Provably Efficient Multi-Task Reinforcement Learning with Model Transfer

Chicheng Zhang and Zhi Wang

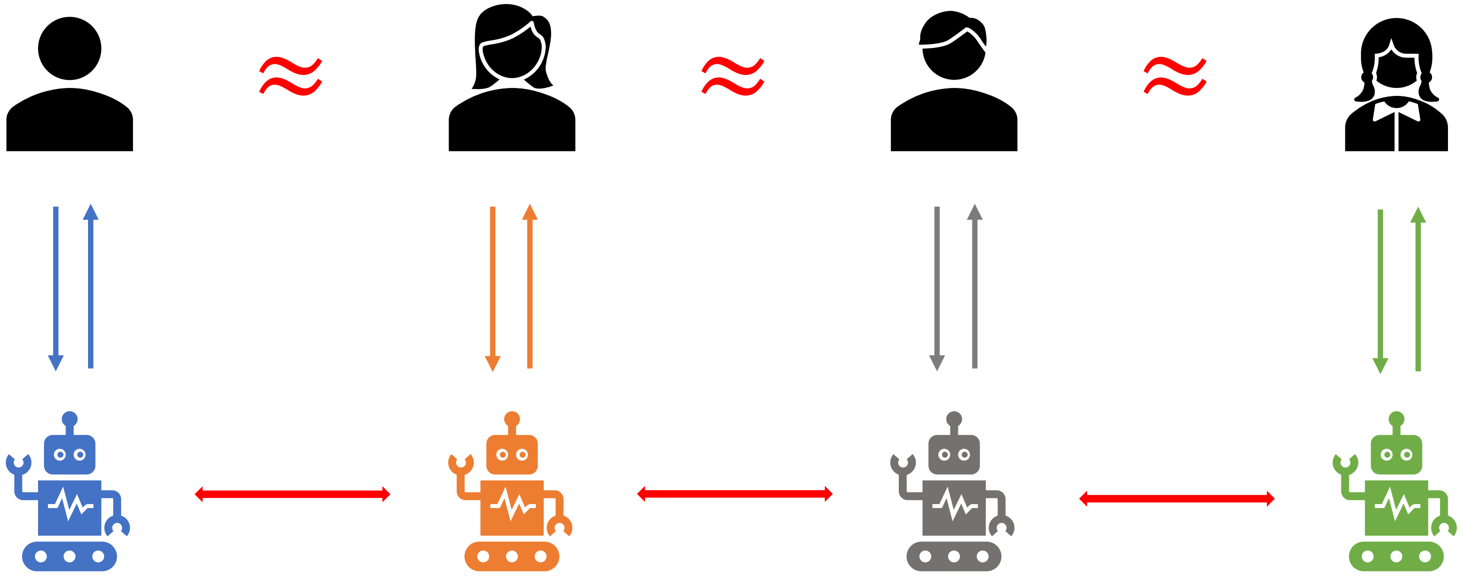
NeurIPS 2021

THE UNIVERSITY OF ARIZONA

UC San Diego
JACOBS SCHOOL OF ENGINEERING

Heterogenous Multi-Task Online Reinforcement Learning (RL)

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

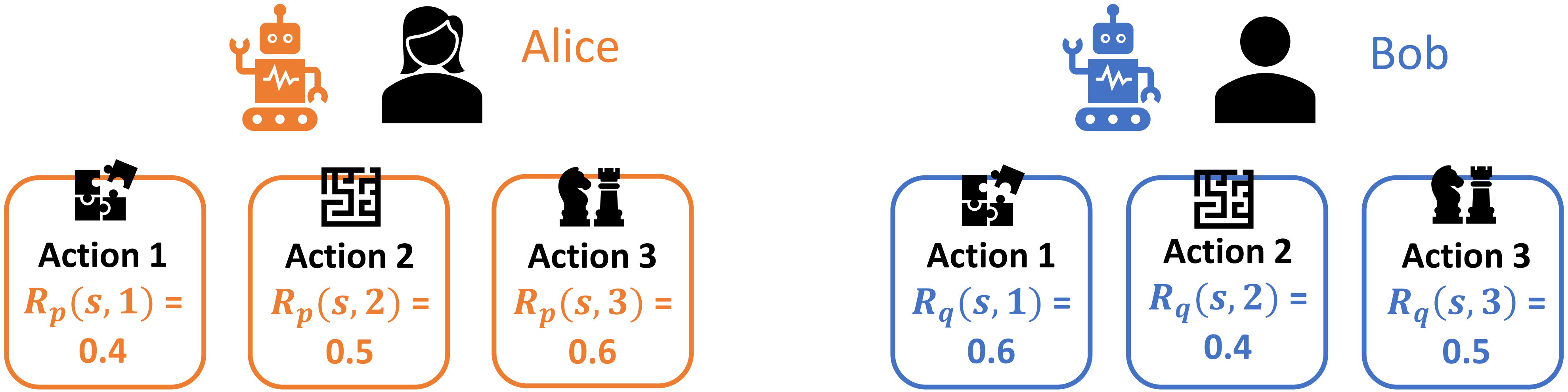


- A group of assistive robots deployed to provide personalized healthcare services (Kubota et al., 2020).
- Question: If the robots receive similar yet nonidentical feedback, how can they learn to perform their respective tasks faster in an online RL setting?

Multi-Player Episodic RL (MPERL)

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

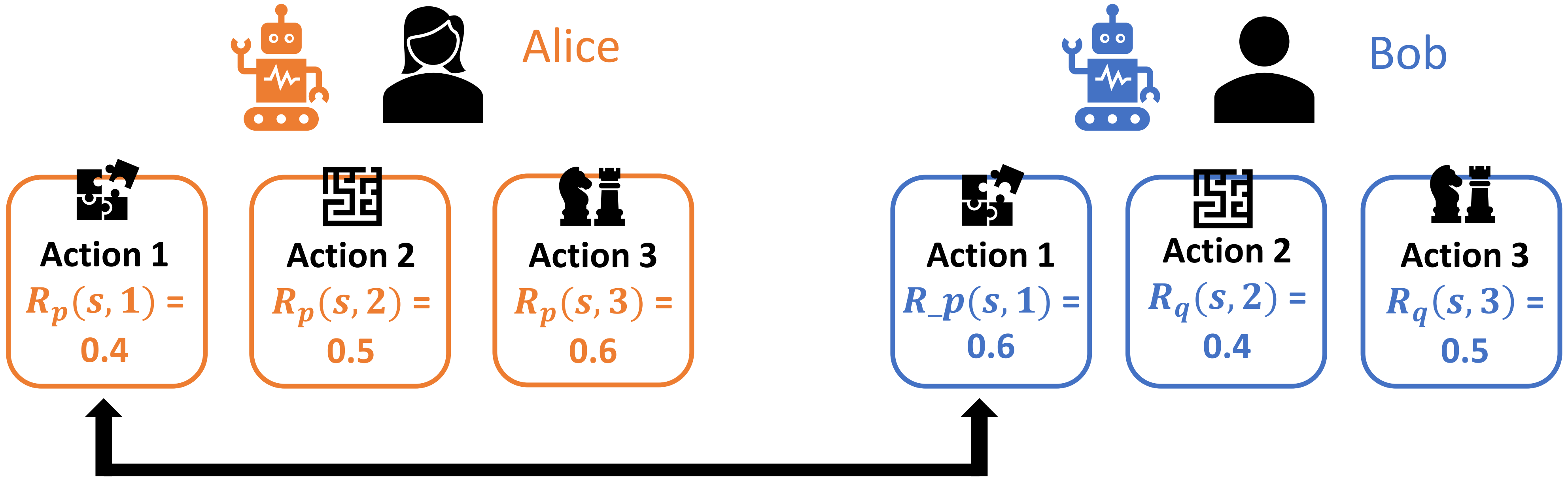
Actions = cognitive activities to recommend



The ε-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



$$\forall\, p, q, s, a:$$

$$\left|R_p(s,a) - R_q(s,a)\right| \leq \varepsilon$$

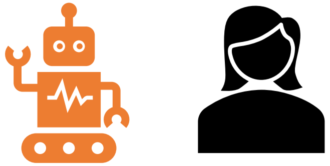
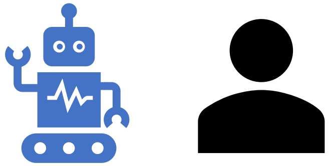
$$\left\|\mathbb{P}_p(\cdot \mid s,a) - \mathbb{P}_q(\cdot \mid s,a)\right\|_1 \leq \varepsilon/H$$

ε: dissimilarity parameter

The ε-MPERL Problem: formal setup

- M episodic, tabular, H-layered MDPs $\left(\mathcal{M}_p\right)_{p=1}^{M}$ with shared state-action spaces, and common initial distribution $\delta(s_0)$

- 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


Alice

Bob

- Collective regret: $\mathrm{Reg}(K) = \sum_{p=1}^{M} \sum_{k=1}^{K} V_p^{\star}(s_0) - V_p^{\pi^k(p)}(s_0)$

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

Baseline: individual single-task learning

- Each player learns separately using a state-of-the-art online tabular RL algorithm, e.g., *Strong-Euler* (Simchowitz and Jamieson, 2019), achieving a collective regret of
  - (Gap-independent bound) $\tilde{O}\left(M\sqrt{H^2SAK}\right)$
  - (Gap-dependent bound)

$$\tilde{O}\left(\sum_{p=1}^{M}\left(\sum_{(s,a)\in Z_{p,\mathrm{opt}}}\frac{H^3\ln K}{\Delta_{p,\min}}+\sum_{(s,a)\notin Z_{p,\mathrm{opt}}}\frac{H^3\ln K}{\Delta_p(s,a)}\right)\right)$$

where $\Delta_p(s,a) := V_p^\star(s) - Q_p^\star(s,a), Z_{p,\mathrm{opt}} = \{(s,a): \Delta_p(s,a) = 0\}$,

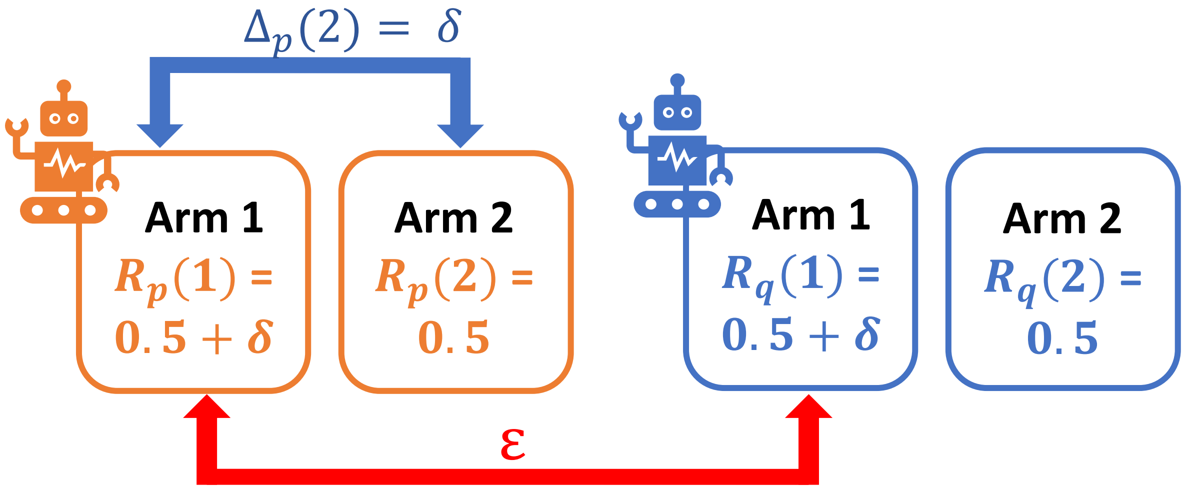
$$\Delta_{p,\min} = \min_{(s,a)\notin Z_{p,\mathrm{opt}}} \Delta_p(s,a)$$

- Can we do better with inter-task information sharing?

The benefit of multi-task learning

- (Wang, Zhang, Singh, Riek, Chaudhuri, 2021): in a multi-task multi-armed bandit setting, information sharing sometimes does not help, *information theoretically*.

- Example: For a fixed ε and $\delta < \varepsilon/4$, consider:



Claim: Any sublinear regret algorithm must have $\Omega\left(\frac{M \ln K}{\delta}\right)$ regret, no better than the individual single-task learning baseline.

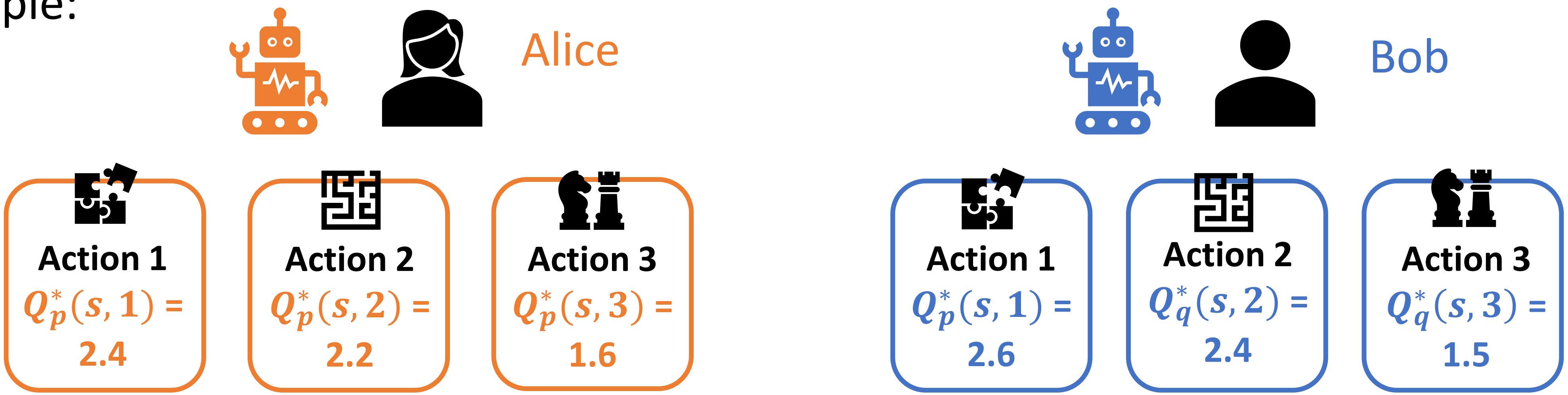
- Key observation: the benefit of multi-task learning depends on the interaction between ε and suboptimality gaps $\Delta_p(s, a)$

Key notion: subpar state-action pairs

- Subpar state-action pairs:

$$\mathcal{I}_\epsilon = \{(s,a): \text{for some } p \in [M], \Delta_p(s,a) \geq \Omega(H\epsilon)\}$$

- Example:



- $(s,3) \in \mathcal{I}_\epsilon$; $(s,2) \notin \mathcal{I}_\epsilon$
- Subpar state-action pairs are those amenable for inter-task information sharing

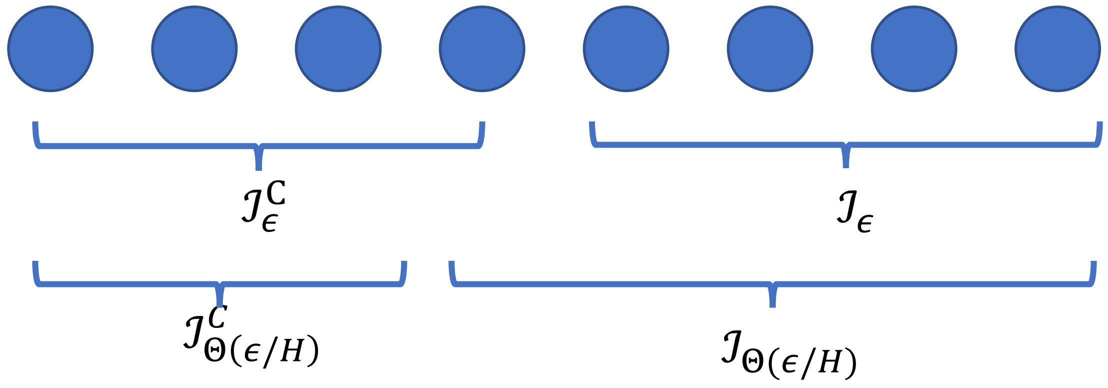
Our results

For ε-MPERL problems, assuming known ε:

- Our algorithm, Multi-Task-Euler(ε), achieves gap-dependent and gap-independent regret upper bounds
- We also show gap-dependent and gap-independent regret lower bounds, that nearly match the upper bounds for constant H

Our results: gap-independent bounds

State-action pairs

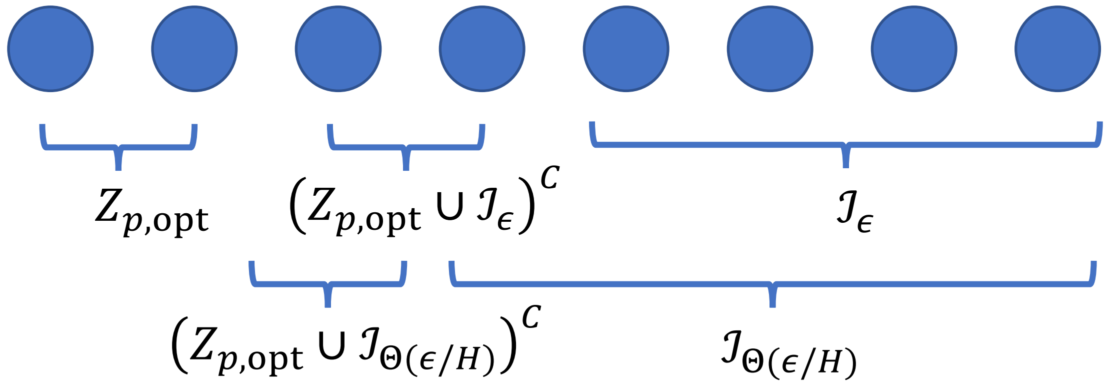


Individual Strong-Euler	$M\sqrt{H^2\lvert\mathcal{J}_\epsilon^C\rvert K}$	+	$M\sqrt{H^2\lvert\mathcal{J}_\epsilon\rvert K}$
Multi-task-Euler(ε)	$M\sqrt{H^2\lvert\mathcal{J}_\epsilon^C\rvert K}$	+	$\sqrt{MH^2\lvert\mathcal{J}_\epsilon\rvert K}$
Lower bound	$M\sqrt{H^2\lvert\mathcal{J}^C_{\Theta(\epsilon/H)}\rvert K}$	+	$\sqrt{MH^2\lvert\mathcal{J}_{\Theta(\epsilon/H)}\rvert K}$

Our results: gap-dependent bounds

For player p's contribution to the collective regret:

State-action pairs



Individual Strong-Euler	$\sum_{s,a}$	$\frac{H^3 \ln K}{\Delta_{p,\min}}$	$\frac{H^3 \ln K}{\Delta_p(s,a)}$	$\frac{H^3 \ln K}{\Delta_p(s,a)}$
Multi-task-Euler(ε)	$\sum_{s,a}$	$\frac{H^3 \ln K}{\Delta_{p,\min}}$	$\frac{H^3 \ln K}{\Delta_p(s,a)}$	$\frac{1}{M} \cdot \frac{H^3 \ln K}{\Delta_p(s,a)}$
Lower bound	$\sum_{s,a}$		$\frac{H^2 \ln K}{\Delta_p(s,a)}$	$\frac{1}{M} \cdot \frac{H^2 \ln K}{\Delta_p(s,a)}$

Multi-task-Euler(ε): main ideas

- For each player p, Multi-Task-Euler(ε):

1. Maintains two model estimates for $\mathcal{M}_p$: (1) an individual estimate $\widehat{\mathcal{M}_p}$ (2) an aggregate model estimate $\widehat{\mathcal{M}}$

2. Performs a ``heterogeneous'' optimistic value iteration using both $\widehat{\mathcal{M}_p}$ and $\widehat{\mathcal{M}}$ to obtain $\widehat{Q_p}$, a tight upper confidence bound of $Q_p^{\star}$, and executes its greedy policy

- Similar algorithmic idea of ``model transfer'' has appeared in prior works, e.g., (Taylor, Jong, & Stone, 2008), (Pazis & Parr, 2016)

Technical overview

- Upper bounds: a new surplus bound in the multi-task setting:

$$\widehat{Q_p}(s,a) - \left(R_p(s,a) + \left\langle \mathbb{P}_p(\cdot \mid s,a), \hat{V}_p \right\rangle\right) \le \tilde{O}\left(\min\left(\sqrt{\frac{1}{n_p(s,a)}}, \epsilon + \sqrt{\frac{1}{n(s,a)}}\right)\right),$$

and combine with the ``clipping trick'' (Simchowitz & Jamieson, 2019)

- Lower bounds: combine the multi-task bandit lower bounds (Wang, Zhang, Singh, Riek, Chaudhuri, 2021) with a standard bandit-to-RL conversion

Conclusion and open problems

- We study ε-MPERL, a new multi-task RL setting; this complements existing multi-task RL settings (e.g., Brunskill & Li, 2013, Liu, Guo, & Brunskill, 2016, Pazis & Parr, 2016)

- We give upper and lower bounds on the collective regret that are nearly matching for constant episode length H

- Open questions:
  - Improve the dependence on H in the collective regret bounds
  - Improve the dependence on $Z_{p,\mathrm{opt}}$, similar to recent works (e.g., Xu, Ma, & Du, 2021)
  - Extensions to RL with function approximation

Thank you!