Thompson Sampling for Robust Transfer in Multi-Task Bandits

Zhi Wang[1], Chicheng Zhang[2], Kamalika Chaudhuri[1]

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

Abstract

- We study multi-task bandits, in which different tasks have *similar but not necessarily identical* reward distributions.
- Our problem setting covers a wide range of transfer learning scenarios, such as multi-player *concurrent* learning and *sequential* transfer, and has applications in healthcare robotics, etc.
- We design and analyze a Thompson sampling-type algorithm that robustly aggregates and utilizes data collected from similar sources.
- We show that our algorithm has near-optimal frequentist regret guarantees *and* superior empirical performance in comparison with Upper confidence bound (UCB)-based algorithms.

Problem Formulation

The ϵ-multi-player multi-armed bandit (ϵ-MPMAB) problem [1]:

- M players, labeled as elements in $[M]$;
- K arms, labeled as elements in $[K]$;
- Each player p and arm i associated with an unknown reward distribution with support $[0,1]$ and mean μ_i^p;
- ϵ: (reward) dissimilarity parameter.

$$\forall i \in [K],\ p,q \in [M], \quad |\mu_i^p - \mu_i^q| \le \epsilon.$$

Interaction protocol (see also Hong et al., 2022).
In each round $t \in [T]$:

- A set of active players $\mathcal{P}_t \subseteq [M]$ is chosen (by an oblivious adversary);
- Each active player pulls an arm and observes a reward;
- Decisions & rewards shared with all players at the end of the round.

Special cases:

- $|\mathcal{P}_t| = 1$ for all t: sequential transfer (e.g., Cesa-Bianchi et al., 2013);
- $|\mathcal{P}_t| = [M]$ for all t: concurrent interaction (e.g., [1]).

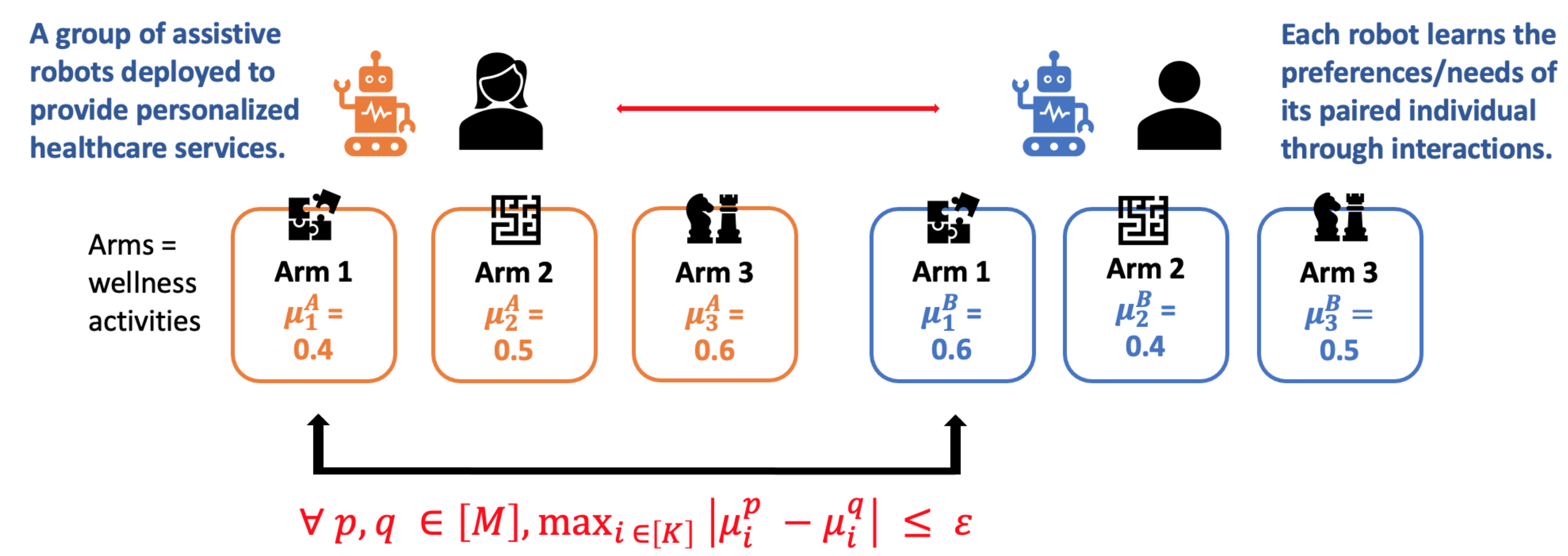
Objective: To minimize the *expected collective regret*,

$$\mathbb{E}[\mathcal{R}(T)] = \sum_{p\in[M]} \sum_{i\in[K]} \Delta_i^p \cdot \mathbb{E}[n_i^p(T)], \text{ where}$$

- $\Delta_i^p = \max_{i\in[K]} \mu_i^p - \mu_i^p \ge 0$ is the *suboptimality gap*, and
- $n_i^p(t)$ is the number of pulls of arm i by player p after t rounds.

[1] Wang et al. "Multitask bandit learning through heterogeneous feedback aggregation." AISTATS 2021.

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



Auxiliary Data: Always Helpful?

Auxiliary data from transfer learning is **not** always helpful!

The **utility of auxiliary data** depends on

- the *dissimilarities* between the player-dependent reward distributions, as indicated by ϵ, and
- the *intrinsic difficulty* of the bandit problem each player faces individually, as indicated by the gaps Δ_i^p's.

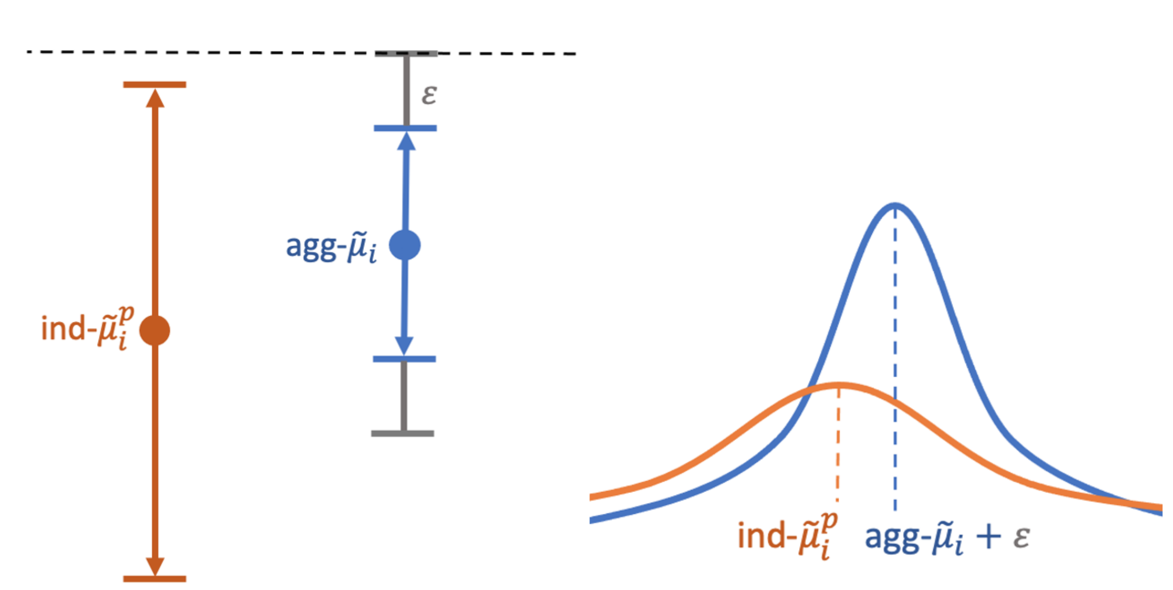
Data aggregation is only provably beneficial on $\mathcal{O}(\epsilon)$-subpar arms:

- The set of α-subpar arms is defined as
$$\mathcal{I}_\alpha = \{i : \exists p \in [M], \Delta_i^p > \alpha\}.$$
- "Easier" arms for which transfer learning can be effective.

Robust Transfer in ϵ-MPMAB

Bias-variance trade-off: utilizing auxiliary data may

- reduce variance of estimations, and
- introduce bias due to dissimilarity of reward distributions.



- ind-$\tilde{\mu}_i^p$: empirical mean reward of i based on p's own data;
- agg-$\tilde{\mu}_i^p$: empirical mean reward of i based on all players' data.

Upper confidence bound (UCB)-based RobustAgg(ϵ) [1]:
For each arm i and player p, compute adaptive weighting of data to minimize width of confidence intervals.

\+ Near-optimal regret guarantees & fallback guarantee;
— Underwhelming empirical performance (too conservative).

Thompson sampling (TS)-type RobustAgg-TS (ϵ):

For each i and p, maintain two posteriors:

- an individual Gaussian posterior for i based on p's own data:
$$\mathcal{N}\left(\text{ind-}\tilde{\mu}_i^p, \mathcal{O}\left(1/n_i^p\right)\right);$$
- an aggregate Gaussian posterior for i using all players' data:
$$\mathcal{N}\left(\text{agg-}\tilde{\mu}_i + \epsilon, \mathcal{O}\left(1/\textstyle\sum_p n_i^p\right)\right).$$

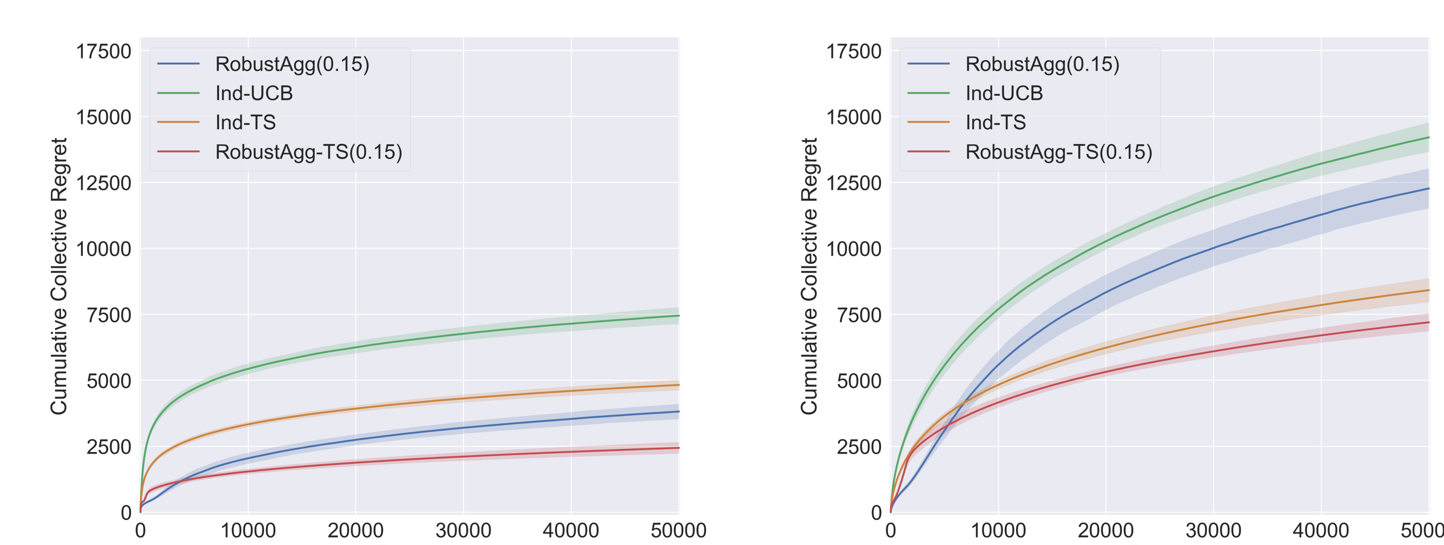
In each round, choose posterior by comparing n_i^p to a threshold in terms of ϵ, and draw sample from chosen posterior.

\+ Near-optimal (slightly weaker) regret guarantees & fallback guarantee;
\+ Superior empirical performance;
— Much harder to analyze.

Regret bound comparison (gap-dependent):

Ind-UCB/Ind-TS	$\mathcal{O}\left(\sum_{i\in[K]} \sum_{p\in[M]:\Delta_i^p>0} \frac{\ln T}{\Delta_i^p}\right)$
RobustAgg(ϵ) [1]	$\tilde{\mathcal{O}}\left(\frac{1}{M} \sum_{i\in\mathcal{I}_{5\epsilon}} \sum_{p\in[M]} \frac{\ln T}{\Delta_i^p} + \sum_{i\in\mathcal{I}_{5\epsilon}^C} \sum_{p\in[M]:\Delta_i^p>0} \frac{\ln T}{\Delta_i^p}\right)$
RobustAgg-TS (ϵ)	$\tilde{\mathcal{O}}\left(\frac{1}{M} \sum_{i\in\mathcal{I}_{10\epsilon}} \sum_{p\in[M]} \frac{\ln T}{\Delta_i^p} + \sum_{i\in\mathcal{I}_{10\epsilon}^C} \sum_{p\in[M]:\Delta_i^p>0} \frac{\ln T}{\Delta_i^p}\right)$
Lower Bound [1]	$\Omega\left(\frac{1}{M} \sum_{i\in\mathcal{I}_{\epsilon/4}} \sum_{p\in[M]:\Delta_i^p>0} \frac{\ln T}{\Delta_i^p} + \sum_{i\in\mathcal{I}_{\epsilon/4}^C} \sum_{p\in[M]:\Delta_i^p>0} \frac{\ln T}{\Delta_i^p}\right)$

Empirical validation:



(a) $|\mathcal{I}_\epsilon| = 8$ (b) $|\mathcal{I}_\epsilon| = 5$

Figure 1: Average performance in randomly generated Bernoulli 0.15-MPMAB problem instances with $K = 10$ and $M = 20$.