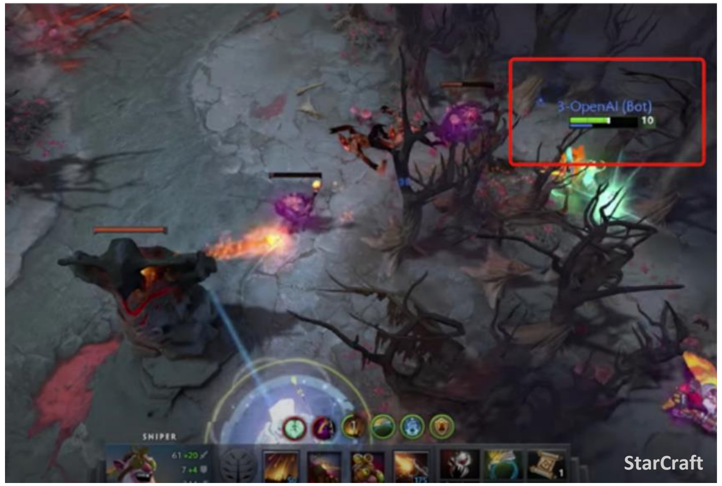
KUKA

Sequential Decision Making

Sequential Interaction with the environment

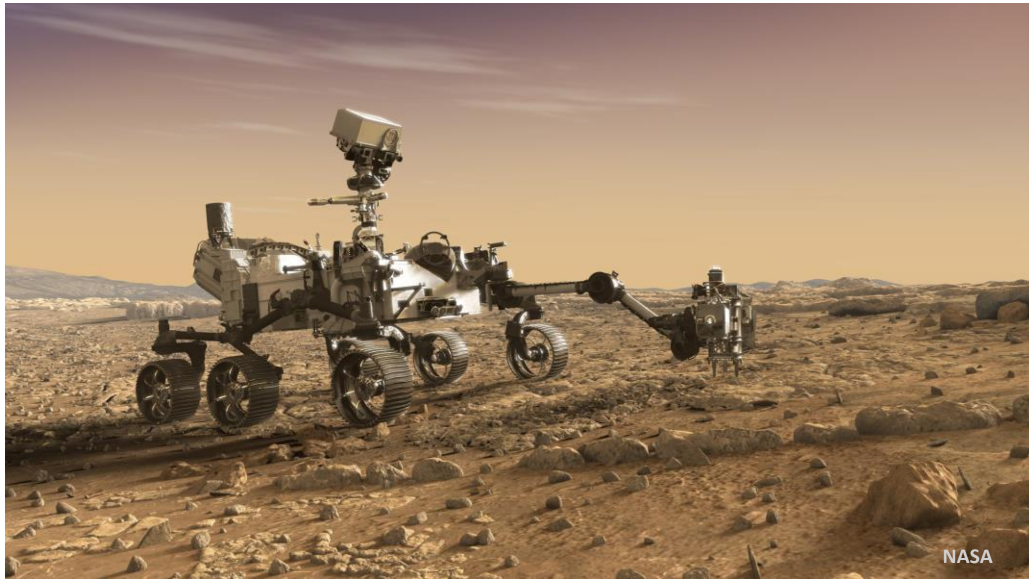
Learning from a fixed reward

Offline: access to a lot of data

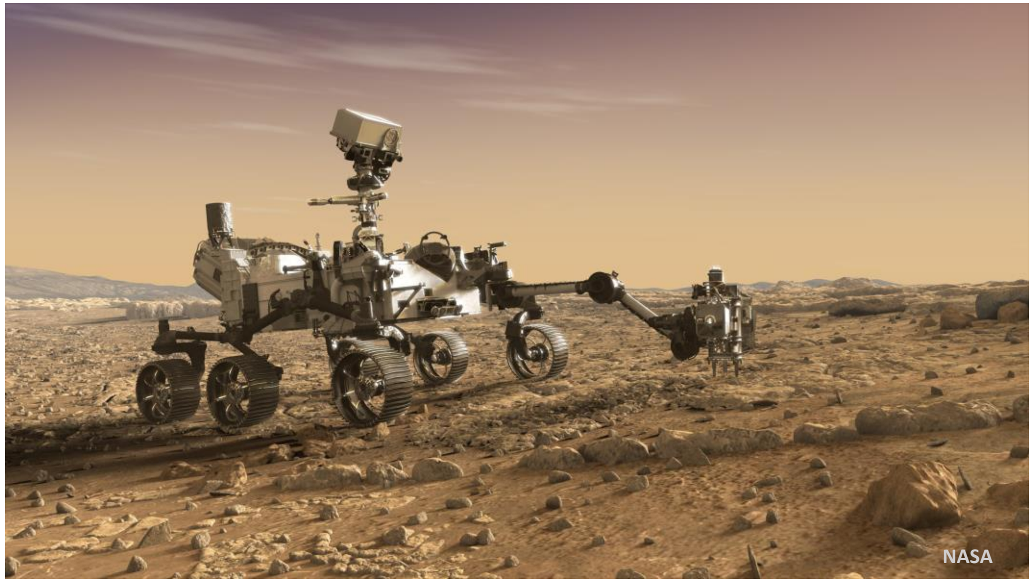


KUKA

Sequential Decision Making with Varying Tasks

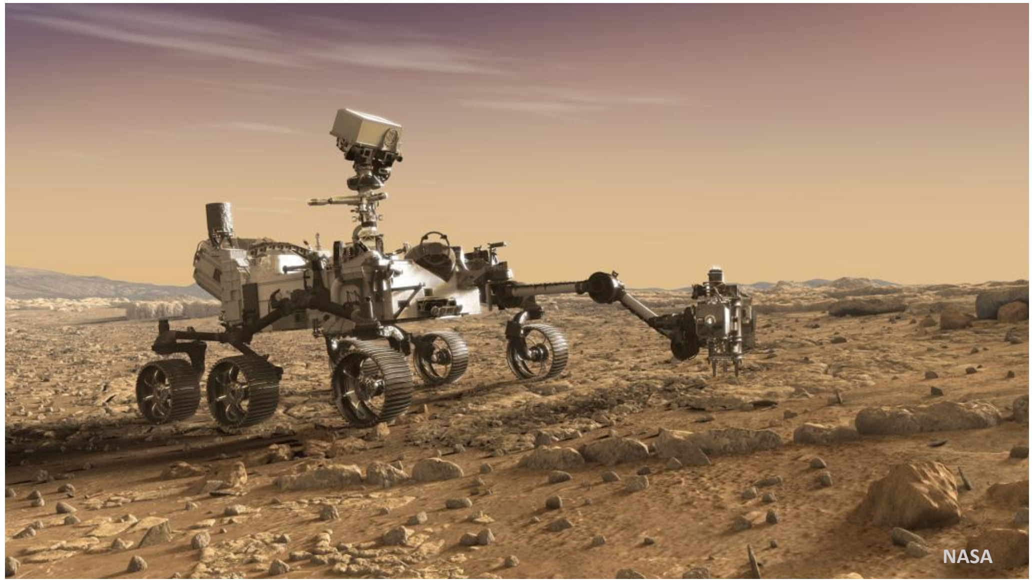


Sequential Decision Making with Varying Tasks



Evolving environment and task

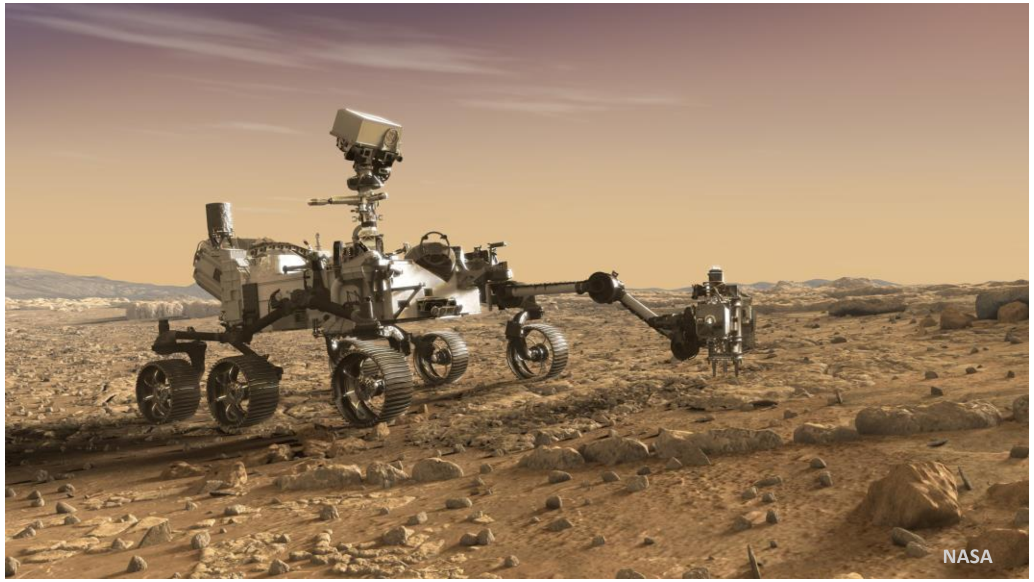
Sequential Decision Making with Varying Tasks



Evolving environment and task

Safety-critical operation

Sequential Decision Making with Varying Tasks

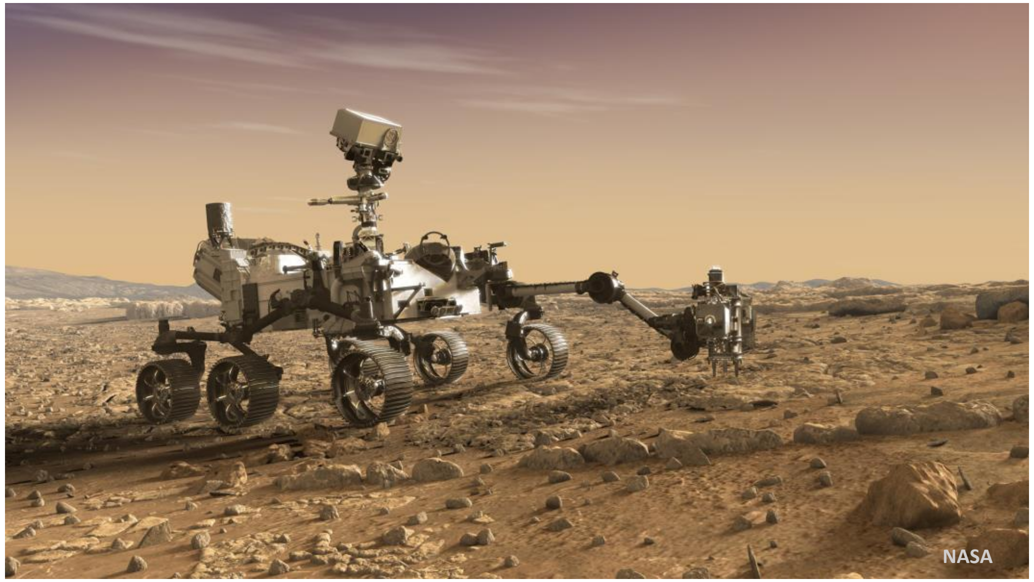


Evolving environment and task

Safety-critical operation

Limited feedback from the environment

Sequential Decision Making with Varying Tasks



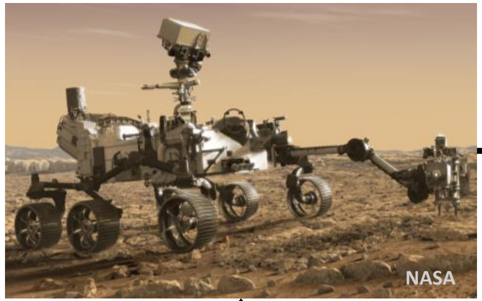
Evolving environment and task

Safety-critical operation

Limited feedback from the environment

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

Online Learning with Implicit Exploration for Varying Tasks

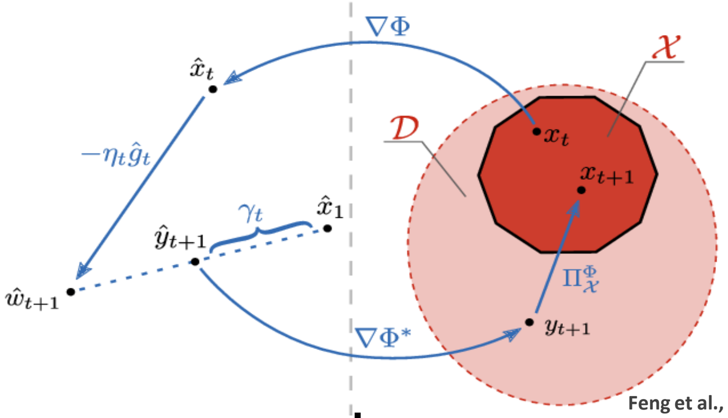


Play an action

Limited, time-varying feedback

Reward estimator

Estimated reward



Feng et al., 2020

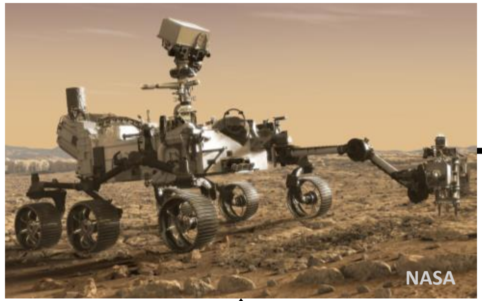
Online mirror descent (OMD)

Occupancy measure

Policy update

New policy

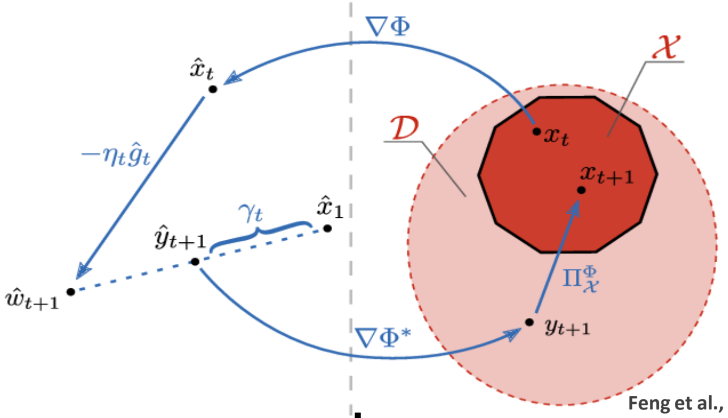
Online Learning with Implicit Exploration for Varying Tasks



Limited, time-varying feedback

Reward estimator

Estimated reward



Feng et al., 2020

Online mirror descent (OMD)

Play an action

New policy

Policy update

Occupancy measure

Contributions:

- A novel optimistically-biased reward estimator for implicit exploration
- Policy search using online mirror descent (OMD)
- Minimax optimal 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

Adversarial Markov Decision Process (A-MDP)

finite state space

finite action space

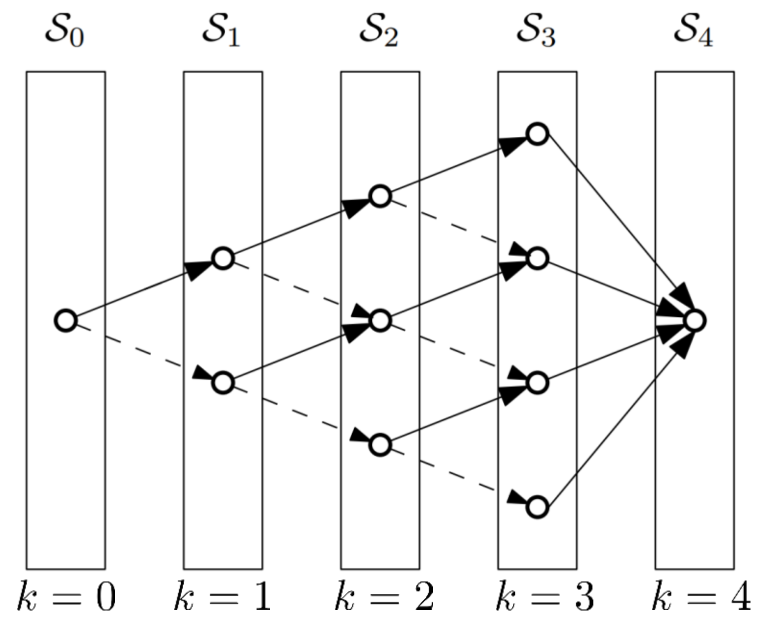
$$\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, \ell_t)$$

probabilistic transition function

time-varying loss function

Loop-free episodic A-MDP:

- States are partitioned into layers
- Transition only exists from one layer to the next



Neu et al., 2020

Adversarial Markov Decision Process (A-MDP)

finite state space

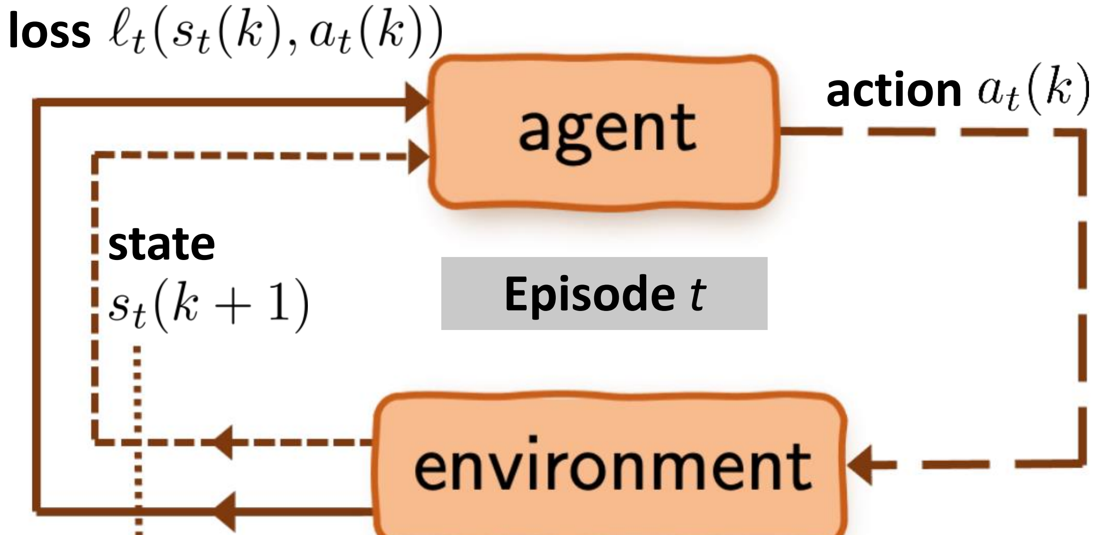
finite action space

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

probabilistic transition function

time-varying loss function

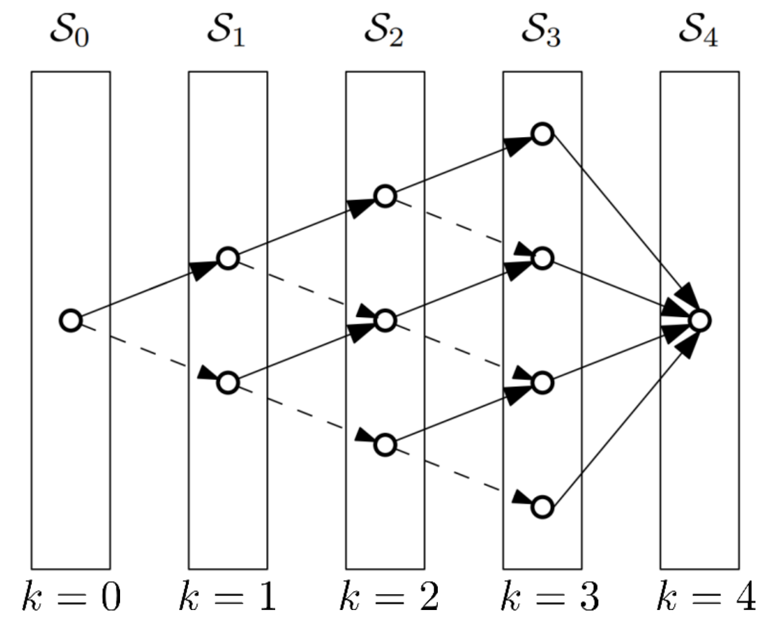
Bandit feedback



Chen et al., 2019

Loop-free episodic A-MDP:

- States are partitioned into layers
- Transition only exists from one layer to the next



Neu et al., 2020

Agent's Policy Representation via Occupancy Measure

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

Agent's Policy Representation via Occupancy Measure

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

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

$$\rho^{\pi}(s, a) = \Pr(\mathbf{s}_{k(s)} = s, \mathbf{a}_{k(s)} = a | \pi)$$

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

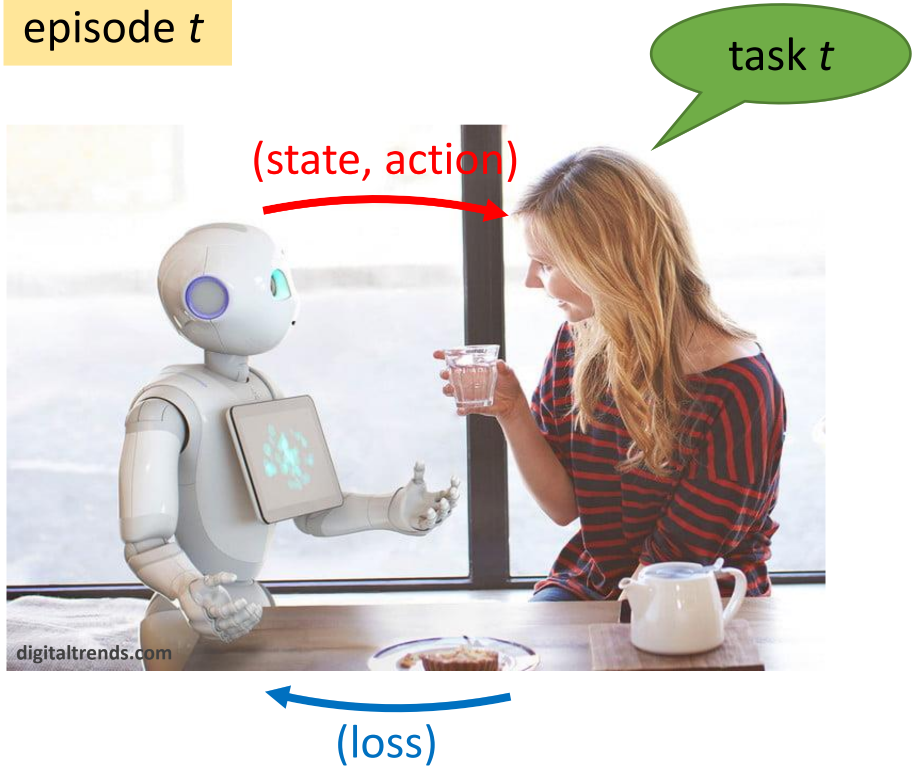
$$\rho^{\pi}(s, a) = \Pr(\mathbf{s}_{k(s)} = s, \mathbf{a}_{k(s)} = a | \pi)$$

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

episode t

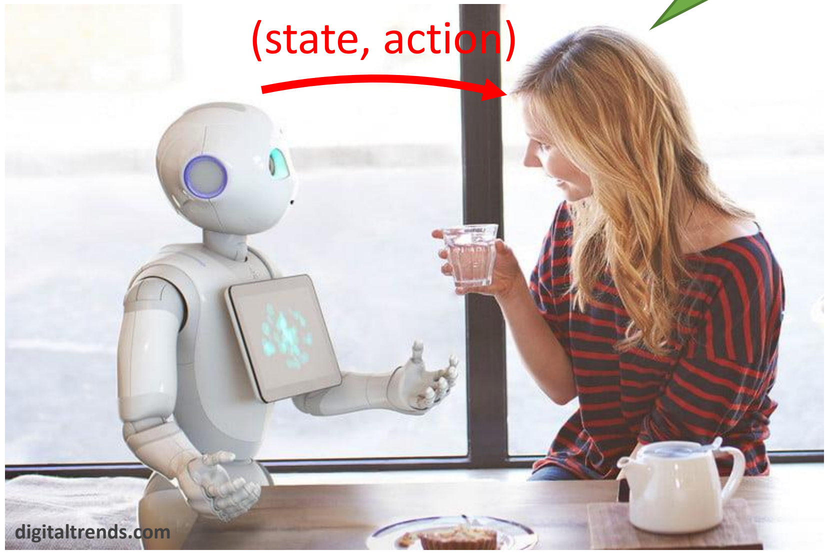


Regret Minimization

episode t

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

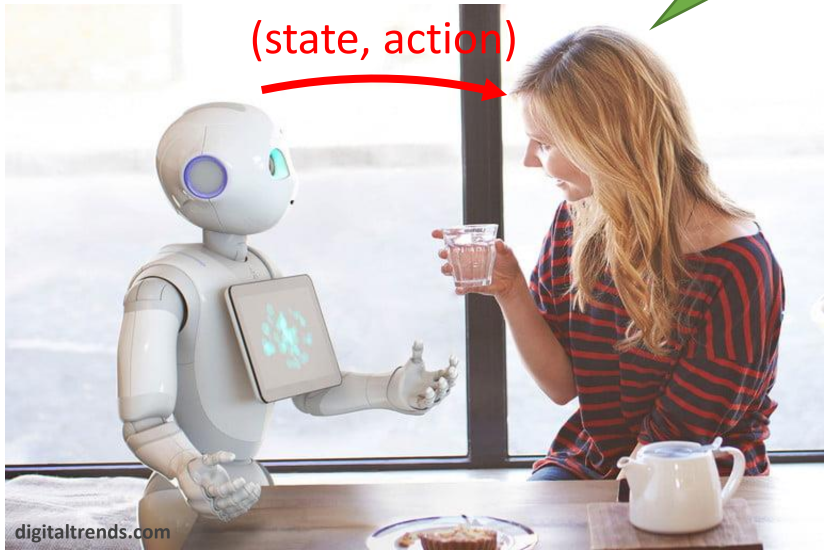
$$\ell_t(s_t(k), a_t(k))$$

Regret Minimization

episode t

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

$$\ell_t(s_t(k), a_t(k))$$

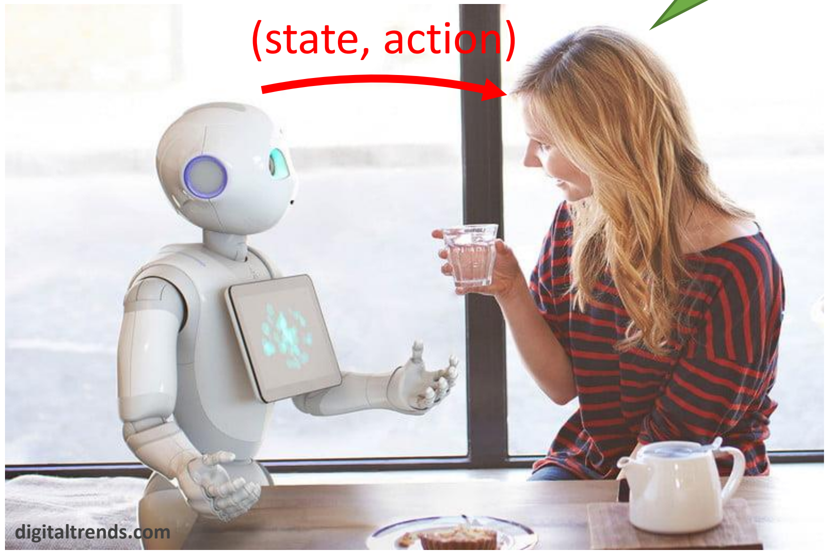
Learn a policy with sublinear regret:

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

best fixed policy in hindsight

Regret Minimization

episode *t*



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

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

Bandit feedback

$$\ell_t(s_t(k), a_t(k))$$

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

Optimistic Loss Estimator

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

Goal: Obtain a low-variance loss estimator

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:

history at current episode

$$\hat{\boldsymbol{\ell}}_t(s,a) = \frac{\ell_t(s,a)}{\boldsymbol{\rho}_t(s,a) + \gamma}\mathbb{I}\{(s,a) \in \mathbf{h}(t)\}$$

exploration parameter

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:

$$\hat{\boldsymbol{\ell}}_t(s,a) = \frac{\ell_t(s,a)}{\boldsymbol{\rho}_t(s,a) + \gamma}\mathbb{I}\{(s,a) \in \mathbf{h}(t)\}$$

history at current episode

exploration parameter

Optimistically biased

$$\mathbb{E}\left[\hat{\boldsymbol{\ell}}_t(s,a)|\mathbf{h}(t-1)\right] \leq \ell_t(s,a)$$

→ Implicit exploration

Policy Optimization via Online Mirror Descent

Goal: Compute a new policy from the estimated loss function

Policy Optimization via Online Mirror Descent

Goal: Compute a new policy from the estimated loss function

An OMD algorithm utilizing the proposed loss estimator:

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

Policy Optimization via Online Mirror Descent

Goal: Compute a new policy from the estimated loss function

An OMD algorithm utilizing the proposed loss estimator:

learning rate　　　　　　unnormalized KL divergence

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

Constrained optimization → Two-step procedure

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

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

No-Regret Learning with High-Probability

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

No-Regret Learning with High-Probability

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

Theorem: (high-probability regret bound)

Let $\delta \in (0,1)$. If

$$\eta = \gamma = \sqrt{L \frac{\log(|\mathcal{S}||\mathcal{A}|/L)}{2T|\mathcal{S}||\mathcal{A}|}},$$

with probability at least $1-\delta$,

$$\text{regret} \leq \mathcal{O}(\sqrt{LT|\mathcal{A}||\mathcal{S}|\log(|\mathcal{S}||\mathcal{A}|/L)}\log\tfrac{1}{\delta}).$$

episode length number of episodes number of states number of actions

No-Regret Learning with High-Probability

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

Theorem: (high-probability regret bound)

Let $\delta \in (0,1)$. If

$$\eta = \gamma = \sqrt{L\frac{\log(|\mathcal{S}||\mathcal{A}|/L)}{2T|\mathcal{S}||\mathcal{A}|}},$$

with probability at least $1-\delta$,

$$\text{regret} \leq \mathcal{O}(\sqrt{LT|\mathcal{A}||\mathcal{S}|\log(|\mathcal{S}||\mathcal{A}|/L)}\log\tfrac{1}{\delta}).$$

episode length number of episodes number of states number of actions

Minimax optimal regret (up to logarithmic terms)

No-Regret Learning for Uniformly Ergodic MDPs

No-Regret Learning for Uniformly Ergodic MDPs

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

No-Regret Learning for Uniformly Ergodic MDPs

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

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

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

$$\text{regret} \leq CT^{\frac{2}{3}}\tau^{\frac{2}{3}}|\mathcal{S}|^{\frac{2}{3}}|\mathcal{A}|^{\frac{2}{3}}\sqrt{\log(|\mathcal{S}||\mathcal{A}|)\log T\log\tfrac{1}{\delta}} + C'\tau\log T.$$

time horizon   mixing time   number of states   number of actions

Conclusion and Future Work

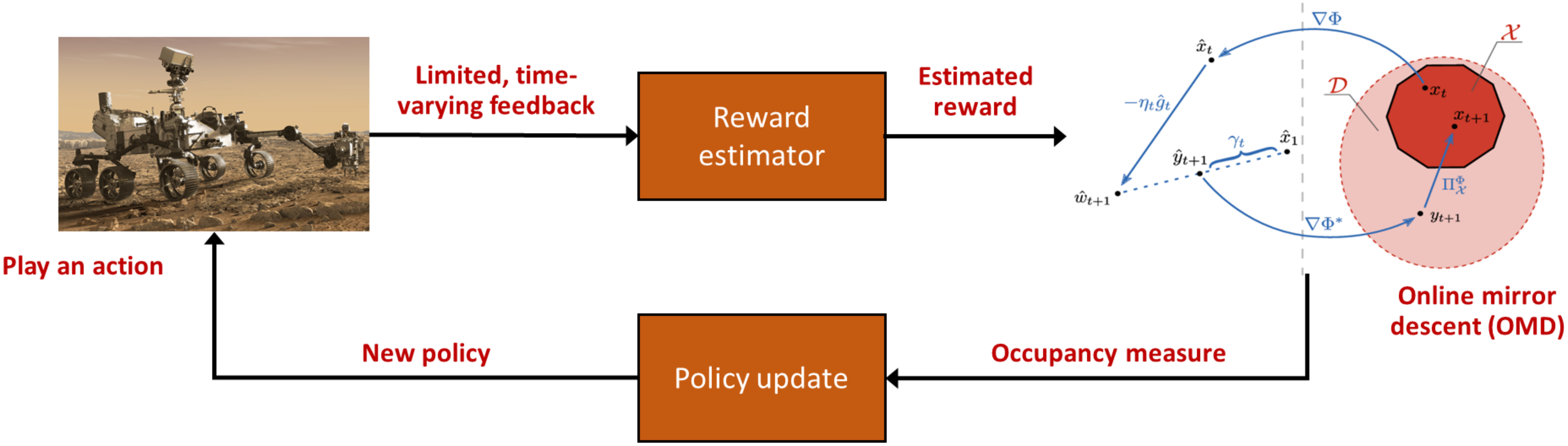
- Studied the problem of learning unknown and varying tasks in adversarial environments
- Proposed an online learning framework that achieves a minimax optimal regret bound with high probability
- Extended our framework to the class of general A-MDPs

Conclusion and Future Work

- Studied the problem of learning unknown and varying tasks in adversarial environments
- Proposed an online learning framework that achieves a minimax optimal regret bound with high probability
- Extended our framework to the class of general A-MDPs

Future Directions

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



Online Policy Learning for Unknown and Varying Tasks in Adversarial Environments

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