Multi-task bandit and reinforcement learning through heterogeneous feedback aggregation

Chicheng Zhang

Department of Computer Science

THE UNIVERSITY OF ARIZONA

Joint work with Zhi Wang, Manish Kumar Singh, Laurel Riek, and Kamalika Chaudhuri (UC San Diego)

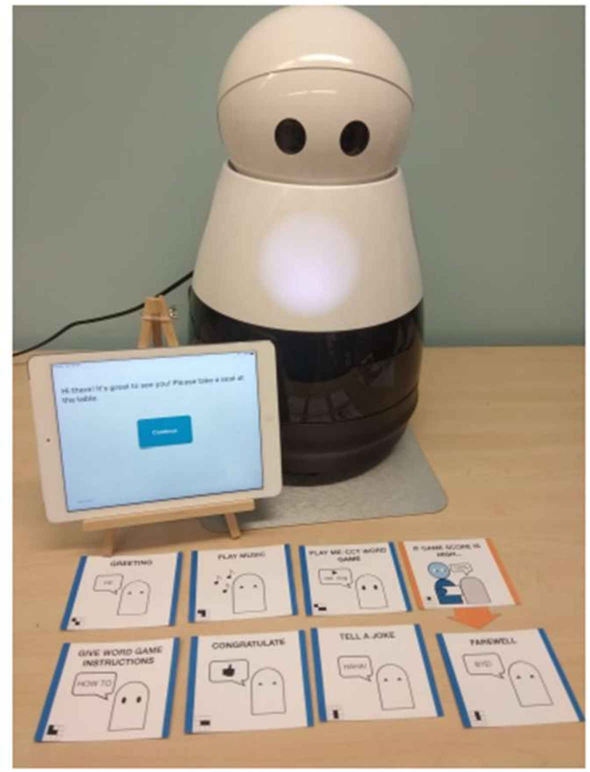
UA Math 586B presentation

Outline

- Motivation
- The ε-multiplayer multi-armed bandit problem
- Our algorithms: Upper Confidence Bound and Thompson Sampling
- Experimental evaluation
- The ε-multiplayer episodic reinforcement learning problem

Motivation 1: healthcare robotics (Kubota et al., 2020)

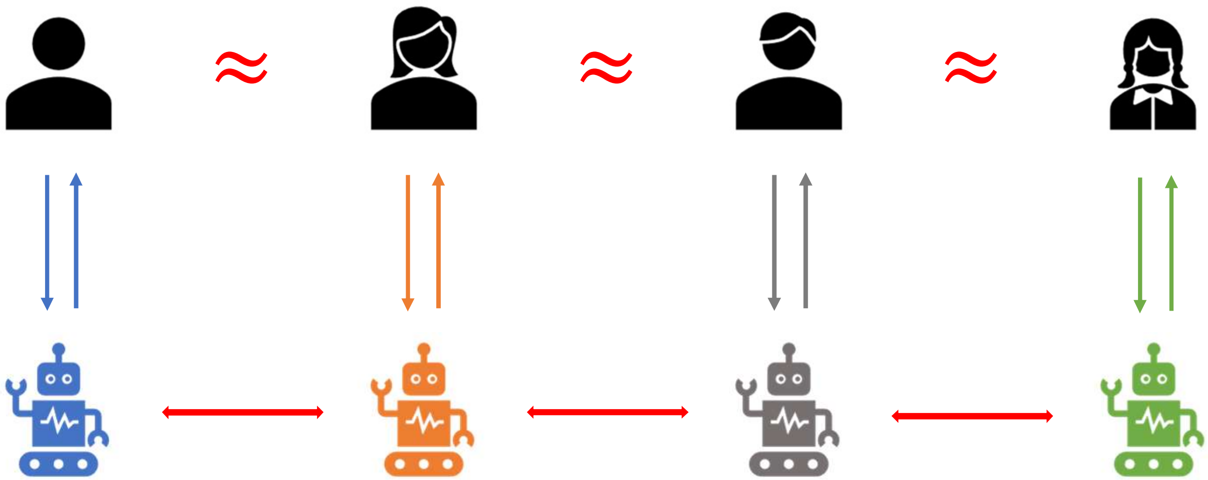
- A group of assistive robots deployed to provide personalized healthcare services.
- Robots can recommend cognitive training activities to patients
  - E.g. chess, maze, puzzle…
- Goal: recommend activities that satisfy all patients' preferences



https://cseweb.ucsd.edu/~lriek/papers/kubota-peterson-rajendren-kress-gazit-riek-hri20.pdf

Motivation 1: healthcare robotics (Kubota et al., 2020)

Each robot learns the preferences of its paired individuals through interactions.



- Question: If the robots receive similar yet nonidentical feedback, how can they cooperatively learn to perform their respective tasks well online?

Motivation 2: movie recommendation (e.g. Qian et al, 2013)

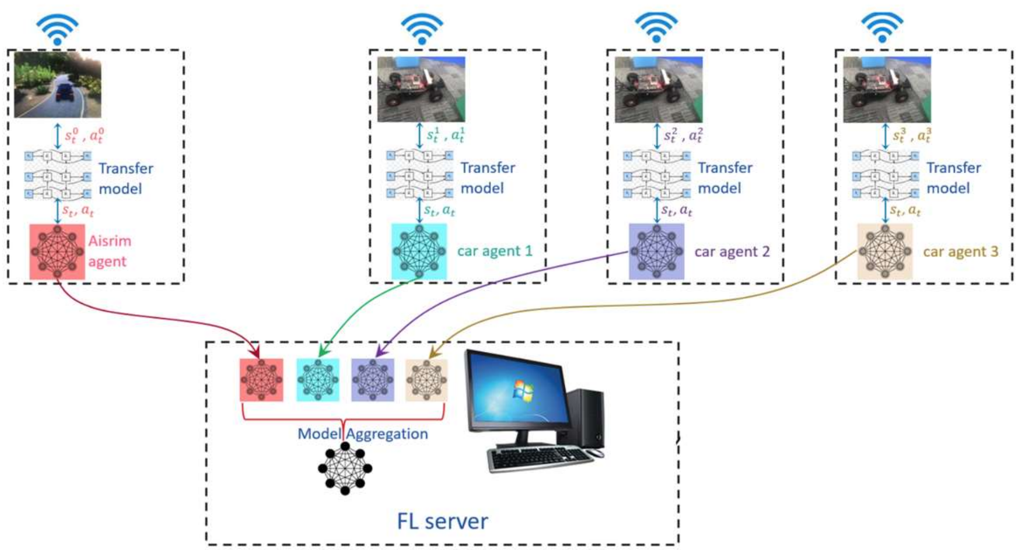
- Recommendation system serves a set of users, many of whom have similar yet nonidentical preferences

- How can we make recommendations to maximize the overall user satisfaction?

https://research.netflix.com/research-area/recommendations

Motivation 3: autonomous driving (Liang et al, 2019)

- A set of self-driving agents, operating on different car make / model / wear & tear conditions



- How can we learn (customized) autonomous driving agents faster, by sharing information among them?

Outline

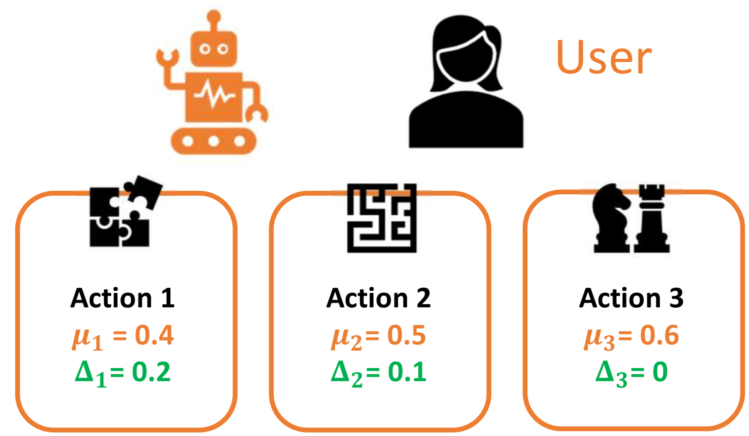
- Motivation
- The ε-multiplayer multi-armed bandit problem
- Our algorithms: Upper Confidence Bound and Thompson Sampling
- Experimental evaluation
- The ε-multiplayer episodic reinforcement learning problem

Background: the multi-armed bandit problem

- Initially: no knowledge about user's preferences on actions
- For round $t \in [T] = \{1, \dots, T\}$:
  - Take action (arm) $a_t \in [K]$
  - Receive reward $r_t \sim \nu_{a_t}$, where each ν_a has mean μ_a
- Goal: maximize $\mathrm{E}[\sum_{t=1}^{T} r_t]$, which is equivalent to minimize regret:

$$\mathrm{Reg}(T) = T\mu^* - \mathrm{E}\left[\sum_{t=1}^{T} \mu_{a_t}\right] \qquad \mu^* = \max_a \mu_a$$
$$= \sum_a \Delta_a \, \mathrm{E}[n_a(T)]$$

$\Delta_a \coloneqq \mu^* - \mu_a$, $n_a(t)$: # times a is taken up to round t

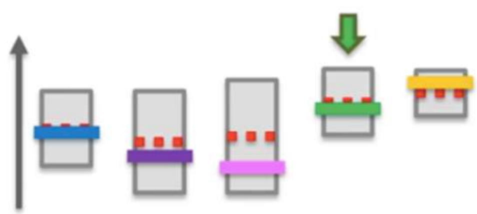


Action = cognitive training activities to recommend

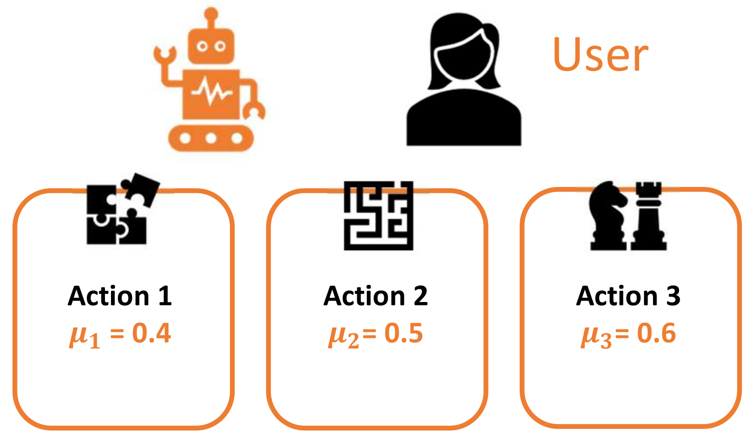
- Many applications: medical treatment, telecommunication, pricing, …

Background: the multi-armed bandit problem (cont'd)

- Challenge: balance exploration vs. exploitation
- Representative approach: the upper confidence bound (UCB) algorithm (Auer et al, 2002)
- At every round $t \in [T]$:
  - Construct upper confidence bounds for μ_1, μ_2, μ_3
  - Take action that maximizes its reward upper confidence bound



- Near-optimal regret guarantees: $\sum_{i:\Delta_i>0} \frac{\ln T}{\Delta_i}$

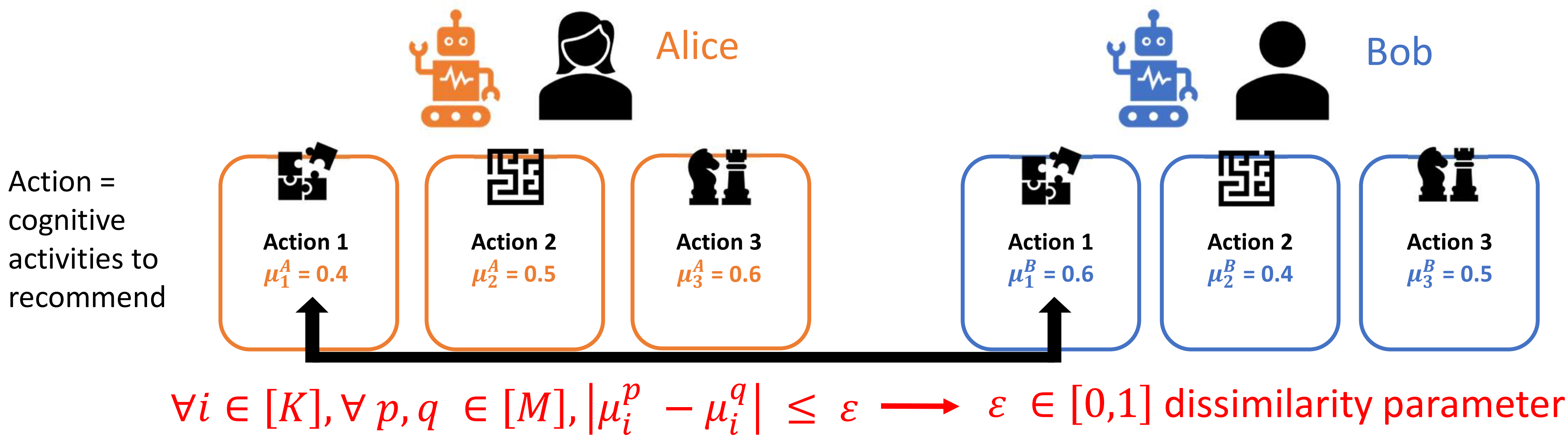


Action = cognitive training activities to recommend

https://rpubs.com/markloessi/501899

The ε-multiplayer multi-armed bandit problem

- A set of M players (robots) concurrently interact with their respective environments (tasks), using K available actions



- How to model the similarity between tasks?
- This work: ε-dissimilarity

The ε-multiplayer multi-armed bandit problem

- **Interaction Protocol:**

  For each round $t \in [T]$:

  For every player $p \in [M]$:

  p takes an action, and observes an independently-drawn reward.

  Players share information at the end of each round.

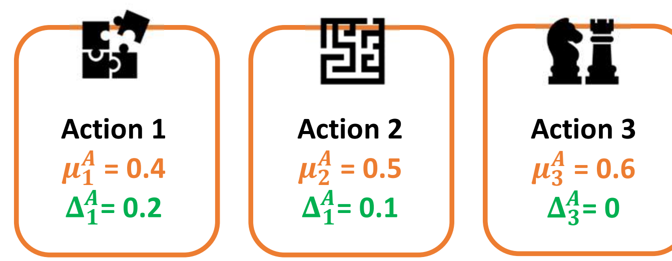
- **Objective:**

  Minimize the collective regret

  $$\mathrm{Reg}(T) = \sum_p \sum_i \Delta_i^p \, \mathrm{E}\left[n_i^p(T)\right]$$

  where $\Delta_i^p = \mu_*^p - \mu_i^p \geq 0$ is the suboptimality gap

  and $n_i^p(t)$ is the number of times action i taken by player p after t rounds.



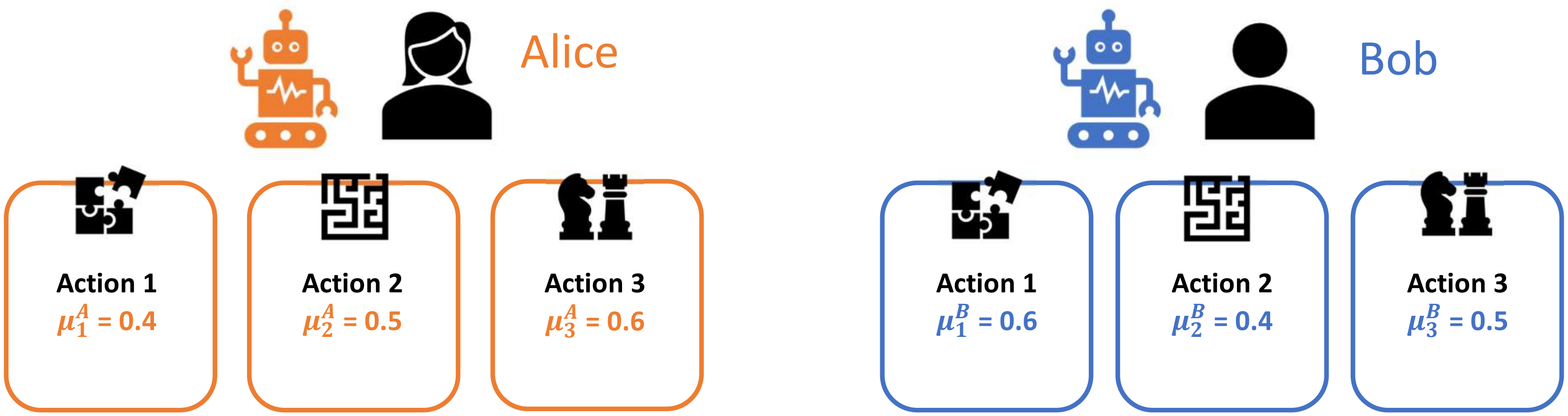
Baseline 1: Individual single-task learning

- Each player runs a bandit algorithm individually (e.g. UCB, Thompson Sampling)
- Single-task optimal learning guarantee ⇒ player p incurs a regret $\sum_{i:\Delta_i^p>0} \frac{\ln T}{\Delta_i^p}$
- Collective regret: $\sum_p \sum_{i:\Delta_i^p>0} \frac{\ln T}{\Delta_i^p}$

- Can we design algorithms with better collective regret, by sharing information across players?

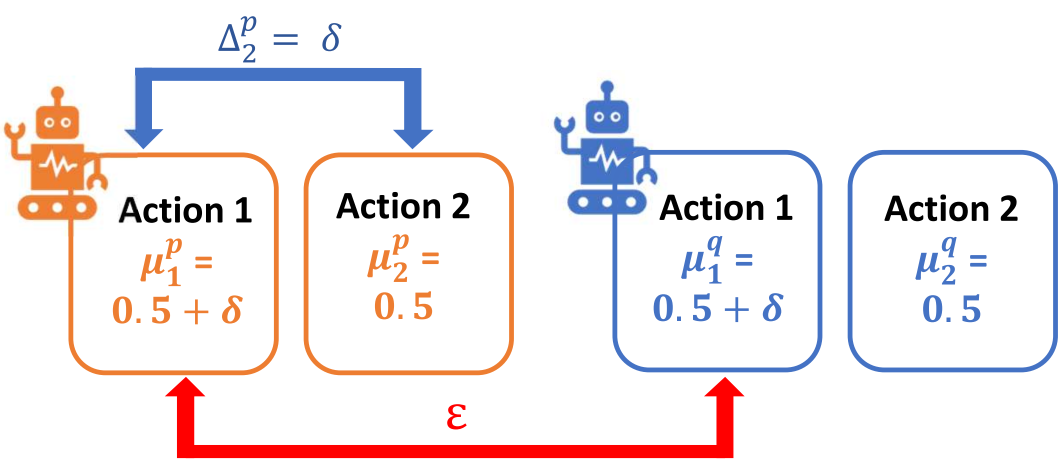
Baseline 2: naïve data aggregation

- Idea: pretend that all M tasks are the same, and maintain only one reward model for decision making
  - Drawback: does not "personalize"
  - OK if $\varepsilon = 0$, but fail if $\varepsilon > 0$
  - Well known as the "negative transfer" issue (Rosenstein et al '05)



Fundamental limits of knowledge transfer

- The utility of cross-task knowledge transfer depends on
  - ε, the dissimilarities between the player-dependent reward distributions
  - the gaps Δ_i^p's, the intrinsic difficulty of each multi-armed bandit problem each player faces individually
- Example: let $\delta < \varepsilon/4$, consider:



Claim: Any "reasonable" algorithm must have $\Omega(\frac{M \ln T}{\delta})$ regret in this case, matching Individual-UCB baseline's regret bound.

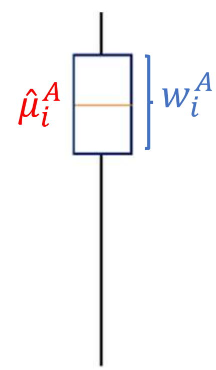
Outline

- Motivation
- The ε-multiplayer multi-armed bandit problem
- Our algorithms: Upper Confidence Bound and Thompson Sampling
- Experimental evaluation
- The ε-multiplayer episodic reinforcement learning problem

Algorithmic principle: optimism in the face of uncertainty

- Key idea: when you are uncertain, act according to the *best plausible* world $\widehat{W}$ (reward-wise)
  - If $\widehat{W}$ is correct => no regret => exploitation
  - If $\widehat{W}$ is wrong => learn useful information => exploration

- $\widehat{W}$ in the multi-player bandit problem:
  - For every p, i, what is the best plausible value of μ_i^p?
  - UCB_i^p: upper confidence bound on μ_i^p



- Algorithm: for every p, choose action $i = \mathrm{argmax}_j \mathrm{UCB}_j^p$

https://rlgammazero.github.io/docs/2020_AAAI_tut_part2.pdf

Naïve construction of reward UCBs

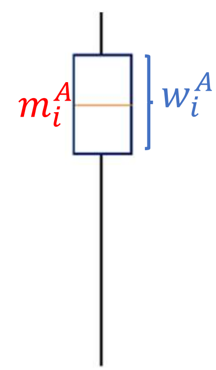
- UCB_i^A: upper confidence bound on μ_i^A
- Alice has observed $n = n_i^A$ rewards from arm i

  $x_1, x_2, \dots, x_n$ iid with sample mean m_i^A

- Confidence interval for μ_i^A:

$$[m_i^A - w_i^A, \underbrace{m_i^A + w_i^A}_{\mathrm{UCB}_i^A}],$$

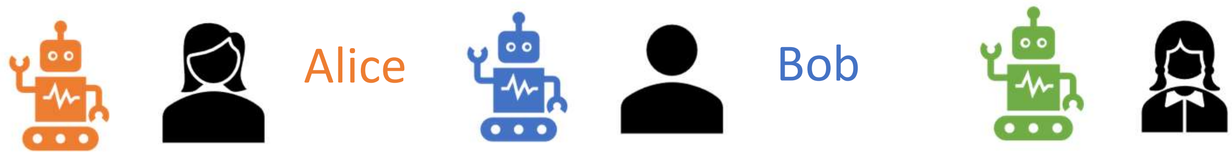
where $w_i^A \propto \sqrt{\frac{\ln T}{n_i^A}}$



- This results in the individual-UCB baseline

Our algorithm: RobustAgg-UCB

- Key idea: robustly estimate upper confidence bounds on μ_i^A's using a weighted combination of Alice's own data and other players' data



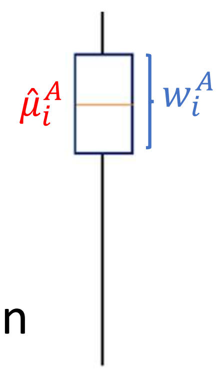
- Let $\mathrm{UCB}_i^A := \min_{\lambda \in [0,1]} \left(\hat{\mu}_i^A(\lambda) + w_i^A(\lambda) \right)$
- Center $\hat{\mu}_i^A(\lambda) := (1-\lambda) m_i^A + \lambda\, m_i^{-A}$

Mean reward of arm i played by others

- Width $w_i^A(\lambda) := (1-\lambda)\sqrt{\frac{\ln T}{n_i^A}} + \lambda\left(\sqrt{\frac{\ln T}{n_i^{-A}}} + \epsilon\right)$

Accounting for bias in other players' data

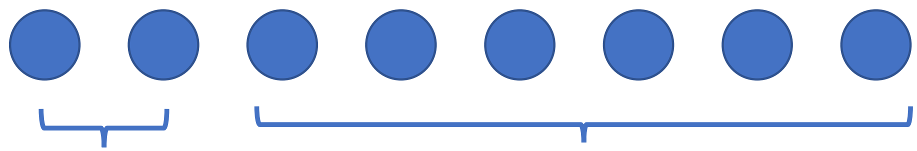
- Tighter UCB than the individual-UCB baseline



RobustAgg-UCB: performance guarantees

- For player p's contribution to collective regret:

Actions



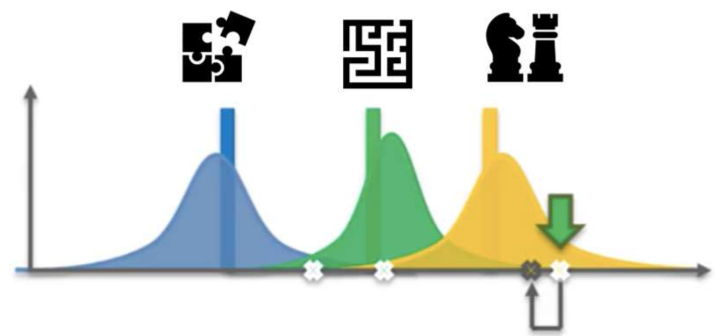
		$\mathcal{J}_\epsilon^C$	subpar arms $\mathcal{J}_\epsilon = \{i:\ \forall p, \Delta_i^p \geq \Omega(\epsilon)\}$
Individual-UCB	$\sum_i$	$\frac{\ln T}{\Delta_i^p}$	$\frac{\ln T}{\Delta_i^p}$
RobustAgg-UCB	$\sum_i$	$\frac{\ln T}{\Delta_i^p}$	$\frac{1}{M} \cdot \frac{\ln T}{\Delta_i^p}$

- Key takeaway: for subpar arms $\mathcal{J}_\epsilon$, players share information to explore less
- **Matching lower bound:** RobustAgg-UCB's regret is essentially unimprovable

Alternative algorithmic principle: Thompson Sampling

- Key idea (Thompson'33):
  - maintain a *posterior distribution* of the world $p(W)$
  - Sample $\widehat{W} \sim p$ and act according to $\widehat{W}$

- $p(W)$ in multi-player bandits:
  - For every p, i, has a separate component over μ_i^p



- Algorithm: for every p:
  - For every i, sample θ_i^p from posterior
  - Choose action $i = \mathrm{argmax}_j \theta_j^p$

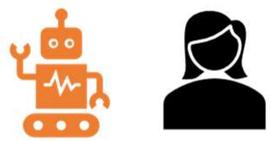
- Strong empirical performance (Chapelle and Li, 2011; Scott, 2010)

Our second algorithm: RobustAgg-TS (Thompson Sampling)

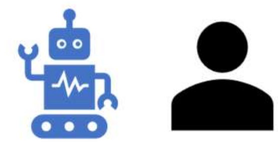
- Challenge: no explicit probabilistic assumptions on the task similarity

$$\forall i \in [K], \forall\, p, q \in [M], \left|\mu_i^p - \mu_i^q\right| \leq \varepsilon$$

How to define posterior?



Alice



Bob

- Workaround: sample θ_i^A's instead from the following "optimistic-posterior":

Mean reward / #times of arm i chosen by all players

$$\theta_i^A \sim \begin{cases} \mathcal{N}\left(m_i + \varepsilon, \frac{1}{n_i}\right), n_i^A \leq O\left(\frac{\ln T}{\varepsilon^2}\right) \\ \mathcal{N}\left(m_i^A, \frac{1}{n_i^A}\right), n_i^A > O\left(\frac{\ln T}{\varepsilon^2}\right) \end{cases}$$

- Same optimality guarantee as RobustAgg-UCB

Outline

- Motivation
- The ε-multiplayer multi-armed bandit problem
- Our algorithms: Upper Confidence Bound and Thompson Sampling
- Experimental evaluation
- The ε-multiplayer episodic reinforcement learning problem

Experiments

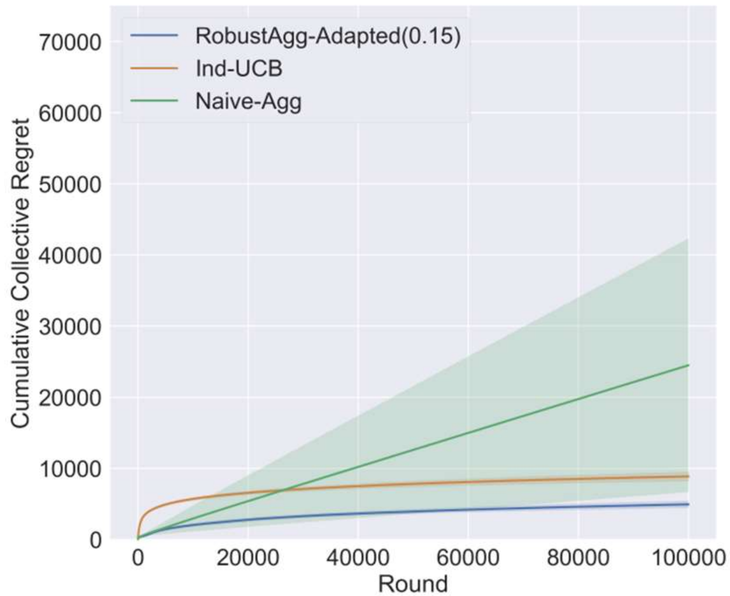
- Key question 1: Are our algorithms resistant to negative transfer?
- Key question 2: Does the notion of subpar arms $\mathcal{I}_\varepsilon$ characterize the difficulty of ε-multi-player muti-armed bandit problems in practice?
- Experimental setup:
  - 20-player 10-armed bandit environments with different values of $|\mathcal{I}_\varepsilon|$, with $\varepsilon = 0.15$
  - Algorithms evaluated:
    - Naïve-Aggregation
    - Individual-UCB
    - Individual-TS
    - RobustAgg-UCB (ours)
    - RobustAgg-TS (ours)

Experiment 1: resistance to negative transfer

- $|\mathcal{I}_{\varepsilon}| = 8$

- Naïve-Aggregation suffers a linear regret

- Both Individual-UCB and RobustAgg-UCB have sublinear regret, with the latter performing better

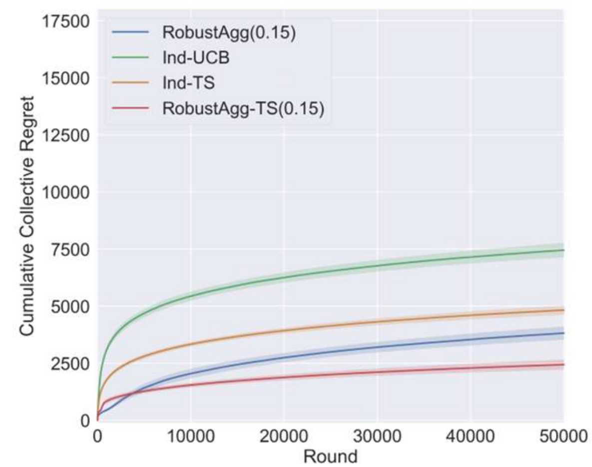


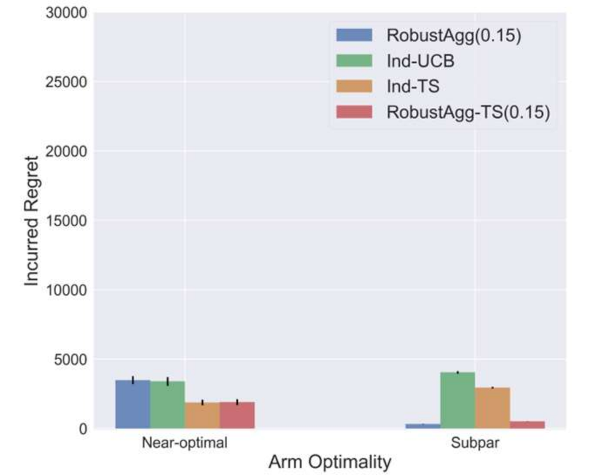
Experiment 2: effect of subpar arms

- $|\mathcal{I}_\varepsilon| = 8$

- RobustAgg-UCB and RobustAgg-TS outperform the two individual single-task baselines
  - Regret from subpar arms is much lower

- Thompson sampling-based algorithms outperforms their UCB counterparts



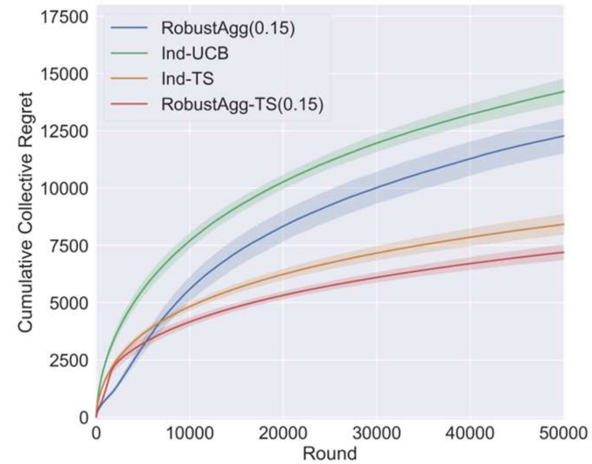


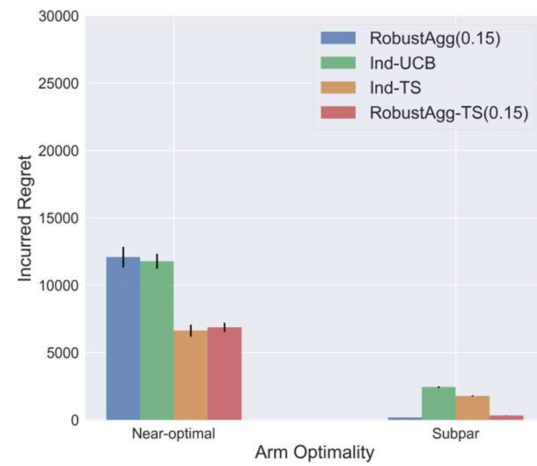
Experiment 2: effect of subpar arms

- $|\mathcal{I}_\varepsilon| = 5$

- The gaps between our robust aggregation algorithms and the individual single-task baselines are smaller

- Contribution of regret from near-optimal arms increases





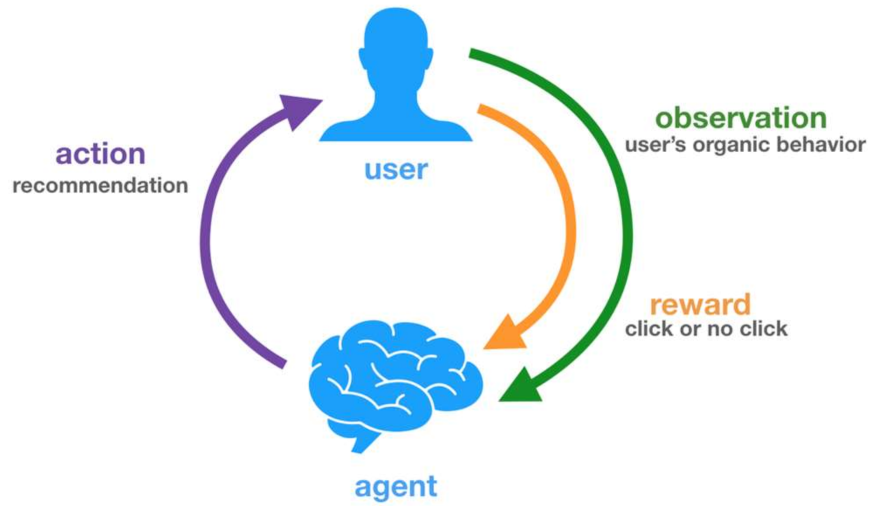
Outline

- Motivation
- The ε-multiplayer multi-armed bandit problem
- Our algorithms: Upper Confidence Bound and Thompson Sampling
- Experimental evaluation
- The ε-multiplayer episodic reinforcement learning problem

Background: episodic reinforcement learning

- Markov decision process (MDP) environment $\mathcal{M}$
- Generalizes multi-armed bandits: environment's state s (e.g user's moood)

- For episodes $k = 1,2, \ldots, K$:
  - Deploy a policy π^k
  - For steps $h = 1,2, \ldots, H$:
    - Observe state s_h
    - Take action a_h
    - Receive reward $r_h = R(s_h, a_h)$
    - Transition to state $s_{h+1} \sim \mathbb{P}(\cdot| \, s_h, a_h)$



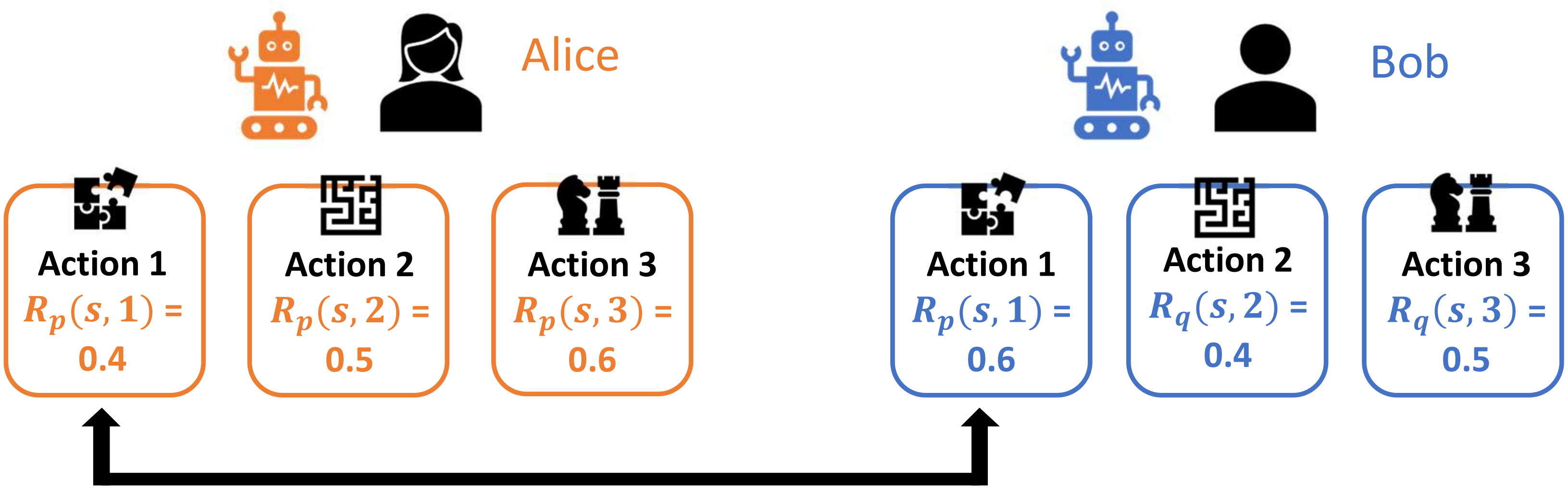
- Goal: maximize cumulative reward $\mathrm{E}\left[\sum_{k=1}^{K} V^{\pi^k}\right]$

V^{π}: expected reward of policy π in $\mathcal{M}$

The $\boldsymbol{\varepsilon}$-Multi-Player Episodic RL ($\boldsymbol{\varepsilon}$-MPERL) Problem

- A set of M players (robots) concurrently interact with their respective environments, each represented as an Episodic MDP.

Actions = cognitive activities to recommend



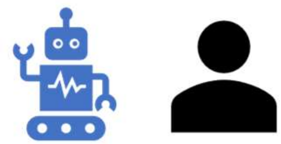
$$\forall\, p, q, s, a:$$
$$|R_p(s,a) - R_q(s,a)| \le \varepsilon$$
$$\|\mathbb{P}_p(\cdot \mid s,a) - \mathbb{P}_q(\cdot \mid s,a)\|_1 \le \varepsilon / H$$

ε: dissimilarity parameter

The $\boldsymbol{\varepsilon}$-MPERL Problem: formal setup

- M episodic MDPs $\left(\mathcal{M}_p\right)_{p=1}^{M}$ with identical state-action spaces

Alice



Bob

- For episodes $k = 1,2, \dots, K$:
  - For players $p = 1,2, \dots, M$:
    - Player p interacts with $\mathcal{M}_p$ with policy $\pi^k(p)$ for one episode, obtaining trajectory τ_p^k
  - All M trajectories $\left(\tau_p^k\right)_{p=1}^{M}$ are shared among the players

- Collective regret: $\mathrm{Reg}(K) = \sum_{p=1}^{M} \sum_{k=1}^{K} \mathrm{E}\left[V_p^\star - V_p^{\pi^k(p)}\right]$

Optimal value of player p Value of player p executing $\pi^k(p)$

Our algorithm: Multi-task-Euler(ε) and guarantees

For player p's contribution to the collective regret:

State-action pairs		$Z_{p,\mathrm{opt}}$	$\left(Z_{p,\mathrm{opt}} \cup \mathcal{J}_\epsilon\right)^C$	$\mathcal{J}_\epsilon$
Individual Single-task baseline	$\sum_{s,a}$	$\frac{H^3 \ln K}{\Delta_{p,\min}}$	$\frac{H^3 \ln K}{\Delta_p(s,a)}$	$\frac{H^3 \ln K}{\Delta_p(s,a)}$
Multi-task-Euler(ε)	$\sum_{s,a}$	$\frac{H^3 \ln K}{\Delta_{p,\min}}$	$\frac{H^3 \ln K}{\Delta_p(s,a)}$	$\frac{1}{M} \cdot \frac{H^3 \ln K}{\Delta_p(s,a)}$

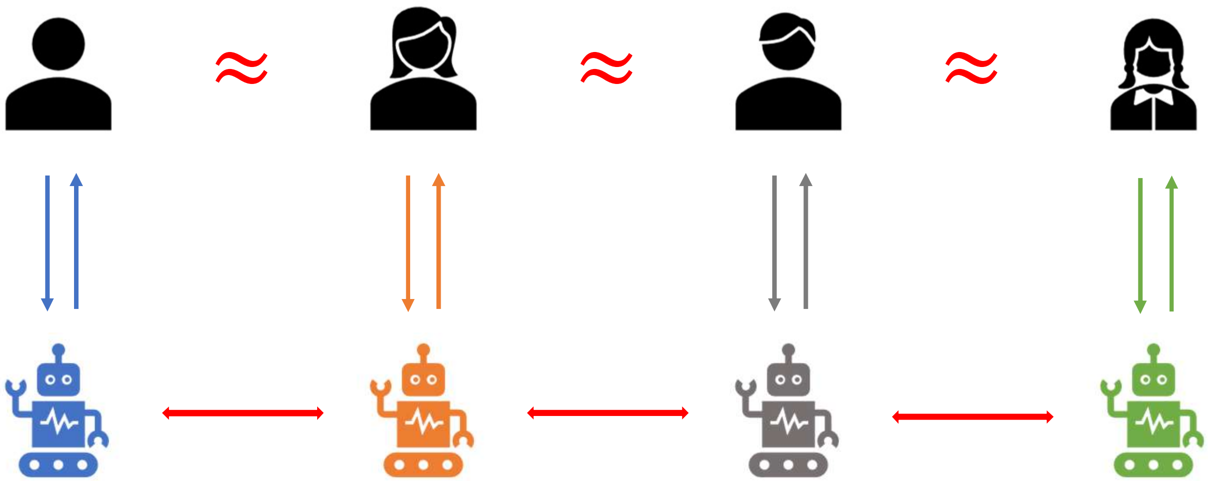
$\mathcal{J}_\epsilon = \left\{(s,a): \forall p \in [M], \Delta_p(s,a) \geq \Omega(H\epsilon)\right\}$ for some generalized notion of suboptimality gap $\Delta_p(s,a)$

Conclusions and open problems

- We study multi-task bandit and reinforcement learning where the tasks are similar but not necessarily identical

- Our algorithms provably avoid "negative transfer"

- Open problem:
  - Are there other practical and interesting notions of task similarity beyond ε-dissimilarity?
    - E.g. recent works on representation transfer in RL (e.g. Yang et al, 2020, Agarwal et al, 2022)



Thank you!

https://arxiv.org/abs/2010.15390
https://arxiv.org/abs/2107.08622
https://arxiv.org/abs/2206.08556