Thompson Sampling for Robust Transfer in Multi-Task Bandits

Zhi Wang [1] Chicheng Zhang [2] Kamalika Chaudhuri [1 3]

Abstract

We study the problem of online multi-task learning where the tasks are performed within similar but not necessarily identical multi-armed bandit environments. In particular, we study how a learner can improve its overall performance across multiple related tasks through robust transfer of knowledge. While an upper confidence bound (UCB)-based algorithm has recently been shown to achieve nearly-optimal performance guarantees in a setting where all tasks are solved concurrently, it remains unclear whether Thompson sampling (TS) algorithms, which have superior empirical performance in general, share similar theoretical properties. In this work, we present a TS-type algorithm for a more general online multi-task learning protocol, which extends the concurrent setting. We provide its frequentist analysis and prove that it is also nearly-optimal using a novel concentration inequality for multi-task data aggregation at random stopping times. Finally, we evaluate the algorithm on synthetic data and show that the TS-type algorithm enjoys superior empirical performance in comparison with the UCB-based algorithm and a baseline algorithm that performs TS for each individual task without transfer.

1. Introduction

We study multi-task transfer learning in a multi-armed bandit (MAB) setting. In practice, auxiliary data from different but related sources are often available, although it is also often less clear how they should be utilized. If properly managed, such data can serve an important role in accelerating learning; in particular, in online learning, auxiliary data may be used to avoid costs associated with unnecessary exploration. In this work, we study how data collected from similar sources can be robustly aggregated and utilized.

We consider a generalization of the ϵ-multi-player multi-armed bandit (ϵ-MPMAB) problem recently proposed by Wang et al. (2021), which can be used to model multi-task bandits. In the ϵ-MPMAB problem[1], a set of players sequentially and potentially concurrently interact with a common set of arms that have player-dependent reward distributions. Each player and its associated reward distributions (data sources) are thereby regarded as a task. Furthermore, we consider the reward distributions that the players face for each arm to be *similar* but not necessarily identical, and the level of (dis)similarity is specified by a parameter $\epsilon \in [0, 1]$.

The ϵ-MPMAB problem can be used to model important real-world applications. For example, in healthcare robotics, a set of robots, which correspond to players, can be paired with people with dementia to provide personalized cognitive training and wellness activities (Kubota et al., 2020). Each training/wellness activity corresponds to an arm in the ϵ-MPMAB problem, and people with similar preferences or symptoms may exhibit similar interests or needs—this is modeled via similarity in reward distributions of each arm (Wang et al., 2021). Another example can be seen in recommendation systems where learning agents are assigned to people within a social network, who may have similar interests due to inter-network influence (Qian et al., 2013).

Despite the similarity in its reward distributions, the ϵ-MPMAB problem is still challenging for two reasons: on the one hand, misusing auxiliary data can lead to negative transfer and substantially impair a player's performance (Rosenstein et al., 2005); on the other hand, while auxiliary data are often immediately accessible in their entirety in offline transfer learning settings, in the ϵ-MPMAB problem, the available auxiliary data grow in time and depend on the interactions between the players and the environments.

An upper confidence bound (UCB)-based algorithm, RobustAgg(ϵ), has been proposed for the ϵ-MPMAB problem (Wang et al., 2021). It achieves strong, near-optimal theoretical guarantees through robust data aggregation. Nevertheless, RobustAgg(ϵ)'s empirical performance can, unfortunately, be underwhelming.

Meanwhile, Thompson sampling (TS) algorithms (Thomp-

[1] University of California San Diego [2] University of Arizona [3] Facebook AI Research. Correspondence to: Zhi Wang <zhiwang@ucsd.edu>.

Proceedings of the 39^th International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022. Copyright 2022 by the author(s).

[1] We shall still refer to the generalized problem as the ϵ-MPMAB problem.

son, 1933), another family of bandit algorithms, have been shown superior empirically in comparison with UCB-based algorithms in standard single-task settings (e.g., Chapelle & Li, 2011). In fact, we show in Section 7 that, for the ϵ-MPMAB problem, a baseline algorithm which employs TS for each task individually without transfer learning can outperform RobustAgg(ϵ) in many cases.

In spite of the encouraging signs from the empirical evaluations, the theoretical study of TS have lagged behind, especially in terms of *frequentist* analyses (Agrawal & Goyal, 2017; Kaufmann et al., 2012) for data aggregation and transfer learning in the multi-task setting[2]. It is therefore imperative to design multi-task TS-type algorithms that have superior empirical performance *and* strong theoretical guarantees. Our contributions in this work are:

1. Inspired by prior works (Cesa-Bianchi et al., 2013; Gentile et al., 2014; Hong et al., 2021), we generalize the ϵ-MPMAB problem (Wang et al., 2021) to model a wider class of multi-task bandit learning scenarios so that it covers sequential and concurrent multi-task learning as special cases.
2. We design a TS-type algorithm, RobustAgg-TS(ϵ), for the ϵ-MPMAB problem and provide a frequentist analysis with near-optimal performance guarantees.
3. We empirically evaluate RobustAgg-TS(ϵ) on synthetic data and show that it outperforms the UCB-based RobustAgg(ϵ) and a baseline algorithm that runs TS for each individual task without data sharing.
4. Technical highlight: frequentist analyses of Thompson sampling can be much harder to conduct than those of UCB-based algorithms (see Remark 5.2); a concentration inequality loose in logarithmic factors can result in a polynomial increase in regret guarantee (see Remark 5.7). To cope with this challenge, we prove a novel concentration inequality for multi-task data aggregation at random stopping times (Lemma 5.6), which leads to tight performance guarantees for RobustAgg-TS(ϵ). Our technique may be of independent interest for analyzing other multi-task sequential learning problems.

2. Preliminaries

In this section, we first present the problem formulation and some important known results. We then introduce a new baseline algorithm based on TS.

Notations. Throughout, we use $[n]$ to denote the set $\{1, 2, \ldots, n\}$. Let $\mathcal{N}(\mu, \sigma^2)$ denote the Gaussian distribution with mean μ and variance σ^2. Let $a \vee b = \max(a, b)$. For a set $A \subseteq U$, denote by $A^C = U \setminus A$ the complement of A in the universe U. We use $\tilde{O}$ to hide logarithmic factors.

2.1. Problem Formulation

We consider and generalize the ϵ-MPMAB problem introduced by Wang et al. (2021). An ϵ-MPMAB problem instance comprises M players, K arms, and a dissimilarity parameter $\epsilon \in [0, 1]$. Let $[M]$ denote the set of players and $[K]$ the set of arms. For each player $p \in [M]$ and each arm $i \in [K]$, there is an initially-unknown reward distribution $\mathcal{D}_i^p$, which has support $[0, 1]$ and has mean μ_i^p.

Reward dissimilarity. The reward distributions for each arm are assumed to be *similar but not necessarily identical* for different players; specifically,

$$\forall i \in [K],\ p, q \in [M], \quad \left|\mu_i^p - \mu_i^q\right| \le \epsilon. \tag{1}$$

Protocol. In the work of Wang et al. (2021), the players interact with the arms in rounds, and within each round, all players take an action concurrently. In this paper, inspired by the problem setup of Hong et al. (2021), we generalize the interaction protocol such that it allows any subset of the players to take an action. In each round $t \in [T]$, where $T > \max(K, M)$ is the time horizon of learning, a subset of players $\mathcal{P}_t \subseteq [M]$ is chosen (called the *active player set* at round t) by an oblivious adversary; each active player $p \in \mathcal{P}_t$ then pulls an arm $i_t^p \in [K]$ and observes an independently-drawn reward $r_t^p \sim \mathcal{D}_{i_t^p}^p$. At the end of round t, the active players communicate their decisions, $\left\{i_t^p : p \in \mathcal{P}_t\right\}$, as well as their observed rewards, $\left\{r_t^p : p \in \mathcal{P}_t\right\}$, to all players. Note that, when $|\mathcal{P}_t| = 1$ for all t, the problem setting resembles the one in (Cesa-Bianchi et al., 2013) and captures a sequential transfer bandit learning setting (e.g., Azar et al., 2013); when $\mathcal{P}_t = [M]$ for all t, we recover the setting in the work of Wang et al. (2021).

Performance metric. The goal of the players is to minimize their expected collective regret, which we define shortly. For each player $p \in [M]$, let $\mu_*^p = \max_{j \in [K]} \mu_j^p$ denote the mean reward of an optimal arm for p; then, for each arm $i \in [K]$, let $\Delta_i^p = \mu_*^p - \mu_i^p \ge 0$ denote the (suboptimality) gap of arm i for player p. In addition, let $n_i^p(t) = \sum_{s \le t} \mathbb{1}\left\{p \in \mathcal{P}_s, i_s^p = i\right\}$ denote the number of pulls of arm i by player p after t rounds. Then, the individual expected regret of any player p is defined as

$$\mathrm{Reg}^p(T) = \mathbb{E}\left[\sum_{\substack{t \in [T]: \\ p \in \mathcal{P}_t}} \mu_*^p - \mu_{i_t^p}^p\right] = \sum_{i \in [K]} \mathbb{E}\left[n_i^p(T)\right] \Delta_i^p.$$

[2] See Section 6 for a discussion on related work.

Finally, the *expected collective regret* is defined as the sum of individual expected regret over all the players, i.e.,

$$\mathrm{Reg}(T) = \sum_{p\in[M]} \mathrm{Reg}^p(T) = \sum_{i\in[K]}\sum_{p\in[M]} \mathbb{E}\left[n_i^p(T)\right]\Delta_i^p. \tag{2}$$

Does one need to know ϵ? In this work, we focus on the case where ϵ is *known* to the players in the ϵ-MPMAB problem. This is because Wang et al. (2021) prove that, unfortunately, not much can be done when ϵ is unknown to the players—a lower bound (Theorem 11 therein) shows that no sublinear-regret algorithms can effectively take advantage of inter-task data aggregation for every $\epsilon \in [0, 1]$ to achieve improved regret upper bounds.

2.2. Existing Results

In the concurrent setting ($\mathcal{P}_t = [M]$ for all t), Wang et al. (2021) show that, whether data aggregation can be provably beneficial for an arm i depends on how its associated suboptimality gaps, Δ_i^p's, compare with the dissimilarity parameter, ϵ.

Subpar arms. Specifically, the problem complexity is captured by a notion called *subpar arms*. The set of α-subpar arms is defined as:

$$\mathcal{I}_\alpha = \left\{i \in [K] : \exists p,\ \Delta_i^p > \alpha\right\}. \tag{3}$$

Regret guarantees. The upper and lower bounds provided in (Wang et al., 2021) characterize that, informally, the collective performance of the players can be improved by a factor of M (resp. $\sqrt{M}$) for each $\mathcal{O}(\epsilon)$-subpar arm in the (suboptimality) gap-dependent (resp. gap-independent) bounds, where we recall that M is the number of players.

This improvement is in comparison with baseline algorithms in which each player runs their own instance of a bandit algorithm individually. Let IND-UCB be a baseline in which each player runs the UCB-1 algorithm (Auer et al., 2002). Its collective regret guarantees are obtained by simply summing over individual gap-dependent and gap-independent regret bounds, respectively: $\mathcal{O}\left(\sum_{p\in[M]}\sum_{i\in[K]:\Delta_i^p>0}\frac{\ln T}{\Delta_i^p}\right)$ and $\tilde{\mathcal{O}}\left(M\sqrt{KT}\right)$.

In contrast, through leveraging auxiliary data from inter-player communication, the UCB-based algorithm, ROBUSTAGG(ϵ), proposed by Wang et al. (2021) has gap-dependent and gap-independent regret bounds of

$$\mathcal{O}\left(\underbrace{\frac{1}{M}\sum_{i\in\mathcal{I}_{5\epsilon}}\sum_{\substack{p\in[M]\\ \Delta_i^p>0}}\frac{\ln T}{\Delta_i^p}}_{(*)} + \sum_{i\in\mathcal{I}_{5\epsilon}^C}\sum_{\substack{p\in[M]\\ \Delta_i^p>0}}\frac{\ln T}{\Delta_i^p} + MK\right) \text{ and}$$

$$\tilde{\mathcal{O}}\left(\underbrace{\sqrt{M|\mathcal{I}_{5\epsilon}|T}}_{(*)} + M\sqrt{\left(|\mathcal{I}_{5\epsilon}^C| - 1\right)T} + MK\right),$$

respectively[3]. These guarantees exhibit a factor of $\frac{1}{M}$ and $\frac{1}{\sqrt{M}}$ improvement in the respective $(*)$ terms, for the set of $\mathcal{O}(\epsilon)$-subpar arms, $\mathcal{I}_{5\epsilon}$, and is nearly optimal.

In Appendix D, we give a brief recap of ROBUSTAGG(ϵ)—we show that with a few small modifications, it can be extended to work in the generalized ϵ-MPMAB setting, and achieve generalized regret guarantees (see Theorem D.2).

Lower bounds. In the setting where the dissimilarity parameter ϵ is known, a lower bound in (Wang et al., 2021) shows that, for any algorithm that has a sublinear-regret guarantee, when facing a large class of ϵ-MPMAB problem instances, it must have regret at least

$$\Omega\left(\sum_{i\in\mathcal{I}_{\epsilon/4}:\Delta_i^{\min}>0}\frac{\ln T}{\Delta_i^{\min}} + \sum_{i\in\mathcal{I}_{\epsilon/4}^C}\sum_{p\in[M]:\Delta_i^p>0}\frac{\ln T}{\Delta_i^p}\right),$$

where $\Delta_i^{\min} = \min_{p\in[M]}\Delta_i^p$. This lower bound shows that, data aggregation cannot be effective for the arm set $\mathcal{I}_{\epsilon/4}^C \subseteq \mathcal{I}_\epsilon^C$.

In addition, Wang et al. (2021) also show a gap-independent lower bound: for any algorithm, there exists an ϵ-MPMAB instance, in which the algorithm has regret at least

$$\Omega\left(\sqrt{M|\mathcal{I}_{5\epsilon}|T} + M\sqrt{\left(|\mathcal{I}_{5\epsilon}^C| - 1\right)T}\right),$$

in the setting where $\mathcal{P}_t = [M]$ for all $t \in [T]$.

2.3. Baseline: IND-TS

In this work, we consider another baseline algorithm, IND-TS, in which each player runs the standard TS algorithm with Gaussian priors. We now describe the TS algorithm. At a high level, every learner (player) p begins with some prior belief on the mean reward of each arm, and through interactions with the environment, the learner updates its posterior belief. Specifically, we consider TS with Gaussian product priors—a learner maintains one Gaussian posterior distribution for each arm, beginning with $\mathcal{N}(0, 1)$. In each round t, the learner draws an independent sample $\theta_i^p(t)$ for each arm i from its corresponding posterior distribution, which is of form $\mathcal{N}\left(\bar{\mu}_i^p, \frac{1}{n_i^p(t-1)\vee 1}\right)$, where $\bar{\mu}_i^p = \frac{1}{n_i^p(t-1)\vee 1}\sum_{s<t: p\in\mathcal{P}_s, i_s^p = i} r_s^p$ is the empirical mean reward of player p pulling arm i. The learner then pulls the arm $i_t^p = \mathrm{argmax}_i\, \theta_i^p(t)$, receives a reward $r_t^p \sim \mathcal{D}_{i_t^p}^p$, and updates the posterior distribution for arm i.

[3] The results may appear different from (Wang et al., 2021) at a glance because we use a slightly notation for subpar arms.

Based on the results of Agrawal & Goyal (2017), we obtain the regret guarantees of IND-TS by summing over individual bounds: $\mathcal{O}\left(\sum_{p\in[M]}\sum_{i\in[K]:\Delta_i^p>0}\frac{\ln T}{\Delta_i^p}\right)$ and $\tilde{\mathcal{O}}\left(M\sqrt{KT}\right)$.

In Appendix D, we briefly recap the guarantees of IND-UCB and IND-TS in the generalized ϵ-MPMAB setting, where $\mathcal{P}_t$'s are not necessarily $[M]$ in every round.

Algorithm 1 ROBUSTAGG-TS (ϵ)

```
1:  Input: Dissimilarity parameter $\epsilon \in [0,1]$, universal constants $c_1, c_2 > 0$.
2:  Initialization: For every $i \in [K]$ and $p \in [M]$, set $n_i^p = 0$, ind-$\hat{\mu}_i^p = 0$, ind-$\text{var}_i^p = c_2$, agg-$\hat{\mu}_i^p = 0$, and agg-$\text{var}_i^p = c_2$; for every $i \in [K]$, set $n_i = 0$.
3:  for round $t \in [T]$ do
4:    Receive active set of players $\mathcal{P}_t$.
5:    for active player $p \in \mathcal{P}_t$ do
6:      for arm $i \in [K]$ do
7:        if $n_i^p \geq \frac{c_1 \ln T}{\epsilon^2} + 2M$ then
8:          $\hat{\mu}_i^p \leftarrow$ ind-$\hat{\mu}_i^p$, $\text{var}_i^p \leftarrow$ ind-$\text{var}_i^p$;
                                          ▷ Use the individual posterior
9:        else
10:         $\hat{\mu}_i^p \leftarrow$ agg-$\hat{\mu}_i^p$, $\text{var}_i^p \leftarrow$ agg-$\text{var}_i^p$;
                                          ▷ Use the aggregate posterior
11:       end if
12:       $\theta_i^p(t) \sim \mathcal{N}(\hat{\mu}_i^p, \text{var}_i^p)$
13:     end for
14:     Player $p$ pulls arm $i_t^p = \text{argmax}_{i\in[K]}\, \theta_i^p(t)$ and observes reward $r_t^p$.
15:   end for
16:   for active player $p \in \mathcal{P}_t$ do
17:     Let $i = i_t^p$. Update $n_i^p \leftarrow n_i^p + 1$ and $n_i \leftarrow n_i + 1$.
18:   end for
19:   for active player $p \in \mathcal{P}_t$ do
20:     Let $i = i_t^p$.
        ▷ Only update posteriors associated with $p$ and $i_t^p$
21:     Update
          $\text{ind-}\hat{\mu}_i^p \leftarrow \frac{1}{n_i^p \vee 1}\sum_{s\leq t}\mathbb{1}\{p\in\mathcal{P}_s, i_s^p = i\}\, r_s^p,$
          $\text{ind-var}_i^p \leftarrow \frac{c_2}{n_i^p \vee 1};$
22:
          $\text{agg-}\hat{\mu}_i^p \leftarrow \frac{1}{n_i \vee 1}\sum_{s\leq t}\sum_{q\in\mathcal{P}_s}\mathbb{1}\{i_s^q = i\}\, r_s^q + \epsilon,$
          $\text{agg-var}_i^p \leftarrow \frac{c_2}{(n_i - M)\vee 1}.$
23:   end for
24: end for
```

3. Algorithm

In this section, we present a TS-type randomized exploration algorithm, ROBUSTAGG-TS(ϵ) (Algorithm 1), which can robustly leverage data collected by all the players.

In each round t, for each active player $p \in \mathcal{P}_t$ and arm i, ROBUSTAGG-TS(ϵ) maintains two Gaussian "posterior" distributions. As a standard single-task TS algorithm with Gaussian priors would normally maintain (e.g. Agrawal & Goyal, 2017), $\mathcal{N}\left(\text{ind-}\hat{\mu}_i^p, \text{ind-var}_i^p\right)$, the *individual posterior* is solely based on player p's own interactions with arm i, with ind-$\hat{\mu}_i^p$ and ind-var_i^p defined in line 21. In contrast, the *aggregate posterior*, $\mathcal{N}\left(\text{agg-}\hat{\mu}_i^p, \text{agg-var}_i^p\right)$, is unique to the multi-task setting—its mean, agg-$\hat{\mu}_i^p$, is the sum of the empirical mean of all players' observed rewards for arm i and a bonus term ϵ, and its variance, agg-var_i^p, is based on the total number of pulls of arm i by all players (line 22).

The algorithm chooses one of the posterior distributions (lines 7 to 11), i.e., decides whether to utilize data shared by other players, by balancing a bias-variance trade-off (Ben-David et al., 2010; Soare et al., 2014; Wang et al., 2021): while an inclusion of n_i reward samples collected by all players leads to a variance, agg-var_i^p, which can be much smaller than ind-var_i^p, it may also cause agg-$\hat{\mu}_i^p$ to be biased as the reward distributions for different players may be different. The algorithm then independently draws a sample, $\theta_i^p(t)$, from the chosen posterior distribution (line 12) and pulls the arm with the largest $\theta_i^p(t)$ for player p (line 14).

Specifically, in round t, for player $p \in \mathcal{P}_t$ and arm $i \in [K]$, the algorithm chooses a posterior distribution by comparing n_i^p, the number of pulls of i by p at the beginning of round t, to a threshold in terms of the dissimilarity parameter, i.e., $\frac{c_1 \ln T}{\epsilon^2} + 2M$ (line 7), where $c_1 > 0$ is some numerical constant. Intuitively, when ϵ is smaller, each player stays longer on using the aggregate posterior to perform randomized exploration, which indicates a higher degree of trust on data from other tasks.

After all players in $\mathcal{P}_t$ obtain rewards for their arm pulls, they compute and *update* their posteriors with new data. In principle, data from one player can affect the aggregate posteriors of all players. We make the design choice that this effect gets delayed: the algorithm only updates the posteriors for player p and arm i in round t, if $p \in \mathcal{P}_t$ and $i = i_t^p$ (line 20). Although our current analysis (see Sections 4 and 5 below) relies on this property to establish sharp regret guarantees, we conjecture that similar regret guarantees can be shown even if the algorithm updates the posteriors of all players and all arms in every round[4].

[4] In Section E.1 of the appendix, we show that this variation induces little effect on the empirical performance of the algorithm.

4. Main Results

We now present gap-dependent and gap-independent regret upper bounds of RobustAgg-TS(ϵ). Recall that $\mathcal{I}_\alpha = \{i \in [K] : \exists p,\ \Delta_i^p > \alpha\}$ is the set of α-subpar arms.

Theorem 4.1 (Gap-dependent bound)**.** *There exists a setting of $c_1, c_2 > 0$, such that, the expected collective regret of* RobustAgg-TS(ϵ) *after $T > \max(K, M)$ rounds satisfies:* $\mathrm{Reg}(T) \leq$

$$\mathcal{O}\left(\frac{1}{M}\sum_{i \in \mathcal{I}_{10\epsilon}} \sum_{\substack{p \in [M] \\ \Delta_i^p > 0}} \frac{\ln T}{\Delta_i^p} + \sum_{i \in \mathcal{I}_{10\epsilon}^C} \sum_{\substack{p \in [M] \\ \Delta_i^p > 0}} \frac{\ln T}{\Delta_i^p} + M^2 K\right).$$

Theorem 4.2 (Gap-independent bound)**.** *There exists a setting of $c_1, c_2 > 0$, such that, the expected collective regret of* RobustAgg-TS(ϵ) *after $T > \max(K, M)$ rounds satisfies:*

$$\mathrm{Reg}(T) \leq \tilde{\mathcal{O}}\left(\sqrt{|\mathcal{I}_{10\epsilon}| P} + \sqrt{M\left(|\mathcal{I}_{10\epsilon}^C| - 1\right)P} + M^2 K\right),$$

where $P = \sum_{t=1}^{T} |\mathcal{P}_t|$.

The proofs of Theorems 4.1 and 4.2 can be found in Appendix C; in Section 5, we also highlight several technical challenges and proof ingredients in our analysis.

Guarantees in the generalized ϵ-MPMAB setting. Our guarantees for RobustAgg-TS(ϵ) hold under the generalized ϵ-MPMAB setting, in that $\mathcal{P}_t$'s at each round can change over time. Observe that the regret bound given by Theorem 4.1 does not depend on $\mathcal{P}_t$'s, and the regret bound given by Theorem 4.2 has the highest value when $P = MT$. In addition, recall that near-matching gap-dependent and gap-independent lower bounds have been shown by Wang et al. (2021) in the $\mathcal{P}_t \equiv [M]$ setting (Section 2.2). These lower bounds indicate the near-optimality of RobustAgg-TS(ϵ)'s guarantees, modulo an additive lower-order term $O(M^2 K)$ which does not depend on T.

Furthermore, the gap-independent guarantee in Theorem 4.2 adapts to the value of P. This shows the flexibility of RobustAgg-TS(ϵ). Specifically, if $|\mathcal{P}_t| = 1$ (similar to the settings of Cesa-Bianchi et al. 2013; Gentile et al. 2014), we have $P = T$, and $\mathrm{Reg}(T) \leq$

$$\tilde{\mathcal{O}}\left(\sqrt{|\mathcal{I}_{10\epsilon}| T} + \sqrt{M\left(|\mathcal{I}_{10\epsilon}^C| - 1\right)T} + M^2 K\right).$$

Similarly, if $\mathcal{P}_t = [M]$ for all t (Wang et al., 2021), then $P = MT$, and $\mathrm{Reg}(T) \leq$

$$\tilde{\mathcal{O}}\left(\sqrt{M|\mathcal{I}_{10\epsilon}| T} + M\sqrt{\left(|\mathcal{I}_{10\epsilon}^C| - 1\right)T} + M^2 K\right).$$

Comparison with baselines. In comparison with the guarantees of the UCB-based algorithm RobustAgg(ϵ) in Appendix D.2, we see that RobustAgg-TS(ϵ) has competitive guarantees, except that the set of arms which benefits from data aggregation changes from $\mathcal{I}_{5\epsilon}$ to $\mathcal{I}_{10\epsilon}$.

In comparison with the guarantees of Ind-UCB and Ind-TS, the regret guarantees of RobustAgg-TS(ϵ) are never worse (modulo lower-order terms), and save factors of $\frac{1}{M}$ and $\frac{1}{\sqrt{M}}$ in $\mathcal{I}_{10\epsilon}$'s contribution in the gap-dependent and gap-independent regret guarantees, respectively.

5. Proof Ingredients

In this section, we highlight some of the novel proof ingredients used in our analysis of Algorithm 1, which are unique to the *multi-task* setting[5].

We begin by decomposing the regret in terms of subpar arms and non-subpar arms. It follows from Eq. (2) that

$$\mathrm{Reg}(T) = \mathcal{O}\left(\sum_{i \in \mathcal{I}_{10\epsilon}} \mathbb{E}\left[n_i(T)\right] \Delta_i^{\min} + \sum_{i \in \mathcal{I}_{10\epsilon}^C} \sum_{p \in [M]} \mathbb{E}\left[n_i^p(T)\right] \Delta_i^p\right),$$

where we let $n_i(T) = \sum_{p=1}^{M} n_i^p(T)$ be the number of pulls of arm i by all players after T rounds; we recall that $\Delta_i^{\min} = \min_{p \in [M]} \Delta_i^p$; and we use the fact that for any subpar arm $i \in \mathcal{I}_{10\epsilon}$ and any player $p \in [M]$, $\Delta_i^p \leq 2\Delta_i^{\min}$ (Fact A.24).

In the interest of space, we focus on the analysis for subpar arms and defer the discussion on non-subpar arms to the appendix. The following lemma provides an upper bound on $\mathbb{E}\left[n_i(T)\right]$ for $i \in \mathcal{I}_{10\epsilon}$, which can be subsequently used to derive the upper bounds on the expected collective regret incurred by the 10ϵ-subpar arms in Section 4.

Lemma 5.1. *For any arm $i \in \mathcal{I}_{10\epsilon}$,*

$$\mathbb{E}\left[n_i(T)\right] \leq \mathcal{O}\left(\frac{\ln T}{(\Delta_i^{\min})^2} + M\right).$$

While a similar lemma can be found in (Wang et al., 2021, Lemma 20) for the UCB-based algorithm, RobustAgg(ϵ), proving Lemma 5.1 requires new ingredients that we present in the rest of this section.

Let us fix an arm $i \in \mathcal{I}_{10\epsilon}$. To control $\mathbb{E}\left[n_i(T)\right] = \mathbb{E}\left[\sum_{t \in [T]} \sum_{p \in \mathcal{P}_t} \mathbb{1}\left\{i_t^p = i\right\}\right]$, we begin by generalizing a

[5] Our analysis involves various proofs by cases. Figure 2 in the appendix provides an overview illustrating the case division rules used in our proofs.

technique introduced by Agrawal & Goyal (2017) for standard TS to the multi-task setting. In each round t and for each active player p, we consider two cases: (1) player p pulls arm i (namely, $i_t^p = i$), and $\theta_i^p(t)$ (line 12 in Algorithm 1) is greater than some threshold $y_i^p \in (\mu_i^p, \mu_*^p)$ to be defined shortly, and (2) $i_t^p = i$ and $\theta_i^p(t) \leq y_i^p$. We have

$$\mathbb{E}\left[n_i(T)\right] = \underbrace{\mathbb{E}\left[\sum_{t\in[T]}\sum_{p\in\mathcal{P}_t} \mathbb{1}\left\{i_t^p = i, \theta_i^p(t) > y_i^p, \mathcal{E}_t\right\}\right]}_{(A)}$$
$$+ \underbrace{\mathbb{E}\left[\sum_{t\in[T]}\sum_{p\in\mathcal{P}_t} \mathbb{1}\left\{i_t^p = i, \theta_i^p(t) \leq y_i^p, \mathcal{E}_t\right\}\right]}_{(B)} + \mathcal{O}\left(1\right),$$

where $\mathcal{E}_t$, informally, is a high-probability "clean" event in which $\hat{\mu}_i^p$'s maintained by Algorithm 1 in round t for each i and p concentrate towards their respective expected values.

Term (A) can be controlled because, as more pulls of arm i are made, $\left\{\theta_i^p(t) > y_i^p\right\}$ is unlikely to happen, as $\hat{\mu}_i^p$ concentrates towards a value smaller than y_i^p, and var_i^p decreases. See Lemma C.6 in the appendix for a detailed proof.

In what follows, we focus on bounding term (B). Observe that the event $\left\{i_t^p = i, \theta_i^p(t) \leq y_i^p\right\}$ in (B) happens only if $\forall j \in [K]$, $\theta_j^p(t) \leq y_i^p$, including the optimal arm(s) for player p. Since in an ϵ-MPMAB problem instance, different players may have different optimal arms, we consider a common near-optimal arm $\dagger \in \mathcal{I}_{2\epsilon}^C$—see Fact A.24 in the appendix for the existence of such an arm. It can be easily verified that, for any arm $i \in \mathcal{I}_{10\epsilon}$ and player $p \in [M]$, $\delta_i^p := \mu_\dagger^p - \mu_i^p > 0$ (see Fact C.4). In other words, while $\dagger$ may not necessarily be an optimal arm for every player, it has a larger mean reward than any $i \in \mathcal{I}_{10\epsilon}$. We can now define $y_i^p := \mu_i^p + \frac{1}{2}\delta_i^p \in (\mu_i^p, \mu_\dagger^p) \subset (\mu_i^p, \mu_*^p)$.

Using a technique first introduced in (Agrawal & Goyal, 2017), we will show that $\theta_\dagger^p(t)$ converges to a value greater than y_i^p *fast* enough so that $\left\{\forall j \in [K], \theta_j^p(t) \leq y_i^p\right\}$ will unlikely happen soon enough and thus (B) can be controlled.

Remark 5.2 (Comparison with UCB-based analyses)**.** We note that controlling term (B) is often not required in the analyses of UCB-based algorithms. Colloquially, this term concerns the event in which arm i is pulled even when its sample/index value is smaller than y_i^p; such an event would unlikely happen for UCB-based algorithms as the *optimism in the face of uncertainty* principle ensures that, with high probability, the UCB index of an optimal arm for player p is greater than or equal to $\mu_*^p \geq \mu_\dagger^p > y_i^p$.

Before we formalize the above-mentioned intuition for bounding term (B) in Lemma 5.3, we first lay out a few helpful definitions. We define $\{\mathcal{F}_t\}_{t=0}^T$ to be a filtration such that $\mathcal{F}_t = \sigma\left(\left\{i_s^q, r_s^q : s \leq t, q \in \mathcal{P}_s\right\}\right)$ is the σ-algebra generated by interactions of all players up until round t. Then, let $\phi_{i,t}^p = \Pr\left(\theta_\dagger^p(t) > y_i^p \mid \mathcal{F}_{t-1}\right)$. Observe that if $\phi_{i,t}^p$ is large, the event $\left\{i_t^p = i, \theta_i^p(t) \leq y_i^p\right\}$ will unlikely happen.

Lemma 5.3.

$$(B) \leq \underbrace{\sum_{t\in[T]}\sum_{p\in\mathcal{P}_t} \mathbb{E}\left[\left(\frac{1}{\phi_{i,t}^p} - 1\right)\mathbb{1}\left\{i_t^p = \dagger, \mathcal{E}_t\right\}\right]}_{(B*)}.$$

See Lemma C.11 and its proof in the appendix for details. We now consider the following two cases: in any round t and for any active player p that pulls arm $\dagger$, i.e., $i_t^p = \dagger$, p uses *either* the individual *or* the aggregate posterior distribution associated with arm $\dagger$ (lines 7 to 11 in Algorithm 1). Let $H_\dagger^p(t)$ be the event that p uses the individual posterior distribution and $\overline{H_\dagger^p(t)}$ be the event that p uses the aggregate posterior (see Definition A.13 in the appendix for the formal definitions). We can then decompose $(B*)$ as follows:

$$(B*) = \underbrace{\sum_{t\in[T]}\sum_{p\in\mathcal{P}_t} \mathbb{E}\left[\left(\frac{1}{\phi_{i,t}^p} - 1\right)\mathbb{1}\left\{i_t^p = \dagger, \mathcal{E}_t, H_\dagger^p(t)\right\}\right]}_{(b1)}$$
$$+ \underbrace{\sum_{t\in[T]}\sum_{p\in\mathcal{P}_t} \mathbb{E}\left[\left(\frac{1}{\phi_{i,t}^p} - 1\right)\mathbb{1}\left\{i_t^p = \dagger, \mathcal{E}_t, \overline{H_\dagger^p(t)}\right\}\right]}_{(b2)}.$$

Let $m_\dagger^p(t)$ denote the aggregate number of pulls of arm $\dagger$ maintained by player p after t rounds (see Definition A.9 in the appendix). Note that, by the design choice of Algorithm 1 (line 20), $m_\dagger^p(t)$ is not necessarily the same as $n_\dagger(t)$. With foresight, let $L = \Theta\left(\frac{\ln T}{(\Delta_i^{\min})^2} + M\right)$, and let $G_t^p = \left\{i_t^p = \dagger, \mathcal{E}_t, \overline{H_\dagger^p(t)}\right\}$. We have

$$(b2) =$$
$$\underbrace{\sum_{t\in[T]}\sum_{p\in\mathcal{P}_t} \mathbb{E}\left[\left(\frac{1}{\phi_{i,t}^p} - 1\right)\mathbb{1}\left\{G_t^p, m_\dagger^p(t-1) < L\right\}\right]}_{(b2.1)}$$
$$+ \underbrace{\sum_{t\in[T]}\sum_{p\in\mathcal{P}_t} \mathbb{E}\left[\left(\frac{1}{\phi_{i,t}^p} - 1\right)\mathbb{1}\left\{G_t^p, m_\dagger^p(t-1) \geq L\right\}\right]}_{(b2.2)}.$$

Both $(b1)$ and $(b2.2)$ can be bounded by $\mathcal{O}\left(M\right)$, because, informally speaking, either player p has pulled arm $\dagger$ many times when the individual posterior is used (term $(b1)$) or the players collectively have pulled † many times when the aggregate posterior is used (term $(b2.2)$), and $\frac{1}{\phi^p_{i,t}} - 1$ can therefore be upper bounded by $\frac{1}{T}$. See Lemma C.13 and Lemma C.18 and their proofs for details.

The main challenge in bounding $\mathbb{E}\left[n_i(T)\right]$ lies in term $(b2.1)$, for which we show the following lemma.

Lemma 5.4 (Bounding term $(b2.1)$)**.**

$$(b2.1) \le \mathcal{O}\left(L\right) \le \mathcal{O}\left(\frac{\ln T}{(\Delta_i^{\min})^2} + M\right).$$

Proving Lemma 5.4 is *central* to our analysis and as we will see, requires special care. We begin by introducing the following notion. For any arm $j \in [K]$ and $k \in [TM]$, let

$$\tau_k(j) = \min\left\{T+1, \min\left\{t : n_j(t) \ge k\right\}\right\}$$

be the round in which arm j is pulled the k-th time by any player. Furthermore, let $\tau_0(j) = 0$ by convention. For any $j \in [K]$ and $k \in [TM]$, it is easy to verify that $\tau_k(j)$ is a stopping time with respect to $\{\mathcal{F}_t\}_{t=0}^T$. In what follows, when circumstances permit, we abuse the notation and denote $\tau_k(\dagger)$ by τ_k.

Invariant property. By the construction of Algorithm 1, in any round t, a player only updates the posteriors associated with an arm if the player pulls the arm in the round t (line 20). This design choice induces an invariant property: for any arm and player, certain random variables associated with them stay invariant between consecutive pulls of the arm by the player (see Definition A.20 and a few examples in the appendix).

The invariant property allows us to bound $(b2.1)$ as follows in terms of the stopping times τ_k's (See Lemma C.14 and Lemma C.38 in the appendix):

$$(b2.1) \le \sum_{p=1}^{M} \mathbb{E}\left[\left(\frac{1}{\phi^p_{i,1}} - 1\right) \mathbb{1}\left\{\overline{H^p_\dagger(1)}\right\}\right] + \sum_{k=1}^{L-1} \mathbb{E}\left[\left(\frac{1}{\phi^{p_k}_{i,\tau_k+1}} - 1\right) \mathbb{1}\left\{\tau_k \le T, \overline{H^p_\dagger(\tau_k+1)}\right\}\right],$$

where $p_k := p_k(\dagger)$ is the player that makes the k-th pull of arm † (Definition A.17).

Using basic Gaussian tail bounds, we can show that $\mathbb{E}\left[\left(\frac{1}{\phi^p_{i,1}} - 1\right) \mathbb{1}\left\{\overline{H^p_\dagger(1)}\right\}\right] \le \mathcal{O}\left(1\right)$ for any player p. Then, the following lemma suffices to prove Lemma 5.4.

Lemma 5.5. *For any $k \in [TM]$,*

$$\mathbb{E}\left[\left(\frac{1}{\phi^{p_k}_{i,\tau_k+1}} - 1\right) \mathbb{1}\left\{\tau_k \le T, \overline{H^p_\dagger(\tau_k+1)}\right\}\right] \le \mathcal{O}\left(1\right).$$

Technical highlight. Lemma 5.5 generalizes Agrawal & Goyal (2017, Lemma 2.13) for standard TS to the *multi-task* setting. A complete proof can be found in the appendix, which uses anti-concentration bounds of Gaussian random variables (Gordon, 1941) as well as a *novel* concentration inequality for multi-task data aggregation at random stopping times $\tau_k(\dagger)$'s, which we highlight here[6]. For any arm j, let

$$\text{agg-}\hat{\mu}_j(t) = \frac{1}{n_j(t) \vee 1} \sum_{s \le t} \sum_{q \in \mathcal{P}_s} \mathbb{1}\left\{i^q_s = j\right\} r^q_s + \epsilon$$

be the aggregate mean reward estimate of j constructed using data by all players after t rounds, offset by ϵ.

Lemma 5.6. *For any arm $j \in [K]$ and $k \in [TM] \cup \{0\}$, denote by $\tau_k = \tau_k(j)$. Then, for any $\delta \in (0, 1]$, with probability at least $1 - \delta$, one of the following events happens:*

1. $\tau_k = T + 1$;
2. $\forall p \in [M]$, $\mu^p_j - \text{agg-}\hat{\mu}_j(\tau_k) \le \sqrt{\frac{2\ln\left(\frac{2}{\delta}\right)}{\left(n_j(\tau_k) - M\right) \vee 1}}$.

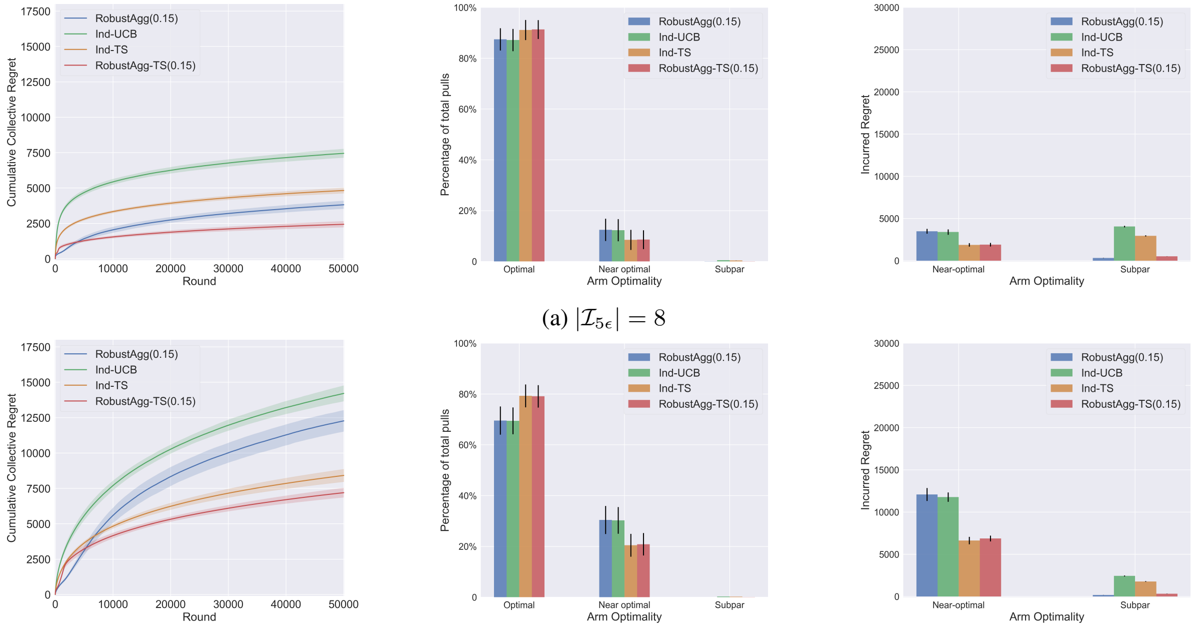
Remark 5.7. We note that Lemma 5.6 is critical to the tight performance guarantee in Lemma 5.5 and subsequently the near-optimal regret guarantees. This result is non-trivial, as it is a concentration bound for a sequence of random variables whose length, $n_j(\tau_k(j))$, is also a random variable. Furthermore, since $\tau_k(j)$ is the round in which arm j is pulled the k-th time by any player, $n_j(\tau_k(j))$ can potentially take any integer value in $[k, k + M - 1]$ because there can be up to M pulls of arm j in round $\tau_k(j)$. We note that using the Azuma-Hoeffding inequality together with a union bound or Freedman's inequality (similar to Wang et al., 2021, Lemma 17) can lead to extra $\mathcal{O}\left(M\right)$ or $\mathcal{O}\left(\ln T\right)$ terms for Lemma 5.5, respectively (see Remark C.17 in the appendix for details).

To our best knowledge, we are not aware of any similar tight concentration bounds for data aggregation in multi-task bandits, and our technique may be of independent interest for analyzing other multi-task sequential learning problems.

6. Related Work

There exist many prior works that study multi-player or multi-task bandits with heterogeneous reward distributions.

[6] In the single-task case ($M = 1$), our proof technique (Lemma C.36) also simplifies the proof of the first case of Agrawal & Goyal (2017, Lemma 2.13).



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

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

Figure 1. Compares the average performance of the algorithms on 30 randomly generated problem instances with $|\mathcal{I}_{5\epsilon}| = 8$ and $|\mathcal{I}_{5\epsilon}| = 5$ in a horizon of $T = 50000$ rounds. Figures in the left column plot the cumulative collective regret over time; figures in the middle column demonstrate the percentages of pulls of optimal arms, non-subpar yet non-optimal arms (referred to as near-optimal arms), and subpar arms; figures in the right column then show the incurred cumulative regret by arm optimality.

For example, Cesa-Bianchi et al. (2013) use Laplacian-based regularization to learn a network of bandit problem instances such that connected problems have similar parameters; Gentile et al. (2014), among others, study clustering of bandit problem instances. The ϵ-MPMAB problem studied in this paper is introduced by Wang et al. (2021); see Appendix A thereof for a detailed comparison with related work. More recently, Zhang & Wang (2021) generalize the ϵ-MPMAB problem to episodic, tabular Markov decision processes. We note that while the methods in the above-mentioned works are UCB-based, we study TS-type algorithms in this work.

TS is initially proposed by Thompson (1933) decades ago, but its frequentist analysis has not emerged until recent years (e.g., Agrawal & Goyal, 2012; Kaufmann et al., 2012). Jin et al. (2021) present the first minimax optimal TS-type algorithm. Our proof techniques in this paper are mostly inspired by the work of Agrawal & Goyal (2017).

TS algorithms have been studied in multi-task Bayesian bandits. For example, several recent works study the setting of interacting with a sequence of M bandit problem instances (tasks) sampled from a common, unknown prior distribution, with a goal of minimizing the M-instance Bayesian regret (Bastani et al., 2021; Kveton et al., 2021; Peleg et al., 2021; Basu et al., 2021). The recent work of Hong et al. (2021) proposes a hierarchical Bayesian bandit problem that generalizes many multi-task bandit settings, and analyzes the Bayes regret. In contrast, we use frequentist regret as our performance metric, and we do not assume a shared prior distribution over the players' problem instances/tasks. Wan et al. (2021) study multi-task TS in a hierarchical Bayesian model and assume knowledge of metadata of each task; while they provide a frequentist regret bound, we study the ϵ-MPMAB problem which models task relations differently.

Similar models on *sequential* transfer between problem instances have also been studied by Azar et al. (2013) and Soare et al. (2014). Zhang & Bareinboim (2017); Zhang et al. (2019); Sharma et al. (2020) investigate warm-starting bandits from misaligned data. In this work, we focus on a more general interaction protocol, under which the players may interact with the environment concurrently.

7. Empirical Evaluation

In this section, we present an empirical evaluation of RobustAgg-TS(ϵ) on synthetic data[7]. We focus on the concurrent setting ($\mathcal{P}_t = [M]$ for all t), which is the setting studied in the experiments of (Wang et al., 2021). Our goal is to address the following two questions:

[7] Our code is available at https://github.com/zhiwang123/eps-MPMAB-TS.

(1) How does RobustAgg-TS(ϵ) perform in comparison with the UCB-based algorithm, RobustAgg(ϵ), and the baseline algorithms without transfer learning?

(2) Does the notion of subpar arms characterize the performance of the algorithms in practice?

Experimental Setup. We compared the performance of 4 algorithms: (1) RobustAgg-TS(ϵ) with constants $c_1 = \frac{1}{2}$ and $c_2 = 1$; (2) RobustAgg(ϵ) (Wang et al., 2021, Section 6.1); (3) Ind-TS, the baseline algorithm that runs TS with Gaussian priors for each player individually; and (4) Ind-UCB, the baseline algorithm that runs UCB-1 for each player individually.

The algorithms were evaluated on randomly generated 0.15-MPMAB problem instances with different numbers of subpar arms. To stay consistent with the work of Wang et al. (2021), we followed the same instance generation procedure and considered $\mathcal{I}_{5\epsilon}$ to be the set of subpar arms—we set the number of players $M = 20$ and the number of arms $K = 10$; then, for each integer value $v \in [0, 9]$, we generated 30 0.15-MPMAB problem instances with Bernoulli reward distributions and $|\mathcal{I}_{5\epsilon}| = v$. We ran the algorithms on each instance for a horizon of $T = 50,000$ rounds.

Results and Discussion. Figure 1 compares the average performance of the algorithms on instances with $|\mathcal{I}_{5\epsilon}| = 8$ and 5. We defer the rest of the results to Appendix E.

From the left column, we first observe that, while the UCB-based algorithm, RobustAgg(ϵ), outperforms its counterpart, Ind-UCB, in the cumulative collective regret ($\sum_{t\in[T]} \sum_{p\in\mathcal{P}_t} \mu_*^p - \mu_{i_t^p}^p$), its empirical performance is underwhelming in comparison with TS algorithms. In particular, even on instances with half of the arms *subpar* ($|\mathcal{I}_{5\epsilon}| = 5$), RobustAgg($\epsilon$) is outperformed by the Ind-TS baseline without transfer learning. Importantly, we note that RobustAgg-TS(ϵ) shows a superior performance than the other algorithms.

The figures in the middle and right columns illustrate the arm selection of each algorithm. We categorize all arms into three groups: optimal arms, subpar arms, and near-optimal arms which are neither subpar nor optimal. Comparing the TS-type algorithms with the UCB-based algorithms, we observe that the former algorithms perform better mainly because they pull near-optimal arms a smaller number of times and incur less regret on these arms.

Furthermore, we observe that RobustAgg(ϵ) and RobustAgg-TS(ϵ), when compared with their counterparts (Ind-UCB and Ind-TS, respectively), incur a similar amount of regret from near-optimal arms. Meanwhile, they make fewer pulls on subpar arms. This may be less obvious from the plots on the percentage of total pulls because none of the algorithms pull subpar arms extensively over the horizon. However, since the suboptimality gaps of subpar arms are large, we see from the figures in the right column that RobustAgg(ϵ) and RobustAgg-TS(ϵ) incur far less regret on subpar arms. These results thereby demonstrate that the notion of subpar arms can capture the amenability of transfer learning in subpar arms but not near-optimal arms.

In addition, the results show that, empirically, our proposed algorithm RobustAgg-TS(ϵ) can robustly leverage transfer for arms in $\mathcal{I}_{5\epsilon} \supseteq \mathcal{I}_{10\epsilon}$—this suggests that our upper bounds may be improved; we leave this as future work.

8. Conclusion

In this work, we studied transfer learning in multi-task bandits under the framework of a generalized version of the ϵ-MPMAB problem (Wang et al., 2021). We proposed a TS-type algorithm, RobustAgg-TS(ϵ), which can robustly leverage auxiliary data collected for other tasks. We showed that RobustAgg-TS(ϵ) is empirically superior when evaluated on synthetic data, and also near-optimal in gap-dependent and gap-independent frequentist guarantees. In our analysis, we also proved a novel concentration inequality for multi-task data aggregation, which can be of independent interest in the analysis of other multi-task online learning problems. For future work, we are interested in improving the lower-order terms in our regret bounds and evaluating our algorithm in real-world applications.

9. Acknowledgements

We thank Geelon So for insightful discussions. ZW and KC thank the National Science Foundation under IIS 1915734 and CCF 1719133 for research support. CZ acknowledges startup funding support from the University of Arizona. We also thank the anonymous reviewers for their constructive feedback.

References

Agrawal, S. and Goyal, N. Analysis of thompson sampling for the multi-armed bandit problem. In *Conference on learning theory*, pp. 39–1. JMLR Workshop and Conference Proceedings, 2012.

Agrawal, S. and Goyal, N. Near-optimal regret bounds for thompson sampling. *J. ACM*, 64(5), sep 2017. ISSN 0004-5411. doi: 10.1145/3088510. URL https://doi.org/10.1145/3088510.

Auer, P., Cesa-Bianchi, N., and Fischer, P. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 2002.

Azar, M. G., Lazaric, A., and Brunskill, E. Sequential transfer in multi-armed bandit with finite set of models. In Burges, C. J. C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems 26*, pp. 2220–2228. Curran Associates, Inc., 2013.

Bastani, H., Simchi-Levi, D., and Zhu, R. Meta dynamic pricing: Transfer learning across experiments. *Management Science*, 2021.

Basu, S., Kveton, B., Zaheer, M., and Szepesvári, C. No regrets for learning the prior in bandits. *arXiv preprint arXiv:2107.06196*, 2021.

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Vaughan, J. W. A theory of learning from different domains. *Machine learning*, 79(1-2):151–175, 2010.

Cesa-Bianchi, N., Gentile, C., and Zappella, G. A gang of bandits. In *Advances in Neural Information Processing Systems*, pp. 737–745, 2013.

Chapelle, O. and Li, L. An empirical evaluation of thompson sampling. *Advances in neural information processing systems*, 24:2249–2257, 2011.

Gentile, C., Li, S., and Zappella, G. Online clustering of bandits. In *International Conference on Machine Learning*, pp. 757–765, 2014.

Gordon, R. D. Values of mills' ratio of area to bounding ordinate and of the normal probability integral for large values of the argument. *The Annals of Mathematical Statistics*, 12(3):364–366, 1941.

Hong, J., Kveton, B., Zaheer, M., and Ghavamzadeh, M. Hierarchical bayesian bandits. *arXiv preprint arXiv:2111.06929*, 2021.

Jin, T., Xu, P., Shi, J., Xiao, X., and Gu, Q. Mots: Minimax optimal thompson sampling. In *International Conference on Machine Learning*, pp. 5074–5083. PMLR, 2021.

Kaufmann, E., Korda, N., and Munos, R. Thompson sampling: An asymptotically optimal finite-time analysis. In *International conference on algorithmic learning theory*, pp. 199–213. Springer, 2012.

Kubota, A., Peterson, E. I., Rajendren, V., Kress-Gazit, H., and Riek, L. D. Jessie: Synthesizing social robot behaviors for personalized neurorehabilitation and beyond. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, pp. 121–130, 2020.

Kveton, B., Konobeev, M., Zaheer, M., Hsu, C.-W., Mladenov, M., Boutilier, C., and Szepesvari, C. Meta-thompson sampling. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5884–5893. PMLR, 18–24 Jul 2021.

Peleg, A., Pearl, N., and Meir, R. Metalearning linear bandits by prior update. *arXiv preprint arXiv:2107.05320*, 2021.

Qian, X., Feng, H., Zhao, G., and Mei, T. Personalized recommendation combining user interest and social circle. *IEEE transactions on knowledge and data engineering*, 26(7):1763–1777, 2013.

Rosenstein, M. T., Marx, Z., Kaelbling, L. P., and Dietterich, T. G. To transfer or not to transfer. In *NIPS 2005 workshop on transfer learning*, 2005.

Sharma, N., Basu, S., Shanmugam, K., and Shakkottai, S. Warm starting bandits with side information from confounded data. *arXiv preprint arXiv:2002.08405*, 2020.

Soare, M., Alsharif, O., Lazaric, A., and Pineau, J. Multitask linear bandits. *NIPS2014 Workshop on Transfer and Multi-task Learning : Theory meets Practice*, 2014.

Thompson, W. R. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.

Wan, R., Ge, L., and Song, R. Metadata-based multi-task bandits with bayesian hierarchical models. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, 2021.

Wang, Z., Zhang, C., Singh, M. K., Riek, L., and Chaudhuri, K. Multitask bandit learning through heterogeneous feedback aggregation. In *International Conference on Artificial Intelligence and Statistics*, pp. 1531–1539. PMLR, 2021.

Zhang, C. and Wang, Z. Provably efficient multi-task reinforcement learning with model transfer. *arXiv preprint arXiv:2107.08622*, 2021.

Zhang, C., Agarwal, A., Iii, H. D., Langford, J., and Negahban, S. Warm-starting contextual bandits: Robustly combining supervised and bandit feedback. In *International Conference on Machine Learning*, pp. 7335–7344, 2019.

Zhang, J. and Bareinboim, E. Transfer learning in multi-armed bandit: a causal approach. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*, pp. 1778–1780, 2017.

Outline. The structure of this appendix is as follows.

- In Section A, we introduce some basic definitions, facts and additional notations that are used in our analysis.
- In Section B, we formally present and prove the concentration bounds used in our proofs, including our novel concentration inequality for multi-task data aggregation at stopping times.
- In Section C, we prove Theorem 4.1 and Theorem 4.2.
- In Section D, we discuss the performance guarantees of the baseline algorithms in the ϵ-MPMAB problem, which include IND-UCB, IND-TS, and ROBUSTAGG(ϵ).
- Finally, we provide additional experimental results in Section E.

A. Basic Definitions and Facts

In this section, we revisit and introduce a few basic definitions, facts and additional notations that are useful in our proofs.

Definition A.1 (Constants used in the analysis)**.** In the analysis, we set

$$c_1 = 40, c_2 = 4$$

to be the constants used in Algorithm 1.[8]

Definition A.2 (Number of pulls)**.** Recall that

$$n_i^p(t) = \sum_{s \le t} \mathbb{1}\left\{p \in \mathcal{P}_s, i_s^p = i\right\}$$

is the number of pulls of arm i by player p after t rounds. We define

$$n_i(t) = \sum_{p \in [M]} n_i^p(t)$$

to be the total of number of pulls of arm i by all the players after t rounds.

Definition A.3 (Individual mean estimate)**.** For any $i \in [K]$, $p \in [M]$, and $t \in [T] \cup \{0\}$, let

$$\text{ind-}\hat{\mu}_i^p(t) = \frac{1}{n_i^p(t) \vee 1} \sum_{s \le t} \mathbb{1}\left\{p \in \mathcal{P}_s, i_s^p = i\right\} r_s^p$$

be the empirical mean computed for arm i using player p's own data from the first t rounds.

Definition A.4. Define

$$\text{ind-var}_i^p(t) = \frac{4}{n_i^p(t) \vee 1}.$$

Remark A.5 (mean and variance of the individual posteriors)**.** By the construction of Algorithm 1, we have that, in any round $t \in [T]$, for any active player $p \in \mathcal{P}_t$ and arm i, $\text{ind-}\hat{\mu}_i^p(t-1)$ and $\text{ind-var}_i^p(t-1)$ are the mean and variance of the individual posterior associated with arm i and player p in round t, respectively.

Definition A.6 (Aggregate mean estimate)**.** For any $i \in [K]$ and $t \in [T] \cup \{0\}$, let

$$\text{agg-}\hat{\mu}_i(t) = \frac{1}{n_i(t) \vee 1} \sum_{s \le t} \sum_{q : q \in \mathcal{P}_s} \mathbb{1}\left\{i_s^q = i\right\} r_s^q + \epsilon$$

be the empirical mean computed for arm i using all players' data from the first t rounds, offset by the dissimilarity parameter ϵ. Note that the definition of $\text{agg-}\hat{\mu}_i(t)$ does not depend on the identity of a specific player p.

[8] If we choose c_1 to some other positive number, we can still show guarantees similar to Theorems 4.1 and 4.2, except that $\mathcal{I}_{10\epsilon}$ needs to be changed to $\mathcal{I}_{\mathcal{O}\left(\sqrt{\frac{1}{c_1}}\epsilon\right)}$—the analysis of case $(A1)$ needs to be changed accordingly. On the other hand, it is also possible to change c_2 to any constant > 1 and establish similar regret guarantees, by tightening the exponents of the concentration inequalities (Corollaries B.4 and B.6) and Lemma C.36. We leave the details to interested readers.

Definition A.7 (Most recent pull)**.** In any round $t \in [T] \cup \{0\}$, for any player $p \in [M]$ and arm $i \in [K]$, we define

$$u_i^p(t) = \begin{cases} \max\left\{s \le t : p \in \mathcal{P}_s, i_s^p = i\right\}, & n_i^p(t) > 0 \\ 0, & n_i^p(t) = 0 \end{cases}$$

to be the round in which player p most recently pulled arm i (including round t); we let $u_i^p(t) = 0$ by convention if player p has not yet pulled arm i.

Definition A.8 (Aggregate mean estimate maintained by player p)**.** For any $t \in [T] \cup \{0\}$, $p \in [M]$, and $i \in [K]$, define

$$\text{agg-}\hat{\mu}_i^p(t) = \text{agg-}\hat{\mu}_i(u_i^p(t)).$$

Note that the superscript p differentiates this player-dependent aggregate mean estimate from $\text{agg-}\hat{\mu}_i(t)$ in Definition A.6, which does not depend on any individual player.

Definition A.9 (Aggregate number of pulls maintained by player p)**.** For any $t \in [T] \cup \{0\}$, $p \in [M]$, and $i \in [K]$, define

$$m_i^p(t) = n_i(u_i^p(t))$$

to be the total number of pulls of arm i by all the players until the round in which player p last pulled arm i.

Definition A.10. Define

$$\text{agg-var}_i^p(t) = \frac{4}{(m_i^p(t) - M) \vee 1}.$$

Remark A.11 (mean and variance of the aggregate posteriors)**.** By the construction of Algorithm 1, in any round $t \in [T]$, for any active player $p \in \mathcal{P}_t$ and arm i, we have that $\text{agg-}\hat{\mu}_i^p(t-1)$ and $\text{agg-var}_i^p(t-1)$ are the mean and variance of the aggregate posterior associated with arm i and player p in round t, respectively.

Definition A.12 (Filtration)**.** Let $\{\mathcal{F}_t\}_{t=0}^T$ be a filtration such that

$$\mathcal{F}_t = \sigma\left(\{i_s^q, r_s^q : s \le t, q \in \mathcal{P}_s\}\right)$$

is the σ-algebra generated by interactions of all players up until and including round t.

Definition A.13. Let

$$H_i^p(t) = \left\{n_i^p(t-1) \ge \frac{40 \ln T}{\epsilon^2} + 2M\right\}$$

be the event that in round t, for arm i, player p uses the *individual* posterior distribution; correspondingly, let

$$\overline{H_i^p(t)} = \left\{n_i^p(t-1) < \frac{40 \ln T}{\epsilon^2} + 2M\right\}$$

be the event that in round t, for arm i, player p uses the *aggregate* posterior distribution. See lines 7 to 11 in Algorithm 1.

Remark A.14. With the above notations,

$$\hat{\mu}_i^p(t-1) = \text{agg-}\hat{\mu}_i^p(t-1) \cdot \mathbb{1}(\overline{H_i^p(t)}) + \text{ind-}\hat{\mu}_i^p(t-1) \cdot \mathbb{1}(H_i^p(t)),$$

and

$$\text{var}_i^p(t-1) = \text{agg-var}_i^p(t-1) \cdot \mathbb{1}(\overline{H_i^p(t)}) + \text{ind-var}_i^p(t-1) \cdot \mathbb{1}(H_i^p(t)).$$

Stopping times. In our analysis, we will frequently use the following notions of stopping times:

Definition A.15. For any arm $i \in [K]$ and $k \in [TM]$, let

$$\tau_k(i) = \min\Big\{T + 1, \min\big\{t : n_i(t) \ge k\big\}\Big\}$$

be the round in which arm i is pulled the k-th time by any player. Furthermore, as a convention, let $\tau_0(i) = 0$.

Remark A.16. For any $i \in [K]$ and $k \in [TM]$, $\tau_k(i)$ is a stopping time with respect to $\{\mathcal{F}_t\}_{t=0}^T$. Indeed, for any $t \le T$,

$$\big\{\tau_k(i) \le t\big\} = \left\{\sum_{s \in [t]} \sum_{p : p \in \mathcal{P}_s} \mathbb{1}\left\{i_s^p = i\right\} \ge k\right\} \in \mathcal{F}_t.$$

Definition A.17. For any arm $i \in [K]$ and $k \in [TM]$, such that $\tau_k(i) \le T$, let $p_k(i)$ be the unique $p \in [M]$ such that $i^p_{\tau_k(i)} = i$ and

$$\sum_{s=1}^{\tau_k(i)-1} \sum_{q \in \mathcal{P}_s} \mathbb{1}\left\{i^q_s = i\right\} + \sum_{q \in \mathcal{P}_{\tau_k(i)} : q \le p} \mathbb{1}\left\{i^q_s = i\right\} = k.$$

In words, $p_k(i)$ is the player that makes the k-th pull of arm i, where arm pulls within a round are ordered by the indices of active players in that round.

Definition A.18. For any arm $i \in [K]$, player $p \in [M]$, and $k \in [T]$, let

$$\pi_k(i,p) = \min\Big\{T+1, \min\big\{t : n^p_i(t) \ge k\big\}\Big\}$$

be the round in which arm i is pulled the k-th time by player p. In addition, let $\pi_0(i,p) = 0$ by convention.

Remark A.19. For any $i \in [K]$ and $k \in [T]$, $\pi_k(i,p)$ is a stopping time with respect to $\{\mathcal{F}_t\}_{t=0}^T$. Indeed, for any $t \le T$,

$$\big\{\pi_k(i,p) \le t\big\} = \left\{ \sum_{s \in [t] : p \in \mathcal{P}_s} \mathbb{1}\left\{i^p_s = i\right\} \ge k \right\} \in \mathcal{F}_t.$$

The following property, namely, the invariant property, will also be useful for our analysis.

Definition A.20 (Invariant property)**.** We say that:

1. a set of random variables $\big\{g_t : t \in [T]\big\}$ satisfies the *invariant property with respect to arm* $i \in [K]$ *and player* $p \in [M]$, if g_t stays constant/invariant between two consecutive pulls of arm i by player p, i.e., for any $s \in [T]$ such that $\pi_s(i,p) \le T$, g_t is constant for all $t \in [\pi_{s-1}(i,p)+1, \pi_s(i,p)]$. In other words, for any $s \in [T]$ such that $\pi_s(i,p) \le T$,
$$g_{\pi_{s-1}(i,p)+1} = g_{\pi_{s-1}(i,p)+2} = \ldots = g_{\pi_s(i,p)}.$$
2. a set of random variables $\big\{f^p_t : t \in [T], p \in [M]\big\}$ satisfies the *invariant property with respect to arm* $i \in [K]$, if for every player $p \in [M]$, $\big\{f^p_t : t \in [T]\big\}$ satisfy the invariant property with respect to (i,p).

Example A.21. By the construction of Algorithm 1, in any round t, a player only updates the posteriors associated with an arm if the player pulls the arm in round t (line 20). It is easy to verify that for any arm $i \in [K]$ and $p \in [M]$, $\big\{H^p_i(t) : t \in [T]\big\}$ satisfies the invariant property with respect to (i,p). Specifically, for any $s \in [T]$ such that $\pi_s(i,p) \le T$,

$$H^p_i(\pi_{s-1}(i,p)+1) = H^p_i(\pi_{s-1}(i,p)+2) = \ldots = H^p_i(\pi_s(i,p)).$$

Consequently, $\big\{H^p_i(t) : t \in [T], p \in [M]\big\}$ satisfies the invariant property with respect to i.

Example A.22. For any arm $i \in [K]$ and any player $p \in [M]$, $\big\{n^p_i(t-1) : t \in [T]\big\}$ and $\big\{m^p_i(t-1) : t \in [T]\big\}$ both satisfy the invariant property with respect to (i,p) (see Definition A.2 and Definition A.9, respectively). Specifically, for any player p and any $s \in [T]$ such that $\pi_s(i,p) \le T$,

$$n^p_i(\pi_{s-1}(i,p)) = n^p_i(\pi_{s-1}(i,p)+1) = \ldots = n^p_i(\pi_s(i,p)-1) = s-1,$$
$$m^p_i(\pi_{s-1}(i,p)) = m^p_i(\pi_{s-1}(i,p)+1) = \ldots = m^p_i(\pi_s(i,p)-1) = n_i(\pi_{s-1}(i,p))$$

However, $\big\{n^p_i(t) : t \in [T]\big\}$ and $\big\{m^p_i(t) : t \in [T]\big\}$ do *not* necessarily satisfy the invariant property with respect to i. Similarly, $\big\{\text{ind-}\hat{\mu}^p_i(t-1) : t \in [T]\big\}$, $\big\{\text{ind-var}^p_i(t-1) : t \in [T]\big\}$, $\big\{\text{agg-}\hat{\mu}^p_i(t-1) : t \in [T]\big\}$, $\big\{\text{agg-var}^p_i(t-1) : t \in [T]\big\}$ all satisfy the invariant property with respect to (i,p).

Example A.23. For any arm $i \in [K]$ and any player $p \in [M]$, $\big\{\hat{\mu}^p_i(t-1) : t \in [T]\big\}$ satisfy the invariant property with respect to (i,p). This follows from Eq. (A.14), and the above two examples that $\big\{\text{ind-}\hat{\mu}^p_i(t-1) : t \in [T]\big\}$, $\big\{\text{agg-}\hat{\mu}^p_i(t-1) : t \in [T]\big\}$, $\big\{H^p_i(t) : t \in [T]\big\}$ all satisfy the invariant property with respect to (i,p).

Following a similar reasoning, $\big\{\text{var}^p_i(t-1) : t \in [T]\big\}$ satisfy the invariant property with respect to (i,p).

Facts about Subpar Arms. We now present some facts about subpar arms.

Fact A.24 (Properties of subpar arms, see also Wang et al. 2021, Fact 15)**.** The following are true:

1. for any $i \in [K]$ and $p, q \in [M]$, $\left|\Delta_i^p - \Delta_i^q\right| \le 2\epsilon$ (Wang et al., 2021, Fact 14);
2. For any $i \in \mathcal{I}_{10\epsilon}$ and $p \in [M]$, $\Delta_i^p > 8\epsilon$, which means that $\Delta_i^{\min} > 8\epsilon$.
3. $\left|\mathcal{I}_{2\epsilon}^C\right| \ge 1$;
4. Let $\Delta_i^{\max} = \max_{p\in[M]} \Delta_i^p$. For any $i \in \mathcal{I}_{10\epsilon} \subseteq \mathcal{I}_{5\epsilon}$, $\Delta_i^{\max} \le 2\Delta_i^{\min}$; furthermore, $\frac{1}{\Delta_i^{\min}} \le \frac{2}{M}\sum_{p\in[M]}\frac{1}{\Delta_i^p}$ (Wang et al., 2021, Fact 15).

Proof. For item 2, by the definition of $\mathcal{I}_{10\epsilon}$, there exists p such that $\Delta_i^p > 10\epsilon$. Then, for all $q \in [M]$, we have $\Delta_i^q > 8\epsilon$ by item 1.

For item 3, using a similar argument, we have, for each $i \in \mathcal{I}_{2\epsilon}$ and $p \in [M]$, $\Delta_i^p > 0$. Let j be an optimal for player 1 such that $\Delta_j^p = 0$. Then $j \notin \mathcal{I}_{2\epsilon}$. □

Additional notations.

- Denote by $\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz$ the cumulative distribution function (CDF) of the standard Gaussian distribution.
- Let $\overline{\Phi}(x) = 1 - \Phi(x) = \int_{x}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz$ denote the complementary CDF of the standard Gaussian distribution.
- Denote by $(z)_+ = z \vee 0$.
- For any arm $i \in [K]$, player $p \in [M]$ and $t \in [T] \cup \{0\}$, let
$$\overline{n_i^p}(t) := n_i^p(t) \vee 1,$$
and
$$\overline{m_i^p}(t) := (m_i^p(t) - M) \vee 1.$$

B. Concentration Bounds

B.1. Novel concentration inequality for multi-task data aggregation at random stopping time τ_k's

We begin by introducing the following definition.

Definition B.1 (Mixture expected reward at t)**.** For any arm $i \in [K]$ and $t \in [T]$, define

$$\tilde{\mu}_i(t) = \frac{1}{n_i(t) \vee 1} \sum_{s \le t} \sum_{q \in \mathcal{P}_s} \mathbb{1}\left\{i_s^q = i\right\} \mu_i^q + \epsilon$$

to be the ϵ-offset mixture expected reward of arm i up to round t.

In what follows, we will consider $\tilde{\mu}_i(\tau_k(i))$ for any $i \in [K]$ and $k \in [TM]$, where the definition of $\tau_k(i)$ can be found in Definition A.15.

Lemma B.2. *For any arm $i \in [K]$ and $k \in [TM]$, denote by $\tau_k = \tau_k(i)$. If $\tau_k \le T$, then for every player $p \in [M]$, we have*

$$\begin{aligned}&\text{agg-}\hat{\mu}_i(\tau_k) - \mu_i^p \le \text{agg-}\hat{\mu}_i(\tau_k) - \tilde{\mu}_i(\tau_k) + 2\epsilon; \textit{ and}\\ &\mu_i^p - \text{agg-}\hat{\mu}_i(\tau_k) \le \tilde{\mu}_i(\tau_k) - \text{agg-}\hat{\mu}_i(\tau_k).\end{aligned}$$

Proof. For every $t \in [T]$, observe that

$$\tilde{\mu}_i(t) = \frac{1}{n_i(t) \vee 1} \sum_{s \le t} \sum_{\substack{q \in \mathcal{P}_s: \\ i_s^q = i}} \mu_i^q + \epsilon = \sum_{q \in [M]} \frac{n_i^q(t) \cdot \mu_i^q}{n_i(t) \vee 1} + \epsilon.$$

It can be easily verified that, if $n_i(t) > 0$, for every player $p \in [M]$,

$$\tilde{\mu}_i(t) - \mu_i^p \le 2\epsilon \text{ and } \mu_i^p - \tilde{\mu}_i(t) \le 0,$$

where we note that the asymmetry comes from the additive term ϵ in $\tilde{\mu}_i(t)$. Therefore, for $k \in [TM]$, if $\tau_k \le T$, then $n_i(\tau_k) \ge k > 0$ and we have

$$\tilde{\mu}_i(\tau_k) - \mu_i^p \le 2\epsilon \text{ and } \mu_i^p - \tilde{\mu}_i(\tau_k) \le 0.$$

It then follows that, for every player $p \in [M]$,

$$\begin{aligned} &\text{agg-}\hat{\mu}_i(\tau_k) - \mu_i^p \le \text{agg-}\hat{\mu}_i(\tau_k) - \tilde{\mu}_i(\tau_k) + 2\epsilon, \text{ and} \\ &\mu_i^p - \text{agg-}\hat{\mu}_i(\tau_k) \le \tilde{\mu}_i(\tau_k) - \text{agg-}\hat{\mu}_i(\tau_k). \end{aligned}$$

□

We are now ready to present Lemma B.3, our novel concentration bound (see also Lemma 5.6).

Lemma B.3. *For any arm $i \in [K]$ and $k \in [TM] \cup \{0\}$, denote by $\tau_k = \tau_k(i)$; for $\delta \in (0, 1]$, we have*

$$\Pr\left(\{\tau_k = T+1\} \cup \left\{\{\tau_k \le T\} \cap \left\{\forall p \in [M],\ \text{agg-}\hat{\mu}_i(\tau_k) - \mu_i^p \le \sqrt{\frac{2\ln\left(\frac{2}{\delta}\right)}{\left(n_i(\tau_k) - M\right) \vee 1}} + 2\epsilon\right\}\right\}\right) > 1 - \delta; \quad (4)$$

$$\Pr\left(\{\tau_k = T+1\} \cup \left\{\{\tau_k \le T\} \cap \left\{\forall p \in [M],\ \mu_i^p - \text{agg-}\hat{\mu}_i(\tau_k) \le \sqrt{\frac{2\ln\left(\frac{2}{\delta}\right)}{\left(n_i(\tau_k) - M\right) \vee 1}}\right\}\right\}\right) > 1 - \delta. \quad (5)$$

The following corollary is an equivalent form of Equation (5):

Corollary B.4. *For any arm $i \in [K]$ and $k \in [TM] \cup \{0\}$, denote by $\tau_k = \tau_k(i)$. Equivalently, for any $z \ge 0$, we have*

$$\Pr\left((\tau_k \le T) \wedge \left(\exists p \in [M],\ \mu_i^p - \text{agg-}\hat{\mu}_i(\tau_k) \ge z\sqrt{\frac{4}{(n_i(\tau_k) - M) \vee 1}}\right)\right) \le 2e^{-2z^2}. \quad (6)$$

Proof of Corollary B.4. If $z \le \sqrt{\frac{1}{2}\ln 2}$, Equation (6) holds trivially as $2e^{-2z^2} \ge 1$. Otherwise $z > \sqrt{\frac{1}{2}\ln 2}$. In this case, let $\delta = 2e^{-2z^2} \in (0, 1]$ in Equation (5), and using De Morgan's law, we also obtain Equation (6). □

Proof of Lemma B.3. Fix any arm $i \in [K]$. For $k = 0$, we have $\tau_0 = 0$; both Eq. (4) and Eq. (5) hold trivially because for all $p \in [M]$ and $\delta \in (0, 1]$, $\left|\text{agg-}\hat{\mu}_i(\tau_0) - \mu_i^p\right| \le 1 \le \sqrt{2\ln 2} \le \sqrt{2\ln(\frac{2}{\delta})}$.

We now focus on $k \in [TM]$. By Lemma B.2, it suffices to show that

$$\Pr\left(\{\tau_k = T+1\} \cup \left\{\{\tau_k \le T\} \cap \left\{\text{agg-}\hat{\mu}_i(\tau_k) - \tilde{\mu}_i(\tau_k) \le \sqrt{\frac{2\ln\left(\frac{2}{\delta}\right)}{\left(n_i(\tau_k) - M\right) \vee 1}}\right\}\right\}\right) > 1 - \delta; \text{ and,} \quad (7)$$

$$\Pr\left(\{\tau_k = T+1\} \cup \left\{\{\tau_k \le T\} \cap \left\{\tilde{\mu}_i(\tau_k) - \text{agg-}\hat{\mu}_i(\tau_k) \le \sqrt{\frac{2\ln\left(\frac{2}{\delta}\right)}{\left(n_i(\tau_k) - M\right) \vee 1}}\right\}\right\}\right) > 1 - \delta.$$

To avoid redundancy, we only prove Eq. (7); the other inequality follows by symmetry.

Now, for $t \in [T] \cup \{0\}$, consider $Z_t = \sum_{s=1}^{t} \sum_{p \in \mathcal{P}_s} \mathbb{1}\{i_s^p = i\}(r_s^p - \mu_i^p)$. Furthermore, for $t \in [T] \cup \{0\}$ and $\lambda > 0$, let

$$w_t(\lambda) = \exp\left(\lambda Z_t - n_i(t)\frac{\lambda^2}{8}\right).$$

We now show that $\{w_t(\lambda)\}_{t=0}^{T}$ is a nonnegative supermartingale with respect to $\{\mathcal{F}_t\}_{t=0}^{T}$ for all $\lambda > 0$. Since $\mathbb{E}\left[|w_t(\lambda)|\right] < \infty$ and $w_t(\lambda) \geq 0$ for all $t \in [T] \cup \{0\}$, it suffices to show that, for all $t \in [T]$,

$$\begin{aligned}
&\mathbb{E}\left[w_t(\lambda) \mid \mathcal{F}_{t-1}\right] \\
=&\mathbb{E}\left[\exp\left(\sum_{s\in[t]}\sum_{p\in\mathcal{P}_s} \mathbb{1}\{i_s^p = i\}\left(\lambda(r_s^p - \mu_i^p) - \frac{\lambda^2}{8}\right)\right) \mid \mathcal{F}_{t-1}\right] \\
=&\mathbb{E}\left[\exp\left(\sum_{s\in[t-1]}\sum_{p\in\mathcal{P}_s} \mathbb{1}\{i_s^p = i\}\left(\lambda(r_s^p - \mu_i^p) - \frac{\lambda^2}{8}\right)\right)\exp\left(\sum_{p\in\mathcal{P}_t} \mathbb{1}\{i_t^p = i\}\left(\lambda(r_t^p - \mu_i^p) - \frac{\lambda^2}{8}\right)\right) \mid \mathcal{F}_{t-1}\right] \\
=&\exp\left(\sum_{s\in[t-1]}\sum_{p\in\mathcal{P}_s} \mathbb{1}\{i_s^p = i\}\left(\lambda(r_s^p - \mu_i^p) - \frac{\lambda^2}{8}\right)\right)\mathbb{E}\left[\exp\left(\sum_{p\in\mathcal{P}_t} \mathbb{1}\{i_t^p = i\}\left(\lambda(r_t^p - \mu_i^p) - \frac{\lambda^2}{8}\right)\right) \mid \mathcal{F}_{t-1}\right] \\
=&w_{t-1}(\lambda)\cdot\mathbb{E}\left[\exp\left(\lambda\sum_{p\in\mathcal{P}_t} \mathbb{1}\{i_t^p = i\}(r_t^p - \mu_i^p)\right)\exp\left(-\sum_{p\in\mathcal{P}_t} \mathbb{1}\{i_t^p = i\}\frac{\lambda^2}{8}\right) \mid \mathcal{F}_{t-1}\right] \\
\leq&w_{t-1}(\lambda),
\end{aligned}$$

where the last inequality uses the law of iterated expectation along with Hoeffding's lemma, i.e.,

$$\begin{aligned}
&\mathbb{E}\left[\exp\left(\lambda\sum_{p\in\mathcal{P}_t} \mathbb{1}\{i_t^p = i\}(r_t^p - \mu_i^p)\right)\cdot\exp\left(-\sum_{p\in\mathcal{P}_t} \mathbb{1}\{i_t^p = i\}\frac{\lambda^2}{8}\right) \mid \mathcal{F}_{t-1}\right] \\
\leq&\mathbb{E}\left[\mathbb{E}\left[\exp\left(\lambda\sum_{p\in\mathcal{P}_t} \mathbb{1}\{i_t^p = i\}(r_t^p - \mu_i^p)\right) \mid \mathcal{F}_{t-1}, (i_t^p)_{p\in\mathcal{P}_t}\right]\cdot\exp\left(-\sum_{p\in\mathcal{P}_t} \mathbb{1}\{i_t^p = i\}\frac{\lambda^2}{8}\right) \mid \mathcal{F}_{t-1}\right] \\
\leq&\mathbb{E}\left[\prod_{p\in\mathcal{P}_t}\exp\left(\frac{\lambda^2\cdot\left(\mathbb{1}\{i_t^p = i\}\right)^2}{8}\right)\cdot\exp\left(-\sum_{p\in\mathcal{P}_t} \mathbb{1}\{i_t^p = i\}\frac{\lambda^2}{8}\right) \mid \mathcal{F}_{t-1}\right] \leq 1
\end{aligned}$$

Recall from Remark A.16 that τ_k is a stopping time with respect to $\{\mathcal{F}_t\}_{t=0}^{T}$ and $\tau_k \leq T + 1 < \infty$ almost surely, it follows that, by the optional sampling theorem, for all $\lambda > 0$,

$$\mathbb{E}\left[\mathbb{1}\{\tau_k \leq T\}\cdot w_{\tau_k}(\lambda)\right] \leq \mathbb{E}\left[w_0(\lambda)\right] = 1. \tag{8}$$

Rewriting Eq. (8), we have

$$\mathbb{E}\left[\mathbb{1}\{\tau_k \leq T\}\cdot\exp\left(\lambda Z_{\tau_k} - n_i(\tau_k)\frac{\lambda^2}{8}\right)\right] \leq 1.$$

It then follows that, by Markov's inequality, for any $\delta > 0$,

$$\Pr\left(\mathbb{1}\left\{\tau_k \le T\right\} \cdot \exp\left(\lambda Z_{\tau_k} - n_i(\tau_k)\frac{\lambda^2}{8}\right) \ge \frac{1}{\delta}\right) \le \frac{\mathbb{E}\left[\mathbb{1}\left\{\tau_k \le T\right\} \cdot \exp\left(\lambda Z_{\tau_k} - n_i(\tau_k)\frac{\lambda^2}{8}\right)\right]}{\frac{1}{\delta}} \le \delta;$$

therefore,

$$\Pr\left(\left\{\tau_k \le T\right\} \cap \left\{\exp\left(\lambda Z_{\tau_k} - n_i(\tau_k)\frac{\lambda^2}{8}\right) \ge \frac{1}{\delta}\right\}\right) \le \delta.$$

Rearranging the terms in the above inequality, we have, for any $\lambda > 0$,

$$\Pr\left(\left\{\tau_k = T+1\right\} \cup \left\{\left\{\tau_k \le T\right\} \cap \left\{\frac{1}{n_i(\tau_k)} Z_{\tau_k} - \frac{\lambda}{8} < \frac{\ln\left(\frac{1}{\delta}\right)}{n_i(\tau_k)\cdot\lambda}\right\}\right\}\right) > 1-\delta,$$

where we use the elementary fact that for sets A and B, $\neg(A \cap B) = \neg A \cup (A \cap \neg B)$.

Choosing $\lambda = \sqrt{\frac{\ln(\frac{1}{\delta})}{k}}$ and using the fact that $n_i(\tau_k) \ge k$, we have

$$\Pr\left(\left\{\tau_k = T+1\right\} \cup \left\{\left\{\tau_k \le T\right\} \cap \left\{\frac{1}{n_i(\tau_k)} Z_{\tau_k} < \sqrt{\frac{2\ln(\frac{1}{\delta})}{k}}\right\}\right\}\right) > 1-\delta;$$

it then follows that

$$\Pr\left(\left\{\tau_k = T+1\right\} \cup \left\{\left\{\tau_k \le T\right\} \cap \left\{\frac{1}{n_i(\tau_k)} Z_{\tau_k} < \sqrt{\frac{2\ln(\frac{2}{\delta})}{k}}\right\}\right\}\right) > 1-\delta. \tag{9}$$

We now consider two cases:

1. $n_i(\tau_k) \le M$. We have $\frac{1}{n_i(\tau_k)} Z_{\tau_k} \le 1 < \sqrt{\frac{2\ln(\frac{2}{\delta})}{(n_i(\tau_k)-M)\vee 1}} = \sqrt{2\ln(\frac{2}{\delta})}$ trivially for $\delta \in (0,1]$.

2. $n_i(\tau_k) \ge M+1$. Since $k \ge n_i(\tau_k) - M$, we have $\sqrt{\frac{2\ln(\frac{2}{\delta})}{k}} \le \sqrt{\frac{2\ln(\frac{2}{\delta})}{n_i(\tau_k)-M}} = \sqrt{\frac{2\ln(\frac{2}{\delta})}{(n_i(\tau_k)-M)\vee 1}}$.

Eq. (7) then follows from Eq. (9) and the elementary fact that $A \subseteq B$ if $(A \cap C) \subseteq B$ and $(A \cap \neg C) \subseteq B$. This completes the proof. □

B.2. Other concentration bounds

Recall the definition of stopping times $\pi_k(i,p)$ for any arm i and player p (see Definition A.18).

Lemma B.5. *For any $i \in [K]$, $p \in [M]$, $k \in [T] \cup \{0\}$, and $\delta \in (0,1]$, we have*

$$\Pr\left(\left\{\pi_k(i,p) = T+1\right\} \cup \left\{\left\{\pi_k(i,p) \le T\right\} \cap \left\{\left|\text{ind-}\hat{\mu}_i^p(\pi_k(i,p)) - \mu_i^p\right| \le \sqrt{\frac{2\ln\left(\frac{4}{\delta}\right)}{n_i^p(\pi_k(i,p)) \vee 1}}\right\}\right\}\right) > 1-\delta. \tag{10}$$

Corollary B.6. *For any $i \in [K]$, $p \in [M]$, $k \in [T] \cup \{0\}$, and $z \geq 0$, we have*

$$\Pr\left(\left(\pi_k(i,p) \leq T\right) \wedge \left(\left| \mu_i^p - \text{ind-}\hat{\mu}_i^p(\pi_k(i,p)) \right| \geq z\sqrt{\frac{4}{n_i^p(\pi_k(i,p)) \vee 1}} \right) \right) \leq 4e^{-2z^2}. \tag{11}$$

Proof of Corollary B.6. If $z \leq \sqrt{\frac{1}{2}\ln 4}$, Equation (11) holds trivially as $4e^{-2z^2} \geq 1$. Otherwise $z > \sqrt{\frac{1}{2}\ln 4}$. In this case, let $\delta = 4e^{-2z^2} \in (0,1]$ in Equation (10), and using De Morgan's law, we also obtain Equation (11). □

Proof of Lemma B.5. The proof of Lemma B.5 is largely similar to the one for Lemma B.3. Therefore, we omit some details here to avoid redundancy. See the proof of Lemma B.3 for full details.

Let us fix any arm $i \in [K]$ and player $p \in [M]$. Throughout this proof, to ease the exposition, we use π_k to denote $\pi_k(i,p)$.

We first observe that when $k = 0$, we have $\pi_k = 0$, ind-$\hat{\mu}_i^p(0) = 0$, and $n_i^p(0) = 0$. It follows that $\left|\text{ind-}\hat{\mu}_i^p(\pi_k) - \mu_i^p\right| \leq 1 \leq \sqrt{2\ln\left(\frac{4}{\delta}\right)}$ trivially.

It then suffices to only consider the case when $k \in [T]$. Note that $n_i^p(\pi_k) = k \geq 1$. We will show that

$$\Pr\left(\{\pi_k = T+1\} \cup \left\{ \{\pi_k \leq T\} \cap \left\{ \text{ind-}\hat{\mu}_i^p(\pi_k) - \mu_i^p \leq \sqrt{\frac{2\ln\left(\frac{2}{\delta}\right)}{n_i^p(\pi_k)}} \right\} \right\} \right) > 1 - \delta. \tag{12}$$

For $t \in [T] \cup \{0\}$, let $X_t = \sum_{s \in [t]} \mathbb{1}\{p \in \mathcal{P}_s, i_s^p = i\}(r_s^p - \mu_i^p)$; and for $\lambda > 0$, further define $\xi_t(\lambda) = \exp\left(\lambda X_t - n_i^p(t)\frac{\lambda^2}{8}\right)$. It can be verified that $\left\{\xi_t(\lambda)\right\}_{t=0}^T$ is a nonnegative supermartingale with respect to $\{\mathcal{F}_t\}_{t=0}^T$ for all $\lambda > 0$:

1. $\mathbb{E}\left[\left|\xi_t(\lambda)\right|\right] < \infty$ for all $t \in [T] \cup \{0\}$;
2. $\xi_t(\lambda) \geq 0$ for all $t \in [T] \cup \{0\}$;
3. $\mathbb{E}\left[\xi_t(\lambda) \mid \mathcal{F}_{t-1}\right] \leq \xi_{t-1}(\lambda)$ for all $t \in [T]$.

Item 3 is true because

$$\begin{aligned}
&\mathbb{E}\left[\xi_t(\lambda) \mid \mathcal{F}_{t-1}\right] \\
=\ &\exp\left(\sum_{s=1}^{t-1} \mathbb{1}\{p \in \mathcal{P}_s, i_s^p = i\}\left(\lambda(r_s^p - \mu_i^p) - \frac{\lambda^2}{8}\right)\right) \mathbb{E}\left[\exp\left(\mathbb{1}\{p \in \mathcal{P}_t, i_t^p = i\}\left(\lambda(r_t^p - \mu_i^p) - \frac{\lambda^2}{8}\right)\right) \mid \mathcal{F}_{t-1}\right] \\
=\ &\xi_{t-1}(\lambda) \cdot \mathbb{E}\left[\exp\left(\lambda \cdot \mathbb{1}\{p \in \mathcal{P}_t, i_t^p = i\}(r_t^p - \mu_i^p)\right) \exp\left(-\mathbb{1}\{p \in \mathcal{P}_t, i_t^p = i\}\frac{\lambda^2}{8}\right) \mid \mathcal{F}_{t-1}\right] \\
=&\xi_{t-1}(\lambda) \cdot \mathbb{E}\left[\mathbb{E}\left[\exp\left(\lambda \cdot \mathbb{1}\{p \in \mathcal{P}_t, i_t^p = i\}(r_t^p - \mu_i^p)\right) \mid \mathcal{F}_{t-1}, i_t^p\right] \cdot \exp\left(-\mathbb{1}\{p \in \mathcal{P}_t, i_t^p = i\}\frac{\lambda^2}{8}\right) \mid \mathcal{F}_{t-1}\right] \\
\leq\ &\xi_{t-1}(\lambda),
\end{aligned}$$

where we use the law of total expectation, the observation that $\xi_{t-1}(\lambda)$ is $\mathcal{F}_{t-1}$-measurable, and Hoeffding's Lemma.

Recall from Remark A.19 that π_k is a stopping time with respect to $\{\mathcal{F}_t\}_{t=0}^T$ and $\pi_k \le T+1 < \infty$ almost surely. Then, by the optional sampling theorem, for all $\lambda > 0$,

$$\mathbb{E}\left[\mathbb{1}\left\{\pi_k \le T\right\} \cdot \xi_{\pi_k}(\lambda)\right] \le \mathbb{E}\left[\xi_0(\lambda)\right] = 1. \tag{13}$$

In other words,

$$\mathbb{E}\left[\mathbb{1}\left\{\pi_k \le T\right\} \cdot \exp\left(\lambda X_{\pi_k} - n_i^p(\pi_k)\frac{\lambda^2}{8}\right)\right] \le 1.$$

By Markov's inequality, we have

$$\Pr\left(\mathbb{1}\left\{\pi_k \le T\right\} \cdot \exp\left(\lambda X_{\pi_k} - n_i^p(\pi_k)\frac{\lambda^2}{8}\right) \ge \frac{1}{\delta}\right) \le \delta;$$

and thus,

$$\Pr\left(\left\{\pi_k \le T\right\} \cap \left\{\exp\left(\lambda X_{\pi_k} - n_i^p(\pi_k)\frac{\lambda^2}{8}\right) \ge \frac{1}{\delta}\right\}\right) \le \delta.$$

Using the elementary fact that for sets A and B, $\neg(A \cap B) = \neg A \cup (A \cap \neg B)$, we have, for any $\lambda > 0$,

$$\Pr\left(\left\{\pi_k = T+1\right\} \cup \left\{\left\{\pi_k \le T\right\} \cap \left\{\frac{1}{n_i^p(\pi_k)}X_{\pi_k} - \frac{\lambda}{8} < \frac{\ln\left(\frac{1}{\delta}\right)}{n_i^p(\pi_k)\cdot\lambda}\right\}\right\}\right) > 1-\delta,$$

where we slightly rearrange the terms.

Choose $\lambda = \sqrt{\frac{\ln(\frac{1}{\delta})}{k}}$ and observe that $n_i^p(\pi_k) = k$. It follows that

$$\Pr\left(\left\{\pi_k = T+1\right\} \cup \left\{\left\{\pi_k \le T\right\} \cap \left\{\frac{1}{n_i^p(\pi_k)}X_{\pi_k} < \sqrt{\frac{2\ln(\frac{1}{\delta})}{n_i^p(\pi_k)}}\right\}\right\}\right) > 1-\delta.$$

Eq. (12) follows trivially by the observation that $\ln(\frac{2}{\delta}) > \ln(\frac{1}{\delta})$. By symmetry, it can also be shown that the following inequality is true:

$$\Pr\left(\left\{\pi_k = T+1\right\} \cup \left\{\left\{\pi_k \le T\right\} \cap \left\{\mu_i^p - \text{ind-}\hat{\mu}_i^p(\pi_k) \le \sqrt{\frac{2\ln\left(\frac{2}{\delta}\right)}{n_i^p(\pi_k)}}\right\}\right\}\right) > 1-\delta.$$

The proof is then completed by applying the union bound. □

Definition B.7. For any $\delta \in (0,1]$, let

$$E_{\text{agg}}(\delta) = \left\{\forall i \in [K], \forall k \in [TM] \cup \{0\}, \left(\tau_k(i) = T+1\right) \vee \left(\left(\tau_k(i) \le T\right) \wedge \right.\right.$$

$$\left.\left.\left(\forall p \in [M],\ \text{agg-}\hat{\mu}_i(\tau_k(i)) - \mu_i^p \le \sqrt{\frac{2\ln\left(\frac{2}{\delta}\right)}{\left(n_i(\tau_k(i)) - M\right)\vee 1}} + 2\epsilon, \mu_i^p - \text{agg-}\hat{\mu}_i(\tau_k(i)) \le \sqrt{\frac{2\ln\left(\frac{2}{\delta}\right)}{\left(n_i(\tau_k(i)) - M\right)\vee 1}}\right)\right)\right\},$$

and

$$E_{\text{ind}}(\delta) = \Bigg\{ \forall i \in [K], \forall p \in [M], \forall k \in [T] \cup \{0\}, \big(\pi_k(i,p) = T+1\big) \vee \left(\big(\pi_k(i,p) \le T\big) \wedge \left(\left| \text{ind-}\hat{\mu}_i^p(\pi_k(i,p)) - \mu_i^p \right| \le \sqrt{\frac{2\ln(\frac{4}{\delta})}{n_i^p(\pi_k(i,p)) \vee 1}} \right) \right) \Bigg\}.$$

Furthermore, let

$$E(\delta) = E_{\text{agg}}(\delta) \cap E_{\text{ind}}(\delta).$$

Corollary B.8. *For $\delta \in (0, 1]$,*

$$\Pr(E(\delta)) \ge 1 - 6T^3\delta.$$

Proof. By the union bound, Lemma B.3, Lemma B.5, and the assumption that $T \ge \max(K, M)$, we have

$$\Pr(E_{\text{agg}}(\delta)) \ge 1 - K(TM+1)(2\delta) \ge 1 - 4T^3\delta.$$
$$\Pr(E_{\text{ind}}(\delta)) \ge 1 - KM(T+1)\delta \ge 1 - 2T^3\delta.$$

The corollary then follows by the union bound. □

B.3. Clean Event

We now define our notion of "clean" event for each t.

Definition B.9. For any $t \in [T+1]$, let

$$\mathcal{E}_t = \Bigg\{ \forall p \in [M], \forall i \in [K], \ \left| \text{ind-}\hat{\mu}_i^p(t-1) - \mu_i^p \right| \le \sqrt{\frac{10\ln T}{\overline{n_i^p}(t-1)}}, \quad \text{agg-}\hat{\mu}_i^p(t-1) - \mu_i^p \le \sqrt{\frac{10\ln T}{\overline{m_i^p}(t-1)}} + 2\epsilon, \quad \mu_i^p - \text{agg-}\hat{\mu}_i^p(t-1) \le \sqrt{\frac{10\ln T}{\overline{m_i^p}(t-1)}} \Bigg\},$$

where we recall that $\overline{n_i^p}(t-1) = n_i^p(t-1) \vee 1$, $\overline{m_i^p}(t-1) = (m_i^p(t-1) - M) \vee 1$. Furthermore, let $\overline{\mathcal{E}_t}$ denote the complement of $\mathcal{E}_t$.

The following lemma shows that the clean event happens with high probability.

Lemma B.10.

$$\Pr(\mathcal{E}_t) > 1 - \frac{24}{T^2}.$$

Proof. The proof of Lemma B.10 follows from Corollary B.8. It suffices to show that, for any t, $E(\frac{4}{T^5}) \subseteq \mathcal{E}_t$. To this end, we will show that if $E(\frac{4}{T^5})$ happens, then $\mathcal{E}_t$ must happen.

For any $t \in [T+1]$, $i \in [K]$, $p \in [M]$, let $u = u_i^p(t-1)$ be the round in which player p last pulls arm i (see Definition A.7). In addition, let $s = n_i^p(u) \in ([T] \cup \{0\})$ and $k = n_i(u) \in ([TM] \cup \{0\})$. Note that $\pi_s(i,p) = u \le T$ and $\tau_k(i) = u \le T$.

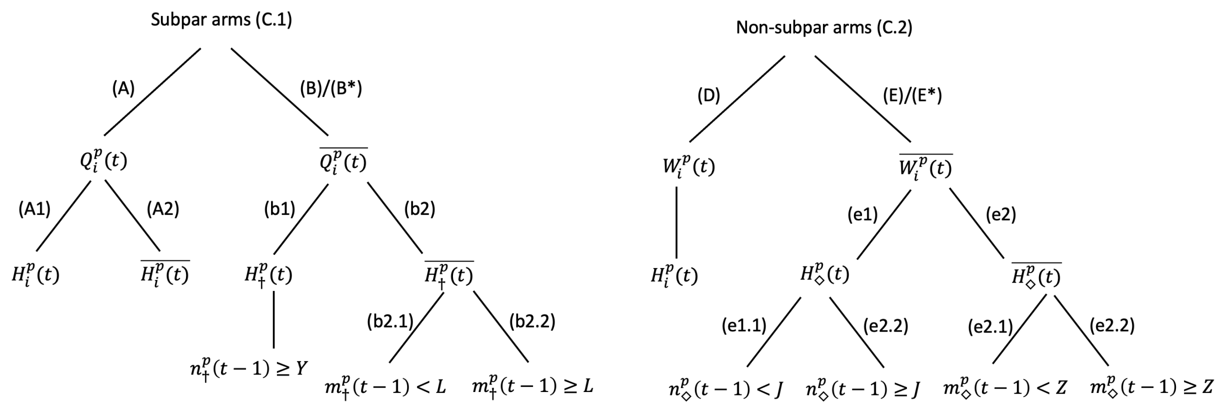
It then follows by definition that,

$$\begin{aligned}\text{ind-}\hat{\mu}_i^p(t-1) = \text{ind-}\hat{\mu}_i^p(\pi_s(i,p)), \quad n_i^p(t-1) = n_i^p(\pi_s(i,p));\\ \text{agg-}\hat{\mu}_i^p(t-1) = \text{agg-}\hat{\mu}_i(\tau_k(i)), \quad m_i^p(t-1) = n_i(\tau_k(p)).\end{aligned}$$

The proof is then completed straightforwardly by the definition of $E(\frac{4}{T^5})$, which indicates that for all $s \in [T] \cup \{0\}$ and $k \in [TM] \cup \{0\}$,

$$\begin{aligned}\left|\text{ind-}\hat{\mu}_i^p(\pi_s(i,p)) - \mu_i^p\right| &\leq \sqrt{\frac{10\ln T}{n_i^p(\pi_s(i,p)) \vee 1}},\\ \text{agg-}\hat{\mu}_i(\tau_k(i)) - \mu_i^p &\leq \sqrt{\frac{10\ln T}{\big(n_i(\tau_k(p)) - M\big) \vee 1}} + 2\epsilon, \text{ and}\\ \mu_i^p - \text{agg-}\hat{\mu}_i(\tau_k(i)) &\leq \sqrt{\frac{10\ln T}{\big(n_i(\tau_k(p)) - M\big) \vee 1}}.\end{aligned}$$

□



(a) Subpar arms (Section C.1)

(b) Non-subpar arms (Section C.2)

Figure 2. Illustrations of the case division rules used in the proofs of Theorem 4.1 and Theorem 4.2, respectively. Formal definitions of the notions used in the figure can be found in Section A, Section C.1 and Section C.2.

C. Proofs of Theorem 4.1 and Theorem 4.2

The following lemmas are central to our proofs of Theorem 4.1 and Theorem 4.2. In Section C.1, we prove Lemma C.1. In Section C.2, we prove Lemma C.2. We then conclude our proofs in Section C.3.

Lemma C.1 (Subpar arms)**.** *For any arm $i \in \mathcal{I}_{10\epsilon}$,*

$$\mathbb{E}\left[n_i(T)\right] \leq \mathcal{O}\left(\frac{\ln T}{(\Delta_i^{\min})^2} + M\right),$$

where we recall that $\Delta_i^{\min} = \min_{p \in [M]} \Delta_i^p$.

Lemma C.2 (Non-subpar arms)**.** *For any arm $i \in \mathcal{I}_{10\epsilon}^C$ and player $p \in [M]$,*

$$\mathbb{E}\left[n_i^p(T)\right] \leq \mathcal{O}\left(\frac{\ln T}{(\Delta_i^p)^2} + M\right).$$

Our analysis in the following Section C.1 and Section C.2 involve various proofs by cases. Figure 2 provides an overview of the case division rules used in our analysis.

C.1. Subpar Arms

In this section, we prove Lemma C.1.

Fix any subpar arm $i \in \mathcal{I}_{10\epsilon}$ and an arm $\dagger \in \mathcal{I}_{2\epsilon}^C$. See Fact A.24 for the existence of such an arm. We first consider the following definitions.

Definition C.3. For any arm $i \in \mathcal{I}_{10\epsilon}$ and any player p, let

$$\delta_i^p = \mu_\dagger^p - \mu_i^p > 0.$$

Fact C.4. For any $i \in \mathcal{I}_{10\epsilon}$ and player $p \in [M]$,

$$\frac{3}{4}\Delta_i^p < \delta_i^p \leq \Delta_i^p.$$

Proof. For any player $p \in [M]$, since $\dagger \in \mathcal{I}_{2\epsilon}^C$, we have $\Delta_\dagger^p = \mu_*^p - \mu_\dagger^p \leq 2\epsilon$ by the definition of $\mathcal{I}_{2\epsilon}^C$. Furthermore, for any $i \in \mathcal{I}_{10\epsilon}$, $\Delta_i^p = \mu_*^p - \mu_i^p > 8\epsilon$. Therefore, we have

1. $\delta_i^p = \mu_\dagger^p - \mu_i^p \leq \mu_*^p - \mu_i^p = \Delta_i^p$;

2. Note that $\frac{\mu_*^p - \mu_\dagger^p}{\mu_*^p - \mu_i^p} \leq \frac{2\epsilon}{8\epsilon} \leq \frac{1}{4}$. This implies that $\frac{\delta_i^p}{\Delta_i^p} = 1 - \frac{\mu_*^p - \mu_\dagger^p}{\mu_*^p - \mu_i^p} \geq \frac{3}{4}$. □

Definition C.5. For any player p, let $y_i^p = \mu_i^p + \frac{1}{2}\delta_i^p$ be a threshold; in any round t, further define

$$Q_i^p(t) = \left\{\theta_i^p(t) > y_i^p\right\}$$

to be the event that the sample $\theta_i^p(t)$ from the posterior distribution associated with arm i and player p in round t is greater than the threshold y_i^p. In addition, let $\overline{Q_i^p(t)} = \left\{\theta_i^p(t) \leq y_i^p\right\}$.

C.1.1. Subpar Arms—Decomposition

We can then decompose $\mathbb{E}\left[n_i(T)\right]$ as follows.

$$\begin{aligned}
\mathbb{E}\left[n_i(T)\right] =& \mathbb{E}\left[\sum_{t\in[T]}\sum_{p\in\mathcal{P}_t} \mathbb{1}\left\{i_t^p = i\right\}\right] \\
\leq& \mathbb{E}\left[\sum_{t\in[T]}\sum_{p\in\mathcal{P}_t} \mathbb{1}\left\{i_t^p = i, Q_i^p(t), \mathcal{E}_t\right\}\right] + \mathbb{E}\left[\sum_{t\in[T]}\sum_{p\in\mathcal{P}_t} \mathbb{1}\left\{i_t^p = i, \overline{Q_i^p(t)}, \mathcal{E}_t\right\}\right] + \mathbb{E}\left[\sum_{t\in[T]}\sum_{p\in\mathcal{P}_t} \mathbb{1}\left\{\overline{\mathcal{E}_t}\right\}\right] \\
\leq& \underbrace{\mathbb{E}\left[\sum_{t\in[T]}\sum_{p\in\mathcal{P}_t} \mathbb{1}\left\{i_t^p = i, Q_i^p(t), \mathcal{E}_t\right\}\right]}_{(A)} + \underbrace{\mathbb{E}\left[\sum_{t\in[T]}\sum_{p\in\mathcal{P}_t} \mathbb{1}\left\{i_t^p = i, \overline{Q_i^p(t)}, \mathcal{E}_t\right\}\right]}_{(B)} + \mathcal{O}\left(1\right),
\end{aligned} \quad (14)$$

where the second inequality follows from Lemma B.10. In the following two subsections, we bound term (A) and (B), respectively.

C.1.2. Bounding Term (A)

The following lemma provides an upper bound on term (A).

Lemma C.6.

$$(A) \leq \mathcal{O}\left(\frac{\ln T}{(\Delta_i^{\min})^2} + M\right), \quad (15)$$

where we recall that $\Delta_i^{\min} = \min_{p\in[M]} \Delta_i^p$.

Proof of Lemma C.6. Recall the definition of $\mathcal{E}_t$ in Definition B.9 and the definition of $H_i^p(t)$ in Definition A.13, we have

$$(A) = \underbrace{\sum_{t\in[T]}\sum_{p\in\mathcal{P}_t} \mathbb{E}\left[\mathbb{1}\left\{i_t^p = i, Q_i^p(t), \mathcal{E}_t, H_i^p(t)\right\}\right]}_{(A1)} + \underbrace{\sum_{t\in[T]}\sum_{p\in\mathcal{P}_t} \mathbb{E}\left[\mathbb{1}\left\{i_t^p = i, Q_i^p(t), \mathcal{E}_t, \overline{H_i^p(t)}\right\}\right]}_{(A2)}.$$

We first consider term $(A1)$. Recall that, for simplicity, we let $\overline{n_i^p}(t-1)$ denote $n_i^p(t-1) \vee 1$; also recall that $\overline{\Phi}(\cdot)$ is the complementary CDF of the standard Gaussian distribution, and $(z)_+ = z \vee 0$. We have

$$
\begin{aligned}
(A1) &\leq \sum_{t\in[T]} \sum_{p\in\mathcal{P}_t} \mathbb{E}\left[\mathbb{1}\left\{Q_i^p(t), \mathcal{E}_t, H_i^p(t)\right\}\right] \\
&= \sum_{t\in[T]} \sum_{p\in\mathcal{P}_t} \mathbb{E}\left[\mathbb{E}\left[\mathbb{1}\left\{Q_i^p(t), \mathcal{E}_t, H_i^p(t)\right\} \mid \mathcal{F}_{t-1}\right]\right] \\
&= \sum_{t\in[T]} \sum_{p\in\mathcal{P}_t} \mathbb{E}\left[\mathbb{1}\left\{\mathcal{E}_t, H_i^p(t)\right\} \cdot \mathbb{E}\left[\mathbb{1}\left\{\theta_i^p(t) > y_i^p\right\} \mid \mathcal{F}_{t-1}\right]\right] \\
&= \sum_{t\in[T]} \sum_{p\in\mathcal{P}_t} \mathbb{E}\left[\mathbb{1}\left\{\mathcal{E}_t, H_i^p(t)\right\} \cdot \overline{\Phi}\left(\sqrt{\overline{n_i^p}(t-1)/4}\left(y_i^p - \text{ind-}\hat{\mu}_i^p(t-1)\right)\right)\right] \\
&\leq \sum_{t\in[T]} \sum_{p\in\mathcal{P}_t} \mathbb{E}\left[\mathbb{1}\left\{\mathcal{E}_t, H_i^p(t)\right\} \cdot \exp\left(-\frac{n_i^p(t-1)(y_i^p - \text{ind-}\hat{\mu}_i^p(t-1))_+^2}{8}\right)\right] \\
&\leq \sum_{t\in[T]} \sum_{p\in\mathcal{P}_t} \mathbb{E}\left[\mathbb{1}\left\{\mathcal{E}_t, H_i^p(t)\right\} \cdot \exp\left(-\frac{n_i^p(t-1)(\mu_i^p + \frac{3}{8}\Delta_i^p - \mu_i^p - \frac{1}{16}\Delta_i^p)_+^2}{8}\right)\right] \\
&\leq \sum_{t\in[T]} \sum_{p\in\mathcal{P}_t} \mathbb{E}\left[\mathbb{1}\left\{\mathcal{E}_t, H_i^p(t)\right\} \cdot \exp\left(-\frac{n_i^p(t-1)(\Delta_i^p)^2}{8(16)}\right)\right] \\
&\leq \sum_{t\in[T]} \sum_{p\in\mathcal{P}_t} \frac{1}{T^2} = \mathcal{O}\left(1\right).
\end{aligned}
$$

where the first inequality drops the indicator $\mathbb{1}\left\{i_t^p = i\right\}$; the first equality uses the law of total expectation; the second equality follows from the observation that $\mathcal{E}_t$ and $H_i^p(t)$ are $\mathcal{F}_{t-1}$-measurable; the third equality follows from the observation that when $H_i^p(t)$ happens, $\mathbb{E}\left[\mathbb{1}\left\{\theta_i^p(t) > y_i^p\right\} \mid \mathcal{F}_{t-1}\right] = \mathbb{P}\left(\theta_i^p(t) > y_i^p \mid \mathcal{F}_{t-1}\right) = \overline{\Phi}\left(\frac{y_i^p - \text{ind-}\hat{\mu}_i^p(t-1)}{\sqrt{4/\overline{n_i^p}(t-1)}}\right)$; the second inequality is from Lemma C.35 and that $\overline{n_i^p}(t-1) \geq n_i^p(t-1)$; the third inequality follows from the facts that when $\mathcal{E}_t$ and $H_i^p(t)$ happen,

1. $\overline{n_i^p}(t-1) \geq n_i^p(t-1) \geq \frac{40 \ln T}{\epsilon^2} \geq \frac{2560 \ln T}{(\Delta_i^p)^2}$ (see Fact A.24),

2. $\text{ind-}\hat{\mu}_i^p(t-1) \leq \mu_i^p + \sqrt{\frac{10 \ln T}{\overline{n_i^p}(t-1)}} \leq \mu_i^p + \frac{1}{16}\Delta_i^p$ (see Definition B.9), and

3. $y_i^p = \mu_i^p + \frac{1}{2}\delta_i^p > \mu_i^p + \frac{3}{8}\Delta_i^p$ (see Fact C.4);

the fourth inequality is by algebra; and the fifth inequality again uses the observation that when $H_i^p(t)$ happens, $n_i^p(t-1) \geq \frac{2560 \ln T}{(\Delta_i^p)^2}$.

We now turn our attention to term $(A2)$. With foresight, let $l = \frac{10240 \ln T}{\left(\Delta_i^{\min}\right)^2} + M$. We have

$$\begin{aligned}
(A2) &= \sum_{t\in[T]} \sum_{p\in\mathcal{P}_t} \mathbb{E}\left[\mathbb{1}\left\{i_t^p = i, Q_i^p(t), \mathcal{E}_t, \overline{H_i^p(t)}\right\}\right] \\
&\leq \sum_{t\in[T]} \sum_{p\in\mathcal{P}_t} \mathbb{E}\left[\mathbb{1}\left\{i_t^p = i, Q_i^p(t), \mathcal{E}_t, \overline{H_i^p(t)}, m_i^p(t-1) < l\right\}\right] \\
&\qquad\qquad + \sum_{t\in[T]} \sum_{p\in\mathcal{P}_t} \mathbb{E}\left[\mathbb{1}\left\{i_t^p = i, Q_i^p(t), \mathcal{E}_t, \overline{H_i^p(t)}, m_i^p(t-1) \geq l\right\}\right] \\
&\leq (l+M) + \sum_{t\in[T]} \sum_{p\in\mathcal{P}_t} \mathbb{E}\left[\mathbb{1}\left\{i_t^p = i, Q_i^p(t), \mathcal{E}_t, \overline{H_i^p(t)}, m_i^p(t-1) \geq l\right\}\right]. \qquad (16)
\end{aligned}$$

To see why Eq. (16) is true, it suffices to show that, with probability 1,

$$\sum_{t\in[T]} \sum_{p\in\mathcal{P}_t} \mathbb{1}\left\{i_t^p = i, m_i^p(t-1) < l\right\} \leq l + M.$$

Indeed, let us define $\iota = \min\left\{t : n_i(t) = \sum_{s\in[t]} \sum_{p\in\mathcal{P}_s} \mathbb{1}\left\{i_s^p = i\right\} \geq l\right\}$. The above summation can be simplified as

$$\begin{aligned}
&\sum_{t=1}^{T} \sum_{p\in\mathcal{P}_t} \mathbb{1}\left\{i_t^p = i, m_i^p(t-1) < l\right\} \\
=&\sum_{t=1}^{\iota-1} \sum_{p\in\mathcal{P}_t} \mathbb{1}\left\{i_t^p = i, m_i^p(t-1) < l\right\} + \sum_{t=\iota}^{T} \sum_{p\in\mathcal{P}_t} \mathbb{1}\left\{i_t^p = i, m_i^p(t-1) < l\right\} \\
\leq&\sum_{t=1}^{\iota-1} \sum_{p\in\mathcal{P}_t} \mathbb{1}\left\{i_t^p = i\right\} + \sum_{p\in[M]} \sum_{t\geq\iota : p\in\mathcal{P}_t} \mathbb{1}\left\{i_t^p = i, m_i^p(t-1) < l\right\} \\
\leq&(l-1) + M,
\end{aligned}$$

where the $\sum_{p\in[M]} \sum_{t\geq\iota : p\in\mathcal{P}_t} \mathbb{1}\left\{i_t^p = i, m_i^p(t-1) < l\right\} \leq M$ follows from the observation that, once the total number of pulls of arm i by all players has reached l, any player p cannot pull arm i more than once before the aggregate number of pulls of i maintained by p is updated to a value $\geq l$ (see Definition A.9).

Remark C.7. Eq. (16) can also be deducted from the more general Lemma C.38 in Section C.4, by taking $f_t^p = 1$ for all t, p.

Now, recall that we denote $\left(m_i^p(t-1) - M\right) \vee 1$ by $\overline{m_i^p}(t-1)$. And again, recall that $\overline{\Phi}(\cdot)$ is the complementary CDF of the standard Gaussian distribution, and $(z)_+ = z \vee 0$. It follows from Eq. (16) that

$$
\begin{aligned}
(A2) &\le (l+M) + \sum_{t\in[T]}\sum_{p\in\mathcal{P}_t} \mathbb{E}\left[\mathbb{E}\left[\mathbb{1}\left\{Q_i^p(t), \mathcal{E}_t, \overline{H_i^p(t)}, m_i^p(t-1) \ge l\right\} \mid \mathcal{F}_{t-1}\right]\right] \\
&= (l+M) + \sum_{t\in[T]}\sum_{p\in\mathcal{P}_t} \mathbb{E}\left[\mathbb{1}\left\{\mathcal{E}_t, \overline{H_i^p(t)}, m_i^p(t-1) \ge l\right\} \mathbb{E}\left[\mathbb{1}\left\{\theta_i^p(t) > y_i^p\right\} \mid \mathcal{F}_{t-1}\right]\right] \\
&= (l+M) + \sum_{t\in[T]}\sum_{p\in\mathcal{P}_t} \mathbb{E}\left[\mathbb{1}\left\{\mathcal{E}_t, \overline{H_i^p(t)}, m_i^p(t-1) \ge l\right\} \cdot \overline{\Phi}\left(\sqrt{\overline{m_i^p}(t-1)/4}\left(y_i^p - \text{agg-}\hat{\mu}_i^p(t-1)\right)\right)\right] \\
&\le (l+M) + \sum_{t\in[T]}\sum_{p\in\mathcal{P}_t} \mathbb{E}\left[\mathbb{1}\left\{\mathcal{E}_t, \overline{H_i^p(t)}, m_i^p(t-1) \ge l\right\} \cdot \exp\left(-\frac{\overline{m_i^p}(t-1)\left(y_i^p - \text{agg-}\hat{\mu}_i^p(t-1)\right)_+^2}{8}\right)\right] \\
&\le (l+M) + \sum_{t\in[T]}\sum_{p\in\mathcal{P}_t} \mathbb{E}\left[\mathbb{1}\left\{\mathcal{E}_t, \overline{H_i^p(t)}, m_i^p(t-1) \ge l\right\} \cdot \exp\left(-\frac{\overline{m_i^p}(t-1)\left(\mu_i^p + \frac{3}{8}\Delta_i^p - \mu_i^p - \frac{9}{32}\Delta_i^p\right)_+^2}{8}\right)\right] \\
&\le (l+M) + \sum_{t\in[T]}\sum_{p\in\mathcal{P}_t} \mathbb{E}\left[\mathbb{1}\left\{\mathcal{E}_t, \overline{H_i^p(t)}, m_i^p(t-1) \ge l\right\} \exp\left(-\frac{\overline{m_i^p}(t-1)\left(\Delta_i^{\min}\right)^2}{(8)(256)}\right)\right] \\
&\le (l+M) + \sum_{t\in[T]}\sum_{p\in\mathcal{P}_t} \frac{1}{T^2} \\
&= \mathcal{O}\left(\frac{\ln T}{\left(\Delta_i^{\min}\right)^2} + M\right),
\end{aligned}
$$

where the first inequality is from Eq. (16), dropping the indicator $\mathbb{1}\left\{i_t^p = i\right\}$ and using the law of total expectation; the first equality follows from the observation that $\mathcal{E}_t$, $\overline{H_i^p(t)}$, and $\left\{m_i^p(t-1) \ge l\right\}$ are $\mathcal{F}_{t-1}$-measurable; the second equality follows from the observation that when $\overline{H_i^p(t)}$ happens, $\mathbb{E}\left[\mathbb{1}\left\{\theta_i^p(t) > y_i^p\right\} \mid \mathcal{F}_{t-1}\right] = \mathbb{P}\left(\theta_i^p(t) > y_i^p \mid \mathcal{F}_{t-1}\right) = \overline{\Phi}\left(\frac{y_i^p - \text{agg-}\hat{\mu}_i^p(t-1)}{\sqrt{4/\overline{m_i^p}(t-1)}}\right)$; the second inequality follows from Lemma C.35; the third inequality uses the facts that

1. when $\left\{m_i^p(t-1) \ge l\right\}$ happens, $\overline{m_i^p}(t-1) \ge m_i^p(t-1) - M \ge l - M = \frac{10240 \ln T}{\left(\Delta_i^{\min}\right)^2}$,
2. $y_i^p = \mu_i^p + \frac{1}{2}\delta_i^p > \mu_i^p + \frac{3}{8}\Delta_i^{\min}$ (see Fact C.4), and
3. when $\mathcal{E}_t$ happens, agg-$\hat{\mu}_i^p(t-1) \le \mu_i^p + \sqrt{\frac{10\ln T}{\overline{m_i^p}(t-1)}} + 2\epsilon < \mu_i^p + \frac{1}{32}\Delta_i^{\min} + \frac{1}{4}\Delta_i^{\min} = \mu_i^p + \frac{9}{32}\Delta_i^{\min}$ (see Definition B.9 and Fact A.24);

the fourth inequality is by algebra; and the fifth inequality again uses the fact that when $\left\{m_i^p(t-1) \ge l\right\}$ happens, $\overline{m_i^p}(t-1) \ge m_i^p(t-1) - M \ge \frac{10240 \ln T}{\left(\Delta_i^{\min}\right)^2}$.

In summary, we have

$$
(A) \le (A1) + (A2) + \mathcal{O}(1) \le \mathcal{O}\left(\frac{\ln T}{\left(\Delta_i^{\min}\right)^2} + M\right). \qquad \square
$$

C.1.3. Bounding Term (B)

We now bound term (B) in Eq. (14).

Lemma C.8.

$$(B) \leq \mathcal{O}\left(\frac{\ln T}{(\Delta_i^{\min})^2} + M\right).$$

Proof. Lemma C.8 follows from Lemmas C.11 and C.12, which we present shortly. □

Consider the following definition.

Definition C.9. In any round $t \in [T]$, for any active player $p \in \mathcal{P}_t$, define

$$\phi_{i,t}^p = \Pr\left(\theta_\dagger^p(t) > y_i^p \mid \mathcal{F}_{t-1}\right).$$

Remark C.10. Recall that $\overline{\Phi}(\cdot)$ denotes the complementary CDF of the standard Gaussian distribution; and recall $\overline{n_i^p}(t-1) = n_i^p(t-1) \vee 1$, and $\overline{m_i^p}(t-1) = (m_i^p(t-1) - M) \vee 1$. $\phi_{i,t}^p$ can be explicitly written as:

$$\phi_{i,t}^p = \overline{\Phi}\left(\frac{y_i^p - \hat{\mu}_\dagger^p(t-1)}{\sqrt{\mathrm{var}_\dagger^p(t-1)}}\right) \tag{17}$$

$$= \overline{\Phi}\left((y_i^p - \text{ind-}\hat{\mu}_\dagger^p(t-1))\sqrt{\overline{n_\dagger^p}(t-1)/4}\right) \cdot \mathbb{1}\left\{H_\dagger^p(t)\right\} + \overline{\Phi}\left((y_i^p - \text{agg-}\hat{\mu}_\dagger^p(t-1))\sqrt{\overline{m_\dagger^p}(t-1)/4}\right) \cdot \mathbb{1}\left\{\overline{H_\dagger^p(t)}\right\}. \tag{18}$$

Proof of Remark C.10. We have

$$\begin{aligned}
\phi_{i,t}^p =& \Pr\left(\theta_\dagger^p(t) > y_i^p \mid \hat{\mu}_\dagger^p(t-1), \mathrm{var}_\dagger^p(t-1)\right) \\
=& 1 - \Pr\left(\theta_\dagger^p(t) \leq y_i^p \mid \hat{\mu}_\dagger^p(t-1), \mathrm{var}_\dagger^p(t-1)\right) \\
=& 1 - \Phi\left(\frac{y_i^p - \hat{\mu}_\dagger^p(t-1)}{\sqrt{\mathrm{var}_\dagger^p(t-1)}}\right) = \overline{\Phi}\left(\frac{y_i^p - \hat{\mu}_\dagger^p(t-1)}{\sqrt{\mathrm{var}_\dagger^p(t-1)}}\right).
\end{aligned}$$

Eq. (18) now follows by observing that:

1. if $H_\dagger^p(t)$ happens, then $\hat{\mu}_\dagger^p(t-1) = \text{ind-}\hat{\mu}_\dagger^p(t-1)$ and $\mathrm{var}_\dagger^p(t-1) = \frac{4}{n_\dagger^p(t-1) \vee 1}$;

2. if $\overline{H_\dagger^p(t)}$ happens, then $\hat{\mu}_\dagger^p(t-1) = \text{agg-}\hat{\mu}_\dagger^p(t-1)$ and $\mathrm{var}_\dagger^p(t-1) = \frac{4}{(m_\dagger^p(t-1) - M) \vee 1}$. □

We now present the following lemma, which is inspired by a technique introduced in the work of (Agrawal & Goyal, 2017).

Lemma C.11.

$$(B) \leq \underbrace{\sum_{t \in [T]} \sum_{p \in \mathcal{P}_t} \mathbb{E}\left[\left(\frac{1 - \phi_{i,t}^p}{\phi_{i,t}^p}\right) \mathbb{1}\left\{i_t^p = \dagger, \mathcal{E}_t\right\}\right]}_{(B*)}.$$

Proof. In any round t and for any active player $p \in \mathcal{P}_t$, consider

$$\begin{aligned}
&\Pr\left(i_t^p = i, \overline{Q_i^p(t)}, \mathcal{E}_t \mid \mathcal{F}_{t-1}\right) \\
={}& \Pr\left(i_t^p = i, \theta_i^p(t) \leq y_i^p \mid \mathcal{F}_{t-1}\right) \cdot \mathbb{1}\left\{\mathcal{E}_t\right\} \\
\leq{}& \Pr\left(i_t^p = \dagger \mid \mathcal{F}_{t-1}\right) \cdot \frac{\Pr\left(\theta_\dagger^p(t) \leq y_i^p \mid \mathcal{F}_{t-1}\right)}{\Pr\left(\theta_\dagger^p(t) > y_i^p \mid \mathcal{F}_{t-1}\right)} \cdot \mathbb{1}\left\{\mathcal{E}_t\right\} \\
={}& \left(\frac{1-\phi_{i,t}^p}{\phi_{i,t}^p}\right) \cdot \Pr\left(i_t^p = \dagger \mid \mathcal{F}_{t-1}\right) \cdot \mathbb{1}\left\{\mathcal{E}_t\right\} \\
={}& \left(\frac{1-\phi_{i,t}^p}{\phi_{i,t}^p}\right) \Pr\left(i_t^p = \dagger, \mathcal{E}_t \mid \mathcal{F}_{t-1}\right),
\end{aligned} \tag{19}$$

where the first equality follows from the definition of $Q_i^p(t)$ and that $\mathcal{E}_t$ is $\mathcal{F}_{t-1}$-measurable; the first inequality uses Lemma C.40 with $l = \dagger$ and $z = y_i^p$; the second equality inequality is from the definition of $\phi_{i,t}^p$; and the last equality is again because $\mathcal{E}_t$ is $\mathcal{F}_{t-1}$-measurable.

Finally, we have

$$\begin{aligned}
\mathbb{E}\left[\mathbb{1}\left\{i_t^p = i, \overline{Q_i^p(t)}, \mathcal{E}_t\right\}\right] &= \mathbb{E}\left[\Pr\left(i_t^p = i, \overline{Q_i^p(t)}, \mathcal{E}_t \mid \mathcal{F}_{t-1}\right)\right] \\
&\leq \mathbb{E}\left[\left(\frac{1-\phi_{i,t}^p}{\phi_{i,t}^p}\right) \Pr\left(i_t^p = \dagger, \mathcal{E}_t \mid \mathcal{F}_{t-1}\right)\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\left(\frac{1-\phi_{i,t}^p}{\phi_{i,t}^p}\right) \mathbb{1}\left\{i_t^p = \dagger, \mathcal{E}_t\right\} \mid \mathcal{F}_{t-1}\right]\right] \\
&= \mathbb{E}\left[\left(\frac{1-\phi_{i,t}^p}{\phi_{i,t}^p}\right) \mathbb{1}\left\{i_t^p = \dagger, \mathcal{E}_t\right\}\right],
\end{aligned}$$

where we use the law of total expectation and Eq. (19). The lemma follows by summing over all t, p's. □

With foresight, let $L = \frac{2560 \ln T}{\left(\Delta_i^{\min}\right)^2} + M$. We further decompose term $(B*)$ as follows.

$$\begin{aligned}
(B*) &= \sum_{t\in[T]} \sum_{p\in\mathcal{P}_t} \mathbb{E}\left[\left(\frac{1-\phi_{i,t}^p}{\phi_{i,t}^p}\right) \mathbb{1}\left\{i_t^p = \dagger, \mathcal{E}_t\right\}\right] \\
&= \underbrace{\sum_{t\in[T]} \sum_{p\in\mathcal{P}_t} \mathbb{E}\left[\left(\frac{1-\phi_{i,t}^p}{\phi_{i,t}^p}\right) \mathbb{1}\left\{i_t^p = \dagger, \mathcal{E}_t, H_\dagger^p(t)\right\}\right]}_{(b1)} + \underbrace{\sum_{t\in[T]} \sum_{p\in\mathcal{P}_t} \mathbb{E}\left[\left(\frac{1-\phi_{i,t}^p}{\phi_{i,t}^p}\right) \mathbb{1}\left\{i_t^p = \dagger, \mathcal{E}_t, \overline{H_\dagger^p(t)}\right\}\right]}_{(b2)}, \\
&= (b1) + \underbrace{\sum_{t\in[T]} \sum_{p\in\mathcal{P}_t} \mathbb{E}\left[\left(\frac{1-\phi_{i,t}^p}{\phi_{i,t}^p}\right) \mathbb{1}\left\{i_t^p = \dagger, \mathcal{E}_t, \overline{H_\dagger^p(t)}, m_\dagger^p(t-1) < L\right\}\right]}_{(b2.1)} \\
&\qquad + \underbrace{\sum_{t\in[T]} \sum_{p\in\mathcal{P}_t} \mathbb{E}\left[\left(\frac{1-\phi_{i,t}^p}{\phi_{i,t}^p}\right) \mathbb{1}\left\{i_t^p = \dagger, \mathcal{E}_t, \overline{H_\dagger^p(t)}, m_\dagger^p(t-1) \geq L\right\}\right]}_{(b2.2)}.
\end{aligned} \tag{20}$$

where the inequality uses Lemma C.11.

Lemma C.12.

$$(B*) \le \mathcal{O}\left(\frac{\ln T}{\left(\Delta_i^{\min}\right)^2} + M\right).$$

Proof. Lemma C.12 follows directly from Eq. (20) and the following Lemma C.13, Lemma C.14 and Lemma C.18, which provide upper bounds on terms $(b1)$, $(b2.1)$ and $(b2.2)$, respectively. □

Lemma C.13 (Bounding term $(b1)$)**.**

$$(b1) \le \mathcal{O}\left(M\right).$$

Proof of Lemma C.13. For any player $p \in [M]$ and $t \in [T]$, recall that $\overline{n^p_\dagger}(t-1) = n^p_\dagger(t-1) \vee 1$ and $(z)_+ = z \vee 0$. When $\mathcal{E}_t$ and $H^p_\dagger(t)$ happen, $n^p_\dagger(t-1) \ge \frac{40 \ln T}{\epsilon^2} =: Y$; we have:

$$\begin{aligned}
&1 - \phi^p_{i,t} \\
=&\Pr\left(\theta^p_\dagger(t) \le y^p_i \mid \mathcal{F}_{t-1}\right) \\
=&\Phi\left((y^p_i - \text{ind-}\hat{\mu}^p_\dagger(t-1))\sqrt{\overline{n^p_\dagger}(t-1)/4}\right) \\
\le&\exp\left(-\frac{\overline{n^p_\dagger}(t-1)(\text{ind-}\hat{\mu}^p_\dagger(t-1) - y^p_i)^2_+}{8}\right) \\
\le&\exp\left(-\frac{n^p_\dagger(t-1)(\mu^p_\dagger - \frac{1}{4}\Delta^p_i - \mu^p_\dagger + \frac{3}{8}\Delta^p_i)^2_+}{8}\right) \\
\le&\exp\left(-\frac{n^p_\dagger(t-1)(\Delta^p_i)^2}{8(64)}\right) \\
\le&\frac{1}{T+1},
\end{aligned}$$

where the second equality uses Remark C.10; the first inequality uses Lemma C.35; the second inequality follows from the observations that, when $\mathcal{E}_t$ and $H^p_\dagger(t)$ happen:

1. $\overline{n^p_\dagger}(t-1) \ge n^p_\dagger(t-1) \ge Y = \frac{40 \ln T}{\epsilon^2} \ge \frac{2560 \ln T}{(\Delta^p_i)^2}$ (see Fact C.4),

2. $\text{ind-}\hat{\mu}^p_\dagger(t-1) \ge \mu^p_\dagger - \sqrt{\frac{10 \ln T}{n^p_\dagger(t-1)}} \ge \mu^p_\dagger - \frac{1}{4}\Delta^p_i$ (see Definition B.9), and

3. $y^p_i = \mu^p_\dagger - \frac{1}{2}\delta^p_i < \mu^p_\dagger - \frac{3}{8}\Delta^p_i$;

the third inequality is by algebra; and the last inequality follows because, again, when $H^p_\dagger(t)$ happens, $n^p_\dagger(t-1) \ge Y = \frac{40 \ln T}{\epsilon^2} \ge \frac{2560 \ln T}{(\Delta^p_i)^2} \ge \frac{1280 \ln(T+1)}{(\Delta^p_i)^2}$ for $T > 1$.

It follows that, when $\mathcal{E}_t$ and $H^p_\dagger(t)$ happen, $\phi^p_{i,t} \ge \frac{T}{T+1}$ and $\frac{1-\phi^p_{i,t}}{\phi^p_{i,t}} \le \frac{1}{T}$. Hence,

$$(b1) \le \sum_{p \in [M]} \sum_{t : p \in \mathcal{P}_t} \mathbb{E}\left[\left(\frac{1 - \phi^p_{i,t}}{\phi^p_{i,t}}\right) \mathbb{1}\left\{\mathcal{E}_t, H^p_\dagger(t)\right\}\right] \le M.$$

□

Lemma C.14 (Bounding term $(b2.1)$)**.**

$$(b2.1) \leq \mathcal{O}\left(\frac{\ln T}{\left(\Delta_i^{\min}\right)^2} + M\right).$$

The remark below is useful for proving Lemma C.14.

Remark C.15 (Invariant property)**.** Recall from Example A.21 that $\left\{H_\dagger^p(t) : t \in [T], p \in [M]\right\}$ satisfies the invariant property with respect to †.

Moreover, the construction of Algorithm 1 enforces that $\left\{\phi_{i,t}^p : t \in [T], p \in [M]\right\}$ satisfies the invariant property with respect to † (note that it does not necessarily satisfy the invariant property with respect to i). Indeed, this follows from Eq. (17), along with Example A.23 that shows that the posterior parameters, $\left\{(\hat{\mu}_\dagger^p(t-1), \mathrm{var}_\dagger^p(t-1)) : t \in [T], p \in [M]\right\}$, satisfy the invariant property with respect to †.

Combining the two observations above, $\left\{\left(\frac{1}{\phi_{i,t}^p} - 1\right) \mathbb{1}\left\{\overline{H_\dagger^p(t)}\right\} : t \in [T], p \in [M]\right\}$ also satisfies the invariant property with respect to arm †.

Proof of Lemma C.14. Proving Lemma C.14 requires more special care. Recall that

$$\begin{aligned}(b2.1) &= \sum_{t \in [T]} \sum_{p \in \mathcal{P}_t} \mathbb{E}\left[\left(\frac{1 - \phi_{i,t}^p}{\phi_{i,t}^p}\right) \mathbb{1}\left\{i_t^p = \dagger, \mathcal{E}_t, \overline{H_\dagger^p(t)}, m_\dagger^p(t-1) < L\right\}\right] \\ &\leq \sum_{t \in [T]} \sum_{p \in \mathcal{P}_t} \mathbb{E}\left[\left(\frac{1 - \phi_{i,t}^p}{\phi_{i,t}^p}\right) \mathbb{1}\left\{i_t^p = \dagger, \overline{H_\dagger^p(t)}, m_\dagger^p(t-1) < L\right\}\right].\end{aligned}$$

Also recall the definition of stopping time $\tau_k(\dagger)$ (Definition A.15), the round in which † is pulled the k-th time by any player. To ease exposition, we abuse the notation and denote $\tau_k(\dagger)$ by τ_k. Similarly, let $p_k := p_k(\dagger)$ denote the player that issues the k-th pull of arm † (recall Definition A.17).

Since $\left\{\left(\frac{1}{\phi_{i,t}^p} - 1\right) \mathbb{1}\left\{\overline{H_\dagger^p(t)}\right\} : t \in [T], p \in [M]\right\}$ satisfies the invariant property with respect to arm †, by Lemma C.38, we have

$$(b2.1) \leq \sum_{p=1}^{M} \mathbb{E}\left[\left(\frac{1}{\phi_{i,1}^p} - 1\right) \mathbb{1}\left\{\overline{H_\dagger^p(1)}\right\}\right] + \sum_{k=1}^{L-1} \mathbb{E}\left[\left(\frac{1}{\phi_{i,\tau_k+1}^{p_k}} - 1\right) \mathbb{1}\left\{\tau_k \leq T, \overline{H_\dagger^p(\tau_k + 1)}\right\}\right], \tag{21}$$

where we also use the linearity of expectations.

Since the variance of the aggregate posteriors are initialized as the constant $c_2 = 4$ in RobustAgg-TS(ϵ), we have that $\left(\frac{1}{\phi_{i,1}^p} - 1\right) \mathbb{1}\left\{\overline{H_\dagger^p(1)}\right\} \leq \mathcal{O}(1)$ with probability 1. Therefore,

$$\sum_{p=1}^{M} \mathbb{E}\left[\left(\frac{1}{\phi_{i,1}^p} - 1\right) \mathbb{1}\left\{\overline{H_\dagger^p(1)}\right\}\right] \leq \mathcal{O}(M). \tag{22}$$

It then suffices to bound the second term in Eq. (21)—it follows straightforwardly from Lemma C.16, which we present shortly, that the second term is bounded by $\mathcal{O}(L)$. It then follows from Eq. (21), Eq. (22), and Lemma C.16 that $(b2.1) \leq \mathcal{O}\left(\frac{\ln T}{\left(\Delta_i^{\min}\right)^2} + M\right)$. □

Lemma C.16. *For any $k \in [TM]$,*

$$\mathbb{E}\left[\left(\frac{1}{\phi^{p_k}_{i,\tau_k+1}} - 1\right)\mathbb{1}\left\{\tau_k \le T, \overline{H^p_\dagger(\tau_k+1)}\right\}\right] \le \mathcal{O}\left(1\right),$$

where we recall that $\tau_k = \tau_k(\dagger)$ and $p_k = p_k(\dagger)$ is the player that issues the k-th pull of arm $\dagger$.

Proof. Using Remark C.10, we observe that

$$\phi^{p_k}_{i,\tau_k+1} = \left[\overline{\Phi}\left(\frac{y^p_i - \text{ind-}\hat{\mu}^p_\dagger(\tau_k)}{2}\sqrt{\left(n^p_\dagger(\tau_k)\right) \vee 1}\right)\right] \cdot \mathbb{1}\left\{H^p_\dagger(\tau_k+1)\right\}$$
$$+ \left[\overline{\Phi}\left(\frac{y^p_i - \text{agg-}\hat{\mu}^p_\dagger(\tau_k)}{2}\sqrt{\left(m^p_\dagger(\tau_k) - M\right) \vee 1}\right)\right] \cdot \mathbb{1}\left\{\overline{H^p_\dagger(\tau_k+1)}\right\}. \tag{23}$$

We have

$$\mathbb{E}\left[\left(\frac{1}{\phi^{p_k}_{i,\tau_k+1}} - 1\right)\mathbb{1}\left\{\tau_k \le T, \overline{H^{p_k}_\dagger(\tau_k+1)}\right\}\right]$$
$$= \mathbb{E}\left[\left(\frac{1}{\overline{\Phi}\left(\left(y^{p_k}_i - \text{agg-}\hat{\mu}^{p_k}_\dagger(\tau_k)\right)\sqrt{\left(\left(m^{p_k}_\dagger(\tau_k) - M\right) \vee 1\right)/4}\right)} - 1\right)\mathbb{1}\left\{\tau_k \le T, \overline{H^{p_k}_\dagger(\tau_k+1)}\right\}\right]$$
$$\le \mathbb{E}\left[\frac{1}{\overline{\Phi}\left(\left(\mu^{p_k}_\dagger - \text{agg-}\hat{\mu}_\dagger(\tau_k)\right)\sqrt{\left(\left(n_\dagger(\tau_k) - M\right) \vee 1\right)/4}\right)}\mathbb{1}\left\{\tau_k \le T\right\}\right], \tag{24}$$

where the last inequality uses the observations that $y^{p_k}_i \le \mu^{p_k}_\dagger$, $\text{agg-}\hat{\mu}^{p_k}_\dagger(\tau_k) = \text{agg-}\hat{\mu}_\dagger(\tau_k)$ and $m^{p_k}_\dagger(\tau_k) = n_\dagger(\tau_k)$, as well as the monotonic increasing property of $z \mapsto \frac{1}{\overline{\Phi}(z)}$.

Observe that, from Corollary B.4, for any $z \ge 1$,

$$\Pr\left((\tau_k \le T) \wedge \left(\mu^{p_k}_\dagger - \text{agg-}\hat{\mu}_\dagger(\tau_k) \ge z\sqrt{\frac{4}{(n_\dagger(\tau_k) - M) \vee 1}}\right)\right)$$
$$\le \Pr\left((\tau_k \le T) \wedge \left(\exists p \in [M],\ \mu^p_\dagger - \text{agg-}\hat{\mu}_\dagger(\tau_k) \ge z\sqrt{\frac{4}{(n_\dagger(\tau_k) - M) \vee 1}}\right)\right)$$
$$\le 2e^{-2z^2},$$

Applying Lemma C.36 with $X = \left(\text{agg-}\hat{\mu}_\dagger(\tau_k) - \mu^{p_k}_\dagger\right)\sqrt{\left(\left(n_\dagger(\tau_k) - M\right) \vee 1\right)/4}$ and $E = \{\tau_k \le T\}$, we conclude the proof. □

Remark C.17. Note that it follows from our novel concentration inequality (Corollary B.4) that

$$\Pr\left(\tau_k \le T, \mu_\dagger^p - \text{agg-}\hat{\mu}_\dagger(\tau_k) > \sqrt{\frac{2\ln\left(\frac{2}{\delta}\right)}{\left(n_\dagger(\tau_k) - M\right)\vee 1}}\right) < \delta;$$

this tight bound enables us to bound Eq. (24) by $\mathcal{O}\left(1\right)$, which is essential to our proof of Lemma C.16.

Since $n_i(\tau_k) \le [k, k+M-1]$, using the Azuma-Hoeffding inequality and the union bound, one can obtain

$$\Pr\left(\tau_k \le T, \mu_\dagger^p - \text{agg-}\hat{\mu}_\dagger(\tau_k) > \mathcal{O}\left(\sqrt{\frac{\ln\left(\frac{M}{\delta}\right)}{\left(n_\dagger(\tau_k) - M\right)\vee 1}}\right)\right) < \delta;$$

and using Freedman's inequality (e.g., Wang et al., 2021, Lemma 17), one can obtain

$$\Pr\left(\tau_k \le T, \mu_\dagger^p - \text{agg-}\hat{\mu}_\dagger(\tau_k) > \mathcal{O}\left(\sqrt{\frac{\ln\left(\frac{\ln T}{\delta}\right)}{\left(n_\dagger(\tau_k) - M\right)\vee 1}}\right)\right) < \delta.$$

However, naively combining the above bounds with Lemma C.36, one needs to set C_1 in Lemma C.36 to be $\mathcal{O}\left(M\right)$ or $\mathcal{O}\left(\ln T\right)$, which incurs extra (undesirable) $\mathcal{O}\left(M\right)$ or $\mathcal{O}\left(\ln T\right)$ factors for bounding Eq. (24).

Lemma C.18 (Bounding term $(b2.2)$)**.**

$$(b2.2) \le \mathcal{O}\left(M\right).$$

Proof of Lemma C.18. For any player $p \in [M]$ and $t \in [T]$, recall that $\overline{m_\dagger^p}(t-1) = (m_\dagger^p(t-1) - M) \vee 1$ and $(z)_+ = z \vee 0$. When $\mathcal{E}_t$, $\left\{m_\dagger^p(t-1) \ge L\right\}$ and $\overline{H_\dagger^p(t)}$ happen,

$$\begin{aligned}
&1 - \phi_{i,t}^p \\
=&\Pr\left(\theta_\dagger^p(t) \le y_i^p \mid \mathcal{F}_{t-1}\right) \\
=&\Phi\left((y_i^p - \text{agg-}\hat{\mu}_\dagger^p(t-1))\sqrt{\overline{m_\dagger^p}(t-1)/4}\right) \\
\le&\exp\left(-\frac{\overline{m_\dagger^p}(t-1)(\text{agg-}\hat{\mu}_\dagger^p(t-1) - y_i^p)_+^2}{8}\right) \\
\le&\exp\left(-\frac{\overline{m_\dagger^p}(t-1)(\mu_\dagger^p - \frac{1}{4}\Delta_i^p - \mu_\dagger^p + \frac{3}{8}\Delta_i^p)_+^2}{8}\right) \\
\le&\exp\left(-\frac{\overline{m_\dagger^p}(t-1)(\Delta_i^p)^2}{8(64)}\right) \\
\le&\frac{1}{T+1},
\end{aligned}$$

where the second equality uses Remark C.10; the first inequality uses Lemma C.35; the second inequality follows from the observations that, when $\mathcal{E}_t$, $\left\{m_\dagger^p(t-1) \ge L\right\}$ and $\overline{H_\dagger^p(t)}$ happen:

1. $\overline{m_\dagger^p}(t-1) \ge m_\dagger^p(t-1) - M \ge L - M \ge \frac{2560\ln T}{(\Delta_i^p)^2}$,

2. agg-$\hat{\mu}^p_\dagger(t-1) \geq \mu^p_\dagger - \sqrt{\frac{10 \ln T}{\overline{m^p_\dagger}(t-1)}} \geq \mu^p_\dagger - \frac{1}{4}\Delta^p_i$ (see Definition B.9), and

3. $y^p_i = \mu^p_\dagger - \frac{1}{2}\delta^p_i < \mu^p_\dagger - \frac{3}{8}\Delta^p_i$;

the third inequality is by algebra; and the last inequality follows from the observation that $\overline{m^p_\dagger}(t-1) \geq m^p_\dagger(t-1) - M \geq L - M \geq \frac{2560 \ln T}{(\Delta^p_i)^2} \geq \frac{1280 \ln(T+1)}{(\Delta^p_i)^2}$ for $T > 1$.

It follows that, when $\mathcal{E}_t$, $\left\{m^p_\dagger(t-1) \geq L\right\}$ and $\overline{H^p_\dagger(t)}$ happen, $\phi^p_{i,t} \geq \frac{T}{T+1}$ and $\frac{1-\phi^p_{i,t}}{\phi^p_{i,t}} \leq \frac{1}{T}$. Hence,

$$(b2.2) \leq \sum_{p \in [M]} \sum_{t: p \in \mathcal{P}_t} \mathbb{E}\left[\left(\frac{1-\phi^p_{i,t}}{\phi^p_{i,t}}\right) \mathbb{1}\left\{\mathcal{E}_t, \overline{H^p_\dagger(t)}, m^p_\dagger(t-1) \geq L\right\}\right] \leq M. \qquad \square$$

C.2. Non-subpar Arms

In this section, we provide a proof for Lemma C.2.

Let us fix any player $p \in [M]$ and any suboptimal arm $i \in \mathcal{I}^C_{10\epsilon}$ for player p such that $\Delta^p_i > 0$. In the rest of this section, let us also fix an optimal arm for player p, $\diamond_p$, and we abbreviate it by $\diamond$. We have $\mu^p_\diamond = \mu^p_* = \max_{j \in [K]} \mu^p_j$.

Definition C.19. Let $z^p_i = \mu^p_i + \frac{1}{2}\Delta^p_i$ be a threshold. In any round t, define

$$W^p_i(t) = \left\{\theta^p_i(t) > z^p_i\right\}$$

to be the event that the sample $\theta^p_i(t)$ from the posterior distribution associated with arm i and player p in round t is greater than the threshold z^p_i. Therefore, $\overline{W^p_i(t)} = \left\{\theta^p_i(t) \leq z^p_i\right\}$.

C.2.1. Non-subpar Arms—Decomposition

We consider the following decomposition.

$$\begin{aligned}
&\mathbb{E}\left[n^p_i(T)\right] \\
=&\mathbb{E}\left[\sum_{t: p \in \mathcal{P}_t} \mathbb{1}\left\{i^p_t = i\right\}\right] \\
=&\mathbb{E}\left[\sum_{t: p \in \mathcal{P}_t} \mathbb{1}\left\{i^p_t = i, W^p_i(t), \mathcal{E}_t\right\}\right] + \mathbb{E}\left[\sum_{t: p \in \mathcal{P}_t} \mathbb{1}\left\{i^p_t = i, \overline{W^p_i(t)}, \mathcal{E}_t\right\}\right] + \sum_{t: p \in \mathcal{P}_t} \mathbb{E}\left[\mathbb{1}\left\{i^p_t = i, \overline{\mathcal{E}_t}\right\}\right] \\
\leq&\underbrace{\mathbb{E}\left[\sum_{t: p \in \mathcal{P}_t} \mathbb{1}\left\{i^p_t = i, W^p_i(t), \mathcal{E}_t\right\}\right]}_{(D)} + \underbrace{\mathbb{E}\left[\sum_{t: p \in \mathcal{P}_t} \mathbb{1}\left\{i^p_t = i, \overline{W^p_i(t)}, \mathcal{E}_t\right\}\right]}_{(E)} + \mathcal{O}(1),
\end{aligned} \tag{25}$$

where the last inequality follows from the observation that $\mathbb{E}\left[\mathbb{1}\left\{i^p_t = i, \overline{\mathcal{E}_t}\right\}\right] \leq \mathbb{E}\left[\mathbb{1}\left\{\overline{\mathcal{E}_t}\right\}\right]$ and Lemma B.10.

Following this decomposition, Lemma C.2 is proved straightforwardly given Lemma C.20 and Lemma C.21 which we present in what follows.

C.2.2. Bounding Term (D)

We first bound term (D) in Eq. (25).

Lemma C.20.

$$(D) \leq \mathcal{O}\left(\frac{\ln T}{(\Delta^p_i)^2} + M\right).$$

Proof of Lemma C.20. With foresight, let $h = \frac{4000 \ln T}{(\Delta_i^p)^2} + 2M$. Recall that $H_i^p(t)$ is the event that the individual posterior is used in round t by active player p for arm i (see Definition A.13). We have

$$
\begin{aligned}
(D) &= \mathbb{E}\left[\sum_{t:p\in\mathcal{P}_t} \mathbb{1}\left\{i_t^p = i, W_i^p(t), \mathcal{E}_t\right\}\right] \\
&\le h + \sum_{t:p\in\mathcal{P}_t} \mathbb{E}\left[\mathbb{1}\left\{i_t^p = i, W_i^p(t), \mathcal{E}_t, n_i^p(t-1) \ge h\right\}\right] \\
&= h + \underbrace{\sum_{t:p\in\mathcal{P}_t} \mathbb{E}\left[\mathbb{1}\left\{i_t^p = i, W_i^p(t), \mathcal{E}_t, H_i^p(t), n_i^p(t-1) \ge h\right\}\right]}_{(d)},
\end{aligned}
$$

where the last equality follows from the observation that $\left\{n_i^p(t-1) \ge h\right\}$ implies that $H_i^p(t)$ happening. To see why this is true, recall that $H_i^p(t) = \left\{n_i^p(t-1) \ge \frac{40 \ln T}{\epsilon^2} + 2M\right\}$; and observe that for non-subpar arm $i \in \mathcal{I}_{10\epsilon}^C$ and player p, $\left\{n_i^p(t-1) \ge h = \frac{4000 \ln T}{(\Delta_i^p)^2} + 2M\right\}$ implies $\left\{n_i^p(t-1) \ge \frac{40 \ln T}{\epsilon^2} + 2M\right\}$ because $\Delta_i^p \le 10\epsilon$.

It therefore suffices to bound term (d). We have

$$
\begin{aligned}
(d) &\le \sum_{t:p\in\mathcal{P}_t} \mathbb{E}\left[\mathbb{1}\left\{W_i^p(t), \mathcal{E}_t, H_i^p(t), n_i^p(t-1) \ge h\right\}\right] \\
&= \sum_{t:p\in\mathcal{P}_t} \mathbb{E}\left[\mathbb{1}\left\{\mathcal{E}_t, H_i^p(t), n_i^p(t-1) \ge h\right\} \mathbb{E}\left[\mathbb{1}\left\{W_i^p(t)\right\} \mid \mathcal{F}_{t-1}\right]\right] \\
&= \sum_{t:p\in\mathcal{P}_t} \mathbb{E}\left[\mathbb{1}\left\{\mathcal{E}_t, H_i^p(t), n_i^p(t-1) \ge h\right\} \overline{\Phi}\left((z_i^p - \text{ind-}\hat{\mu}_i^p(t-1))\sqrt{\overline{n_i^p}(t-1)/4}\right)\right] \\
&\le \sum_{t:p\in\mathcal{P}_t} \mathbb{E}\left[\mathbb{1}\left\{\mathcal{E}_t, H_i^p(t), n_i^p(t-1) \ge h\right\} \exp\left(-\frac{\overline{n_i^p}(t-1)(z_i^p - \text{ind-}\hat{\mu}_i^p(t-1))_+^2}{8}\right)\right] \\
&\le \sum_{t:p\in\mathcal{P}_t} \mathbb{E}\left[\mathbb{1}\left\{\mathcal{E}_t, H_i^p(t), n_i^p(t-1) \ge h\right\} \exp\left(-\frac{n_i^p(t-1)(\mu_i^p + \frac{1}{2}\Delta_i^p - \mu_i^p - \frac{1}{16}\Delta_i^p)_+^2}{8}\right)\right] \\
&\le \sum_{t:p\in\mathcal{P}_t} \mathbb{E}\left[\mathbb{1}\left\{\mathcal{E}_t, H_i^p(t), n_i^p(t-1) \ge h\right\} \exp\left(-\frac{n_i^p(t-1)(\Delta_i^p)^2}{8(16)}\right)\right] \\
&\le \mathcal{O}(1).
\end{aligned}
$$

where the first inequality drops the indicator $\mathbb{1}\left\{i_t^p = i\right\}$; the first equality uses the law of total expectation and the observation that $\mathcal{E}_t$, $H_i^p(t)$ and $\left\{n_i^p(t-1) \ge h\right\}$ are $\mathcal{F}_{t-1}$-measurable; the second inequality follows from Lemma C.35; the third inequality is from the observations that when $\mathcal{E}_t$ and $H_i^p(t)$ happen:

1. $\overline{n_i^p}(t-1) \ge n_i^p(t-1) \ge h = \frac{4000 \ln T}{(\Delta_i^p)^2} + 2M$,
2. $\text{ind-}\hat{\mu}_i^p(t-1) \le \mu_i^p + \sqrt{\frac{10 \ln T}{\overline{n_i^p}(t-1)}} \le \mu_i^p + \frac{1}{16}\Delta_i^p$ (see Definition B.9), and
3. $z_i^p = \mu_i^p + \frac{1}{2}\Delta_i^p$;

the fourth inequality is by algebra; and the last inequality is from the observation that when $n_i^p(t-1) \ge h$, $\exp\left(-\frac{n_i^p(t-1)(\Delta_i^p)^2}{8(16)}\right) \le \frac{1}{T}$.

In summary, $(D) \le h + (d) \le \mathcal{O}\left(\frac{\ln T}{(\Delta_i^p)^2} + M\right)$. □

C.2.3. Bounding Term (E)

We now bound (E) in Eq. (25):

Lemma C.21.

$$(E) \leq \mathcal{O}\left(\frac{\ln T}{(\Delta_i^p)^2} + M\right).$$

Proof. Lemma C.21 follows from Lemma C.24, Eq. (29), Lemma C.25, and Lemma C.30 which we present shortly. □

We begin with the following definition, similar to the notion of $\phi_{i,t}^p$ used for subpar arms.

Definition C.22. Recall that p is a fixed player, i is a fixed suboptimal arm for p, and $\diamond$ is a fixed optimal arm for p. In any round t, define

$$\psi_{i,t}^p = \Pr\left(\theta_\diamond^p(t) > z_i^p \mid \mathcal{F}_{t-1}\right).$$

Remark C.23. Recall that $\overline{n_\diamond^p}(t-1) = n_\diamond^p(t-1) \vee 1$ and $\overline{m_\diamond^p}(t-1) = \big(m_\diamond^p(t-1) - M\big) \vee 1$. $\psi_{i,t}^p$ can be explicitly written as:

$$\psi_{i,t}^p = \overline{\Phi}\left(\frac{z_i^p - \hat{\mu}_\diamond^p(t-1)}{\sqrt{\mathrm{var}_\diamond^p(t-1)}}\right) \tag{26}$$

$$= \overline{\Phi}\left((z_i^p - \text{ind-}\hat{\mu}_\diamond^p(t-1))\sqrt{\overline{n_\diamond^p}(t-1)/4}\right) \cdot \mathbb{1}\left\{H_\diamond^p(t)\right\} + \overline{\Phi}\left((z_i^p - \text{agg-}\hat{\mu}_\diamond^p(t-1))\sqrt{\overline{m_\diamond^p}(t-1)/4}\right) \cdot \mathbb{1}\left\{\overline{H_\diamond^p}(t)\right\}. \tag{27}$$

The proof for the above remark is omitted, as it is very similar to that of Remark C.10.

We now present the following lemma.

Lemma C.24.

$$(E) = \mathbb{E}\left[\sum_{t: p \in \mathcal{P}_t} \mathbb{1}\left\{i_t^p = i, \overline{W_i^p(t)}, \mathcal{E}_t\right\}\right] \leq \underbrace{\sum_{t: p \in \mathcal{P}_t} \mathbb{E}\left[\left(\frac{1 - \psi_{i,t}^p}{\psi_{i,t}^p}\right) \mathbb{1}\left\{i_t^p = \diamond, \mathcal{E}_t\right\}\right]}_{(E*)}$$

Proof. The proof largely follows the same outline as that of Lemma C.11.

In any round t and such that $p \in \mathcal{P}_t$, consider

$$\begin{aligned}
&\Pr\left(i_t^p = i, \overline{Q_i^p(t)}, \mathcal{E}_t \mid \mathcal{F}_{t-1}\right) \\
=& \Pr\left(i_t^p = i, \theta_i^p(t) \leq z_i^p \mid \mathcal{F}_{t-1}\right) \cdot \mathbb{1}\left\{\mathcal{E}_t\right\} \\
\leq& \Pr\left(i_t^p = \diamond \mid \mathcal{F}_{t-1}\right) \cdot \frac{\Pr\left(\theta_\diamond^p(t) \leq z_i^p \mid \mathcal{F}_{t-1}\right)}{\Pr\left(\theta_\diamond^p(t) > z_i^p \mid \mathcal{F}_{t-1}\right)} \cdot \mathbb{1}\left\{\mathcal{E}_t\right\} \\
=& \left(\frac{1 - \psi_{i,t}^p}{\psi_{i,t}^p}\right) \cdot \Pr\left(i_t^p = \diamond \mid \mathcal{F}_{t-1}\right) \cdot \mathbb{1}\left\{\mathcal{E}_t\right\} \\
=& \left(\frac{1 - \psi_{i,t}^p}{\psi_{i,t}^p}\right) \Pr\left(i_t^p = \diamond, \mathcal{E}_t \mid \mathcal{F}_{t-1}\right),
\end{aligned} \tag{28}$$

where the first equality follows from the definition of $Q_i^p(t)$ and that $\mathcal{E}_t$ is $\mathcal{F}_{t-1}$-measurable; the first inequality uses Lemma C.40 with $l = \diamond$ and $z = z_i^p$; and the second equality inequality is from the definition of $\psi_{i,t}^p$; the last equality is again because $\mathcal{E}_t$ is $\mathcal{F}_{t-1}$-measurable.

Finally, we have

$$\begin{aligned}
\mathbb{E}\left[\mathbb{1}\left\{i_t^p = i, \overline{Q_i^p(t)}, \mathcal{E}_t\right\}\right] &= \mathbb{E}\left[\Pr\left(i_t^p = i, \overline{Q_i^p(t)}, \mathcal{E}_t \mid \mathcal{F}_{t-1}\right)\right] \\
&\leq \mathbb{E}\left[\left(\frac{1-\psi_{i,t}^p}{\psi_{i,t}^p}\right) \Pr\left(i_t^p = \diamond, \mathcal{E}_t \mid \mathcal{F}_{t-1}\right)\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\left(\frac{1-\psi_{i,t}^p}{\psi_{i,t}^p}\right) \mathbb{1}\left\{i_t^p = \diamond, \mathcal{E}_t\right\} \mid \mathcal{F}_{t-1}\right]\right] \\
&= \mathbb{E}\left[\left(\frac{1-\psi_{i,t}^p}{\psi_{i,t}^p}\right) \mathbb{1}\left\{i_t^p = \diamond, \mathcal{E}_t\right\}\right],
\end{aligned}$$

where we use the law of total expectation and Eq. (28). The lemma follows by summing over all t's. □

Let us further decompose $(E*)$ as follows.

$$(E*) = \underbrace{\sum_{t:p\in\mathcal{P}_t} \mathbb{E}\left[\left(\frac{1-\psi_{i,t}^p}{\psi_{i,t}^p}\right) \mathbb{1}\left\{i_t^p = \diamond, \mathcal{E}_t, H_\diamond^p(t)\right\}\right]}_{(e1)} + \underbrace{\sum_{t:p\in\mathcal{P}_t} \mathbb{E}\left[\left(\frac{1-\psi_{i,t}^p}{\psi_{i,t}^p}\right) \mathbb{1}\left\{i_t^p = \diamond, \mathcal{E}_t, \overline{H_\diamond^p(t)}\right\}\right]}_{(e2)}. \tag{29}$$

We first consider term $(e1)$.

Lemma C.25.

$$(e1) \leq \mathcal{O}\left(\frac{\ln T}{(\Delta_i^p)^2}\right).$$

Proof of Lemma C.25. With foresight, let $J = \frac{640 \ln T}{(\Delta_i^p)^2}$. We have

$$(e1) = \underbrace{\sum_{t:p\in\mathcal{P}_t} \mathbb{E}\left[\left(\frac{1-\psi_{i,t}^p}{\psi_{i,t}^p}\right) \mathbb{1}\left\{i_t^p = \diamond, \mathcal{E}_t, H_\diamond^p(t), n_\diamond^p(t-1) < J\right\}\right]}_{(e1.1)} +$$

$$\underbrace{\sum_{t:p\in\mathcal{P}_t} \mathbb{E}\left[\left(\frac{1-\psi_{i,t}^p}{\psi_{i,t}^p}\right) \mathbb{1}\left\{i_t^p = \diamond, \mathcal{E}_t, H_\diamond^p(t), n_\diamond^p(t-1) \geq J\right\}\right]}_{(e1.2)}.$$

Lemma C.25 follows straightforwardly from Lemma C.26 and Lemma C.29, which bound $(e1.1)$ and $(e1.2)$, respectively. □

Lemma C.26.

$$(e1.1) \leq \mathcal{O}\left(\frac{\ln T}{(\Delta_i^p)^2}\right).$$

To prove Lemma C.26, we first present the following Remark C.15.

Remark C.27 (Invariant Property)**.** Similar to Remark C.15, by the construction of Algorithm 1, we have that for any arm $i \in [K]$, and player $p \in [M]$, $\left\{\psi_{i,t}^p : t \in [T]\right\}$ and $\left\{H_\diamond^p(t) : t \in [T]\right\}$ satisfy the invariant property with respect to $(\diamond, p)$ (Definition A.20). Indeed, the former follows from Eq. (26), along with Example A.23 that shows that the posterior parameters, $\left\{(\hat{\mu}_\diamond^p(t-1), \mathrm{var}_\diamond^p(t-1)) : t \in [T]\right\}$, satisfy the invariant property with respect to $(\diamond, p)$; and the latter is from Example A.21.

Proof of Lemma C.26. We start by rewriting $(e1.1)$ as follows, where we drop $\mathcal{E}_t$.

$$\begin{aligned}(e1.1) \leq & \mathbb{E}\left[\sum_{t:p\in\mathcal{P}_t}\left(\frac{1-\psi^p_{i,t}}{\psi^p_{i,t}}\right)\mathbb{1}\left\{i^p_t=\diamond, H^p_\diamond(t), n^p_\diamond(t-1)<J\right\}\right]\\ =&\mathbb{E}\left[\sum_{t:p\in\mathcal{P}_t} g_t\mathbb{1}\left\{i^p_t=\diamond, n^p_\diamond(t-1)<J\right\}\right],\end{aligned}$$

where in the second line, we introduce the notation $g_t := \left(\frac{1-\psi^p_{i,t}}{\psi^p_{i,t}}\right)\mathbb{1}\left\{H^p_\diamond(t)\right\}$;

We now focus on the sum inside the expectation. Recall that $\pi_s(\diamond,p)$ is the round in which player p pulls arm $\diamond$ the s-th time. Here, we abuse the notation and denote $\pi_s(\diamond,p)$ by π_s. By Remark C.27, $\left\{g_t : t\in[T]\right\}$ satisfies the invariant property with respect to $(\diamond,p)$. Applying Lemma C.37 on $\left\{g_t : t\in[T]\right\}$'s, we have that the term inside the above expectation is at most:

$$\sum_{s=1}^{J-1}\left(\frac{1}{\psi^p_{i,\pi_s+1}}-1\right)\mathbb{1}\left\{\pi_s\leq T, H^p_\diamond(\pi_s+1)\right\},$$

where we also use the observation that $\left(\frac{1}{\psi^p_{i,1}}-1\right)\mathbb{1}\left\{H^p_\diamond(1)\right\}=0$.

Therefore, by the linearity of expectation, we have

$$(e1.1)\leq\sum_{s=1}^{J-1}\mathbb{E}\left[\left(\frac{1}{\psi^p_{i,\pi_s+1}}-1\right)\mathbb{1}\left\{\pi_s\leq T, H^p_\diamond(\pi_s+1)\right\}\right].$$

Therefore, the following Lemma C.28 suffices to prove Lemma C.26, which we prove next. □

Lemma C.28. *For any $s\in[T]$,*

$$\mathbb{E}\left[\left(\frac{1}{\psi^p_{i,\pi_s+1}}-1\right)\mathbb{1}\left\{\pi_s\leq T, H^p_\diamond(\pi_s+1)\right\}\right]\leq\mathcal{O}(1),$$

where we recall that $\pi_s=\pi_s(\diamond,p)$ is the round in which player p pulls arm $\diamond$ the s-th time.

Proof of Lemma C.28. We note that this proof is similar to that of Lemma C.16. We have

$$\begin{aligned}&\mathbb{E}\left[\left(\frac{1}{\psi^p_{i,\pi_s+1}}-1\right)\mathbb{1}\left\{\pi_s\leq T, H^p_\diamond(\pi_s+1)\right\}\right]\\ =&\mathbb{E}\left[\left(\frac{1}{\overline{\Phi}\left(\left(z^p_i-\text{ind-}\hat{\mu}^p_\diamond(\pi_s)\right)\sqrt{n^p_\diamond(\pi_s)/4}\right)}-1\right)\mathbb{1}\left\{\pi_s\leq T, H^p_\diamond(\pi_s+1)\right\}\right]\\ \leq&\mathbb{E}\left[\frac{1}{\overline{\Phi}\left(\left(\mu^p_\diamond-\text{ind-}\hat{\mu}^p_\diamond(\pi_s)\right)\sqrt{n^p_\diamond(\pi_s)/4}\right)}\mathbb{1}\left\{\pi_s\leq T\right\}\right],\end{aligned}$$

where the inequality drops $H^p_\diamond(\tau_s+1)$ and uses the observation that $z^p_i\leq\mu^p_\diamond$, and the monotonic increasing property of $z\mapsto\frac{1}{\overline{\Phi}(z)}$. Now, using Lemma C.36 and Corollary B.6, we conclude that this is at most $\mathcal{O}(1)$. □

Lemma C.29.

$$(e1.2) \leq \mathcal{O}(1).$$

Proof. Recall that

$$(e1.2) = \sum_{t:p\in\mathcal{P}_t} \mathbb{E}\left[\left(\frac{1-\psi^p_{i,t}}{\psi^p_{i,t}}\right) \mathbb{1}\left\{i^p_t = \diamond, \mathcal{E}_t, H^p_\diamond(t), n^p_\diamond(t-1) \geq J\right\}\right].$$

Dropping $\mathbb{1}\left\{i^p_t = i^p_\diamond\right\}$, we have

$$(e1.2) \leq \sum_{t:p\in\mathcal{P}_t} \mathbb{E}\left[\left(\frac{1-\psi^p_{i,t}}{\psi^p_{i,t}}\right) \mathbb{1}\left\{\mathcal{E}_t, H^p_\diamond(t), n^p_\diamond(t-1) \geq J\right\}\right],$$

When $\mathcal{E}_t$, $H^p_\diamond(t)$, and $\left\{n^p_\diamond(t-1) \geq J\right\}$ happen, we have

$$\begin{aligned}
&1-\psi^p_{i,t}\\
=&\Pr\left(\theta^p_\diamond(t) \leq z^p_i \mid \mathcal{F}_{t-1}\right)\\
=&\Phi\left(\left(z^p_i - \text{ind-}\hat{\mu}^p_\diamond(t-1)\right)\sqrt{\overline{n^p_\diamond}(t-1)/4}\right)\\
\leq&\exp\left(-\frac{\overline{n^p_\diamond}(t-1)\left(\text{ind-}\hat{\mu}^p_\diamond(t-1) - z^p_i\right)^2_+}{8}\right)\\
\leq&\exp\left(-\frac{n^p_\diamond(t-1)\left(\mu^p_\diamond - \frac{1}{4}\Delta^p_i - \mu^p_\diamond + \frac{1}{2}\Delta^p_i\right)^2_+}{8}\right)\\
\leq&\exp\left(-\frac{n^p_\diamond(t-1)(\Delta^p_i)^2}{8(16)}\right)\\
\leq&\frac{1}{T+1},
\end{aligned}$$

where the first inequality uses Lemma C.35; the second inequality uses the observations that, when $\mathcal{E}_t$ and $\left\{n^p_\diamond(t-1) \geq J\right\}$ happen:

1. $\overline{n^p_\diamond}(t-1) \geq n^p_\diamond(t-1) \geq J = \frac{640\ln T}{(\Delta^p_i)^2}$,

2. $\text{ind-}\hat{\mu}^p_\diamond(t-1) \geq \mu^p_\diamond - \sqrt{\frac{10\ln T}{\overline{n^p_\diamond}(t-1)}} \geq \mu^p_\diamond - \frac{1}{4}\Delta^p_i$ (see Definition B.9), and

3. $z^p_i = \mu^p_\diamond - \frac{1}{2}\Delta^p_i$;

the third inequality is by algebra; and the last inequality follows because when $\left\{n^p_\diamond(t-1) \geq J\right\}$ happens, $n^p_\diamond(t-1) \geq \frac{640\ln T}{(\Delta^p_i)^2} \geq \frac{320\ln(T+1)}{(\Delta^p_i)^2}$ for $T > 1$.

It follows that, when $\mathcal{E}_t$ and $\left\{n^p_\diamond(t-1) \geq J\right\}$ happen, $\psi^p_{i,t} \geq \frac{T}{T+1}$ and $\frac{1-\psi^p_{i,t}}{\psi^p_{i,t}} \leq \frac{1}{T}$. Hence, $(e1.2) \leq 1$. □

We now consider term $(e2)$. Recall that

$$(e2) = \sum_{t:p\in\mathcal{P}_t} \mathbb{E}\left[\left(\frac{1-\psi^p_{i,t}}{\psi^p_{i,t}}\right) \mathbb{1}\left\{i^p_t = \diamond, \mathcal{E}_t, \overline{H^p_\diamond(t)}\right\}\right]$$

Lemma C.30.

$$(e2) \leq \mathcal{O}\left(\frac{\ln T}{(\Delta_i^p)^2} + M\right).$$

Proof of Lemma C.30. With foresight, let $Z = \frac{640 \ln T}{(\Delta_i^p)^2} + M$. We have

$$(e2) = \underbrace{\sum_{t:p\in\mathcal{P}_t} \mathbb{E}\left[\left(\frac{1-\psi_{i,t}^p}{\psi_{i,t}^p}\right)\mathbb{1}\left\{i_t^p = \diamond, \mathcal{E}_t, \overline{H_\diamond^p(t)}, m_\diamond^p(t-1) < Z\right\}\right]}_{(e2.1)} + \underbrace{\sum_{t:p\in\mathcal{P}_t} \mathbb{E}\left[\left(\frac{1-\psi_{i,t}^p}{\psi_{i,t}^p}\right)\mathbb{1}\left\{i_t^p = \diamond, \mathcal{E}_t, \overline{H_\diamond^p(t)}, m_\diamond^p(t-1) \geq Z\right\}\right]}_{(e2.2)}.$$

The proof follows straightforwardly from Lemma C.31 and Lemma C.33 which we present subsequently. □

Lemma C.31.

$$(e2.1) \leq \mathcal{O}\left(\frac{\ln T}{(\Delta_i^p)^2} + M\right).$$

Proof of Lemma C.31. We have

$$\begin{aligned}
(e2.1) &\leq \mathbb{E}\left[\sum_{t:p\in\mathcal{P}_t}\left(\frac{1}{\psi_{i,t}^p} - 1\right)\mathbb{1}\left\{i_t^p = \diamond, \overline{H_\diamond^p(t)}, m_\diamond^p(t-1) < Z\right\}\right] \\
&\leq \mathbb{E}\left[\sum_{t:p\in\mathcal{P}_t}\frac{1}{\psi_{i,t}^p}\mathbb{1}\left\{i_t^p = \diamond, \overline{H_\diamond^p(t)}, m_\diamond^p(t-1) < Z\right\}\right],
\end{aligned}$$

where we drop $\mathcal{E}_t$ and use the observation that $\frac{1}{\psi_{i,t}^p} - 1 \leq \frac{1}{\psi_{i,t}^p}$.

We now focus on sum inside the expectation. We denote $\tau_k(\diamond)$ by τ_k and the player that makes the k's pull of $\diamond$ by $p_k := p_k(\diamond)$. Recall that we use $\overline{m_\diamond^p}(t-1)$ to denote $\big(m_\diamond^p(t-1) - M\big) \vee 1$. We have

$$\begin{aligned}
&\sum_{t:p\in\mathcal{P}_t}\frac{1}{\psi_{i,t}^p}\mathbb{1}\left\{i_t^p = \diamond, \overline{H_\diamond^p(t)}, m_\diamond^p(t-1) < Z\right\} \\
&= \sum_{t:p\in\mathcal{P}_t}\frac{1}{\overline{\Phi}\left(\left(z_i^p - \text{agg-}\hat{\mu}_\diamond^p(t-1)\right)\sqrt{\overline{m_\diamond^p}(t-1)/4}\right)}\mathbb{1}\left\{i_t^p = \diamond, \overline{H_\diamond^p(t)}, m_\diamond^p(t-1) < Z\right\} \\
&\leq \sum_{t:p\in\mathcal{P}_t}\frac{1}{\overline{\Phi}\left(\left(\mu_\diamond^p - \text{agg-}\hat{\mu}_\diamond^p(t-1)\right)\sqrt{\overline{m_\diamond^p}(t-1)/4}\right)}\mathbb{1}\left\{i_t^p = \diamond, m_\diamond^p(t-1) < Z\right\} \qquad (30) \\
&\leq \sum_{t\in[T]}\sum_{q\in\mathcal{P}_t}\frac{1}{\overline{\Phi}\left(\left(\mu_\diamond^q - \text{agg-}\hat{\mu}_\diamond^q(t-1)\right)\sqrt{\overline{m_\diamond^q}(t-1)/4}\right)}\mathbb{1}\left\{i_t^q = \diamond, m_\diamond^q(t-1) < Z\right\}, \qquad (31)
\end{aligned}$$

where the first equality uses Remark C.23; the first inequality drops $\overline{H_\diamond^p(t)}$ and uses the observation that $z_i^p \leq \mu_\diamond^p$ (see Definition C.19), along with the monotonic increasing property of $z \mapsto \frac{1}{\overline{\Phi}(z)}$.; the second inequality adds similar terms for other players $q \neq p$.

Now, define $\left\{f_t^q : t \in [T], q \in [M]\right\}$ where $f_t^q = \frac{1}{\overline{\Phi}\left(\left(\mu_\diamond^q - \text{agg-}\hat{\mu}_\diamond^q(t-1)\right)\sqrt{\overline{m_\diamond^q}(t-1)/4}\right)}$; recall from Example A.22 that $\left\{\text{agg-}\hat{\mu}_\diamond^q(t-1) : t \in [T]\right\}$ and $\left\{m_\diamond^q(t-1) : t \in [T]\right\}$ both satisfy the invariant property with respect to $(\diamond, q)$; therefore, $\left\{f_t^q : t \in [T], q \in [M]\right\}$ satisfies the invariant property with respect to $\diamond$. Applying Lemma C.38 to it, we have that

$$(31) \leq \sum_{q \in [M]} \frac{1}{\overline{\Phi}\left(0\right)} + \sum_{k=1}^{Z-1} \frac{1}{\overline{\Phi}\left(\left(\mu_\diamond^{p_k} - \text{agg-}\hat{\mu}_\diamond^{p_k}(\tau_k)\sqrt{\overline{m_\diamond^{p_k}}(\tau_k)/4}\right)\right)} \mathbb{1}\left\{\tau_k \leq T\right\}.$$

Since $\sum_{q \in [M]} \frac{1}{\overline{\Phi}(0)} \leq \mathcal{O}\left(M\right)$, it then suffices to show that for every $k \in \mathbb{N}$,

$$\mathbb{E}\left[\frac{1}{\overline{\Phi}\left(\left(\mu_\diamond^{p_k} - \text{agg-}\hat{\mu}_\diamond^{p_k}(\tau_k)\right)\sqrt{\overline{m_\diamond^{p_k}}(\tau_k)/4}\right)} \mathbb{1}\left\{\tau_k \leq T\right\}\right] \leq \mathcal{O}\left(1\right). \tag{32}$$

Note that $\overline{m_\diamond^{p_k}}(\tau_k) = \left(n_\diamond(\tau_k) - M\right) \vee 1$. Directly applying Corollary B.4 and Lemma C.36 with $X = (\text{agg-}\hat{\mu}_\diamond^{p_k}(\tau_k) - \mu_\diamond^{p_k})\sqrt{\overline{m_\diamond^{p_k}}(\tau_k)/4}$ and $E = \{\tau_k \leq T\}$ proves Eq. (32). □

Remark C.32. In the above proof, we relaxed Eq. (30) to Eq. (31) by adding the corresponding terms for all other players $q \neq p$. Alternatively, we could use the observation that $n_\diamond^p(t-1) \leq m_\diamond^p(t-1)$ to bound Eq. (30) by

$$\sum_{t : p \in \mathcal{P}_t} \frac{1}{\overline{\Phi}\left(\left(\mu_\diamond^p - \text{agg-}\hat{\mu}_\diamond^p(t-1)\right)\sqrt{\overline{m_\diamond^p}(t-1)/4}\right)} \mathbb{1}\left\{i_t^p = \diamond, n_\diamond^p(t-1) < Z\right\},$$

and apply Lemma C.37 and subsequently Lemma C.36. However, right now, we do not have tight-enough concentration inequalities for $\text{agg-}\hat{\mu}_\diamond^p(\pi_k(\diamond, p))$—the best known inequality here is Freedman's inequality, which incurs an undesirable extra $\mathcal{O}\left(\ln T\right)$ factor in the bound for $(e2.1)$.

Lemma C.33.

$$(e2.2) \leq \mathcal{O}\left(1\right).$$

Proof of Lemma C.33. Recall that

$$(e2.2) = \sum_{t : p \in \mathcal{P}_t} \mathbb{E}\left[\left(\frac{1 - \psi_{i,t}^p}{\psi_{i,t}^p}\right) \mathbb{1}\left\{i_t^p = \diamond, \mathcal{E}_t, \overline{H_\diamond^p(t)}, m_\diamond^p(t-1) \geq Z\right\}\right].$$

Recall that $\overline{m^p_\diamond}(t-1) = \big(m^p_\diamond(t-1) - M\big) \vee 1$. When $\mathcal{E}_t, \overline{H^p_\diamond(t)}$ and $\big\{m^p_\diamond(t-1) \geq Z\big\}$ happen simultaneously,

$$\begin{aligned}
&1 - \psi^p_{i,t}\\
=&\Pr\left(\theta^p_\diamond(t) \leq z^p_i \mid \mathcal{F}_{t-1}\right)\\
=&\Phi\left(\left(z^p_i - \text{agg-}\hat{\mu}^p_\diamond(t-1)\right)\sqrt{\overline{m^p_\diamond}(t-1)/4}\right)\\
\leq&\exp\left(-\frac{\overline{m^p_\diamond}(t-1)\left(\text{agg-}\hat{\mu}^p_\diamond(t-1) - z^p_i\right)^2_+}{8}\right)\\
\leq&\exp\left(-\frac{\overline{m^p_\diamond}(t-1)\left(\mu^p_\diamond - \frac{1}{4}\Delta^p_i - \mu^p_\diamond + \frac{1}{2}\Delta^p_i\right)^2_+}{8}\right)\\
\leq&\exp\left(-\frac{\overline{m^p_\diamond}(t-1)(\Delta^p_i)^2}{8(16)}\right)\\
\leq&\frac{1}{T+1},
\end{aligned}$$

where the first inequality uses Lemma C.35; the second inequality uses the observations that when $\mathcal{E}_t, \overline{H^p_\diamond(t)}$ and $\big\{m^p_\diamond(t-1) \geq Z\big\}$ happen:

1. $\overline{m^p_\diamond}(t-1) \geq m^p_\diamond(t-1) - M \geq Z - M \geq \frac{640\ln T}{(\Delta^p_i)^2}$,
2. $\text{agg-}\hat{\mu}^p_\diamond(t-1) \geq \mu^p_\diamond - \sqrt{\frac{10\ln T}{\overline{m^p_\diamond}(t-1)}} \geq \mu^p_\diamond - \frac{1}{4}\Delta^p_i$ (see Definition B.9), and
3. $z^p_i = \mu^p_\diamond - \frac{1}{2}\Delta^p_i$ (see Definition C.19);

the third inequality is by algebra; and the fourth inequality is by the fact that when $m^p_\diamond(t-1) \geq Z, \overline{m^p_\diamond}(t-1) \geq Z - M = \frac{640\ln T}{(\Delta^p_i)^2} \geq \frac{320\ln(T+1)}{(\Delta^p_i)^2}$ for $T > 1$.

It follows that, when $\mathcal{E}_t, \overline{H^p_\diamond(t)}$ and $\big\{m^p_\diamond(t-1) \geq Z\big\}$ happen, $\psi^p_{i,t} \geq \frac{T}{T+1}$ and $\frac{1-\psi^p_{i,t}}{\psi^p_{i,t}} \leq \frac{1}{T}$. As a result, $(e2.2) \leq \mathcal{O}(1)$. □

C.3. Concluding the proofs of Theorems 4.1 and 4.2

Lemma C.34. *Let a generalized ϵ-MPMAB problem instance and $\alpha > 0$ be such that for all $i \in \mathcal{I}_\alpha$ and all $p \in [M]$, $\Delta^p_i \leq 2\Delta^{\min}_i$. If algorithm $\mathcal{A}$ guarantees that when interacting with this problem instance:*

1. *For any arm $i \in \mathcal{I}_\alpha$,*
$$\mathbb{E}\left[n_i(T)\right] \leq \mathcal{O}\left(\frac{\ln T}{(\Delta^{\min}_i)^2} + M\right); \tag{33}$$
2. *For any arm $i \in \mathcal{I}^C_\alpha$ and player $p \in [M]$,*
$$\mathbb{E}\left[n^p_i(T)\right] \leq \mathcal{O}\left(\frac{\ln T}{(\Delta^p_i)^2} + C\right), \tag{34}$$

for some $C \geq 0$, then it has the following regret bounds simultaneously:

1. *gap-dependent regret bound:*
$$\text{Reg}(T) \leq \mathcal{O}\left(\frac{1}{M}\sum_{i\in\mathcal{I}_\alpha}\sum_{p\in[M]:\Delta^p_i>0}\frac{\ln T}{\Delta^p_i} + \sum_{i\in\mathcal{I}^C_\alpha}\sum_{p\in[M]:\Delta^p_i>0}\frac{\ln T}{\Delta^p_i} + MK(1+C)\right), \tag{35}$$

2. *gap-independent regret bound:*

$$\mathrm{Reg}(T) \leq \tilde{\mathcal{O}}\left(\sqrt{|\mathcal{I}_\alpha| P} + \sqrt{M\left(|\mathcal{I}_\alpha^C| - 1\right)P} + MK(1+C) \right), \tag{36}$$

where we recall that $P = \sum_{t=1}^T |\mathcal{P}_t|$.

Proof. We prove the two items respectively. Recall that $\Delta_i^{\min} = \min_{p\in[M]} \Delta_i^p$.

1. Note that for all $i \in \mathcal{I}_\alpha$ and all $p \in [M]$, $\Delta_i^p \leq 2\Delta_i^{\min}$, and $\sum_{p=1}^M \mathbb{E}\left[n_i^p(T)\right] = \mathbb{E}\left[n_i(T)\right]$; as a consequence,

$$\mathrm{Reg}(T) = \sum_{p=1}^{M}\sum_{i=1}^{K} \mathbb{E}\left[n_i^p(T)\right]\Delta_i^p = O\left(\sum_{i\in\mathcal{I}_\alpha} \mathbb{E}\left[n_i(T)\right]\Delta_i^{\min} + \sum_{i\in\mathcal{I}_\alpha^C}\sum_{p\in[M]:\Delta_i^p>0} \mathbb{E}\left[n_i^p(T)\right]\Delta_i^p \right). \tag{37}$$

Using Eq. (33), the first term can be bounded by:

$$\sum_{i\in\mathcal{I}_\alpha} \mathbb{E}\left[n_i(T)\right]\Delta_i^{\min} \leq \mathcal{O}\left(\sum_{i\in\mathcal{I}_\alpha} \frac{\ln T}{\Delta_i^{\min}} + MK \right) \leq \mathcal{O}\left(\frac{1}{M}\sum_{i\in\mathcal{I}_\alpha}\sum_{p\in[M]:\Delta_i^p>0} \frac{\ln T}{\Delta_i^p} + MK \right),$$

where the second inequality follows from the assumption that for all $i \in \mathcal{I}_\alpha$ and $p \in [M]$, $\Delta_i^p \leq 2\Delta_i^{\min}$.

Using Eq. (34), the second term can be bounded by:

$$\sum_{i\in\mathcal{I}_\alpha^C}\sum_{p\in[M]:\Delta_i^p>0} \mathbb{E}\left[n_i^p(T)\right]\Delta_i^p \leq \mathcal{O}\left(\sum_{i\in\mathcal{I}_\alpha^C}\sum_{p\in[M]:\Delta_i^p>0} \frac{\ln T}{\Delta_i^p} + MKC \right).$$

Combining the above two bounds yields Eq. (35).

2. As with the proof of Eq. (36), we continue from Eq. (37), but look at the two terms respectively. For the first term,

$$\begin{aligned}
\sum_{i\in\mathcal{I}_\alpha} \mathbb{E}\left[n_i(T)\right]\Delta_i^{\min} \leq& \mathcal{O}\left(\sum_{i\in\mathcal{I}_\alpha} \min\left(\mathbb{E}\left[n_i(T)\right], \frac{\ln T}{(\Delta_i^{\min})^2} + M \right)\Delta_i^{\min} \right) \\
\leq& \mathcal{O}\left(\sum_{i\in\mathcal{I}_\alpha} \min\left(\mathbb{E}\left[n_i(T)\right]\Delta_i^{\min}, \frac{\ln T}{\Delta_i^{\min}} \right) + MK \right) \\
\leq& \mathcal{O}\left(\sum_{i\in\mathcal{I}_\alpha} \sqrt{\mathbb{E}\left[n_i(T)\right]\ln T} + MK \right) \\
\leq& \mathcal{O}\left(\sqrt{|\mathcal{I}_\alpha| P \ln T} + MK \right)
\end{aligned} \tag{38}$$

where the first inequality is from Eq. (33); the second inequality is by algebra; the third inequality is from the elementary fact that $\min(A, B) \leq \sqrt{AB}$; the last inequality is from Jensen's inequality and the concavity of function $x \mapsto \sqrt{x}$, which implies that $\sum_{i\in\mathcal{I}_\alpha} \sqrt{\mathbb{E}\left[n_i(T)\right]} \leq \sqrt{|\mathcal{I}_\alpha|\left(\sum_{i\in\mathcal{I}_\alpha}\mathbb{E}\left[n_i(T)\right]\right)}$, and the fact that $\sum_{i\in\mathcal{I}_\alpha}\mathbb{E}\left[n_i(T)\right] \leq \sum_{i=1}^M \mathbb{E}\left[n_i(T)\right] \leq P$.

For the second term in Eq. (36), first observe that if $\left|\mathcal{I}_\alpha^C\right| = 1$, then let i^* be the only element in $\mathcal{I}_\alpha^C$; it must be the case that for all $p \in [M]$, i^* is the optimal arm for player p. As a consequence, $\sum_{i\in\mathcal{I}_\alpha^C}\sum_{p=1}^M \mathbb{E}\left[n_i^p(T)\right]\Delta_i^p = 0 = \mathcal{O}(\sqrt{M(\left|\mathcal{I}_\alpha^C\right| - 1)P})$.

Otherwise, $\left|\mathcal{I}_\alpha^C\right| \geq 2$. In this case,

$$\begin{aligned}
\sum_{p\in[M]}\sum_{i\in\mathcal{I}_\alpha^C}\mathbb{E}\left[n_i^p(T)\right]\Delta_i^p \leq & \mathcal{O}\left(\sum_{p\in[M]}\sum_{i\in\mathcal{I}_\alpha^C}\min\left(\mathbb{E}\left[n_i^p(T)\right], \frac{\ln T}{(\Delta_i^p)^2}\right)\Delta_i^p + MKC\right) \\
\leq & \mathcal{O}\left(\sum_{p\in[M]}\sum_{i\in\mathcal{I}_\alpha^C}\min\left(\mathbb{E}\left[n_i^p(T)\right]\Delta_i^p, \frac{\ln T}{\Delta_i^p}\right) + MKC\right) \\
\leq & \mathcal{O}\left(\sum_{p\in[M]}\sum_{i\in\mathcal{I}_\alpha^C}\sqrt{\mathbb{E}\left[n_i^p(T)\right]\ln T} + MKC\right) \\
\leq & \mathcal{O}\left(\sqrt{M\left|\mathcal{I}_\alpha^C\right|P\ln T} + MKC\right) \\
\leq & \mathcal{O}\left(\sqrt{M\left(\left|\mathcal{I}_\alpha^C\right|-1\right)P\ln T} + MKC\right).
\end{aligned}$$

where the first inequality is by Eq. (34) and algebra; the second inequality is by algebra; the third inequality is from the elementary fact that $\min(A,B) \leq \sqrt{AB}$; the fourth inequality is from Jensen's inequality and the concavity of function $x \mapsto \sqrt{x}$, which implies that $\sum_{i\in\mathcal{I}_\alpha}\sqrt{\mathbb{E}\left[n_i(T)\right]} \leq \sqrt{|\mathcal{I}_\alpha|\left(\sum_{i\in\mathcal{I}_\alpha}\mathbb{E}\left[n_i(T)\right]\right)}$, and the fact that $\sum_{i\in\mathcal{I}_\alpha}\mathbb{E}\left[n_i(T)\right] \leq \sum_{i=1}^M\mathbb{E}\left[n_i(T)\right] \leq P$; the last inequality is from the simple observation that $\left|\mathcal{I}_\alpha^C\right| \leq 2(\left|\mathcal{I}_\alpha^C\right|-1)$ when $\left|\mathcal{I}_\alpha^C\right| \geq 2$.

In summary, $\sum_{p=1}^M\sum_{i\in\mathcal{I}_\alpha^C}\mathbb{E}\left[n_i^p(T)\right]\Delta_i^p \leq \mathcal{O}\left(\sqrt{M\left(\left|\mathcal{I}_\alpha^C\right|-1\right)P\ln T}\right) + MKC$. Combining this with Eq. (38), this concludes the proof of Eq. (36). □

Proofs of Theorems 4.1 and 4.2. Combining Lemmas C.1, C.2, C.34 with $C = M$ and $\alpha = 10\epsilon$, Theorems 4.1 and 4.2 follow immediately. □

C.4. Auxiliary Lemmas

Recall that we denote by $\overline{\Phi}(x) = \int_x^\infty \frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}dz$ the complementary CDF of the standard normal distribution.

Lemma C.35. *$\overline{\Phi}$ is monotonically decreasing. In addition, for $z \geq 0$,*

$$\frac{1}{\sqrt{2\pi}}\frac{z}{z^2+1}\exp\left(-\frac{z^2}{2}\right) \leq \overline{\Phi}(z) \leq \exp\left(-\frac{z^2}{2}\right),$$

where the first inequality (anti-concentration) is from (Gordon, 1941). In addition, for any $z \in \mathbb{R}$,

$$\overline{\Phi}(z) \leq \exp\left(-\frac{(z)_+^2}{2}\right), \quad \Phi(z) \leq \exp\left(-\frac{(-z)_+^2}{2}\right),$$

where we recall that $(z)_+ = \max(z,0)$.

The following lemma is useful in bounding $(b2.1)$, $(e1.1)$, $(e2.1)$; it can also be used to provide a simplified proof of the first case of Agrawal & Goyal (2017, Lemma 2.13). Roughly speaking, the lemma shows that a random variable X with a light lower probability tail must have a small value of $\mathbb{E}\left[\frac{1}{\overline{\Phi}(-X)}\right]$; it crucially uses the lower bound on $\overline{\Phi}$ (Gaussian anti-concentration) given in Lemma C.35.

Lemma C.36. *There exists some absolute constants $c_1, c_2 > 0$ such that the following holds. Given a random variable X, an event E and some $C_1 > 0$; if, for every $z \geq 1$, $\mathbb{P}(X \leq -z, E) \leq C_1 \exp(-2z^2)$, such that*

$$\mathbb{E}\left[\frac{1}{\overline{\Phi}(-X)} \mathbb{1}\left\{E\right\}\right] \leq c_1 C_1 + c_2.$$

Proof. Define $Y = -X$; we have $\mathbb{P}(Y \geq z, E) \leq C_1 \exp(-2z^2)$ for all $z \geq 1$.

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{\overline{\Phi}(-X)} \mathbb{1}\left\{E\right\}\right] \\
=&\mathbb{E}\left[\frac{1}{\overline{\Phi}(-X)} \mathbb{1}\left\{E, X \leq -1\right\}\right] + \mathbb{E}\left[\frac{1}{\overline{\Phi}(-X)} \mathbb{1}\left\{E, X \geq -1\right\}\right] \\
\leq&\mathbb{E}\left[\frac{1}{\overline{\Phi}(Y)} \mathbb{1}\left\{E, Y \geq 1\right\}\right] + \frac{1}{\overline{\Phi}(1)} \\
\leq&8\sqrt{2\pi} \cdot \mathbb{E}\left[e^{Y^2} \mathbb{1}\left\{E, Y \geq 1\right\}\right] + \frac{1}{\overline{\Phi}(1)}.
\end{aligned}$$

where the first inequality follows due to the fact that $\frac{1}{\overline{\Phi}(z)}$ increases monotonically as z increases; and the second inequality is based on the observation that for $y \geq 1$, $\frac{1}{\overline{\Phi}(y)} \leq \sqrt{2\pi} \frac{y^2+1}{y} \exp(\frac{y^2}{2}) \leq 8\sqrt{2\pi} e^{y^2}$ (see Lemma C.35).

It suffices to show that $\mathbb{E}\left[e^{Y^2} \mathbb{1}\left\{E, Y \geq 1\right\}\right]$ is bounded by some constant, given the assumption on Y. Define $W = e^{Y^2} \mathbb{1}\left\{E, Y \geq 1\right\}$. We have that for any $w \geq e$,

$$\mathbb{P}(W \geq w) = \mathbb{P}(E, Y \geq \sqrt{\ln w}) \leq \frac{C_1}{w^2}.$$

As a result,

$$\begin{aligned}
\mathbb{E}\left[W\right] =& \int_0^\infty \mathbb{P}(W \geq w) \,\mathrm{d}w \\
=& \int_0^e \mathbb{P}(W \geq w) \,\mathrm{d}w + \int_e^\infty \mathbb{P}(W \geq w) \,\mathrm{d}w \\
\leq& e + \int_e^\infty \frac{C_1}{w^2} \,\mathrm{d}w \\
\leq& e + \frac{C_1}{e},
\end{aligned}$$

Therefore, the lemma holds by taking $c_1 = \frac{8\sqrt{2\pi}}{e}$ and $c_2 = 8\sqrt{2\pi}e + \frac{1}{\overline{\Phi}(1)}$. □

The following two lemmas are useful in bounding $(e1.1)$ (Lemma C.37), as well as $(b2.1)$ and $(e2.1)$ (Lemma C.38), respectively.

Lemma C.37. *Fix any arm $i \in [K]$ and player $p \in [M]$. Let $N \in \mathbb{N}^+$. Suppose $\left\{g_t : t \in [T]\right\}$ satisfies the invariant property with respect to (i, p) (Definition A.20). Then,*

$$\sum_{t: p \in \mathcal{P}_t} g_t \mathbb{1}\left\{i_t^p = i, n_i^p(t-1) < N\right\} \leq g_1 + \sum_{k=1}^{N-1} g_{\pi_k + 1} \mathbb{1}\left\{\pi_k \leq T\right\},$$

where $\pi_k = \pi_k(i, p)$ denotes the round associated with the k-th pull of arm i by player p.

Proof. Let $h_t = g_t \mathbb{1}\left\{n_i^p(t-1) < N\right\}$. As seen in Example A.22, $\left\{n_i^p(t-1) : t \in [T]\right\}$ satisfies the invariant property with respect to (i,p). This, combined with the assumption that $\left\{g_t : t \in [T]\right\}$ satisfies the invariant property with respect to (i,p), implies that $\left\{h_t : t \in [T]\right\}$ is also invariant with respect to (i,p). Applying Lemma C.39 to the above $\left\{h_t : t \in [T]\right\}$, we have

$$\begin{aligned}
\sum_{t:p\in\mathcal{P}_t} g_t \mathbb{1}\left\{i_t^p = i, n_i^p(t-1) < N\right\} = \sum_{t:p\in\mathcal{P}_t} h_t \mathbb{1}\left\{i_t^p = i\right\} \\
&\le h_1 + \sum_{k=1}^{T} h_{\pi_k+1} \mathbb{1}\left\{\pi_k \le T\right\} \\
&= g_1 \mathbb{1}\left\{n_i^p(0) < N\right\} + \sum_{k=1}^{T} g_{\pi_k+1} \mathbb{1}\left\{n_i^p(\pi_k) < N\right\} \mathbb{1}\left\{\pi_k \le T\right\} \\
&= g_1 + \sum_{k=1}^{T} g_{\pi_k+1} \mathbb{1}\left\{k < N\right\} \mathbb{1}\left\{\pi_k \le T\right\} \\
&= g_1 + \sum_{k=1}^{N-1} g_{\pi_k+1} \mathbb{1}\left\{\pi_k \le T\right\},
\end{aligned}$$

where the first inequality is by Equation (40) in Lemma C.39; the second equality is by expanding the definition of h_t's; the third equality is from that $n_i^p(0) = 0$ and $n_i^p(\pi_k) = k$; and the last eqaulity is by algebra. □

Lemma C.38. *Fix any arm $i \in [K]$ and let $N \in \mathbb{N}^+$. Suppose $\left\{f_t^p : t \in [T], p \in [M]\right\}$ satisfies the invariant property with respect to arm i (Definition A.20), then,*

$$\sum_{t\in[T]} \sum_{p\in\mathcal{P}_t} f_t^p \mathbb{1}\left\{i_t^p = i, m_i^p(t-1) < N\right\} \le \sum_{p\in[M]} f_1^p + \sum_{k=1}^{N-1} f_{\tau_k+1}^{p_k} \mathbb{1}\left\{\tau_k \le T\right\},$$

where $(\tau_k, p_k) = (\tau_k(i), p_k(i))$ denote the round and player associated with the k-th pull of arm i by all players.

Proof of Lemma C.38. First, consider any fixed player $p \in [M]$; let $h_t = f_t^p \mathbb{1}\left\{m_i^p(t-1) < N\right\}$. As seen in Example A.22, $\left\{m_i^p(t-1) : t \in [T]\right\}$ satisfies the invariant property with respect to (i,p). This, combined with the assumption that $\left\{f_t^p : t \in [T]\right\}$ satisfies the invariant property with respect to (i,p), implies that $\left\{h_t : t \in [T]\right\}$ is also invariant with respect to (i,p). Applying Lemma C.39 to the above $\left\{h_t : t \in [T]\right\}$, we have

$$\begin{aligned}
\sum_{t:p\in\mathcal{P}_t} f_t^p \mathbb{1}\left\{i_t^p = i, m_i^p(t-1) < N\right\} = \sum_{t:p\in\mathcal{P}_t} h_t \mathbb{1}\left\{i_t^p = i\right\} \\
&\le h_1 + \sum_{t:p\in\mathcal{P}_t} h_{t+1} \mathbb{1}\left\{i_t^p = i\right\} \\
&= f_1^p + \sum_{t:p\in\mathcal{P}_t} f_{t+1}^p \mathbb{1}\left\{i_t^p = i, m_i^p(t) < N\right\} \\
&= f_1^p + \sum_{t:p\in\mathcal{P}_t} f_{t+1}^p \mathbb{1}\left\{i_t^p = i, n_i(t) < N\right\} \qquad (39)
\end{aligned}$$

where the first inequality is from Equation (41) of Lemma C.39; the second equality is by expanding the definition of h_t and noting that $h_1 = \mathbb{1}\left\{m_i^p(0) < N\right\} f_1^p = \mathbb{1}\left\{0 < N\right\} f_1^p = f_1^p$; the third equality is from the observation that, if $i_t^p = i$ and $u_i^p(t) = t$, then $m_i^p(t) = n_i(u_i^p(t)) = n_i(t)$.

Now, summing Equation (39) over all players $p \in [M]$, we have

$$\begin{aligned}
&\sum_{t\in[T]}\sum_{p\in\mathcal{P}_t} f_t^p \mathbb{1}\left\{i_t^p = i, m_i^p(t-1) < N\right\} \\
\le &\sum_{p\in[M]} f_1^p + \sum_{p\in[M]}\sum_{t:p\in\mathcal{P}_t} f_{t+1}^p \mathbb{1}\left\{i_t^p = i, n_i(t) < N\right\} \\
\le &\sum_{p\in[M]} f_1^p + \sum_{k=1}^{N-1} f_{\tau_k+1}^{p_k} \mathbb{1}\left\{\tau_k \le T\right\},
\end{aligned}$$

where the second inequality is from the observation that for every $t \in [T]$, $p \in \mathcal{P}_t$ such that $i_t^p = i$ and $n_i(t) < N$, there must exists some unique $k \in [N-1]$ such that $\tau_k = t$ and $p_k = p$. □

The following auxiliary lemma facilitates the proofs of Lemmas C.37 and C.38.

Lemma C.39. *Fix any arm $i \in [K]$ and player $p \in [M]$. Suppose $\left\{h_t : t \in [T]\right\}$ satisfies the invariant property with respect to (i, p) (Definition A.20). Then,*

$$\begin{aligned}
\sum_{t\in[T]:p\in\mathcal{P}_t} h_t \mathbb{1}\left\{i_t^p = i\right\} \le & h_1 + \sum_{k=1}^{T} h_{\pi_k+1} \mathbb{1}\left\{\pi_k \le T\right\} && (40) \\
= & h_1 + \sum_{t\in[T]:p\in\mathcal{P}_t} h_{t+1} \mathbb{1}\left\{i_t^p = i\right\}, && (41)
\end{aligned}$$

where $\pi_k = \pi_k(i, p)$ denotes the round associated with the k-th pull of arm i by player p.

Proof.

$$\begin{aligned}
\sum_{t\in[T]:p\in\mathcal{P}_t} h_t \mathbb{1}\left\{i_t^p = i\right\} = & \sum_{k=1}^{T} h_{\pi_k} \mathbb{1}\left\{\pi_k \le T\right\} \\
= & \sum_{k=1}^{T} h_{\pi_{k-1}+1} \mathbb{1}\left\{\pi_k \le T\right\} \\
\le & h_1 + \sum_{k=2}^{T} h_{\pi_{k-1}+1} \mathbb{1}\left\{\pi_k \le T\right\} \\
= & h_1 + \sum_{k=1}^{T-1} h_{\pi_k+1} \mathbb{1}\left\{\pi_{k+1} \le T\right\} \\
\le & h_1 + \sum_{k=1}^{T-1} h_{\pi_k+1} \mathbb{1}\left\{\pi_k \le T\right\} \\
= & h_1 + \sum_{t\in[T]:p\in\mathcal{P}_t} h_{t+1} \mathbb{1}\left\{i_t^p = i\right\},
\end{aligned}$$

where the first equality uses the definition of π_k; the second equality uses the invariant property, specifically, $h_{\pi_k} = h_{\pi_{k-1}+1}$; the first inequality uses the observation that the first term $h_{\pi_0+1}\mathbb{1}\left\{\pi_1 \le T\right\} = h_1 \mathbb{1}\left\{\pi_1 \le T\right\} \le h_1$; the third equality shifts the indices in the sum by 1; the second inequality uses the observation that $\pi_{k+1} \le T \implies \pi_k \le T$; and the last equality is again by the definition of π_k. □

The following lemma is largely inspired by Agrawal & Goyal (2017, Lemma 2.8); here we generalize it to the multi-task setting, for reducing bounding (B) and (E) to bounding $(B*)$ and $(E*)$ respectively.

Lemma C.40. *For any player* $p \in [M]$*, time step* $t \in [T]$*, and arm* $i \in [K]$*, we have for any arm* $l \in [K]$ *and any threshold* $z \in \mathbb{R}$*:*

$$\Pr\left(i_t^p = i, \theta_i^p(t) \leq z \mid \mathcal{F}_{t-1}\right) \leq \frac{\Pr\left(\theta_l^p(t) \leq z \mid \mathcal{F}_{t-1}\right)}{\Pr\left(\theta_l^p(t) > z \mid \mathcal{F}_{t-1}\right)} \cdot \Pr\left(i_t^p = l \mid \mathcal{F}_{t-1}\right).$$

Proof. First,

$$\begin{aligned}
&\Pr\left(i_t^p = i, \overline{Q_i^p(t)} \mid \mathcal{F}_{t-1}\right)\\
\leq &\Pr\left(\forall j \in [K],\ \theta_j^p(t) \leq z \mid \mathcal{F}_{t-1}\right)\\
= &\Pr\left(\theta_l^p(t) \leq z \mid \mathcal{F}_{t-1}\right) \cdot \Pr\left(\forall j \neq l,\ \theta_j^p(t) \leq z \mid \mathcal{F}_{t-1}\right),
\end{aligned}$$

where the first inequality follows because the event $\left\{i_t^p = i, \overline{Q_i^p(t)}\right\}$ happens only if $\forall j \in [K],\ \theta_j^p(t) \leq z$; and the second equality follows because conditional on $\mathcal{F}_{t-1}$, the draws $\theta_j^p(t)$'s and $\theta_l^p(t)$ are independent.

Now, observe that

$$\begin{aligned}
&\Pr\left(\forall j \neq l,\ \theta_j^p(t) \leq z \mid \mathcal{F}_{t-1}\right)\\
= &\frac{\Pr\left(\theta_l^p(t) > z, \text{ and } \forall j \neq l,\ \theta_j^p(t) \leq z \mid \mathcal{F}_{t-1}\right)}{\Pr\left(\theta_l^p(t) > z \mid \mathcal{F}_{t-1}\right)}\\
\leq &\frac{\Pr\left(i_t^p = l \mid \mathcal{F}_{t-1}\right)}{\Pr\left(\theta_l^p(t) > z \mid \mathcal{F}_{t-1}\right)}
\end{aligned}$$

where the equality follows, again, by the conditional independence of $\left\{\theta_j^p(t) : j \neq l\right\}$ and $\theta_l^p(t)$; and the inequality follows because the event $\left\{\theta_l^p(t) > z,\ \forall j \neq l,\ \theta_j^p(t) \leq y_i^p\right\}$ implies that $\left\{i_t^p = l\right\}$ happens. The lemma follows from combining the above two inequalities. □

D. Theoretical Guarantees of Baselines

D.1. IND-UCB and IND-TS in the generalized ϵ-MPMAB setting

Theorem D.1. *The expected collective regret of* IND-UCB *and* IND-TS *after T rounds satisfies the following two upper bounds simultaneously:*

$$\begin{aligned} \mathrm{Reg}(T) \leq & \mathcal{O}\left(\sum_{p \in [M]} \sum_{i \in [K] : \Delta_i^p > 0} \frac{\ln T}{\Delta_i^p} \right) && (42) \\ \mathrm{Reg}(T) \leq & \tilde{\mathcal{O}}\left(\sqrt{MKP} \right), && (43) \end{aligned}$$

where we recall that $P = \sum_{t=1}^{T} |\mathcal{P}_t|$.

Proof sketch. For Eq. (42), we note that both IND-UCB and IND-TS guarantees that for every $p \in [M]$,

$$\mathrm{Reg}^p(T) \leq \mathcal{O}\left(\sum_{i \in [K] : \Delta_i^p > 0} \frac{\ln T}{\Delta_i^p} \right);$$

summing over p yields Eq. (42).

For Eq. (43), we note that for every $p \in [M]$,

$$\mathrm{Reg}^p(T) \leq \tilde{\mathcal{O}}\left(\sqrt{K \left| \{ t : p \in \mathcal{P}_t \} \right|} \right).$$

Summing over all $p \in [M]$, we have

$$\mathrm{Reg}(T) = \sum_{p=1}^{M} \mathrm{Reg}^p(T) \leq \tilde{\mathcal{O}}\left(\sum_{p=1}^{M} \sqrt{K \left| \{ t : p \in \mathcal{P}_t \} \right|} \right) \leq \tilde{\mathcal{O}}\left(\sqrt{MK \sum_{p=1}^{M} \left| \{ t : p \in \mathcal{P}_t \} \right|} \right) = \tilde{\mathcal{O}}\left(\sqrt{MKP} \right). \quad \square$$

D.2. ROBUSTAGG(ϵ) and its regret analysis in the generalized ϵ-MPMAB setting

Wang et al. (2021) study a special case of ϵ-MPMAB problem, which can be viewed as ϵ-MPMAB problem defined in Section 2, with active sets of players $\mathcal{P}_t \equiv [M]$. In this specialized setting, they propose ROBUSTAGG(ϵ), a UCB-based algorithm that achieves a gap-dependent and gap-independent regret of

$$\mathcal{O}\left(\frac{1}{M} \sum_{i \in \mathcal{I}_{5\epsilon}} \sum_{p \in [M] : \Delta_i^p > 0} \frac{\ln T}{\Delta_i^p} + \sum_{i \in \mathcal{I}_{5\epsilon}^C} \sum_{p \in [M] : \Delta_i^p > 0} \frac{\ln T}{\Delta_i^p} + MK \right), \tag{44}$$

and

$$\tilde{\mathcal{O}}\left(\sqrt{M |\mathcal{I}_{5\epsilon}| T} + M \sqrt{|\mathcal{I}_{5\epsilon}^C| T} + MK \right), \tag{45}$$

respectively. In this section, we show that, with a few small modifications, their algorithm and analysis can be used in our (more general) ϵ-MPMAB setting, where the active sets $\mathcal{P}_t$ can change over time.

Algorithm 2 ROBUSTAGG(ϵ) for the generalized ϵ-MPMAB setting

```
 1: Input: Dissimilarity parameter $\epsilon \in [0,1]$
 2: Initialization: Set $n_i^p = 0$ for all $p \in [M]$ and all $i \in [K]$.
 3: for $t = 1, 2 \dots, T$ do
 4:     Receive active set of players $\mathcal{P}_t$
 5:     for $p \in \mathcal{P}_t$ do
 6:         for $i \in [K]$ do
 7:             Let $m_i^p = \sum_{q \in [M] : q \neq p} n_i^q$
 8:             Let $\overline{n_i^p} = n_i^p \vee 1$ and $\overline{m_i^p} = m_i^p \vee 1$
 9:             Let
                $\zeta_i^p(t) = \frac{1}{\overline{n_i^p}} \sum_{\substack{s<t: \\ p \in \mathcal{P}_s, i_s^p = i}} r_s^p,\ \eta_i^p(t) = \frac{1}{\overline{m_i^p}} \sum_{s<t} \sum_{\substack{q \in \mathcal{P}_s: \\ q \neq p, i_s^q = i}} r_s^q, \text{and } \kappa_i^p(t,\lambda) = \lambda \zeta_i^p(t) + (1-\lambda)\eta_i^p(t);$
10:             Let $F(\overline{n_i^p}, \overline{m_i^p}, \lambda, \epsilon) = 8\sqrt{13 \ln T \left[\frac{\lambda^2}{\overline{n_i^p}} + \frac{(1-\lambda)^2}{\overline{m_i^p}}\right]} + (1-\lambda)\epsilon$
11:             Compute $\lambda^* = \mathrm{argmin}_{\lambda \in [0,1]} F(\overline{n_i^p}, \overline{m_i^p}, \lambda, \epsilon)$
12:             Compute an upper confidence bound of the reward of arm $i$ for player $p$:
                $\mathrm{UCB}_i^p(t) = \kappa_i^p(t, \lambda^*) + F(\overline{n_i^p}, \overline{m_i^p}, \lambda^*, \epsilon).$
13:         end for
14:         Let $i_t^p = \mathrm{argmax}_{i \in [K]} \mathrm{UCB}_i^p(t)$
15:         Player $p$ pulls arm $i_t^p$ and observes reward $r_t^p$
16:     end for
17:     for active players $p \in \mathcal{P}_t$ do
18:         Let $i = i_t^p$ and set $n_i^p \leftarrow n_i^p + 1$.
19:     end for
20: end for
```

Specifically, Algorithm 2 is our modified version of ROBUSTAGG(ϵ). Recall that ROBUSTAGG(ϵ) performs an UCB-based exploration (Auer et al., 2002): for every player and every arm, it constructs high-probability UCBs on the expected rewards (line 7 to 12); to this end, it makes careful use of both the player and other players' data, and construct a series of UCBs parameterized by λ (line 10), and selects the tightest one (line 11 and 12). Compared to ROBUSTAGG(ϵ), for every round t, Algorithm 2 only computes expected reward UCBs for active players $p \in \mathcal{P}_t$ (line 5), and updates arm pull counts on active players (line 17).

We show that Algorithm 2, when applied to our ϵ-MPMAB setting, has regret guarantees that recover and generalize ROBUSTAGG(ϵ)'s original guarantees. Specifically, in the specialized ϵ-MPMAB setting where $\mathcal{P}_t \equiv [M]$, we recover the regret guarantees of ROBUSTAGG(ϵ) (Equations (44) and (45)).

Theorem D.2. *The expected collective regret of* ROBUSTAGG(ϵ) *after T rounds satisfies the following two upper bounds simultaneously:*

$$\mathrm{Reg}(T) \leq \mathcal{O}\left(\frac{1}{M}\sum_{i \in \mathcal{I}_{5\epsilon}} \sum_{p \in [M] : \Delta_i^p > 0} \frac{\ln T}{\Delta_i^p} + \sum_{i \in \mathcal{I}_{5\epsilon}^C} \sum_{p \in [M] : \Delta_i^p > 0} \frac{\ln T}{\Delta_i^p} + MK\right), \tag{46}$$

$$\mathrm{Reg}(T) \leq \tilde{\mathcal{O}}\left(\sqrt{|\mathcal{I}_{5\epsilon}| P} + \sqrt{M\left(|\mathcal{I}_{5\epsilon}^C| - 1\right) P} + MK\right), \tag{47}$$

where we recall that $P = \sum_{t=1}^T |\mathcal{P}_t|$.

Proof sketch. Even in the general setting where $\mathcal{P}_t$ is not necessarily $[M]$, Freedman's inquality can still be applied to establish the high-probability concentration of the empirically averaged rewards $\zeta_i^p(t)$ and $\eta_i^p(t)$; therefore, Lemma 17 of Wang et al. (2021) still holds in the general setting. As a result, Lemmas 20 and 21 of Wang et al. (2021) carries over; hence, for all $i \in \mathcal{I}_{5\epsilon}$, Algorithm 2 still satisfies that

$$\mathbb{E}[n_i(T)] \leq \mathcal{O}\left(\frac{\ln T}{(\Delta_i^{\min})^2} + M\right), \tag{48}$$

and for all $i \in \mathcal{I}_{5\epsilon}^C$ and all $p \in [M]$,

$$\mathbb{E}[n_i^p(T)] \leq \mathcal{O}\left(\frac{\ln T}{(\Delta_i^p)^2}\right). \tag{49}$$

Equations (46) and (47) now follows directly from applying Lemma C.34 with $C = 0$ and $\alpha = 5\epsilon$. □

E. Additional Experimental Results

In this section, we present the rest of the experimental results. Figures 3, 4, and 5 compare the average performance of ROBUSTAGG-TS(0.15), ROBUSTAGG(0.15), IND-UCB, and IND-TS in randomly generated 0.15-MPMAB problem instances with different numbers of subpar arms.

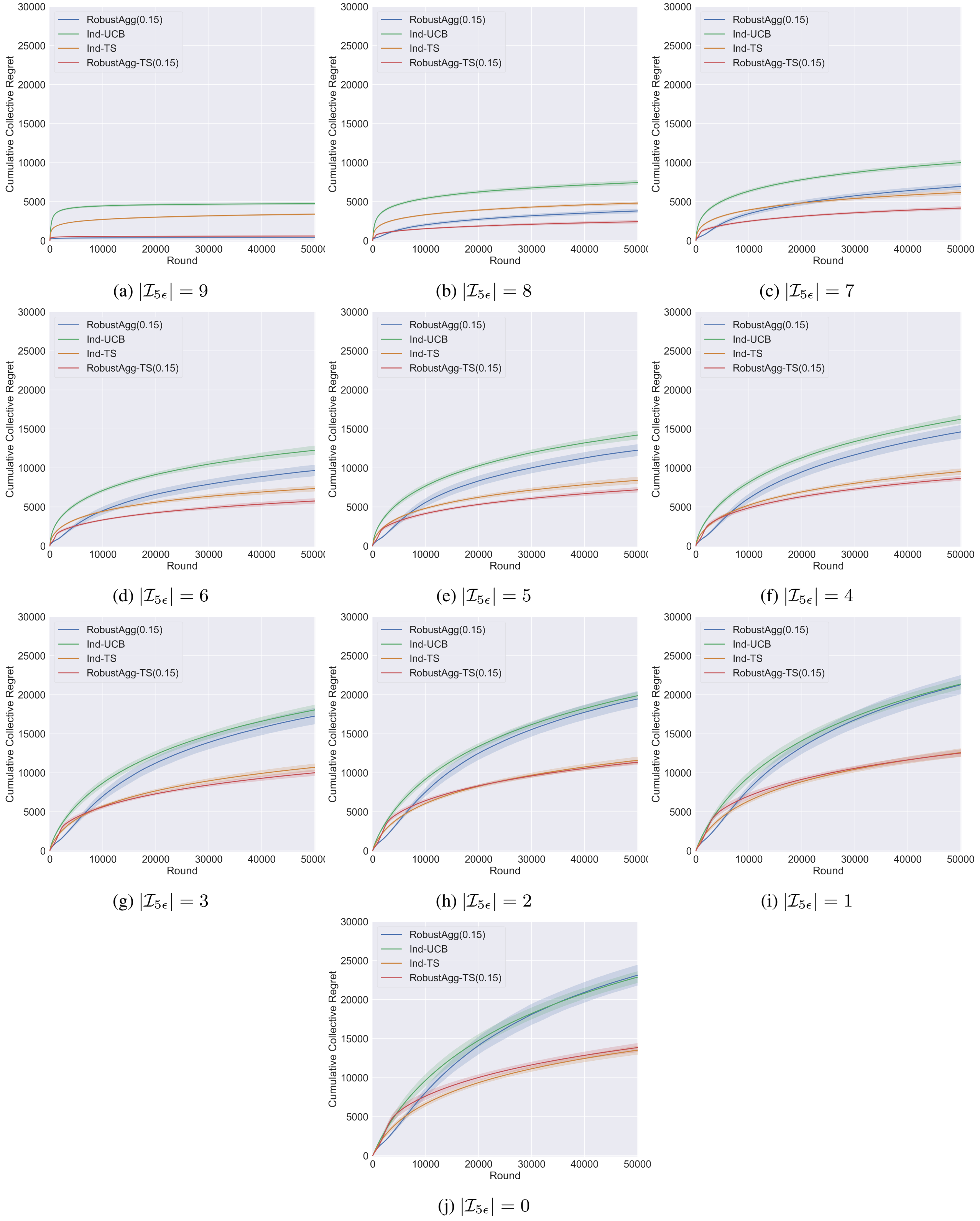
Note that, when $|\mathcal{I}_{5\epsilon}| = 9$, we have $\left|\mathcal{I}_{5\epsilon}^C\right| = 1$ which means that there exists one arm that is optimal to all the players and the other arms are all subpar. In this favorable special case, ROBUSTAGG-TS(0.15) and ROBUSTAGG(0.15) perform significantly better than the baseline algorithms without transfer, as expected.

Furthermore, when $|\mathcal{I}_{5\epsilon}| = 0$, i.e., there is no subpar arm and all the arms have relatively small suboptimality gaps. In this unfavorable special case, ROBUSTAGG-TS(0.15)'s performance is still very competitive in comparison with IND-TS, which demonstrates the robustness of our proposed algorithm.

E.1. Empirical Comparison with ROBUSTAGG-TS-V(ϵ)

We empirically evaluated a variant of Algorithm 1, which we refer to as ROBUSTAGG-TS-V(ϵ). ROBUSTAGG-TS-V(ϵ) differs from ROBUSTAGG-TS(ϵ) (Algorithm 1) in one way: in each round, instead of only updating the posteriors associated with each active player and its pulled arm (i.e., delayed update, line 20 of Algorithm 1), ROBUSTAGG-TS-V(ϵ) updates the posteriors associated with every arm and player. Note that this change only affects the aggregate posteriors, as the individual posteriors associated with a player and an arm remains the same if the player does not pull the arm in this round.

Figure 6 compares the average cumulative regret of ROBUSTAGG-TS(0.15), ROBUSTAGG-TS-V(0.15), ROBUSTAGG(0.15), IND-UCB, and IND-TS in randomly generated 0.15-MPMAB problem instances with different numbers of subpar arms. The instances were generated following the same procedure as the other experiments. Observe that ROBUSTAGG-TS-V(0.15)'s empirical performance is on par with that of ROBUSTAGG-TS(0.15). However, our analysis in this paper takes advantages of the design choice made for ROBUSTAGG-TS(ϵ), i.e., delayed update which leads to the invariant property (Definition A.20 and Examples A.21, A.22 and A.23). It is unclear whether ROBUSTAGG-TS-V(ϵ) enjoys similar near-optimal guarantees.



(a) $|\mathcal{I}_{5\epsilon}| = 9$

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

(c) $|\mathcal{I}_{5\epsilon}| = 7$

(d) $|\mathcal{I}_{5\epsilon}| = 6$

(e) $|\mathcal{I}_{5\epsilon}| = 5$

(f) $|\mathcal{I}_{5\epsilon}| = 4$

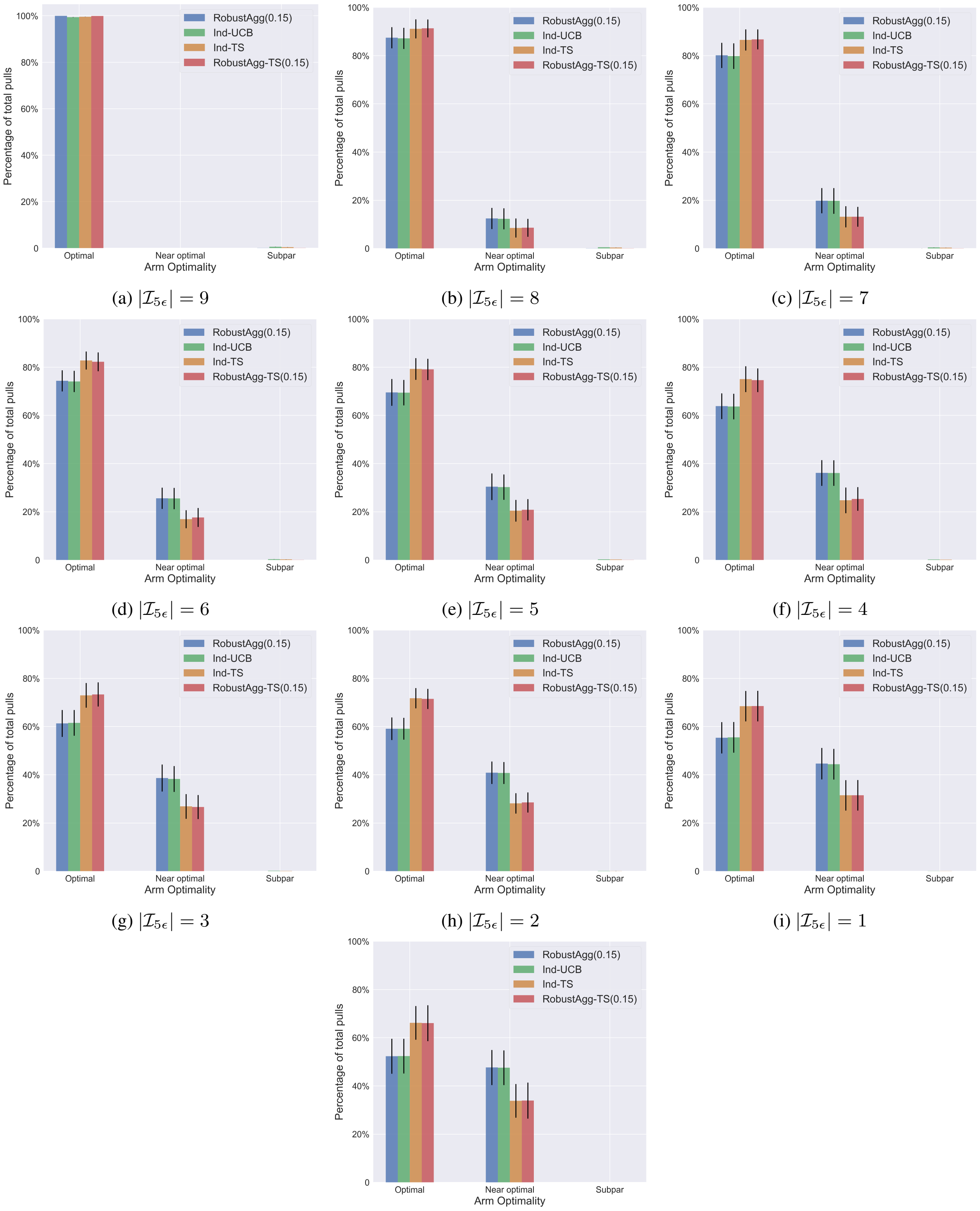
(g) $|\mathcal{I}_{5\epsilon}| = 3$

(h) $|\mathcal{I}_{5\epsilon}| = 2$

(i) $|\mathcal{I}_{5\epsilon}| = 1$

(j) $|\mathcal{I}_{5\epsilon}| = 0$

Figure 3. Compares the cumulative collective regret of the 4 algorithms over a horizon of $T = 50,000$ rounds.



(a) $|\mathcal{I}_{5\epsilon}| = 9$

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

(c) $|\mathcal{I}_{5\epsilon}| = 7$

(d) $|\mathcal{I}_{5\epsilon}| = 6$

(e) $|\mathcal{I}_{5\epsilon}| = 5$

(f) $|\mathcal{I}_{5\epsilon}| = 4$

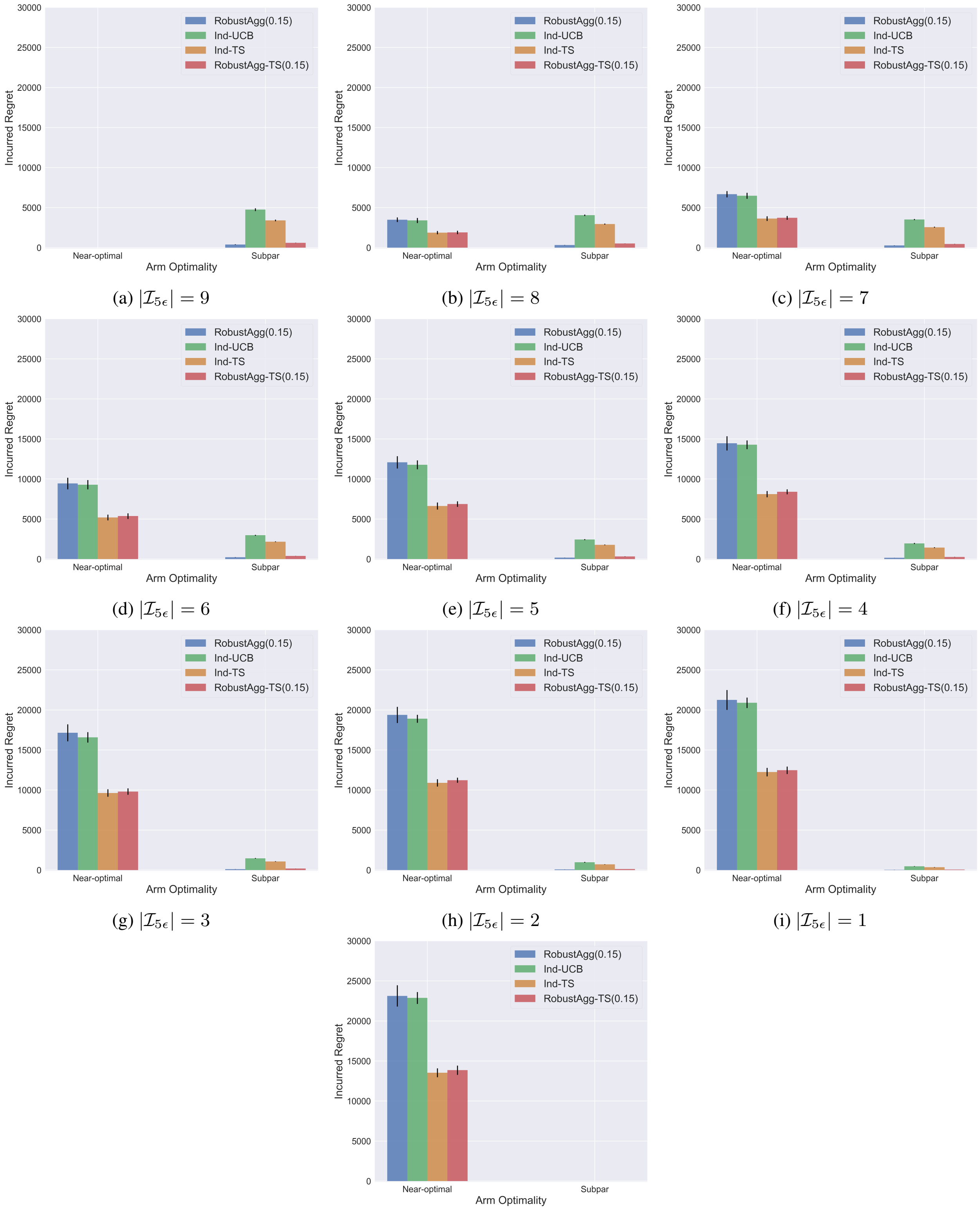
(g) $|\mathcal{I}_{5\epsilon}| = 3$

(h) $|\mathcal{I}_{5\epsilon}| = 2$

(i) $|\mathcal{I}_{5\epsilon}| = 1$

(j) $|\mathcal{I}_{5\epsilon}| = 0$

Figure 4. Compares the percentage of arm pulls by arm optimality for the 4 algorithms in $T = 50,000$ rounds.



(a) $|\mathcal{I}_{5\epsilon}| = 9$

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

(c) $|\mathcal{I}_{5\epsilon}| = 7$

(d) $|\mathcal{I}_{5\epsilon}| = 6$

(e) $|\mathcal{I}_{5\epsilon}| = 5$

(f) $|\mathcal{I}_{5\epsilon}| = 4$

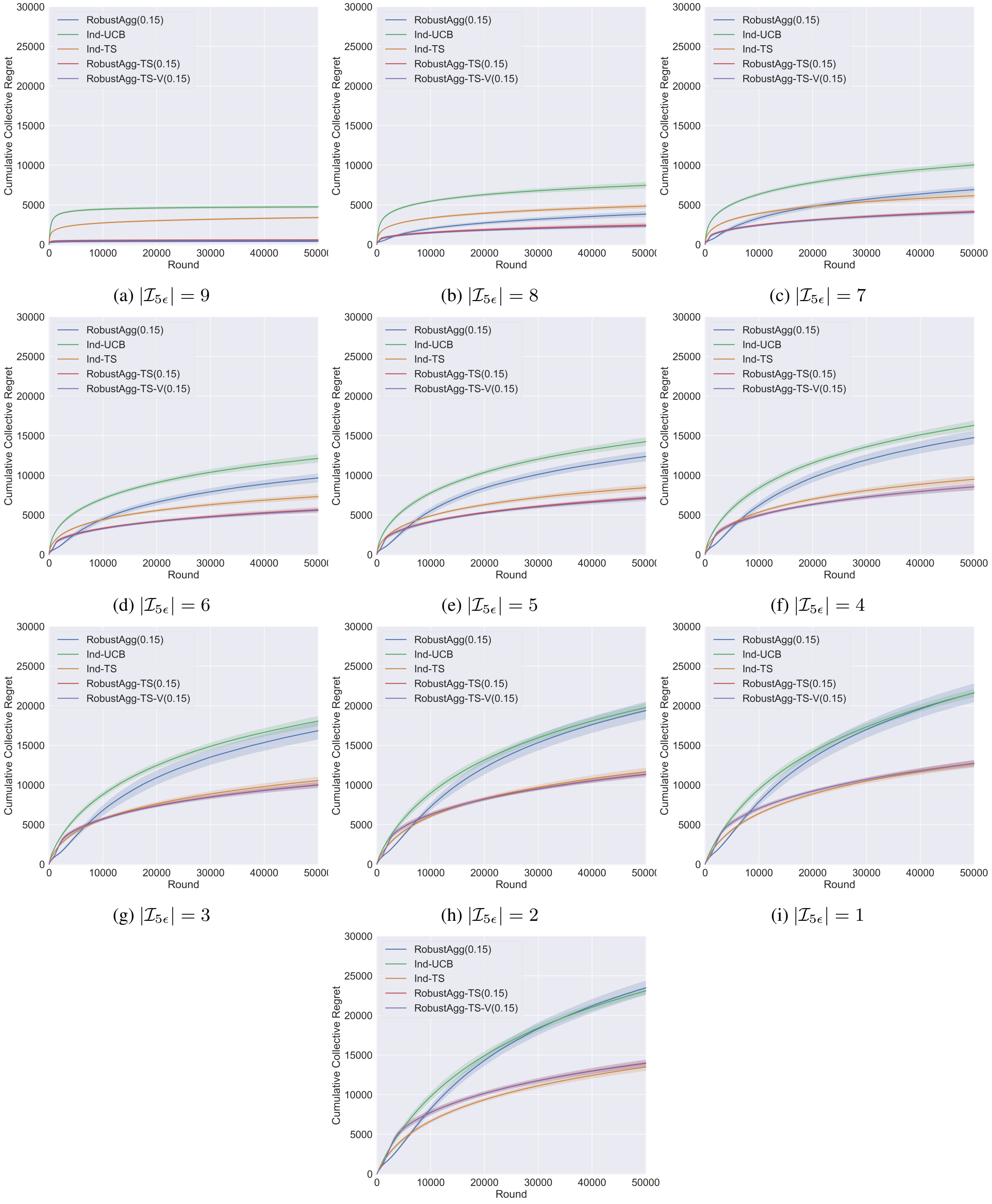
(g) $|\mathcal{I}_{5\epsilon}| = 3$

(h) $|\mathcal{I}_{5\epsilon}| = 2$

(i) $|\mathcal{I}_{5\epsilon}| = 1$

(j) $|\mathcal{I}_{5\epsilon}| = 0$

Figure 5. Compares the cumulative collective regret incurred by arm optimality for the 4 algorithms in $T = 50,000$ rounds.



(a) $|\mathcal{I}_{5\epsilon}| = 9$

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

(c) $|\mathcal{I}_{5\epsilon}| = 7$

(d) $|\mathcal{I}_{5\epsilon}| = 6$

(e) $|\mathcal{I}_{5\epsilon}| = 5$

(f) $|\mathcal{I}_{5\epsilon}| = 4$

(g) $|\mathcal{I}_{5\epsilon}| = 3$

(h) $|\mathcal{I}_{5\epsilon}| = 2$

(i) $|\mathcal{I}_{5\epsilon}| = 1$

(j) $|\mathcal{I}_{5\epsilon}| = 0$

Figure 6. Compares the cumulative collective regret of the 5 algorithms over a horizon of $T = 50,000$ rounds.