Provably Efficient Multi-Task Reinforcement Learning with Model Transfer

Chicheng Zhang[1], Zhi Wang[2]

[1]University of Arizona, [2]University of California San Diego.

Abstract

- We study multi-player reinforcement learning (RL) in *heterogeneous* environments, where the reward distributions and transition probabilities for all players are *similar but not necessarily identical*.
- Our formulation can be used to model *multi-task* RL in application domains such as healthcare robotics.
- We study *when and how* players can improve their collective performance by sharing and aggregating data.
- We provide upper and lower bounds that characterize what can be done and what cannot be done.

Problem Formulation

A multi-player episodic RL (MPERL) problem instance consists of M episodic, layered, tabular MDPs $\{\mathcal{M}_p = (H, \mathcal{S}, \mathcal{A}, d_0, \mathbb{P}_p, R_p)\}_{p=1}^M$, where

- H is an episode length, $\mathcal{S}$ is a finite state space of size S, and $\mathcal{A}$ is a finite action space of size A;
- $d_0 \in \Delta(\mathcal{S})$ is the initial state distributionshared across all players;
- For each player p, $\mathbb{P}_p : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is its transition probability, and $R_p : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is its expected reward.

An MPERL problem instance is said to be ϵ-dissimilar, if for every pair of players $p, q \in [M]$, and $(s, a) \in \mathcal{S} \times \mathcal{A}$,

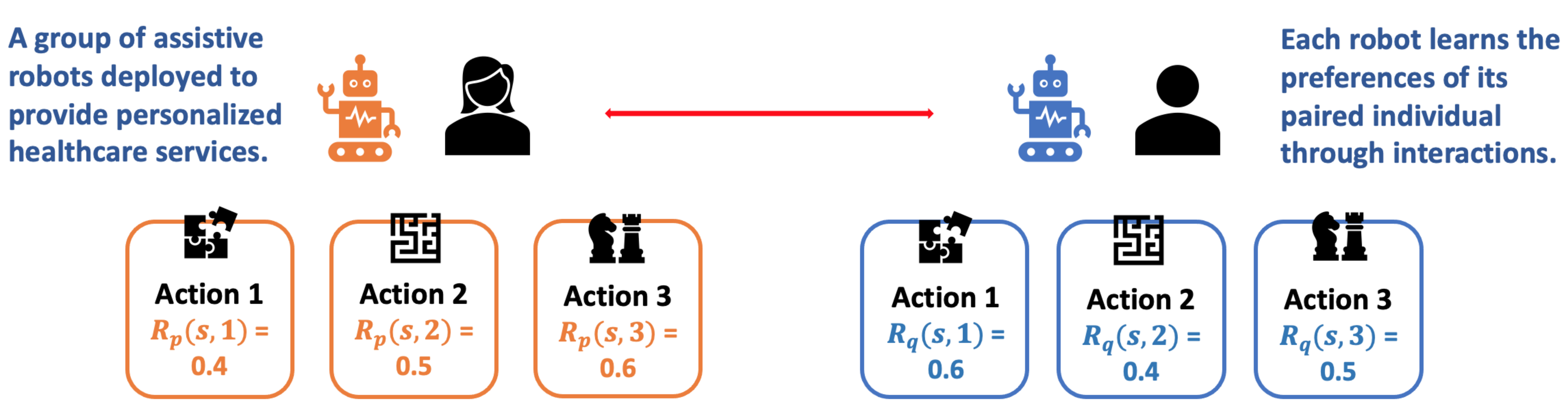
$$|R_p(s,a) - R_q(s,a)| \le \epsilon, \ \|\mathbb{P}_p(\cdot \mid s,a) - \mathbb{P}_q(\cdot \mid s,a)\|_1 \le \frac{\epsilon}{H}.$$

Interaction protocol. In each episode $k \in [K]$, each player $p \in [M]$ interacts with its respective MDP, $\mathcal{M}_p$, and executes a policy, $\pi^k(p)$, generating a trajectory $\tau_p^k = (s_{1,p}^k, a_{1,p}^k, s_{2,p}^k, a_{2,p}^k \ldots, s_{H,p}^k, a_{H,p}^k)$ according to $\mathbb{P}_p$ and R_p. Once all players finish, all M trajectories are shared among the players.

Performance measure. The players seek to minimize their *collective regret*, $\mathrm{Reg}(K) = \sum_{p=1}^M \sum_{k=1}^K \left(V_{0,p}^\star - V_{0,p}^{\pi^k(p)} \right)$, where

- $V_{0,p}^\star = \mathbb{E}_{s_1 \sim d_0} \left[V_{1,p}^\star(s_1) \right]$ is the expected optimal value of player p, and
- $V_{0,p}^{\pi^k(p)} = \mathbb{E}_{s_1 \sim d_0} \left[V_{1,p}^{\pi^k(p)}(s_1) \right]$ is the expected value of player p executing policy $\pi^k(p)$.

Application in healthcare robotics (e.g., Kubota et al, 2020).



Baseline: individual single-task learning. If each player learns separately with a state-of-the art algorithm (e.g. UCBVI-Bernstein (Azar, Osband & Munos, 2017), Euler (Zanette & Brunskill, 2019), Strong-Euler (Simchowitz & Jamieson, 2019)), they can achieve a gap-independent collective regret guarantee of $\mathrm{Reg}(K) \le \tilde{O}(M\sqrt{H^2 SAK})$.

Algorithm: Multi-task-Euler

For each episode k and each player p:

Maintain models:
- Individual estimates of transition probability $\hat{\mathbb{P}}_p$, reward $\hat{R}_p$ and count $n_p(\cdot,\cdot)$ based on player p's experience;
- Aggregate estimates of transition probability $\hat{\mathbb{P}}$, reward $\hat{R}$ and count $n(\cdot,\cdot)$ based on all players' experience.

Optimisic value iteration using heterogeneous data: (recursively) compute upper and lower bound estimates of $Q_p^\star$, namely, $\overline{Q}_p$ and $\underline{Q}_p$, using value iteration; specifically:
- Construct $\underline{\text{agg-}Q}_p$ and $\overline{\text{agg-}Q}_p$ based on aggregate model estimates and an ϵ-aware bonus term;
- Construct $\underline{\text{ind-}Q}_p$ and $\overline{\text{ind-}Q}_p$ based on individual model estimates of player p and a standard bonus term;
- $\overline{Q}_p$ is chosen to be the tighter confidence bound between $\overline{\text{agg-}Q}_p$ and $\overline{\text{ind-}Q}_p$; a similar construction holds for $\underline{Q}_p$.

Execute policy: Execute $\pi^k(p)$, the greedy policy of $\overline{Q}_p$, obtaining trajectory τ_p^k.

Update models: Update individual estimates using τ_p^k, and update aggregate estimates using $\{\tau_q^k\}_{q=1}^M$.

Instance-dependent Regret Upper Bounds

Subpar state-action pairs: state-action pairs that are far from optimal for some player, formally,

$$\mathcal{I}_\epsilon := \left\{ (s,a) \in \mathcal{S} \times \mathcal{A} : \exists p \in [M], \mathrm{gap}_p(s,a) \ge 96H\epsilon \right\},$$

where suboptimality gap $\mathrm{gap}_p(s,a) := V_p^\star(s) - Q_p^\star(s,a)$.

Theorem: If $\{\mathcal{M}_p\}_{p=1}^M$ are ϵ-dissimilar, then for K large enough, Multi-task-Euler satisfies that with probability $1-\delta$,

$$\mathrm{Reg}(K) \le \tilde{O}\left(M\sqrt{H^2 |\mathcal{I}_\epsilon^C| K} + \sqrt{MH^2 |\mathcal{I}_\epsilon| K} \right);$$

see also our full paper for a *gap-dependent* regret upper bound.

Comparison to individual single-task baseline: If $\left|\mathcal{I}_\epsilon^C\right| \ll SA$ and $M \gg 1$, Multi-task-Euler provides a regret bound of lower order than individual Strong-Euler.

Instance-dependent Regret Lower Bounds

Theorem (informal): For any $l, l^C \in \mathbb{N}$ such that $l + l^C = SA$, there exists some ϵ such that for any algorithm Alg, there exists an ϵ-MPERL problem instance with $\left|\mathcal{I}_{\frac{\epsilon}{192H}}\right| \ge l$, and

$$\mathbb{E}\left[\mathrm{Reg}_{\mathrm{Alg}}(K)\right] \ge \Omega\left(M\sqrt{H^2 l^C K} + \sqrt{MH^2 l K} \right);$$

see also our full paper for a *gap-dependent* regret lower bound.

Remark: The upper and lower bounds nearly match for any constant H.