# A Related Work and Comparisons

We review the literature on multi-player bandit problems (see also Landgren (2019, Section 1.3.2) for a survey), and we comment on how existing problem formulations/approaches compare with ours studied in this paper.

**Identical reward distributions.** A large portion of prior studies focuses on the setting where a group of players collaboratively work on one bandit learning problem instance, i.e., for each arm/action, the reward distribution is identical for every player.

For example, Kar et al. (2011) study a networked bandit problem, in which only one agent observes rewards, and the other agents only have access to its sampling pattern. Peer-to-peer networks are explored by Szörényi et al. (2013), in which limited communication is allowed based on an overlay network. Landgren et al. (2016) apply running consensus algorithms to study a distributed cooperative multi-armed bandit problem. Kolla et al. (2018) study collaborative stochastic bandits over different structures of social networks that connect a group of agents. Wang et al. (2019) study communication cost minimization in multi-agent multi-armed bandits. Multi-agent bandit with a gossip-style protocol that has a communication budget is investigated in (Sankararaman et al., 2019; Chawla et al., 2020). Dubey and Pentland (2020a) investigate multi-agent bandits with heavy-tailed rewards. Wang et al. (2020) present an approach with a "parsimonious exploration principle" to minimize regret and communication cost. We note that, in contrast, we study multi-player bandit learning where the reward distributions can be different across players .

**Player-dependent reward distributions.** Multi-agent bandit learning with *heterogeneous feedback* has also been covered by previous studies.

- Cesa-Bianchi et al. (2013) study a network of linear contextual bandit players with heterogeneous rewards, where the players can take advantage of reward similarities hinted by a graph. In (Wu et al., 2016; Wang et al., 2017a,b), reward distributions of each player are *generated* based on social influence, which is modeled using preferences of the player's neighbors in a graph. These papers use regularization-based methods that take advantage of graph structures; in contrast, we study *when and how* to use information from other players based on a dissimilarity parameter.

- Gentile et al. (2014); Bresler et al. (2014); Song et al. (2014); Li et al. (2016); Korda et al. (2016); Li et al. (2019), among others, assume that the players' reward distributions have a cluster structure and players that belong to one cluster share a *common* reward distribution; our paper does not assume such cluster structure.

- Nguyen and Lauw (2014) investigate dynamic clustering of players with independent reward distributions and provides an empirical validation of their algorithm; Zhu et al. (2020) present an algorithm that combines dynamic clustering and Thompson sampling. In contrast, in this paper, we develop a UCB-based approach that has a fallback guarantee[3].

- In the work of Shahrampour et al. (2017), a group of players seek to find the arm with the largest average reward over all players; and, in each round, the players have to reach a consensus and choose the same arm.

- Dubey and Pentland (2020b) assume access to some side information for every player, and learns a reward predictor that takes both player's side information models and action as input. In comparison, our work do not assume access to such side information.

- Further, similarities in reward distributions are explored in the work of Zhang et al. (2019), which studies a *warm-start* scenario, in which data are provided as history (Shivaswamy and Joachims, 2012) for an learning agent to explore faster. Azar et al. (2013); Soare et al. (2014) investigate multitask learning in bandits through *sequential transfer* between tasks that have similar reward distributions. In contrast, we study the multi-player setting, where all players learn continually and concurrently.

[3]In Zhu et al. (2020), it is unclear how to tune the hyper-parameter $\beta$ apriori to ensure a sublinear fall-back regret guarantee, even if the "similarity" parameter $\gamma$ is known.

**Collisions in multi-player bandits.** Multi-player bandit problems with collisions (e.g., Liu and Zhao, 2010; Kalathil et al., 2014; Boursier and Perchet, 2020; Bubeck and Budzinski, 2020; Shi et al., 2020; Bubeck et al., 2020; Wang et al., 2020) are also well-studied. In such models, two players pulling the same arm in the same round *collide* and receive zero reward. These models have a wide range of practical applications (e.g., cognitive radio), and some assume *player-dependent heterogeneous* reward distributions (Bistritz et al., 2020; Boursier et al., 2020); in comparison, collision is not modeled in our paper.

**Side information.** Models in which learning agents observe side information have also been studied in prior works—one can consider data collected by other players in multi-player bandits as side observations (Landgren, 2019). In some models, a player observes side information for some arms that are not chosen in the current round: stochastic models with such side information are studied in (Caron et al., 2012; Buccapatnam et al., 2014; Wu et al., 2015), and adversarial models in (Mannor and Shamir, 2011; Alon et al., 2017); Similarities/closeness among arms in one bandit problem are studied in (Deshmukh et al., 2017; Xu et al., 2017; Wang et al., 2018). We note that our problem formulation is different, because in these models, auxiliary data are from arms in the same bandit problem instance instead of from other players.

Upper and lower bounds on the means of reward distributions are used as side information in (Sharma et al., 2020). Loss predictors (Wei et al., 2020) can also be considered as side information. In contrast, we do not leverage such information. Further, side information can also refer to "context" in contextual bandits (Slivkins, 2014). In comparison, we assume a multi-armed setting and their results do not imply ours.

**Other multi-player bandit learning topics.** Many other multi-player bandit learning topics have also been explored. For example, Awerbuch and Kleinberg (2005); Vial et al. (2020) study multi-player models in which some of the players are malicious. Christakopoulou and Banerjee (2018) study collaborative bandits with applications such as top-$K$ recommendations. Nonstochastic multi-armed bandit models with communicating agents are studied in (Bar-On and Mansour, 2019; Cesa-Bianchi et al., 2019). Privacy protection in decentralized exploration is investigated in (Feraud et al., 2019). We note that, in this paper, our goal does not align closely with these topics.

## B Proof of Claim 3

We first restate Example 2 and Claim 3.

**Example 2.** *For a fixed $\epsilon \in (0, \frac{1}{8})$ and $\delta \le \epsilon/4$, consider the following Bernoulli MPMAB problem instance: for each $p \in [M]$, $\mu_1^p = \frac{1}{2} + \delta$, $\mu_2^p = \frac{1}{2}$. This is a 0-MPMAB instance, hence an $\epsilon$-MPMAB problem instance. Also, note that $\epsilon$ is at least four times larger than the gaps $\Delta_2^p = \delta$.*

**Claim 3.** *For the above example, any sublinear regret algorithm for the $\epsilon$-MPMAB problem must have $\Omega(\frac{M \ln T}{\delta})$ regret on this instance, matching the* Ind-UCB *regret upper bound.*

*Proof of Claim 3.* Suppose $\mathcal{A}$ is a sublinear-regret algorithm for the $\epsilon$-MPMAB problem; i.e., there exist $C > 0$ and $\alpha > 0$ such that $\mathcal{A}$ has $CT^{1-\alpha}$ regret in all $\epsilon$-MPMAB instances.

Recall that we consider the Bernoulli $\epsilon$-MPMAB instance $\mu = (\mu_i^p)_{i\in[2],p\in[M]}$ such that $\mu_1^p = \frac{1}{2} + \delta$ and $\mu_2^p = \frac{1}{2}$ for all $p$. As $\epsilon \in (0, \frac{1}{8})$ and $\delta \le \frac{\epsilon}{4}$, it can be directly verified that all $\mu_i^p$'s are in $[\frac{15}{32}, \frac{17}{32}]$. In addition, since for all $p$, $\Delta_2^p = \delta \le \frac{\epsilon}{4} = 5 \cdot \frac{\epsilon}{20}$, we have $\mathcal{I}_{\epsilon/20} = \emptyset$, i.e., $\mathcal{I}_{\epsilon/20}^C = \{1, 2\}$.

From Theorem 9, we conclude that for this MPMAB instance $\mu$, $\mathcal{A}$ has regret lower bounded as follows:

$$\mathbb{E}_\mu\left[\mathcal{R}(T)\right] \ge \Omega\left(\frac{M \ln(T^\alpha \Delta_2^p / C)}{\Delta_2^p}\right) = \Omega\left(\frac{M \ln(T^\alpha \delta / C)}{\delta}\right) = \Omega\left(\frac{M \ln T}{\delta}\right),$$

for sufficiently large $T$. □

## C Basic properties of $\mathcal{I}_\epsilon$ for $\epsilon$-MPMAB instances

In Section 3 of the paper, we presented the following two facts about properties of $\mathcal{I}_\epsilon$ for $\epsilon$-MPMAB problem instances:

**Fact 4.** *$|\mathcal{I}_\epsilon| \le K-1$. In addition, for each arm $i \in \mathcal{I}_\epsilon$, $\Delta_i^{\min} > 3\epsilon$; in other words, for all players $p$ in $[M]$, $\Delta_i^p = \mu_*^p - \mu_i^p > 3\epsilon$; consequently, arm $i$ is suboptimal for all players $p$ in $[M]$.*

**Fact 6.** *For any $i \in \mathcal{I}_\epsilon$, $\frac{1}{\Delta_i^{\min}} \le \frac{2}{M}\sum_{p\in[M]} \frac{1}{\Delta_i^p}$.*

Here, we will present and prove a more complete collection of facts about the properties of $\mathcal{I}_\epsilon$ which covers every statement in Fact 4 and Fact 6. Before that, we first prove the following fact.

**Fact 14.** *For an $\epsilon$-MPMAB problem instance, for any $i \in [K]$, and $p, q \in [M]$, $|\Delta_i^p - \Delta_i^q| \le 2\epsilon$.*

*Proof.* Fix any player $p \in [M]$, let $j \in [K]$ be an optimal arm for $p$ such that $\mu_j^p = \mu_*^p$. We first show that, for any player $q \in [M]$, $|\mu_*^q - \mu_j^p| \le \epsilon$.

- $\mu_*^q \ge \mu_j^p - \epsilon$ is trivially true because $\mu_j^q \ge \mu_j^p - \epsilon$ by Definition 1 and $\mu_*^q \ge \mu_j^q$ by the definition of $\mu_*^q$;
- $\mu_*^q \le \mu_j^p + \epsilon$ is true because if there exists an arm $k \in [K]$ such that $\mu_k^q > \mu_j^p + \epsilon$, then by Definition 1 we must have $\mu_k^p \ge \mu_k^q - \epsilon > \mu_j^p$ which contradicts with the premise that $j$ is an optimal arm for player $p$.

We have shown that $|\mu_*^q - \mu_*^p| \le \epsilon$. Since $|\mu_i^q - \mu_i^p| \le \epsilon$ by Definition 1, it follows from the triangle inequality that $|\Delta_i^p - \Delta_i^q| \le 2\epsilon$. □

We now present a set of basic properties of $\mathcal{I}_\epsilon$.

**Fact 15** (Basic properties of $\mathcal{I}_\epsilon$)**.** *Let $\Delta_i^{\max} = \max_{p\in[M]} \Delta_i^p$. For an $\epsilon$-MPMAB problem instance, for each arm $i \in \mathcal{I}_\epsilon$,*

*(a) $\Delta_i^p > 3\epsilon$ for all players $p \in [M]$; in other words, $\Delta_i^{\min} > 3\epsilon$;*

*(b) arm $i$ is suboptimal for all players $p \in [M]$, i.e., for any player $p \in [M]$, $\mu_i^p < \mu_*^p$;*

*(c) $\frac{\Delta_i^p}{\Delta_i^q} < 2$ for any pair of players $p, q \in [M]$; consequently, $\frac{\Delta_i^{\max}}{\Delta_i^{\min}} < 2$;*

*(d) $\frac{1}{\Delta_i^{\min}} \le \frac{2}{M}\sum_{p\in[M]} \frac{1}{\Delta_i^p}$;*

*(e) $|\mathcal{I}_\epsilon| \le K-1$.*

*Proof.* We prove each item one by one.

(a) For each arm $i \in \mathcal{I}_\epsilon$, by definition, there exists $p \in [M]$, $\Delta_i^p > 5\epsilon$. It follows from Fact 14 that for any $q \in [M]$, $\Delta_i^q \ge \Delta_i^p - 2\epsilon > 3\epsilon$. $\Delta_i^{\min} > 3\epsilon$ then follows straightforwardly.

(b) For each arm $i \in \mathcal{I}_\epsilon$, it follows from item (a) that for any $p \in [M]$, $\Delta_i^p > 3\epsilon \ge 0$. Therefore, $i$ is suboptimal for all player $p \in [M]$.

(c) By Fact 14, for any $i \in \mathcal{I}_\epsilon \subseteq [K]$ and any $p, q \in [M]$, $\Delta_i^p \le \Delta_i^q + 2\epsilon$, which implies $\frac{\Delta_i^p}{\Delta_i^q} \le 1 + \frac{2\epsilon}{\Delta_i^q}$. Since by item (a), $\Delta_i^q > 3\epsilon$, it follows that $\frac{\Delta_i^p}{\Delta_i^q} \le 1 + \frac{2\epsilon}{\Delta_i^q} < 2$. $\frac{\Delta_i^{\max}}{\Delta_i^{\min}} < 2$ then follows straightforwardly.

(d) For each arm $i \in \mathcal{I}_\epsilon$, it follows from item (c) that for any $p \in [M]$, $\Delta_i^p \in [\Delta_i^{\min}, 2\Delta_i^{\min}]$. Therefore, we have $\frac{1}{\Delta_i^p} \in [\frac{1}{2\Delta_i^{\min}}, \frac{1}{\Delta_i^{\min}}]$, as $\Delta_i^p > 0$ for all $p$. It then follows that $\frac{2}{M}\sum_{p\in[M]} \frac{1}{\Delta_i^p} \ge \frac{2}{M}\sum_{p\in[M]} \frac{1}{2\Delta_i^{\min}} = \frac{1}{\Delta_i^{\min}}$.

(e) Pick an arm $i$ that is optimal with respect to player 1; $i$ cannot be in $\mathcal{I}_\epsilon$ because of item (b). Therefore, $\mathcal{I}_\epsilon \subseteq [K] \setminus \{i\}$, which implies that it has size at most $K-1$. □

# D Proof of Upper Bounds in Section 3

## D.1 Proof Overview

In Appendix D.2 and D.3, we focus on showing that in a "clean" event $\mathcal{E}$ (defined in D.3), the upper confidence bound $\mathrm{UCB}_i^p(t) = \kappa_i^p(t,\lambda) + F(\overline{n_i^p}, \overline{m_i^p}, \lambda, \epsilon)$ (line 10 of Algorithm 1)[4] holds for every $t \in [T], i \in [K], p \in [M]$ and $\lambda \in [0,1]$; and the "clean" event $\mathcal{E}$ occurs with $1 - 4MK/T^4$ probability.

Then, in Appendix D.4, we provide a proof of the gap-dependent upper bound in Theorem 5. In Appendix D.5, we provide a proof of the gap-independent upper bound in Theorem 8.

## D.2 Event $\mathcal{Q}_i(t)$

Recall that $n_i^p(t-1)$ is the number of pulls of arm $i$ by player $p$ after the first $(t-1)$ rounds. Let $m_i^p(t-1) = \sum_{q \in [M]: q \neq p} n_i^q(t-1)$.

We now define the following event.

**Definition 16.** *Let*

$$\mathcal{Q}_i(t) = \left\{ \forall p, \left|\zeta_i^p(t) - \mu_i^p\right| \le 8\sqrt{\frac{3\ln T}{\overline{n_i^p}(t-1)}}, \quad \left|\eta_i^p(t) - \sum_{q \neq p} \frac{n_i^q(t-1)}{\overline{m_i^p}(t-1)} \mu_i^q\right| \le 4\sqrt{\frac{14\ln T}{\overline{m_i^p}(t-1)}} \right\},$$

*where*

$$\zeta_i^p(t) = \frac{\sum_{s=1}^{t-1} \mathbb{1}\{i_t^p = i\} r_t^p}{\overline{n_i^p}(t-1)},$$

*and*

$$\eta_i^p(t) = \frac{\sum_{s=1}^{t-1} \sum_{q=1}^{M} \mathbb{1}\{q \neq p,\ i_t^q = i\} r_t^q}{\overline{m_i^p}(t-1)}.$$

**Lemma 17.**

$$\Pr(\mathcal{Q}_i(t)) \ge 1 - 4MT^{-5}.$$

*Proof.* For any fixed player $p$, we discuss the two inequalities separately. Lemma 17 then follows by a union bound over the two inequalities and over all $p \in [M]$.

We first discuss the concentration of $\zeta_i^p(t)$. We define a filtration $\{\mathcal{B}_t\}_{t=1}^T$, where

$$\mathcal{B}_t = \sigma(\left\{ i_s^{p'}, r_s^{p'} : s \in [t], p' \in [M] \right\} \cup \left\{ i_{t+1}^{p'} : p' \in [M] \right\})$$

is the $\sigma$-algebra generated by the historical interactions up to round $t$ and the arm selection of all players at round $t+1$.

Let random variable $X_t = \mathbb{1}\{i_t^p = i\}\left(r_t^p - \mu_i^p\right)$. We have $\mathbb{E}\left[X_t \mid \mathcal{B}_{t-1}\right] = 0$; in addition, $\mathbb{V}\left[X_t \mid \mathcal{B}_{t-1}\right] = \mathbb{E}\left[(X_t - \mathbb{E}[X_t \mid \mathcal{B}_{t-1}])^2 \mid \mathcal{B}_{t-1}\right] \le \mathbb{E}\left[(\mathbb{1}\{i_t^p = i\} r_t^p)^2 \mid \mathcal{B}_{t-1}\right] \le \mathbb{1}\{i_t^p = i\}$ and $|X_t| \le 1$.

Applying Freedman's inequality (Bartlett et al., 2008, Lemma 2) with $\sigma = \sqrt{\sum_{s=1}^{t-1} \mathbb{V}\left[X_s \mid \mathcal{B}_{s-1}\right]}$ and $b = 1$, and using $\sigma \le \sqrt{\sum_{s=1}^{t-1} \mathbb{1}\{i_s^p = i\}}$, we have that with probability at least $1 - 2T^{-5}$,

$$\left|\sum_{s=1}^{t-1} X_s\right| \le 4\sqrt{\sum_{s=1}^{t-1} \mathbb{1}\{i_s^p = i\} \cdot \ln(T^5 \log_2 T)} + 2\ln(T^5 \log_2 T). \tag{1}$$

We consider two cases:

[4] Recall that $\overline{z} = \max\{z, 1\}$.

1. If $n_i^p(t-1) = \sum_{s=1}^{t-1} \mathbb{1}\{i_s^p = i\} = 0$, we have $\overline{n_i^p}(t-1) = 1$ and $\zeta_i^p(t) = 0$. In this case, we trivially have
$$\left|\zeta_i^p(t) - \mu_i^p\right| \le 1 \le 8\sqrt{\frac{3\ln T}{\overline{n_i^p}(t-1)}}.$$

2. Otherwise, $n_i^p(t-1) \ge 1$. In this case, we have $\overline{n_i^p}(t-1) = n_i^p(t-1)$. Divide both sides of Eq. (1) by $n_i^p(t-1)$, and use the fact that $\log T \le T$, we have
$$\left|\frac{\sum_{s=1}^{t-1}\mathbb{1}\{i_s^p = i\} r_s^p}{n_i^p(t-1)} - \mu_i^p\right| \le 4\sqrt{\frac{6\ln T}{n_i^p(t-1)}} + \frac{12\ln T}{n_i^p(t-1)}.$$
If $\frac{12\ln T}{n_i^p(t-1)} \ge 1$, $\left|\frac{\sum_{s=1}^{t-1}\mathbb{1}\{i_s^p = i\} r_s^p}{n_i^p(t-1)} - \mu_i^p\right| \le 8\sqrt{\frac{3\ln T}{n_i^p(t-1)}}$ is trivially true. Otherwise, $\frac{12\ln T}{n_i^p(t-1)} \le 2\sqrt{\frac{3\ln T}{n_i^p(t-1)}}$, which implies that $\left|\frac{\sum_{s=1}^{t-1}\mathbb{1}\{i_s^p = i\} r_s^p}{n_i^p(t-1)} - \mu_i^p\right| \le (4\sqrt{6}+2\sqrt{3})\sqrt{\frac{\ln T}{n_i^p(t-1)}} \le 8\sqrt{\frac{3\ln T}{n_i^p(t-1)}}$.

In summary, in both cases, with probability at least $1 - 2T^{-5}$, we have
$$\left|\zeta_i^p(t-1) - \mu_i^p\right| \le 8\sqrt{\frac{3\ln T}{\overline{n_i^p}(t-1)}}.$$

A similar application of Freedman's inequality also shows the concentration of $\eta_i^p(t)$. Similarly, we define a filtration $\{\mathcal{G}_{t,q}\}_{t\in[T],q\in[M]}$, where
$$\mathcal{G}_{t,q} = \sigma(\left\{i_s^{p'}, r_s^{p'} : s \in [t], p' \in [M]\right\} \cup \left\{i_{t+1}^{p'} : p' \in [M], p' \le q\right\})$$
is the $\sigma$-algebra generated by the historical interactions up to round $t$ and the arm selection of players $1, 2, \ldots, q$ at round $t+1$. We have
$$\mathcal{G}_{1,1} \subset \mathcal{G}_{1,2} \subset \ldots \subset \mathcal{G}_{1,M} \subset \mathcal{G}_{2,1} \subset \ldots \subset \mathcal{G}_{2,M} \subset \ldots \subset \mathcal{G}_{T,M}.$$
Let random variable $Y_{t,q} = \mathbb{1}\{q \ne p,\ i_t^q = i\}\left(r_t^q - \mu_i^q\right)$. We have $\mathbb{E}\left[Y_{t,q} \mid \mathcal{G}_{t-1,q}\right] = 0$; in addition, $\mathbb{V}\left[Y_{t,q} \mid \mathcal{G}_{t-1,q}\right] = \mathbb{E}\left[Y_{t,q}^2 \mid \mathcal{G}_{t-1,q}\right] \le \mathbb{1}\{q \ne p,\ i_t^q = i\}$, and $\left|Y_{t,q}\right| \le 1$.

Similarly, applying Freedman's inequality (Bartlett et al., 2008, Lemma 2) with $\sigma = \sqrt{\sum_{s=1}^{t-1}\sum_{q=1}^{M}\mathbb{V}\left[Y_{s,q} \mid \mathcal{G}_{s-1,q}\right]}$ and $b = 1$, and using $\sigma \le \sqrt{\sum_{s=1}^{t-1}\sum_{q=1}^{M}\mathbb{1}\{q \ne p, i_s^q = i\}}$, we have that with probability at least $1 - 2T^{-5}$,
$$\left|\sum_{s=1}^{t-1}\sum_{q=1}^{M} Y_{s,q}\right| \le 4\sqrt{\sum_{s=1}^{t-1}\sum_{q=1}^{M}\mathbb{1}\{q \ne p,\ i_s^q = i\} \cdot \ln(T^5\log_2(TM))} + 2\ln(T^5\log_2(TM)). \tag{2}$$

Again, we consider two cases. If $m_i^p(t-1) = 0$, then we have $\eta_i^p(t-1) = 0$ and
$$\left|\eta_i^p(t-1) - \sum_{q\ne p}\frac{n_i^q(t-1)}{\overline{m_i^p}(t-1)}\mu_i^q\right| = 0 \le 4\sqrt{\frac{14\ln T}{\overline{m_i^p}(t-1)}}$$
.

Otherwise, we have $\overline{m_i^p}(t-1) = m_i^p(t-1)$. Divide both sides of Eq. (2) by $m_i^p(t-1)$, and use the fact that $\log_2(TM) \le T^2$, we have
$$\left|\frac{\sum_{s=1}^{t-1}\sum_{q=1}^{M}\mathbb{1}\{q \ne p,\ i_s^q = i\} r_s^q}{m_i^p(t-1)} - \sum_{q\ne p}\frac{n_i^q(t-1)}{m_i^p(t-1)}\mu_i^q\right| \le 4\sqrt{\frac{7\ln T}{m_i^p(t-1)}} + \frac{14\ln T}{m_i^p(t-1)}.$$

If $\frac{14\ln T}{m_i^p(t-1)} \geq 1$, $\left|\frac{\sum_{s=1}^{t-1}\sum_{q=1}^{M}\mathbb{1}\{q\neq p,\ i_s^q=i\}r_s^q}{m_i^p(t-1)} - \sum_{q\neq p}\frac{n_i^q(t-1)}{m_i^p(t-1)}\mu_i^q\right| \leq 4\sqrt{\frac{14\ln T}{m_i^p(t-1)}}$ is trivially true. Otherwise, $\frac{14\ln T}{m_i^p(t-1)} \leq \sqrt{\frac{14\ln T}{m_i^p(t-1)}}$, which implies that

$$\left|\frac{\sum_{s=1}^{t-1}\sum_{q=1}^{M}\mathbb{1}\{q\neq p,\ i_s^q=i\}r_s^q}{m_i^p(t-1)} - \sum_{q\neq p}\frac{n_i^q(t-1)}{m_i^p(t-1)}\mu_i^q\right| \leq (4\sqrt{7}+\sqrt{14})\sqrt{\frac{\ln T}{m_i^p(t-1)}} \leq 4\sqrt{\frac{14\ln T}{m_i^p(t-1)}}.$$

In summary, in both cases, with probability at least $1-2T^{-5}$, we have

$$\left|\eta_i^p(t-1) - \sum_{q\neq p}\frac{n_i^q(t-1)}{\overline{m_i^p}(t-1)}\mu_i^q\right| \leq 4\sqrt{\frac{14\ln T}{\overline{m_i^p}(t-1)}}.$$

The lemma follows by taking a union bound over these two inequalities for each fixed $p$, and over all $p\in[M]$. □

### D.3 Event $\mathcal{E}$

Let $\mathcal{E} = \cap_{t=1}^{T}\cap_{i=1}^{K}\mathcal{Q}_i(t)$. We present the following corollary and lemma regarding event $\mathcal{E}$.

**Corollary 18.** *It follows from Lemma 17 that* $\Pr[\mathcal{E}] \geq 1-\frac{4MK}{T^4}$.

**Lemma 19.** *If $\mathcal{E}$ occurs, we have that for every $t\in[T]$, $i\in[K]$, $p\in[M]$, for all $\lambda\in[0,1]$,*

$$\left|\kappa_i^p(t,\lambda)-\mu_i^p\right| \leq 8\sqrt{13\ln T\left(\frac{\lambda^2}{\overline{n_i^p}(t-1)}+\frac{(1-\lambda)^2}{\overline{m_i^p}(t-1)}\right)} + (1-\lambda)\epsilon,$$

*where $\kappa_i^p(t,\lambda) = \lambda\zeta_i^p(t)+(1-\lambda)\eta_i^p(t)$.*

*Proof.* If $\mathcal{E}$ occurs, for every $t\in[T]$ and $i\in[K]$, by the definition of event $\mathcal{Q}_i(t)$, we have

$$\left|\zeta_i^p(t)-\mu_i^p\right| < 8\sqrt{\frac{3\ln T}{\overline{n_i^p}(t-1)}}, \text{and } \left|\eta_i^p(t) - \sum_{q\neq p}\frac{n_i^q(t-1)}{\overline{m_i^p}(t-1)}\mu_i^q\right| \leq 4\sqrt{\frac{14\ln T}{\overline{m_i^p}(t-1)}}.$$

As $\kappa_i^p(t,\lambda) = \lambda\zeta_i^p(t)+(1-\lambda)\eta_i^p(t)$, we have:

$$\begin{aligned}\left|\kappa_i^p(t,\lambda) - \left[\lambda\mu_i^p + (1-\lambda)\sum_{q\neq p}\frac{n_i^q(t-1)}{\overline{m_i^p}(t-1)}\mu_i^q\right]\right| &\leq 8\lambda\sqrt{\frac{3\ln T}{\overline{n_i^p}(t-1)}} + 4(1-\lambda)\sqrt{\frac{14\ln T}{\overline{m_i^p}(t-1)}} \\ &\leq 8\sqrt{13\ln T\left(\frac{\lambda^2}{\overline{n_i^p}(t-1)}+\frac{(1-\lambda)^2}{\overline{m_i^p}(t-1)}\right)}, \end{aligned} \tag{3}$$

where the second inequality uses the elementary facts that $\sqrt{A}+\sqrt{B}\leq\sqrt{2(A+B)}$.

Furthermore, from Definition 1, we have

$$\left|\sum_{q\neq p}\frac{n_i^q(t-1)}{\overline{m_i^p}(t-1)}\mu_i^q - \mu_i^p\right| \leq \sum_{q\neq p}\frac{n_i^q(t-1)}{\overline{m_i^p}(t-1)}\left|\mu_i^q-\mu_i^p\right| \leq \epsilon.$$

This shows that

$$\left|\mu_i^p - (\lambda\mu_i^p + (1-\lambda)\sum_{q\neq p}\frac{n_i^q(t-1)}{\overline{m_i^p}(t-1)}\mu_i^q)\right| \leq (1-\lambda)\epsilon.$$

Combining the above inequality with Eq. (3), we get

$$\left|\kappa_i^p(t,\lambda) - \mu_i^p\right| \le 8\sqrt{13\ln T\left(\frac{\lambda^2}{\overline{n_i^p}(t-1)} + \frac{(1-\lambda)^2}{\overline{m_i^p}(t-1)}\right)} + (1-\lambda)\epsilon.$$

This completes the proof. □

### D.4 Proof of Theorem 5

We first restate Theorem 5.

**Theorem 5.** *Let* RobustAgg($\epsilon$) *run on an $\epsilon$-MPMAB problem instance for $T$ rounds. Then, its expected collective regret satisfies*

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

Recall that the expected collective regret is defined as $\mathbb{E}[\mathcal{R}(T)] = \sum_{i\in[K]}\sum_{p\in[M]}\Delta_i^p\cdot\mathbb{E}[n_i^p(T)]$. Before we prove Theorem 5, we first present the following two lemmas, which provides an upper bound for (1) the total number of arm pulls for arm $i$, for $i$ in $\mathcal{I}_\epsilon$ and (2) the individual number of arm pulls for arm $i$ and player $p$, for $i$ in $\mathcal{I}_\epsilon^C$, conditioned on $\mathcal{E}$ happening.

**Lemma 20.** *Denote $n_i(T) = \sum_{p\in[M]} n_i^p(T)$ as the total number of pulls of arm $i$ by all the players after $T$ rounds. Let* RobustAgg($\epsilon$) *run on an $\epsilon$-MPMAB problem instance for $T$ rounds. Then, for each $i\in\mathcal{I}_\epsilon$, we have*

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

**Lemma 21.** *Let* RobustAgg($\epsilon$) *run on an $\epsilon$-MPMAB problem instance for $T$ rounds. Then, for each $i\in\mathcal{I}_\epsilon^C$ and player $p\in[M]$ such that $\Delta_i^p>0$, we have*

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

*Proof of Lemma 20.* We first note that it follows from item (b) of Fact 15 that every arm $i\in\mathcal{I}_\epsilon$ is suboptimal for all players $p\in[M]$.

We have

$$\begin{aligned} n_i(T) &= \sum_{t=1}^{T}\sum_{p=1}^{M}\mathbb{1}\{i_t^p = i\} \\ &\le M + \tau + \sum_{t=1}^{T}\sum_{p=1}^{M}\mathbb{1}\{i_t^p = i, n_i(t-1)>\tau\}. \end{aligned} \tag{4}$$

Here, $\tau\ge 1$ is an arbitrary integer. The term $M$ is due to parallel arm pulls in the $\epsilon$-MPMAB problem: Let $s$ be the first round such that after round $s$, the total number of pulls $n_i(s)>\tau$. This implies that $n_i(s-1)\le\tau$. Then in round $s$, there can be up to $M$ pulls of arm $i$ by all the players, which means that in round $(s+1)$ when the third term in Eq. (4) can first start counting, there could have been up to $\tau+M$ pulls of the arm $i$.

It then follows that

$$n_i(T) \le M + \tau + \sum_{t=1}^{T}\sum_{p=1}^{M}\mathbb{1}\{\mathrm{UCB}_{i_*^p}^p(t) \le \mathrm{UCB}_i^p(t),\ n_i(t-1)>\tau\}. \tag{5}$$

Recall that $\Delta_i^{\min} = \min_p \Delta_i^p$, and for each $i \in \mathcal{I}_\epsilon$, we have $\Delta_i^p \geq \Delta_i^{\min} > 3\epsilon$ by item (a) of Fact 15.

With foresight, we choose $\tau = \lceil \frac{3328 \ln T}{(\Delta_i^{\min} - 2\epsilon)^2} \rceil$. Conditional on $\mathcal{E}$, we show that, *for any arm* $i \in \mathcal{I}_\epsilon$, the event $\left\{ \mathrm{UCB}^p_{i^p_*}(t) \leq \mathrm{UCB}^p_i(t), n_i(t-1) > \tau \right\}$ never happens. It suffices to show that if $n_i(t-1) > \tau$,

$$\mathrm{UCB}^p_{i^p_*}(t) \geq \mu^p_*, \tag{6}$$

and

$$\mathrm{UCB}^p_i(t) < \mu^p_* \tag{7}$$

happen simultaneously.

Eq. (6) follows straightforwardly from the definition of $\mathcal{E}$ along with Lemma 19. For Eq. (7), we have the following upper bound on $\mathrm{UCB}^p_i(t)$:

$$\begin{aligned}
\mathrm{UCB}^p_i(t) &= \kappa^p_i(t, \lambda^*) + F(\overline{n^p_i}, \overline{m^p_i}, \lambda^*, \epsilon) \\
&\leq \mu^p_i + 2F(\overline{n^p_i}, \overline{m^p_i}, \lambda^*, \epsilon) \\
&= \mu^p_i + 2\Big[ \min_{\lambda \in [0,1]} 8\sqrt{13 \ln T [\frac{\lambda^2}{\overline{n^p_i}(t-1)} + \frac{(1-\lambda)^2}{\overline{m^p_i}(t-1)}]} + (1-\lambda)\epsilon \Big] \\
&\leq \mu^p_i + 2\Big[ 8\sqrt{\frac{13 \ln T}{\overline{n^p_i}(t-1) + \overline{m^p_i}(t-1)}} + \epsilon \Big] \\
&\leq \mu^p_i + 2\Big[ 8\sqrt{\frac{13 \ln T}{n_i(t-1)}} + \epsilon \Big] \\
&< \mu^p_i + 2\Big[ 8\sqrt{\frac{13 \ln T (\Delta^p_i - 2\epsilon)^2}{3328 \ln T}} + \epsilon \Big] = \mu^p_i + \Delta^p_i = \mu^p_*,
\end{aligned}$$

where the first inequality is from the definition of $\mathcal{E}$ and Lemma 19; the second inequality is from choosing $\lambda = \frac{\overline{n^p_i}(t-1)}{\overline{n^p_i}(t-1) + \overline{m^p_i}(t-1)}$; the third inequality is from the simple facts that $n^p_i(t-1) \leq \overline{n^p_i}(t-1)$, $m^p_i(t-1) \leq \overline{m^p_i}(t-1)$, and $n_i(t-1) = n^p_i(t-1) + m^p_i(t-1)$; the last inequality is from the premise that $n_i(t-1) > \tau \geq \frac{3328 \ln T}{(\Delta^{\min}_i - 2\epsilon)^2} \geq \frac{3328 \ln T}{(\Delta^p_i - 2\epsilon)^2}$.

Continuing Eq. (5), it then follows that, for each $i \in \mathcal{I}_\epsilon$,

$$\mathbb{E}[n_i(T) | \mathcal{E}] \leq \lceil \frac{3328 \ln T}{(\Delta^{\min}_i - 2\epsilon)^2} \rceil + M \leq \frac{3328 \ln T}{(\Delta^{\min}_i - 2\epsilon)^2} + (M+1). \tag{8}$$

Now, by item (a) of Fact 15, for each $i \in \mathcal{I}_\epsilon$, $\Delta^{\min}_i > 3\epsilon$. We then have $\frac{\Delta^{\min}_i}{\Delta^{\min}_i - 2\epsilon} = \frac{\Delta^{\min}_i - 2\epsilon + 2\epsilon}{\Delta^{\min}_i - 2\epsilon} = 1 + \frac{2\epsilon}{\Delta^{\min}_i - 2\epsilon} < 3$. It follows that

$$\frac{3328 \ln T}{(\Delta^{\min}_i - 2\epsilon)^2} = \frac{3328 \ln T}{(\Delta^{\min}_i)^2} \cdot \left( \frac{\Delta^{\min}_i}{\Delta^{\min}_i - 2\epsilon} \right)^2 < \frac{29952 \ln T}{(\Delta^{\min}_i)^2}.$$

Therefore, continuing Eq. (8), for each $i \in \mathcal{I}_\epsilon$, we have

$$\mathbb{E}[n_i(T) | \mathcal{E}] < \frac{29952 \ln T}{(\Delta^{\min}_i)^2} + (M+1) \leq \frac{29952 \ln T}{(\Delta^{\min}_i)^2} + 2M.$$

where the second inequality follows from the fact that $M \geq 1$.

It then follows that

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

This completes the proof of Lemma 20. □

*Proof of Lemma 21.* Let's now turn our attention to arms in $\mathcal{I}_\epsilon^C = [K] \setminus \mathcal{I}_\epsilon$. For each arm $i \in \mathcal{I}_\epsilon^C$ and for each player $p \in [M]$ such that $\mu_i^p < \mu_*^p$, we seek to bound the expected number of pulls of arm $i$ by $p$ in $T$ rounds, under the assumption that the event $\mathcal{E}$ occurs. Since the optimal arm(s) may be different for different players, we treat each player separately.

Fix a player $p \in [M]$ and a suboptimal arm $i \in \mathcal{I}_\epsilon^C$ such that $\Delta_i^p > 0$. Recall that $n_i^p(t-1)$ is the number of pulls of arm $i$ by player $p$ after $(t-1)$ rounds. We have

$$\begin{aligned} n_i^p(T) &= \sum_{t=1}^{T} \mathbb{1}\big\{i_t^p = i\big\} \\ &\le \tau + \sum_{t=\tau+1}^{T} \mathbb{1}\big\{i_t^p = i, n_i^p(t-1) > \tau\big\}, \end{aligned} \tag{9}$$

where $\tau \ge 1$ is an arbitrary integer. It then follows that

$$n_i^p(T) \le \tau + \sum_{t=\tau+1}^{T} \mathbb{1}\big\{\mathrm{UCB}_{i_*^p}^p(t) \le \mathrm{UCB}_i^p(t), n_i^p(t-1) > \tau\big\}.$$

With foresight, let $\tau = \lceil \frac{3328 \ln T}{(\Delta_i^p)^2} \rceil$. Conditional on $\mathcal{E}$, we show that, *for any* $i \in \mathcal{I}_\epsilon^C$ *such that* $\Delta_i^p > 0$, the event $\big\{\mathrm{UCB}_{i_*^p}^p(t) \le \mathrm{UCB}_i^p(t), n_i^p(t-1) > \tau\big\}$ never happens. It suffices to show that if $n_i^p(t-1) > \tau$,

$$\mathrm{UCB}_{i_*^p}^p(t) \ge \mu_*^p, \tag{10}$$

and

$$\mathrm{UCB}_i^p(t) < \mu_*^p \tag{11}$$

happen simultaneously.

Eq. (10) follows straightforwardly from the definition of $\mathcal{E}$ along with Lemma 19. For Eq. (11), we have the following upper bound on $\mathrm{UCB}_i^p(t)$:

$$\begin{aligned} \mathrm{UCB}_i^p(t) &= \kappa_i^p(t, \lambda^*) + F(\overline{n_i^p}, \overline{m_i^p}, \lambda^*, \epsilon) \\ &\le \mu_i^p + 2F(\overline{n_i^p}, \overline{m_i^p}, \lambda^*, \epsilon) \\ &= \mu_i^p + 2\Big[\min_{\lambda \in [0,1]} 8\sqrt{13 \ln T [\frac{\lambda^2}{\overline{n_i^p}(t-1)} + \frac{(1-\lambda)^2}{\overline{m_i^p}(t-1)}]} + (1-\lambda)\epsilon\Big] \\ &\le \mu_i^p + 2\Big[8\sqrt{\frac{13 \ln T}{\overline{n_i^p}(t-1)}}\Big] \\ &\le \mu_i^p + 2\Big[8\sqrt{\frac{13 \ln T}{n_i^p(t-1)}}\Big] \\ &< \mu_i^p + 2\Big[8\sqrt{\frac{13 \ln T (\Delta_i^p)^2}{3328 \ln T}}\Big] = \mu_i^p + \Delta_i^p = \mu_*^p, \end{aligned}$$

where the first inequality is from the definition of event $\mathcal{E}$ and Lemma 19; the second inequality is from choosing $\lambda = 1$; the third inequality uses the basic fact that $n_i^p(t-1) \le \overline{n_i^p}(t-1)$; the fourth inequality is by our premise that $n_i^p(t-1) > \tau \ge \frac{3328 \ln T}{(\Delta_i^p)^2}$.

It follows that conditional on $\mathcal{E}$, the second term in Eq. (9) is always zero, i.e., player $p$ would not pull arm $i$ again. Therefore, for any $i \in \mathcal{I}_\epsilon^C$ such that $\Delta_i^p > 0$, we have

$$\mathbb{E}[n_i^p(T)|\mathcal{E}] \le \lceil \frac{3328 \ln T}{(\Delta_i^p)^2} \rceil \le \frac{3328 \ln T}{(\Delta_i^p)^2} + 1 \le \frac{3328 \ln T}{(\Delta_i^p)^2} \cdot 2 = \frac{6656 \ln T}{(\Delta_i^p)^2}. \tag{12}$$

It then follows that

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

This completes the proof of Lemma 21. □

**Proof of Theorem 5.** We now prove Theorem 5.

*Proof.* We have

$$\begin{aligned}\mathbb{E}[\mathcal{R}(T)] &\leq \mathbb{E}[\mathcal{R}(T)|\mathcal{E}] + \mathbb{E}[\mathcal{R}(T)|\overline{\mathcal{E}}]\Pr[\overline{\mathcal{E}}] \\ &\leq \mathbb{E}[\mathcal{R}(T)|\mathcal{E}] + (TM)\frac{4MK}{T^4} \\ &\leq \mathbb{E}[\mathcal{R}(T)|\mathcal{E}] + O(1) \end{aligned} \tag{13}$$

where the second inequality uses the fact that $\mathbb{E}[\mathcal{R}(T)|\overline{\mathcal{E}}] \leq TM$, as the instantaneous regret for each player in each round is bounded by 1; and the last inequality follows under the premise that $T > \max(M, K)$.

Let $\Delta_i^{\max} = \max_p \Delta_i^p$. We have

$$\begin{aligned}\mathbb{E}[\mathcal{R}(T)|\mathcal{E}] &= \sum_{i\in[K]}\sum_{p\in[M]} \mathbb{E}[n_i^p(T)|\mathcal{E}]\cdot\Delta_i^p \\ &\leq \sum_{i\in\mathcal{I}_\epsilon} \mathbb{E}[n_i(T)|\mathcal{E}]\cdot\Delta_i^{\max} + \sum_{i\in\mathcal{I}_\epsilon^C}\sum_{p\in[M]:\Delta_i^p>0} \mathbb{E}[n_i^p(T)|\mathcal{E}]\cdot\Delta_i^p, \end{aligned} \tag{14}$$

where the inequality holds because the instantaneous regret for any arm $i$ and any player $p$ is bounded by $\Delta_i^{\max}$.

Now, it follows from Lemma 20 that there exists some constant $C_1 > 0$ such that for each $i \in \mathcal{I}_\epsilon$,

$$\mathbb{E}[n_i(T)|\mathcal{E}] \leq C_1\left(\frac{\ln T}{(\Delta_i^{\min})^2} + M\right),$$

and it follows from Lemma 21 that there exists some constant $C_2 > 0$ such that for each $i \in \mathcal{I}_\epsilon^C$ and $p \in [M]$ with $\Delta_i^p > 0$,

$$\mathbb{E}[n_i^p(T)|\mathcal{E}] \leq C_2\left(\frac{\ln T}{(\Delta_i^p)^2}\right).$$

Then, continuing Eq. (14), we have

$$\begin{aligned}\mathbb{E}[\mathcal{R}(T)|\mathcal{E}] &\leq \sum_{i\in\mathcal{I}_\epsilon} C_1\left(\frac{\ln T}{(\Delta_i^{\min})^2} + M\right)\cdot\Delta_i^{\max} + \sum_{i\in\mathcal{I}_\epsilon^C}\sum_{p\in[M]:\Delta_i^p>0} C_2\left(\frac{\ln T}{(\Delta_i^p)^2}\right)\cdot\Delta_i^p \\ &\leq 2C_1\sum_{i\in\mathcal{I}_\epsilon}\left(\frac{\ln T}{\Delta_i^{\min}} + M\Delta_i^{\min}\right) + C_2\sum_{i\in\mathcal{I}_\epsilon^C}\sum_{p\in[M]:\Delta_i^p>0}\frac{\ln T}{\Delta_i^p}, \end{aligned}$$

where the second inequality follows from item (c) of Fact 15 which states that $\forall i \in \mathcal{I}_\epsilon, \Delta_i^{\max} < 2\Delta_i^{\min}$.

It then follows from Eq. (13) that

$$\begin{aligned}\mathbb{E}[\mathcal{R}(T)] &\leq \mathbb{E}[\mathcal{R}(T)|\mathcal{E}] + O(1) \\ &\leq O\left(\sum_{i\in\mathcal{I}_\epsilon}\left(\frac{\ln T}{\Delta_i^{\min}} + M\Delta_i^{\min}\right) + \sum_{i\in\mathcal{I}_\epsilon^C}\sum_{p\in[M]:\Delta_i^p>0}\frac{\ln T}{\Delta_i^p}\right), \end{aligned}$$

This completes the proof of Theorem 5. □

### D.5 Proof of Theorem 8

We first restate Theorem 8.

**Theorem 8.** *Let* RobustAgg($\epsilon$) *run on an $\epsilon$-MPMAB problem instance for $T$ rounds. Then its expected collective regret satisfies*

$$\mathbb{E}[\mathcal{R}(T)] \leq \tilde{O}\left(\sqrt{|\mathcal{I}_\epsilon|\, MT} + M\sqrt{(\left|\mathcal{I}_\epsilon^C\right| - 1)T} + M\,|\mathcal{I}_\epsilon|\right).$$

*Proof.* From the earlier proof of Theorem 5, we have

$$\mathbb{E}[\mathcal{R}(T)] \leq \mathbb{E}[\mathcal{R}(T)|\mathcal{E}] + O(1). \tag{15}$$

Recall that $\Delta_i^{\max} = \max_p \Delta_i^p$. We also have

$$\begin{aligned}\mathbb{E}[\mathcal{R}(T)|\mathcal{E}] &= \sum_{i\in[K]}\sum_{p\in[M]} \mathbb{E}[n_i^p(T)|\mathcal{E}]\cdot\Delta_i^p \\ &\leq \sum_{i\in\mathcal{I}_\epsilon} \mathbb{E}[n_i(T)|\mathcal{E}]\cdot\Delta_i^{\max} + \sum_{i\in\mathcal{I}_\epsilon^C}\sum_{p\in[M]:\Delta_i^p>0} \mathbb{E}[n_i^p(T)|\mathcal{E}]\cdot\Delta_i^p \end{aligned}\tag{16}$$

Again, it follows from Lemma 20 that there exists some constant $C_1 > 0$ such that for each $i \in \mathcal{I}_\epsilon$,

$$\mathbb{E}[n_i(T)|\mathcal{E}] \leq C_1\left(\frac{\ln T}{(\Delta_i^{\min})^2} + M\right), \tag{17}$$

and it follows from Lemma 21 that there exists some constant $C_2 > 0$ such that for each $i \in \mathcal{I}_\epsilon^C$ and $p \in [M]$ with $\Delta_i^p > 0$,

$$\mathbb{E}[n_i^p(T)|\mathcal{E}] \leq C_2\left(\frac{\ln T}{(\Delta_i^p)^2}\right). \tag{18}$$

Now let us bound the two terms in Eq. (16) separately, using the technique from (Lattimore and Szepesvári, 2020, Theorem 7.2).

For the first term, with foresight, let us set $\delta_1 = \sqrt{\frac{C_1|\mathcal{I}_\epsilon|\ln T}{MT}}$. If $|\mathcal{I}_\epsilon| = 0$, we have $\sum_{i\in\mathcal{I}_\epsilon}\mathbb{E}[n_i(T)|\mathcal{E}]\cdot\Delta_i^{\max} = 0$ trivially. Otherwise, $\delta_1 > 0$ because $T > \max(M, K) \geq 1$. Then, we have

$$\begin{aligned}&\sum_{i\in\mathcal{I}_\epsilon} \mathbb{E}[n_i(T)|\mathcal{E}]\cdot\Delta_i^{\max} \\ &\leq 2\sum_{i\in\mathcal{I}_\epsilon} \mathbb{E}[n_i(T)|\mathcal{E}]\cdot\Delta_i^{\min} \\ &\leq 2\left(\sum_{i\in\mathcal{I}_\epsilon:\Delta_i^{\min}\in(0,\delta_1)} \mathbb{E}[n_i(T)|\mathcal{E}]\cdot\Delta_i^{\min} + \sum_{i\in\mathcal{I}_\epsilon:\Delta_i^{\min}\in[\delta_1,1]} \mathbb{E}[n_i(T)|\mathcal{E}]\cdot\Delta_i^{\min}\right) \\ &\leq 2\left(MT\delta_1 + \sum_{i\in\mathcal{I}_\epsilon:\Delta_i^{\min}\in[\delta_1,1]} C_1\left(\frac{\ln T}{(\Delta_i^{\min})^2} + M\right)\Delta_i^{\min}\right) \\ &\leq 2\left(MT\delta_1 + \sum_{i\in\mathcal{I}_\epsilon:\Delta_i^{\min}\in[\delta_1,1]} \frac{C_1\ln T}{\Delta_i^{\min}} + C_1\sum_{i\in\mathcal{I}_\epsilon:\Delta_i^{\min}\in[\delta_1,1]} M\Delta_i^{\min}\right) \\ &\leq 2\left(MT\delta_1 + \frac{C_1|\mathcal{I}_\epsilon|\ln T}{\delta_1} + C_1\sum_{i\in\mathcal{I}_\epsilon} M\Delta_i^{\min}\right) \\ &\leq 4\sqrt{C_1|\mathcal{I}_\epsilon|MT\ln T} + 2C_1M|\mathcal{I}_\epsilon|, \end{aligned}\tag{19}$$

where the first inequality follows from item (c) of Fact 15; the third inequality follows from Eq. (17) and the fact that $\sum_{i\in\mathcal{I}_\epsilon}\mathbb{E}[n_i(T)|\mathcal{E}]\le MT$ as $M$ players each pulls one arm in each of $T$ rounds; and the last inequality follows from our premise that $\delta_1=\sqrt{\frac{C_1|\mathcal{I}_\epsilon|\ln T}{MT}}$.

For the second term, we consider two cases:

**Case 1:** $|\mathcal{I}_\epsilon^C|=1$**.** In this case, as we have discussed in the paper, $\mathcal{I}_\epsilon^C$ is a singleton set $\{i_*\}$ where arm $i_*$ is optimal for all players $p$; that is, $\Delta_{i_*}^p=0$ for all $p\in[M]$. We therefore have

$$\sum_{i\in\mathcal{I}_\epsilon^C}\sum_{p\in[M]:\Delta_i^p>0}\mathbb{E}[n_i^p(T)|\mathcal{E}]\cdot\Delta_i^p=0=4M\sqrt{C_2\big(|\mathcal{I}_\epsilon^C|-1\big)T\ln T}. \tag{20}$$

**Case 2:** $|\mathcal{I}_\epsilon^C|\ge 2$**.** With foresight, let us set $\delta_2=\sqrt{\frac{C_2|\mathcal{I}_\epsilon^C|\ln T}{T}}$.

$$\begin{aligned}
&\sum_{i\in\mathcal{I}_\epsilon^C}\sum_{p\in[M]:\Delta_i^p>0}\mathbb{E}[n_i^p(T)|\mathcal{E}]\cdot\Delta_i^p\\
\le&\sum_{i\in\mathcal{I}_\epsilon^C}\sum_{p\in[M]:\Delta_i^p\in(0,\delta_2)}\mathbb{E}[n_i^p(T)|\mathcal{E}]\cdot\Delta_i^p+\sum_{i\in\mathcal{I}_\epsilon^C}\sum_{p\in[M]:\Delta_i^p\in[\delta_2,1]}\mathbb{E}[n_i^p(T)|\mathcal{E}]\cdot\Delta_i^p\\
\le&MT\delta_2+\sum_{i\in\mathcal{I}_\epsilon^C}\sum_{p\in[M]:\Delta_i^p\in[\delta_2,1]}C_2\left(\frac{\ln T}{(\Delta_i^p)^2}\right)\Delta_i^p\\
\le&MT\delta_2+\sum_{i\in\mathcal{I}_\epsilon^C}\sum_{p\in[M]:\Delta_i^p\in[\delta_2,1]}\left(\frac{C_2\ln T}{\Delta_i^p}\right)\\
\le&MT\delta_2+\frac{C_2M|\mathcal{I}_\epsilon^C|\ln T}{\delta_2}\\
\le&4M\sqrt{C_2(|\mathcal{I}_\epsilon^C|-1)T\ln T},
\end{aligned}\tag{21}$$

where the second inequality follows from Eq. (18) and the fact that $\sum_{i\in\mathcal{I}_\epsilon^C}\mathbb{E}[n_i(T)|\mathcal{E}]\le MT$ as $M$ players each pulls one arm in each of $T$ rounds; and the last inequality follows from our premise that $\delta_2=\sqrt{\frac{C_2|\mathcal{I}_\epsilon^C|\ln T}{T}}$ and $|\mathcal{I}_\epsilon^C|\le 2(|\mathcal{I}_\epsilon^C|-1)$.

In summary, from Eqs. (20) and (21), we have in both cases,

$$\sum_{i\in\mathcal{I}_\epsilon^C}\sum_{p\in[M]:\Delta_i^p>0}\mathbb{E}[n_i^p(T)|\mathcal{E}]\cdot\Delta_i^p\le 4M\sqrt{C_2(|\mathcal{I}_\epsilon^C|-1)T\ln T}. \tag{22}$$

Combining Eq. (19) and Eq. (22), we have

$$\mathbb{E}[\mathcal{R}(T)|\mathcal{E}]\le 4\sqrt{C_1|\mathcal{I}_\epsilon|MT\ln T}+2C_1M|\mathcal{I}_\epsilon|+4M\sqrt{C_2(|\mathcal{I}_\epsilon^C|-1)T\ln T}$$

It then follows from Eq. (15) that

$$\begin{aligned}
\mathbb{E}[\mathcal{R}(T)]&\le 4\sqrt{C_1|\mathcal{I}_\epsilon|MT\ln T}+4M\sqrt{C_2(|\mathcal{I}_\epsilon^C|-1)T\ln T}+2C_1M|\mathcal{I}_\epsilon|+O(1)\\
&\le O\left(\sqrt{|\mathcal{I}_\epsilon|\,MT\ln T}+M\sqrt{(|\mathcal{I}_\epsilon^C|-1)T\ln T}+M\,|\mathcal{I}_\epsilon|\right)\\
&\le \tilde{O}\left(\sqrt{|\mathcal{I}_\epsilon|\,MT}+M\sqrt{(|\mathcal{I}_\epsilon^C|-1)T}+M\,|\mathcal{I}_\epsilon|\right).
\end{aligned}$$

This completes the proof of Theorem 8. □

# E Proof of the lower bounds

## E.1 Gap-independent lower bound with known $\epsilon$

We first restate Theorem 10.

**Theorem 10.** *For any $K \geq 2, M, T \in \mathbb{N}$ such that $T \geq K$, and $l, l^C$ in $\mathbb{N}$ such that $l \leq K-1, l+l^C = K$, there exists some $\epsilon > 0$, such that for any algorithm $\mathcal{A}$, there exists an $\epsilon$-MPMAB problem instance, in which $|\mathcal{I}_\epsilon| = l$, and $\mathcal{A}$ has a collective regret at least $\Omega(M\sqrt{(l^C-1)T} + \sqrt{MlT})$.*

*Proof.* Fix algorithm $\mathcal{A}$. We consider two cases regarding the comparison between $l$ and $Ml^C$.

**Case 1: $l > Ml^C$.** To simplify notations, define $\Delta = \sqrt{\frac{l+1}{24MT}}$. Observe that $\Delta \leq \frac{1}{4}$ as $T \geq K \geq l+1$. We will set $\epsilon = \frac{\Delta}{10}$.

We will now define $l$ different Bernoulli $\epsilon$-MPMAB instances, and show that under at least one of them, $\mathcal{A}$ will have a collective regret at least $\frac{1}{96}\sqrt{MlT} \geq \frac{1}{192}(M\sqrt{(l^C-1)T} + \sqrt{MlT})$.

For $j$ in $[l+1]$, define a Bernoulli MPMAB instance $E_j$ to be such that for all players $p$ in $[M]$, the expected reward of arm $i$,

$$\mu_i^p = \begin{cases} \frac{1}{2} + \Delta & i = j \\ \frac{1}{2} & i \in [l+1] \setminus \{j\} \\ \frac{1}{2} + \frac{3\Delta}{4} & i \notin [l+1] \end{cases}.$$

We first verify that for every instance $E_j$, it (1) is an $\epsilon$-MPMAB instance, and (2) $\mathcal{I}_\epsilon = [l+1] \setminus \{j\}$ and therefore has size $l$:

1. For item (1), observe that for any fixed $i$, we have $\mu_i^p$ share the same value across all player $p$'s. Therefore, the is trivially $\epsilon$-dissimilar.

2. For item (2), first, for all $i$ in $[l+1] \setminus \{j\}$, we have $\Delta_i^p = \Delta > 5\epsilon = \frac{\Delta}{2}$ for all $p$; this implies that $[l+1] \setminus \{j\}$ is a subset of $\mathcal{I}_\epsilon$. On the other hand, for all $i$ in $([K] \setminus [l+1]) \cup \{j\}$, we have that for all $p$, $\Delta_i^p$ is either 0 or $\frac{\Delta}{4}$, both of which are $\leq \frac{\Delta}{2} = 5\epsilon$ for all $p$; this implies that all elements of $([K] \setminus [l+1]) \cup \{j\}$ are outside $\mathcal{I}_\epsilon$.

We will now argue that

$$\mathbb{E}_{j \sim \mathrm{Unif}([l+1])} \mathbb{E}_{E_j} \left[\mathcal{R}(T)\right] \geq \frac{1}{96}\sqrt{MlT}. \tag{23}$$

To this end, it suffices to show

$$\mathbb{E}_{j \sim \mathrm{Unif}([l+1])} \mathbb{E}_{E_j} \left[MT - n_i(T)\right] \geq \frac{MT}{4}. \tag{24}$$

To see why Eq. (24) implies Eq. (23), recall that under instance $E_j$, $j$ is the optimal arm for all players. In this instance, $\mathcal{R}(T) = \sum_{i \neq j} n_i(T)\Delta_i^1$. As under $E_j$, for all $i \neq j$ and all $p$, $\Delta_i^1 \geq \frac{\Delta}{4}$, we have $\mathcal{R}(T) \geq \frac{\Delta}{4} \cdot (MT - n_i(T))$. Eq. (23) follows from combining this inequality with Eq. (24), along with some algebra.

We now come back to the proof of Eq. (24). First, we define a helper instance $E_0$, such that for all players $p$ in $[M]$, the expected reward of arm $i$ is defined as:

$$\mu_i^p = \begin{cases} \frac{1}{2} & i \in [l+1] \\ \frac{1}{2} + \frac{3\Delta}{4} & i \notin [l+1] \end{cases}$$

In addition, for all $i$ in $\{0\} \cup [l+1]$, define $\mathbb{P}_i$ as the joint distribution of the interaction logs (arm pulls and rewards) for all $M$ players over a horizon of $T$; furthermore, denote by $\mathbb{E}_i$ expectation with respect to $\mathbb{P}_i$.

For every $i$ in $[l+1]$, we have

$$\begin{aligned} d_{\mathrm{TV}}(\mathbb{P}_0, \mathbb{P}_i) &= \frac{1}{2}\|\mathbb{P}_0 - \mathbb{P}_i\|_1 \\ &\le \frac{1}{2}\sqrt{2\,\mathrm{KL}(\mathbb{P}_0, \mathbb{P}_i)} \\ &\le \frac{1}{2}\sqrt{2\,\mathrm{KL}(\mathrm{Ber}(0.5, 0.5+\Delta))\mathbb{E}_0[n_i(T)]} \\ &\le \sqrt{\frac{3}{2}\mathbb{E}_0[n_i(T)]\Delta^2} \\ &= \frac{1}{4}\sqrt{\frac{l+1}{MT}\mathbb{E}_0[n_i(T)]} \end{aligned}$$

where the first equality is from $d_{\mathrm{TV}}(\mathbb{P}, \mathbb{Q}) = \frac{1}{2}\|\mathbb{P} - \mathbb{Q}\|_1$ for any two distributions $\mathbb{P}$, $\mathbb{Q}$; the first inequality uses Pinsker's inequality; the second inequality is from the well-known divergence decomposition lemma (e.g. Lattimore and Szepesvári (2020), Lemma 15.1); the third inequality uses Lemma 25; and the last equality is by recalling that $\Delta \in [0, \frac{1}{4}]$ and algebra.

Now, applying Lemma 24 with $m = l + 1 \ge 2$, $N_i = n_i(T)$ for all $i$ in $[l+1]$, and $B = MT$, Eq. (24) is proved. This in turn finishes the proof of the regret lower bound.

**Case 2:** $M(l^C - 1) \ge l$**.** To simplify notations, define $\Delta = \sqrt{\frac{l^C}{24T}} \in [0, \frac{1}{4}]$. Observe that $\Delta \le \frac{1}{4}$ as $T \ge K \ge l^C$. In addition, we must have $l^C \ge 2$ in this case, as if $l^C = 1$, $M(l^C - 1) = 0 < K = l$. We set $\epsilon = \frac{\Delta}{2}$.

We will now define $[l^C]^M$ different Bernoulli $\epsilon$-MPMAB instances, and show that under at least one of them, $\mathcal{A}$ will have a collective regret at least $\frac{1}{24}M\sqrt{l^C T} \ge \frac{1}{192}(M\sqrt{(l^C - 1)T} + \sqrt{MlT})$.

For $i_1, \ldots, i_M \in [l^C]^M$, define Bernoulli MPMAB instance $E_{i_1,\ldots,i_M}$ to be such that for $p$ in $[M]$ and $i$ in $[K]$, the expected reward of player $p$ on pulling arm $i$ is

$$\mu_i^p = \begin{cases} \frac{1}{2} + \Delta & i = i_p \\ \frac{1}{2} & i \in [l^C] \setminus \{i_p\} \\ 0 & i \notin [l^C] \end{cases}$$

We first verify that for every $i_1, \ldots, i_M$, instance $E_{i_1,\ldots,i_M}$ (1) is an $\epsilon$-MPMAB instance, and (2) $\mathcal{I}_\epsilon = [K] \setminus [l^C]$, and therefore has size $l$:

1. For item (1), observe that for all $i$ in $[l^C]$ and all $p$ in $[M]$, $\mu_i^p \in \{\frac{1}{2}, \frac{1}{2} + \Delta\}$; therefore, for every $p, q$, $\left|\mu_i^p - \mu_i^q\right| \le \Delta = \epsilon$. Meanwhile, for all $i$ in $[K] \setminus [l^C]$ and all $p$ in $[M]$, $\mu_i^p = 0$, implying that for every $p, q$, $\left|\mu_i^p - \mu_i^q\right| = 0 \le \epsilon$. Therefore $E_{i_1,\ldots,i_M}$ is $\epsilon$-dissimilar.
2. For item (2), we first observe that for all $i$ in $[l^C]$, for all $p$, $\Delta_i^p$ is 0 or $\Delta$, both of which are $\le \frac{5}{2}\Delta = 5\epsilon$. Therefore all elements of $[l^C]$ are outside $\mathcal{I}_\epsilon$. Meanwhile, for all $i$ in $[K]\setminus[l^C]$ and all $p$, $\Delta_i^p = \frac{1}{2} + \Delta \ge \frac{5}{2}\Delta = 5\epsilon$. This implies that all elements of $[K] \setminus [l^C]$ are in $\mathcal{I}_\epsilon$. Item (2) follows.

We will now argue that

$$\mathbb{E}_{(i_1,\ldots,i_M)\sim\mathrm{Unif}([l^C]^M)}\mathbb{E}_{E_{i_1,\ldots,i_M}}\left[\mathcal{R}(T)\right] \ge \frac{M\sqrt{l^C T}}{24}.$$

As the roles of all $M$ players are the same, by symmetry, it suffices to show that the expected regret of player 1 satisfies

$$\mathbb{E}_{(i_1,\ldots,i_M)\sim\mathrm{Unif}([l^C]^M)}\mathbb{E}_{E_{i_1,\ldots,i_M}}\left[\mathcal{R}^1(T)\right] \ge \frac{\sqrt{l^C T}}{24}. \tag{25}$$

It therefore suffices to show,

$$\mathbb{E}_{(i_1,\ldots,i_M)\sim\mathrm{Unif}([l^C]^M)}\mathbb{E}_{E_{i_1,\ldots,i_M}}\left[T - n_{i_1}^1(T)\right] \ge \frac{T}{4}. \tag{26}$$

This is because, recall that when $i_1$ is the optimal arm for player 1, $\mathcal{R}(T) = \sum_{i=1}^{K} n_i^1(T)\Delta_i^1 = \sum_{i\neq i_1} n_i^1(T)\Delta_i^1$; in addition, for all $i \neq i_1$, $\Delta_i^1 \geq \Delta$. This implies that $\mathcal{R}^1(T) \geq \Delta(T - n_{i_1}^1(T))$. Eq. (25) follows from the above inequality, Eq. (26), and the definition of $\Delta$.

We now come back to the proof of Eq. (26). To this end, we define the following set of "helper" instances to facilitate our reasoning. Given $i_2, \ldots, i_M \in [K]^{M-1}$, define instance $E_{0,i_2,\ldots,i_M}$ such that its reward distribution is identical to $E_{i_1,i_2,\ldots,i_M}$ except for player 1 on arm $i_1$. Formally, it has the following expected reward profile:

$$\text{for } p = 1, \mu_i^1 = \begin{cases} \frac{1}{2} & i \in [l^C] \\ 0 & i \notin [l^C] \end{cases} \quad \text{for } p \neq 1, \mu_i^p = \begin{cases} \frac{1}{2} + \Delta & i = i_p \\ \frac{1}{2} & i \in [l^C] \setminus \{i_p\} \\ 0 & i \notin [l^C] \end{cases}$$

In addition, for all $i_1, \ldots, i_M$ in $(\{0\} \cup [l^C]) \times [l^C]^{M-1}$, define $\mathbb{P}_{i_1,\ldots,i_M}$ as the joint distribution of the interaction logs (arm pulls and rewards) for all $M$ players over a horizon of $T$; furthermore, for $i$ in $(\{0\} \cup [l^C])$, define $\mathbb{P}_i = \frac{1}{(l^C)^{M-1}} \sum_{i_2,\ldots,i_M \in [l^C]^{M-1}} \mathbb{P}_{i,i_2,\ldots,i_M}$, and denote by $\mathbb{E}_i$ expectation with respect to $\mathbb{P}_i$. In this notation, Eq. (26) can be rewritten as

$$\frac{1}{l^C} \sum_{i=1}^{l^C} \mathbb{E}_i[T - n_i^1(T)] \geq \frac{T}{2}.$$

For every $i$ in $[l^C]$,

$$\begin{aligned}
d_{\mathrm{TV}}(\mathbb{P}_0, \mathbb{P}_i) &= \frac{1}{2}\|\mathbb{P}_0 - \mathbb{P}_i\|_1 \\
&= \frac{1}{2}\left\| \frac{1}{l^{C^{M-1}}} \sum_{i_2,\ldots,i_M \in [l^C]^{M-1}} (\mathbb{P}_{0,i_2,\ldots,i_M} - \mathbb{P}_{i,i_2,\ldots,i_M}) \right\|_1 \\
&\leq \frac{1}{(l^C)^{M-1}} \cdot \sum_{i_2,\ldots,i_M \in [l^C]^{M-1}} \frac{1}{2}\|\mathbb{P}_{0,i_2,\ldots,i_M} - \mathbb{P}_{i,i_2,\ldots,i_M}\|_1 \\
&\leq \frac{1}{(l^C)^{M-1}} \cdot \sum_{i_2,\ldots,i_M \in [l^C]^{M-1}} \sqrt{\frac{1}{2}\,\mathrm{KL}(\mathrm{Ber}(0.5, 0.5+\Delta)) \cdot \mathbb{E}_{0,i_2,\ldots,i_M}[N_i^1(T)]} \\
&\leq \frac{1}{(l^C)^{M-1}} \cdot \sum_{i_2,\ldots,i_M \in [l^C]^{M-1}} \sqrt{\frac{3}{2}\Delta^2 \cdot \mathbb{E}_{0,i_2,\ldots,i_M}[N_i^1(T)]} \\
&\leq \sqrt{\frac{3}{2}\Delta^2 \cdot \mathbb{E}_0[N_i^1(T)]}
\end{aligned}$$

where the first equality is from $d_{\mathrm{TV}}(\mathbb{P}, \mathbb{Q}) = \frac{1}{2}\|\mathbb{P} - \mathbb{Q}\|_1$ for any two distributions $\mathbb{P}$, $\mathbb{Q}$; the second equality is from the definition of $\mathbb{P}_i$, $i \in \{0\} \cup [l^C]$; the first inequality is from triangle inequality of $\ell_1$ norm; the second inequality is from Pinsker's inequality, and the divergence decomposition lemma (Lattimore and Szepesvári (2020), Lemma 15.1); the third inequality is from Lemma 25 and recalling that $\Delta \in [0, \frac{1}{4}]$; the last inequality is from Jensen's inequality, and the definition of $\mathbb{P}_0$.

Applying Lemma 24 with $m = l^C \geq 2$, $N_i = n_i^1(T)$ for all $i$ in $[l^C]$ and $B = T$, Eq. (26) is proved. This in turn finishes the proof of the regret lower bound. □

### E.2 Gap-dependent lower bounds with known $\epsilon$

We restate Theorem 9 here with specifications of exact constants in the lower bound.

**Theorem 22** (Restatement of Theorem 9). *Fix $\epsilon \geq 0$ and $\alpha, C > 0$. Let $\mathcal{A}$ be an algorithm such that $\mathcal{A}$ has at most $CT^{1-\alpha}$ regret in all $\epsilon$-MPMAB problem instances. Then, for any Bernoulli $\frac{\epsilon}{2}$-MPMAB instance*

*$\mu = (\mu_i^p)_{i\in[K],p\in[M]}$ such that $\mu_i^p \in [\frac{15}{32}, \frac{17}{32}]$ for all $i$ and $p$, we have:*

$$\mathbb{E}_\mu[\mathcal{R}(T)] \geq \sum_{i\in\mathcal{I}_{\epsilon/20}^C} \sum_{p\in[M]:\Delta_i^p>0} \frac{\ln\left(\Delta_i^p T^\alpha/8C\right)}{12\Delta_i^p} + \sum_{i\in\mathcal{I}_{\epsilon/20}:\Delta_i^{\min}>0} \frac{\ln\left(\Delta_i^{\min} T^\alpha/8C\right)}{12\Delta_i^{\min}}.$$

*Proof.* We will first prove the following two claims:

1. For any $i_0$ in $[K]$ such that $\Delta_{i_0}^{\min} > 0$, $\mathbb{E}_\mu\left[n_{i_0}(T)\right] \geq \frac{\ln\left(\Delta_{i_0}^{\min} T^\alpha/8C\right)}{12(\Delta_{i_0}^{\min})^2}$.

2. For any $i_0$ in $\mathcal{I}_{\epsilon/20}^C$ and any $p_0$ in $[M]$ such that $\Delta_{i_0}^{p_0} > 0$, $\mathbb{E}_\mu\left[n_{i_0}^{p_0}(T)\right] \geq \frac{\ln\left(\Delta_{i_0}^{p_0} T^\alpha/8C\right)}{12(\Delta_{i_0}^{p_0})^2}$.

The proof of these two claims are as follows:

1. Fix $i_0$ in $[K]$ such that $\Delta_{i_0}^{\min} > 0$, i.e., $\Delta_{i_0}^p > 0$ for all $p$ in $[M]$. Define $p_0 = \operatorname{argmin}_{p\in[M]} \Delta_{i_0}^p$.
   We consider a new Bernoulli MPMAB instance, with mean reward defined as follows:

$$\forall p \in [M], \quad \nu_i^p = \begin{cases} \mu_i^p + 2\Delta_{i_0}^{p_0}, & i = i_0, \\ \mu_i^p & \text{otherwise} \end{cases}$$

   We have the following key observations:

   (a) $\nu$ is an $\epsilon$-MPMAB instance; this is because $\nu_i^p - \nu_i^q = \mu_i^p - \mu_i^q$ for any $p, q$ in $[M]$ and $i$ in $[K]$, and $\mu$ is an $\frac{\epsilon}{2}$-MPMAB instance. By our assumption that $\mathcal{A}$ has $CT^{1-\alpha}$ regret on all $\epsilon$-MPMAB environments, we have

$$\mathbb{E}_\mu\left[\mathcal{R}(T)\right] \leq CT^{1-\alpha}, \quad \mathbb{E}_\nu\left[\mathcal{R}(T)\right] \leq CT^{1-\alpha}. \tag{27}$$

   (b) By the divergence decomposition lemma (Lattimore and Szepesvári (2020), Lemma 15.1),

$$\mathrm{KL}(\mathbb{P}_\mu, \mathbb{P}_\nu) = \sum_{p=1}^M \mathbb{E}_\mu\left[n_{i_0}^p(T)\right] \mathrm{KL}\left(\mathrm{Ber}(\mu_{i_0}^p), \mathrm{Ber}(\mu_{i_0}^p + 2\Delta_{i_0}^{p_0})\right), \tag{28}$$

   As for all $p$, $\mu_{i_0}^p \in [\frac{15}{32}, \frac{17}{32}]$, and $\Delta_{i_0}^p \leq \frac{1}{16}$, using Lemma 25, we have that for all $p$,

$$\mathrm{KL}\left(\mathrm{Ber}(\mu_i^p), \mathrm{Ber}(\mu_i^p + 2\Delta_{i_0}^{p_0})\right) \leq 3 \cdot (2\Delta_{i_0}^{p_0})^2 = 12(\Delta_{i_0}^{p_0})^2.$$

   Plugging into Eq. (28), we get

$$\mathrm{KL}(\mathbb{P}_\mu, \mathbb{P}_\nu) \leq \sum_{p=1}^M (\mathbb{E}_\mu\left[n_{i_0}^p(T)\right] \cdot 12(\Delta_{i_0}^{p_0})^2) = 12\mathbb{E}_\mu\left[n_{i_0}(T)\right](\Delta_{i_0}^{p_0})^2. \tag{29}$$

   (c) Under MPMAB instance $\mu$, $\mathcal{R}(T) \geq \mathcal{R}^{p_0}(T) \geq \Delta_{i_0}^{p_0} n_{i_0}^{p_0}(T) \geq \frac{\Delta_{i_0}^{p_0} T}{2} \mathbb{1}\{n_{i_0}^{p_0}(T) \geq \frac{T}{2}\}$. Taking expectations, we get,

$$\mathbb{E}_\mu\left[\mathcal{R}(T)\right] \geq \frac{\Delta_{i_0}^{p_0} T}{2} \mathbb{P}_\mu(n_{i_0}^{p_0}(T) \geq \frac{T}{2}). \tag{30}$$

   Likewise, under MPMAB instance $\nu$, for player $p_0$, $\mathcal{R}(T) \geq \mathcal{R}^{p_0}(T) \geq \Delta_{i_0}^{p_0}(T - n_{i_0}^{p_0}(T)) \geq \frac{\Delta_{i_0}^{p_0} T}{2} \mathbb{1}\{n_{i_0}^{p_0}(T) < \frac{T}{2}\}$. Taking expectations, we get,

$$\mathbb{E}_\nu\left[\mathcal{R}(T)\right] \geq \frac{\Delta_{i_0}^{p_0} T}{2} \mathbb{P}_\nu(n_{i_0}^{p_0}(T) < \frac{T}{2}). \tag{31}$$

Adding up Eq. (30) and Eq. (31), we have

$$\mathbb{E}_\nu\left[\mathcal{R}(T)\right]+\mathbb{E}_\mu\left[\mathcal{R}(T)\right]\geq\frac{\Delta_{i_0}^{p_0}T}{2}\left(\mathbb{P}_\nu(n_{i_0}^{p_0}(T)<\frac{T}{2})+\mathbb{P}_\mu(n_{i_0}^{p_0}(T)\geq\frac{T}{2})\right). \tag{32}$$

From Eq. (27), we have the left hand side is at most $2CT^{1-\alpha}$. By Bretagnolle-Huber inequality (see Lemma 26) and Eq. (29), we have that $\mathbb{P}_\nu(n_{i_0}^{p_0}(T)<\frac{T}{2})+\mathbb{P}_\mu(n_{i_0}^{p_0}(T)\geq\frac{T}{2})\geq\frac{1}{2}\exp(-\mathrm{KL}(\mathbb{P}_\mu,\mathbb{P}_\mu))\geq\frac{1}{2}\exp(-12\mathbb{E}_\mu\left[n_{i_0}(T)\right](\Delta_{i_0}^{p_0})^2)$. Plugging these to Eq. (32), we get

$$2CT^{1-\alpha}\geq\frac{\Delta_{i_0}^{p_0}T}{4}\exp(-12\mathbb{E}_\mu\left[n_{i_0}(T)\right](\Delta_{i_0}^{p_0})^2).$$

Solving for $\mathbb{E}_\mu\left[n_{i_0}(T)\right]$, we conclude that

$$\mathbb{E}_\mu\left[n_{i_0}(T)\right]\geq\frac{1}{12(\Delta_{i_0}^{p_0})^2}\cdot\ln\left(\frac{\Delta_{i_0}^{p_0}T^\alpha}{8C}\right)=\frac{1}{12(\Delta_{i_0}^{\min})^2}\cdot\ln\left(\frac{\Delta_{i_0}^{\min}T^\alpha}{8C}\right).$$

2. Fix $i_0$ in $\mathcal{I}_{\epsilon/20}^C$ and $p_0\in[M]$ such that $\Delta_{i_0}^{p_0}>0$. By definition of $\mathcal{I}_{\epsilon/20}^C$, we also have $\Delta_{i_0}^{p_0}=\mu_*^{p_0}-\mu_{i_0}^{p_0}\leq\epsilon/4$. We consider a new MPMAB environment $\nu$, with mean reward defined as follows:

$$\nu_i^p=\begin{cases}\mu_i^p+2\Delta_{i_0}^{p_0} & i=i_0,p=p_0\\ \mu_i^p & \text{otherwise}\end{cases}$$

Same as before, we have the following three key observations:

(a) $\nu$ is an $\epsilon$-MPMAB instance; this is because $(\nu_i^p-\nu_i^q)-(\mu_i^p-\mu_i^q)\in\left\{-\frac{\epsilon}{2},0,\frac{\epsilon}{2}\right\}$ for any $p,q$ in $[M]$ and $i$ in $[K]$, and $\mu$ is an $\frac{\epsilon}{2}$-MPMAB instance. By our assumption that $\mathcal{A}$ has $CT^{1-\alpha}$ regret on all $\epsilon$-MPMAB problem instances, we have

$$\mathbb{E}_\mu\left[\mathcal{R}(T)\right]\leq CT^{1-\alpha},\quad\mathbb{E}_\nu\left[\mathcal{R}(T)\right]\leq CT^{1-\alpha}. \tag{33}$$

(b) By the divergence decomposition lemma (Lattimore and Szepesvári (2020), Lemma 15.1),

$$\mathrm{KL}(\mathbb{P}_\mu,\mathbb{P}_\nu)=\mathbb{E}_\mu\left[n_{i_0}^{p_0}(T)\right]\mathrm{KL}\left(\mathrm{Ber}(\mu_{i_0}^{p_0}),\mathrm{Ber}(\mu_{i_0}^{p_0}+2\Delta_{i_0}^{p_0})\right)\leq 12\mathbb{E}_\mu\left[n_{i_0}^{p_0}(T)\right](\Delta_{i_0}^{p_0})^2, \tag{34}$$

where the second equality uses the following observation: $\mu_{i_0}^{p_0}\in[\frac{15}{32},\frac{17}{32}]$, and $\Delta_{i_0}^{p_0}\leq\frac{1}{16}$, using Lemma 25, $\mathrm{KL}\left(\mathrm{Ber}(\mu_{i_0}^{p_0}),\mathrm{Ber}(\mu_{i_0}^{p_0}+2\Delta_{i_0}^{p_0})\right)\leq 3\cdot(2\Delta_{i_0}^{p_0})^2=12(\Delta_{i_0}^{p_0})^2$.

(c) Under MPMAB instance $\mu$, $\mathcal{R}(T)\geq\mathcal{R}^{p_0}(T)\geq\Delta_{i_0}^{p_0}n_{i_0}^{p_0}(T)\geq\frac{\Delta_{i_0}^{p_0}T}{2}\mathbb{1}\{n_{i_0}^{p_0}(T)\geq\frac{T}{2}\}$. Taking expectations, we get,

$$\mathbb{E}_\mu\left[\mathcal{R}(T)\right]\geq\frac{\Delta_{i_0}^{p_0}T}{2}\mathbb{P}_\mu(n_{i_0}^{p_0}(T)\geq\frac{T}{2}). \tag{35}$$

Likewise, under MPMAB instance $\nu$, for player $p_0$, $\mathcal{R}(T)\geq\mathcal{R}^{p_0}(T)\geq\Delta_{i_0}^{p_0}(T-n_{i_0}^{p_0}(T))\geq\frac{\Delta_{i_0}^{p_0}T}{2}\mathbb{1}\{n_{i_0}^{p_0}(T)<\frac{T}{2}\}$. Taking expectations, we get,

$$\mathbb{E}_\nu\left[\mathcal{R}(T)\right]\geq\frac{\Delta_{i_0}^{p_0}T}{2}\mathbb{P}_\nu(n_{i_0}^{p_0}(T)<\frac{T}{2}). \tag{36}$$

Same as the proof of item 1, combining Equations (33), (34), (35), (36), and using Bretagnolle-Huber inequality, we get

$$\mathbb{E}_\mu\left[n_{i_0}^{p_0}(T)\right]\geq\frac{1}{12(\Delta_{i_0}^{p_0})^2}\cdot\ln\left(\frac{\Delta_{i_0}^{p_0}T^\alpha}{8C}\right).$$

We now use the above two claims to conclude the proof. Recall that $\mathbb{E}_\mu\left[\mathcal{R}(T)\right] = \sum_{i\in[K]}\sum_{p\in[M]}\Delta_i^p\mathbb{E}_\mu\left[n_i^p(T)\right]$. For $i$ in $\mathcal{I}_{\epsilon/20}$ such that $\Delta_i^{\min} > 0$, item 1 implies:

$$\sum_{p\in[M]}\Delta_i^p\mathbb{E}_\mu\left[n_i^p(T)\right] \geq \Delta_i^{\min}\sum_{p\in[M]}\mathbb{E}_\mu\left[n_i^p(T)\right] \geq \frac{1}{12\Delta_i^{\min}}\cdot\ln\left(\frac{\Delta_i^{\min}T^\alpha}{8C}\right).$$

For $i$ in $\mathcal{I}_{\epsilon/20}^C$, item 2 implies:

$$\sum_{p\in[M]}\Delta_i^p\mathbb{E}_\mu\left[n_i^p(T)\right] \geq \sum_{p\in[M]:\Delta_i^p>0}\frac{1}{12\Delta_i^p}\cdot\ln\left(\frac{\Delta_i^pT^\alpha}{8C}\right).$$

Summing over all $i$ in $[K]$ on the above two inequalities, we have

$$\begin{aligned}\mathbb{E}_\mu\left[\mathcal{R}(T)\right] &= \sum_{i\in[K]}\sum_{p\in[M]}\Delta_i^p\mathbb{E}_\mu\left[n_i^p(T)\right]\\ &\geq \sum_{i\in\mathcal{I}_{\epsilon/20}^C}\sum_{p\in[M]:\Delta_i^p>0}\frac{1}{12\Delta_i^p}\cdot\ln\left(\frac{\Delta_i^pT^\alpha}{8C}\right) + \sum_{i\in\mathcal{I}_{\epsilon/20}:\Delta_i^{\min}>0}\frac{1}{12\Delta_i^{\min}}\cdot\ln\left(\frac{\Delta_i^{\min}T^\alpha}{8C}\right).\end{aligned}$$ □

**Remark.** The above lower bound argument aligns with our intuition that arms that are near-optimal with respect to some players (i.e., those in $\mathcal{I}_{\epsilon/20}^C$) are harder for information sharing: in addition to a lower bound on the collective number of pulls to it across all players (item 1 of the claim), we are able to show a stronger lower bound on the number of pulls to it from *each player* (item 2 of the claim).

### E.3 Gap-dependent lower bounds with unknown $\epsilon$

We restate Theorem 11 here with specifications of exact constants in the lower bound.

**Theorem 23** (Restatement of Theorem 11)**.** *Fix $\alpha, C > 0$. Let $\mathcal{A}$ be an algorithm such that $\mathcal{A}$ has at most $CT^{1-\alpha}$ regret in all MPMAB problem instances. Then, for any Bernoulli MPMAB instance $\mu = (\mu_i^p)_{i\in[K],p\in[M]}$ such that $\mu_i^p \in [\frac{15}{32}, \frac{17}{32}]$ for all $i$ and $p$, we have:*

$$\mathbb{E}_\mu[\mathcal{R}(T)] \geq \sum_{i\in[K]}\sum_{p\in[M]:\Delta_i^p>0}\frac{\ln\left(\Delta_i^pT^\alpha/8C\right)}{12\Delta_i^p}.$$

*Proof.* Recall that $\mathbb{E}_\mu\left[\mathcal{R}(T)\right] = \sum_{i=1}^K\sum_{p=1}^M\Delta_i^p\mathbb{E}_\mu\left[N_i^p(T)\right]$; it suffices to show that for any $i_0$ in $[K]$ and any $p_0$ in $[M]$ such that $\Delta_{i_0}^{p_0} > 0$, $\mathbb{E}_\mu\left[n_{i_0}^{p_0}(T)\right] \geq \frac{\ln\left(\Delta_{i_0}^{p_0}T^\alpha/8C\right)}{12(\Delta_{i_0}^{p_0})^2}$.

The proof of this claim is almost identical to the proof of the the second claim in the previous theorem, except that we have more flexibility to choose the "alternative instances" $\nu$'s, because $\mathcal{A}$ is assumed to have sublinear regret in all MPMAB instances; specifically, $\nu$ no longer needs to be an $\epsilon$-MPMAB instance. We include the argument here for completeness. Fix $i_0$ in $[K]$ and $p_0 \in [M]$ such that $\Delta_{i_0}^{p_0} > 0$. We consider a new Bernoulli MPMAB instance $\nu$, with mean reward defined as follows:

$$\nu_i^p = \begin{cases}\mu_i^p + 2\Delta_{i_0}^{p_0} & i = i_0, p = p_0\\ \mu_i^p & \text{otherwise}\end{cases}$$

We have the following three key observations:

(a) $\nu$ is still a valid Bernoulli MPMAB instance; this is because for all $p$ in $[M]$ and $i$ in $[K]$, $(\nu_i^p - \mu_i^p) \in [-\frac{1}{8}, \frac{1}{8}]$, and $\mu_i^p \in [\frac{15}{32}, \frac{17}{32}]$, implying that $\nu_i^p \in [0, 1]$. By our assumption that $\mathcal{A}$ has $CT^{1-\alpha}$ regret on all Bernoulli MPMAB instances, we have

$$\mathbb{E}_\mu\left[\mathcal{R}(T)\right] \leq CT^{1-\alpha}, \quad \mathbb{E}_\nu\left[\mathcal{R}(T)\right] \leq CT^{1-\alpha}. \tag{37}$$

(b) By the divergence decomposition lemma (Lattimore and Szepesvári (2020), Lemma 15.1),

$$\mathrm{KL}(\mathbb{P}_\mu, \mathbb{P}_\nu) = \mathbb{E}_\mu\left[n_{i_0}^{p_0}(T)\right] \mathrm{KL}\left(\mathrm{Ber}(\mu_{i_0}^{p_0}), \mathrm{Ber}(\mu_{i_0}^{p_0} + 2\Delta_{i_0}^{p_0})\right) \leq 12\mathbb{E}_\mu\left[n_{i_0}^{p_0}(T)\right](\Delta_{i_0}^{p_0})^2, \tag{38}$$

where the second equality uses the following observation: $\mu_{i_0}^{p_0} \in [\frac{15}{32}, \frac{17}{32}]$, and $\Delta_{i_0}^{p_0} \leq \frac{1}{16}$, using Lemma 25, $\mathrm{KL}\left(\mathrm{Ber}(\mu_{i_0}^{p_0}), \mathrm{Ber}(\mu_{i_0}^{p_0} + 2\Delta_{i_0}^{p_0})\right) \leq 3 \cdot (2\Delta_{i_0}^{p_0})^2 = 12(\Delta_{i_0}^{p_0})^2$.

(c) Under MPMAB instance $\mu$, $\mathcal{R}(T) \geq \mathcal{R}^{p_0}(T) \geq \Delta_{i_0}^{p_0} n_{i_0}^{p_0}(T) \geq \frac{\Delta_{i_0}^{p_0} T}{2} \mathbb{1}\{n_{i_0}^{p_0}(T) \geq \frac{T}{2}\}$. Taking expectations, we get,

$$\mathbb{E}_\mu\left[\mathcal{R}(T)\right] \geq \frac{\Delta_{i_0}^{p_0} T}{2} \mathbb{P}_\mu(n_{i_0}^{p_0}(T) \geq \frac{T}{2}). \tag{39}$$

Likewise, under MPMAB instance $\nu$, for player $p_0$, $\mathcal{R}(T) \geq \mathcal{R}^{p_0}(T) \geq \Delta_{i_0}^{p_0}(T - n_{i_0}^{p_0}(T)) \geq \frac{\Delta_{i_0}^{p_0} T}{2} \mathbb{1}\{n_{i_0}^{p_0}(T) < \frac{T}{2}\}$. Taking expectations, we get,

$$\mathbb{E}_\nu\left[\mathcal{R}(T)\right] \geq \frac{\Delta_{i_0}^{p_0} T}{2} \mathbb{P}_\nu(n_{i_0}^{p_0}(T) < \frac{T}{2}). \tag{40}$$

Combining Equations (37), (38), (39), (40), and using Bretagnolle-Huber inequality, we get

$$\mathbb{E}_\mu\left[n_{i_0}^{p_0}(T)\right] \geq \frac{1}{12(\Delta_{i_0}^{p_0})^2} \cdot \ln\left(\frac{\Delta_{i_0}^{p_0} T^\alpha}{8C}\right).$$

This concludes the proof of the claim, and in turn concludes the proof of the theorem. □

### E.4 Auxiliary lemmas

The following lemma is well known for proving gap-independent lower bounds in single player $K$-armed bandits. We will be using the following convention: for probability distribution $\mathbb{P}_i$, denote by $\mathbb{E}_i$ its induced expectation operator.

**Lemma 24.** *Suppose $m, B$ are positive integers and $m \geq 2$; there are $m+1$ probability distributions $\mathbb{P}_0, \mathbb{P}_1, \ldots, \mathbb{P}_m$, and $m$ random variables $N_1, \ldots, N_m$, such that: (1) Under any of the $\mathbb{P}_i$'s, $N_1, \ldots, N_m$ are non-negative and $\sum_{i=1}^m N_i = B$ with probability 1; (2) for all $i$ in $[m]$, $d_{\mathrm{TV}}(\mathbb{P}_0, \mathbb{P}_i) \leq \frac{1}{4}\sqrt{\frac{m}{B} \cdot \mathbb{E}_0[N_i]}$. Then,*

$$\frac{1}{m}\sum_{i=1}^m \mathbb{E}_i[B - N_i] \geq \frac{B}{4}.$$

*Proof.* For every $i$ in $[m]$, as $N_i$ is a random variable that takes values in $[0, B]$, we have,

$$\left|\mathbb{E}_i[N_i] - \mathbb{E}_0[N_i]\right| \leq B d_{\mathrm{TV}}(\mathbb{P}_0, \mathbb{P}_i).$$

By item (2) and algebra, this implies that

$$\mathbb{E}_i[N_i] \leq \mathbb{E}_0[N_i] + \frac{1}{4}\sqrt{mB\mathbb{E}_0[N_i]}.$$

Averaging over $i$ in $[m]$ and using Jensen's inequality, we have

$$\frac{1}{m}\sum_{i=1}^m \mathbb{E}_i[N_i] \leq \frac{1}{m}\sum_{i=1}^m \mathbb{E}_0[N_i] + \frac{1}{4m}\sum_{i=1}^m \sqrt{mB\mathbb{E}_0[N_i]} \leq \frac{1}{m}\sum_{i=1}^m \mathbb{E}_0[N_i] + \frac{1}{4}\sqrt{mB \cdot \left(\frac{1}{m}\sum_{i=1}^m \mathbb{E}_0[N_i]\right)}$$

Noting that item (2) implies $\frac{1}{m}\sum_{i=1}^m \mathbb{E}_0[N_i] = \frac{B}{m}$; plugging this into the above inequality, we have

$$\frac{1}{m}\sum_{i=1}^m \mathbb{E}_i[N_i] \leq \frac{B}{m} + \frac{1}{4}\sqrt{mB \cdot \frac{B}{m}} \leq \frac{B}{2} + \frac{B}{4} = \frac{3B}{4},$$

where the second inequality uses the assumption that $m \geq 2$. The lemma is concluded by negating and adding $B$ on both sides. □

**Lemma 25.** *Suppose $a, b$ are both in $[\frac{1}{4}, \frac{3}{4}]$. Then,* $\mathrm{KL}(\mathrm{Ber}(a), \mathrm{Ber}(b)) \leq 3(b-a)^2$.

*Proof.* Define $h(x) = x \ln \frac{1}{x} + (1-x) \ln \frac{1}{1-x}$. It can be easily verified that $\mathrm{KL}(\mathrm{Ber}(a), \mathrm{Ber}(b)) = a \ln \frac{a}{b} + (1-a) \ln \frac{1-a}{1-b}$, which in turn equals $h(a) - h(b) - h'(b)(a-b)$. By Taylor's theorem, there exists some $\xi \in [a, b] \subseteq [\frac{1}{4}, \frac{3}{4}]$ such that

$$h(a) - h(b) - h'(b)(a-b) = \frac{h''(\xi)}{2}(b-a)^2 = \frac{1}{2\xi(1-\xi)}(b-a)^2.$$

The lemma is concluded by verifying that $\frac{1}{2\xi(1-\xi)} \leq 3$ for $\xi$ in $[\frac{1}{4}, \frac{3}{4}]$. □

**Lemma 26** (Bretagnolle-Huber)**.** *For any two distributions $\mathbb{P}$ and $\mathbb{Q}$ and an event $A$,*

$$\mathbb{P}(A) + \mathbb{Q}(A^C) \geq \frac{1}{2} \exp(-\mathrm{KL}(\mathbb{P}, \mathbb{Q})).$$

## F Upper bounds with unknown $\epsilon$

In this section, we provide a description of RobustAgg-Agnostic, an algorithm that has regret adaptive to $\mathcal{I}_\epsilon$ in all MPMAB environments with unknown $\epsilon$.

To ensure sublinear regret in all MPMAB environments, RobustAgg uses the aggregation-based framework named Corral (Agarwal et al., 2017, see also Lemma 29 below), which we now briefly review. The Corral meta-algorithm allows one to combine multiple online bandit learning algorithms (called base learners) into one master algorithm that has performance competitive with all base learners'. For different environments, different base learners may stand out as the best, and therefore the master algorithm exhibits some degree of adaptivity. We refer readers to (Agarwal et al., 2017) for the full description of Corral.

In the context of MPMAB problems, recall that we have developed RobustAgg($\epsilon$) that has good regret guarantees for all $\epsilon$-MPMAB instances. The central idea of RobustAgg-Agnostic is to apply the Corral algorithm over a series of baser learners, i.e., $\left\{\text{RobustAgg}(\epsilon_b)\right\}_{b=1}^B$, where $E = \{\epsilon_b\}_{b=1}^B$ is a covering of the $[0,1]$ interval. With an appropriate setting of $E$, for any $\epsilon$-MPMAB instance, there exists some $b_0$ in $[B]$ such that $\epsilon_{b_0}$ is not much larger than $\epsilon$, and running RobustAgg($\epsilon_{b_0}$) would achieve regret guarantee competitive to RobustAgg($\epsilon$). As Corral achieves online performance competitive with all RobustAgg($\epsilon_b$)'s (Agarwal et al., 2017), it must be competitive with RobustAgg($\epsilon_{b_0}$), and therefore can inherit the adaptive regret guarantee of RobustAgg($\epsilon_{b_0}$).

We now provide important technical details of RobustAgg-Agnostic:

1. $B = \lceil \log(MT) \rceil + 1$ is the number of base learners, and $E = \left\{\epsilon_b = 2^{-b+1} : b \in [B]\right\}$ is the grid of $\epsilon$ to be aggregated. Corral uses master learning rate $\eta = \frac{1}{M\sqrt{T}}$.

2. For each base learner that runs RobustAgg($\epsilon$) for some $\epsilon$, we require them to take a new parameter $\rho \geq 1$ as input, to accommodate for the fact that it may not be selected by the Corral master all the time. Specifically, it performs bandit learning interaction with an environment whose returned rewards are unbiased but *importance weighted*: at time step $t$, when player $p$ pulls arm $i$, instead of directly receiving reward drawn from $r \sim \mathcal{D}_i^p$, it receives $\hat{r} = \frac{W_t}{w_t} r$, where $w_t \in [1, \frac{1}{\rho}]$ is a random number, and conditioned on $w_t$, $W_t \sim \mathrm{Ber}(w_t)$ is an independently-drawn Bernoulli random variable. Observe that $\hat{r}$ has conditional mean $\mu_i^p$, lies in the interval $[0, \rho]$, and has conditional variance at most $\rho$.

   We call an environment that has the above analytical form a *$\rho$-importance weighted environment*; in the special case of $\rho = 1$, $w_t = 1$ and $W_t = 1$ with probability 1 for all $t$, and therefore a 1-importance weighted environment is the same as the original bandit learning environment.

   Under an $\rho$-importance weighted environment, the rewards are no longer bounded in $[0,1]$, therefore, the constructions of the UCB's of the mean rewards in the original RobustAgg($\epsilon$) becomes invalid. Instead, we will rely on the following lemma (analogue of Lemma 17) for constructing valid UCB's:

**Lemma 27.** *With probability at least* $1 - 4MT^{-5}$*, we have*

$$\left|\zeta_i^p(t-1) - \mu_i^p\right| \le 8\sqrt{\frac{3\rho \ln T}{\overline{n_i^p}(t-1)}},$$

$$\left|\eta_i^p(t-1) - \sum_{q \ne p} \frac{n_i^q(t-1)}{\overline{m_i^p}(t-1)} \mu_i^q\right| \le 4\sqrt{\frac{14\rho \ln T}{\overline{m_i^p}(t-1)}}$$

*holding for all* $p$ *in* $[M]$*, where* $\zeta_i^p(t) = \frac{\sum_{s=1}^{t-1} \mathbb{1}\{i_t^p = i\} \hat{r}_t^p}{\overline{n_i^p}(t-1)}$*, and* $\eta_i^p(t) = \frac{\sum_{s=1}^{t-1} \sum_{q=1}^{M} \mathbb{1}\{q \ne p,\ i_t^q = i\} \hat{r}_t^q}{\overline{m_i^p}(t-1)}$.

According to the above concentration bounds, changing the definition of confidence interval width to $F(\overline{n_i^p}, \overline{m_i^p}, \lambda, \epsilon) = 8\sqrt{13\rho \ln T \left[\frac{\lambda^2}{\overline{n_i^p}} + \frac{(1-\lambda)^2}{\overline{m_i^p}}\right]} + (1-\lambda)\epsilon$ would maintain the validity of the UCB's in $\rho$-importance weighted environments; henceforth, we incorporate this modification in RobustAgg($\epsilon$).

We have the following important analogue of Theorem 8, which establishes a gap-independent regret upper bound when RobustAgg($\epsilon$) is run in a $\rho$-importance weighted $\epsilon$-MPMAB environment. This shows RobustAgg($\epsilon$) enjoys stability: the regret of the algorithm degrades gracefully with increasing $\rho$.[5]

**Lemma 28.** *Let* RobustAgg($\epsilon$) *run on a* $\rho$*-importance weighted* $\epsilon$*-MPMAB problem instance for* $T$ *rounds. Then its expected collective regret satisfies*

$$\mathbb{E}[\mathcal{R}(T)] \le \tilde{O}\left(\sqrt{\rho \left|\mathcal{I}_\epsilon\right| MT} + M\left|\mathcal{I}_\epsilon\right| + \min\left(M\sqrt{\rho \left|\mathcal{I}_\epsilon^C\right| T}, \epsilon MT\right)\right).$$

The proof of Lemmas 27 and 28 can be found at the end of this section.

3. Corral maintains a probability distribution on base learners $q_t = (q_{t,b} : b \in [B])$ over time. At time step $t$, each base learner $b$ proposes their own arm pull decisions $(i_t^p(b) : p \in [M])$; the Corral master chooses a base learner with probability according to $q_t$, that is, $i_t^p = i_t^p(b_t)$ for all $p$, where $b_t \sim q_t$. After the arm pulls, learner $b$ receives feedback $\hat{r}_t^p(b) = \frac{\mathbb{1}\{b_t = b\}}{q_{t,b}} r_t^p$, which is equivalent to interacting with an importance weighted environment discussed before—$q_{t,b}$ and $\mathbb{1}\{b_t = b\}$ correspond to $w_t$ and $W_t$, respectively; when $b_t = b$, $r_t^p$ is drawn from $\mathcal{D}_{i_t^p}^p$ for all $p$ in $[M]$.

   Corral also uses the above feedback to update $q_{t+1}$, its weighting of the base learners: define $\ell_{t,b} = \frac{\mathbb{1}\{b_t = b\}}{q_{t,b}} \mathbb{1}\{b_t = b\}(\sum_{p=1}^{M}(1 - r_t^p))$ to be the importance weighted loss of base learner $b$ at time step $t$; $q_t$ is updated to $q_{t+1}$ using $(\ell_{t,b} : b \in [B])$, with online mirror descent with the log-barrier regularizer and learning rate $\eta > 0$. A small complication of directly applying the existing results of Corral is that Corral originally assumes that the losses suffered by the base learner from each round have range $[0, 1]$. In the multi-player setting, the losses suffered by the base learner is the sum of the losses of all players, which has range $[0, M]$. Nevertheless, we can obtain a similar guarantee. Denote by $\rho_b$ be the final value of $\rho$ of base learner $b$ (see also the next item). A slight modification of Agarwal et al. (2017, Lemma 13) shows that for all base learner $b$,

$$\sum_{t=1}^{T}\sum_{b'=1}^{B} q_{t,b'}\ell_{t,b'} - \sum_{t=1}^{T}\ell_{t,b} \le O\left(\frac{B}{\eta} + \eta M^2 T\right) - \frac{\rho_b}{40\eta \ln T}.$$

   Taking expectation on both sides, and observing that $\mathbb{E}\left[\sum_{t=1}^{T}\sum_{b'=1}^{B} q_{t,b'}\ell_{t,b'}\right] = \mathbb{E}\left[\sum_{t=1}^{T}\sum_{p=1}^{M}(1 - \mu_{i_t^p}^p)\right]$, and $\mathbb{E}\left[\sum_{t=1}^{T}\ell_{t,b}\right] = \mathbb{E}\left[\sum_{t=1}^{T}\sum_{p=1}^{M}(1 - \mu_{i_t^p(b)}^p)\right]$, along with some algebra, we get the following lemma.

[5] See an elegant definition of $(R(T), \alpha)$-(weak) stability for bandit algorithms in (Agarwal et al., 2017). Our guarantee on RobustAgg($\epsilon$) in Lemma 28 is slightly stronger than the $(\tilde{O}\left(\sqrt{|\mathcal{I}_\epsilon| MT} + M|\mathcal{I}_\epsilon| + \min\left(M\sqrt{|\mathcal{I}_\epsilon^C| T}, \epsilon MT\right)\right), \frac{1}{2})$-weak stability, in that the regret bound has terms that are unaffected by $\rho$.

**Lemma 29.** *Suppose* RobustAgg-Agnostic *is run for $T$ rounds. Then, for every $b$ in $[B]$, we have that the regret of the master algorithm with respect to base learner $b$ is bounded by*

$$\mathbb{E}\left[\sum_{t=1}^{T}\sum_{p=1}^{M}\mu^p_{i^p_t(b)} - \sum_{t=1}^{T}\sum_{p=1}^{M}\mu^p_{i^p_t}\right] \le O\left(\frac{B}{\eta} + \eta M^2 T\right) - \frac{\mathbb{E}\left[\rho_b\right]}{40\eta \ln T}.$$

4. Following Agarwal et al. (2017), a doubling trick is used for maintaining the value of $\rho$'s for all base learners over time. Specifically, each base learner $b$ maintains a separate guess of $\rho$, an upper bound of $\max_{s=1}^{t}\frac{1}{q_{s,b}}$; if the upper bound is violated, its $\rho$ gets doubled and the base learner restarts. As Corral initializes $\rho$ as $2B$ for each base learner, and maintains the invariant that $\rho \le BT$, the number of doublings/restarts for each base learner is at most $\lceil \log T \rceil$. For a fixed $b$, summing over the regret guarantees between different restarts of base learner $b$, we have the following regret guarantee.

   **Lemma 30.** *Suppose $\epsilon_b \ge \epsilon$, and* RobustAgg($\epsilon_b$) *is run as a base learner of* RobustAgg-Agnostic*, on a $\epsilon$-MPMAB problem instance for $T$ rounds. Denote by $\rho_b$ the final value of $\rho$. Then its expected collective regret satisfies*

$$\mathbb{E}\left[T\sum_{p=1}^{M}\mu^p_* - \sum_{t=1}^{T}\sum_{p=1}^{M}\mu^p_{i^p_t(b)}\right] \le \tilde{O}\left(\sqrt{\mathbb{E}\left[\rho_b\right]\left|\mathcal{I}_{\epsilon_b}\right| MT} + \min\left(M\sqrt{\mathbb{E}\left[\rho_b\right]\left|\mathcal{I}^C_{\epsilon_b}\right| T}, \epsilon MT\right) + M\left|\mathcal{I}_{\epsilon_b}\right|\right).$$

   The proof of this lemma can be found at the end of this section; we also refer the reader to (Agarwal et al., 2017, Appendix D) for details.

Combining all the lemmas above, we are now ready to prove Theorem 12, restated below for convenience.

**Theorem 12.** *Let* RobustAgg-Agnostic *run on an $\epsilon$-MPMAB problem instance with any $\epsilon \in [0,1]$. Its expected collective regret in a horizon of $T$ rounds satisfies*

$$\mathbb{E}[\mathcal{R}(T)] \le \tilde{O}\left(\left(|\mathcal{I}_{2\epsilon}| + M\left|\mathcal{I}^C_{2\epsilon}\right|\right)\sqrt{T} + M|\mathcal{I}_\epsilon|\right).$$

*Proof of Theorem 12.* Suppose RobustAgg-Agnostic interacts with an $\epsilon$-MPMAB problem instance. Let $b_0 = \max\left\{b \in [B] : \epsilon_b \ge \epsilon\right\}$. From the definition of $E = \left\{1, 2^{-1}, \ldots, 2^{-B+1}\right\}$ and $\epsilon \in [0,1]$, $b_0$ is well-defined.

We present the following technical claim that elucidates the guarantee provided by learner $b_0$ based on Lemma 30; we defer its proof after the proof of the theorem.

**Claim 31.** *Let $b_0$ be defined above.* RobustAgg($\epsilon_{b_0}$) *is run as a base learner of* RobustAgg-Agnostic*, on a $\epsilon$-MPMAB problem instance for $T$ rounds. Denote by $\rho_{b_0}$ the final value of $\rho$. Then its expected collective regret satisfies*

$$\mathbb{E}\left[T\sum_{p=1}^{M}\mu^p_* - \sum_{t=1}^{T}\sum_{p=1}^{M}\mu^p_{i^p_t(b_0)}\right] \le \tilde{O}\left(\sqrt{\mathbb{E}\left[\rho_{b_0}\right] MT(|\mathcal{I}_{2\epsilon}| + M\left|\mathcal{I}^C_{2\epsilon}\right|)} + M\left|\mathcal{I}_\epsilon\right|\right).$$

Combining Claim 31 and Lemma 29 with $b = b_0$, we have the following regret guarantee for RobustAgg-Agnostic:

$$\begin{aligned}
\mathbb{E}\left[\mathcal{R}(T)\right] &= \mathbb{E}\left[T\sum_{p=1}^{M}\mu^p_* - \sum_{t=1}^{T}\sum_{p=1}^{M}\mu^p_{i^p_t}\right] \\
&= \mathbb{E}\left[T\sum_{p=1}^{M}\mu^p_* - \sum_{t=1}^{T}\sum_{p=1}^{M}\mu^p_{i^p_t(b_0)}\right] + \mathbb{E}\left[\sum_{t=1}^{T}\sum_{p=1}^{M}\mu^p_{i^p_t(b_0)} - \sum_{t=1}^{T}\sum_{p=1}^{M}\mu^p_{i^p_t}\right] \\
&\le \tilde{O}\left(\sqrt{\mathbb{E}\left[\rho_{b_0}\right] MT(|\mathcal{I}_{2\epsilon}| + M\left|\mathcal{I}^C_{2\epsilon}\right|)} + M\left|\mathcal{I}_\epsilon\right| + \frac{B}{\eta} + \eta M^2 T\right) - \frac{\mathbb{E}\left[\rho_{b_0}\right]}{40\eta \ln T} \\
&\le \tilde{O}\left(\eta MT(|\mathcal{I}_{2\epsilon}| + M\left|\mathcal{I}^C_{2\epsilon}\right|) + M\left|\mathcal{I}_\epsilon\right| + \frac{B}{\eta} + \eta M^2 T\right),
\end{aligned}$$

where the first inequality is from Claim 31 and Lemma 29; the second inequality is from the AM-GM inequality that $\sqrt{\mathbb{E}\left[\rho_{b_0}\right] MT(|\mathcal{I}_{2\epsilon}| + M\left|\mathcal{I}_{2\epsilon}^C\right|)} \leq O(\eta MT(|\mathcal{I}_{2\epsilon}| + M\left|\mathcal{I}_{2\epsilon}^C\right|) + \frac{\mathbb{E}\left[\rho_{b_0}\right]}{\eta \ln T})$ and algebra (canceling out the second term in the last expression with $-\frac{\mathbb{E}\left[\rho_{b_0}\right]}{40\eta \ln T}$). As RobustAgg-Agnostic chooses Corral's master learning rate $\eta = \frac{1}{M\sqrt{T}}$, and $B = \tilde{O}(1)$, we have that

$$\mathbb{E}\left[\mathcal{R}(T)\right] \leq \tilde{O}\left((M + |\mathcal{I}_{2\epsilon}| + M\left|\mathcal{I}_{2\epsilon}^C\right|)\sqrt{T} + M\left|\mathcal{I}_{\epsilon}\right|\right) \leq \tilde{O}\left((|\mathcal{I}_{2\epsilon}| + M\left|\mathcal{I}_{2\epsilon}^C\right|)\sqrt{T} + M\left|\mathcal{I}_{\epsilon}\right|\right),$$

where the second inequality uses the fact that $\left|\mathcal{I}_{2\epsilon}^C\right| \geq 1$. □

*Proof of Claim 31.* As $\epsilon_b \geq \epsilon$ always holds, $M\left|\mathcal{I}_{\epsilon_{b_0}}\right| \leq M\left|\mathcal{I}_{\epsilon}\right|$. It remains to check by algebra that

$$\sqrt{\mathbb{E}\left[\rho_{b_0}\right]\left|\mathcal{I}_{\epsilon_{b_0}}\right| MT} + \min\left(M\sqrt{\mathbb{E}\left[\rho_{b_0}\right]\left|\mathcal{I}_{\epsilon_{b_0}}^C\right| T}, MT\epsilon_{b_0}\right) = \tilde{O}\left(\sqrt{\mathbb{E}\left[\rho_{b_0}\right] MT(|\mathcal{I}_{2\epsilon}| + M\left|\mathcal{I}_{2\epsilon}^C\right|)}\right). \tag{41}$$

We consider two cases:

1. $\epsilon_{b_0} \leq 2\epsilon$. In this case, we have $\mathcal{I}_{\epsilon_{b_0}} \subset \mathcal{I}_{2\epsilon}$. We have the following derivation:

$$\begin{aligned}\sqrt{\mathbb{E}\left[\rho_{b_0}\right]\left|\mathcal{I}_{\epsilon_{b_0}}\right| MT} + M\sqrt{\mathbb{E}\left[\rho_{b_0}\right]\left|\mathcal{I}_{\epsilon_{b_0}}^C\right| T} &\leq 2\sqrt{\mathbb{E}\left[\rho_{b_0}\right] MT \cdot (\left|\mathcal{I}_{\epsilon_b}\right| + M\left|\mathcal{I}_{\epsilon_b}^C\right|)} \\ &\leq 2\sqrt{\mathbb{E}\left[\rho_{b_0}\right] MT(|\mathcal{I}_{2\epsilon}| + M\left|\mathcal{I}_{2\epsilon}^C\right|)}\end{aligned}$$

   where the first inequality is from the basic fact that $\sqrt{A} + \sqrt{B} \leq 2\sqrt{A+B}$ for positive $A$, $B$; the second inequality is from the fact that $\left|\mathcal{I}_{\epsilon_b}\right| + M\left|\mathcal{I}_{\epsilon_b}^C\right| \leq |\mathcal{I}_{2\epsilon}| + M\left|\mathcal{I}_{2\epsilon}^C\right|$, as $\left|\mathcal{I}_{\epsilon_b}\right| \geq |\mathcal{I}_{2\epsilon}|$, $M \geq 1$, and $|\mathcal{I}_\alpha| + \left|\mathcal{I}_\alpha^C\right| = K$ for any $\alpha$. This verifies Eq. (41).

2. $\epsilon_{b_0} > 2\epsilon$. In this case, $b_0 = B = 1 + \lceil \log(MT) \rceil$ and $\epsilon_{b_0} \leq \frac{1}{MT}$. Although we no longer have $\mathcal{I}_{\epsilon_{b_0}} \subset \mathcal{I}_{2\epsilon}$, we can still upper bound the left hand side as follows.

   First, the second term, $\min\left(M\sqrt{\mathbb{E}\left[\rho_{b_0}\right]\left|\mathcal{I}_{\epsilon_{b_0}}^C\right| T}, MT\epsilon_{b_0}\right) \leq MT \cdot \frac{1}{MT} = 1$.

   Moreover, the first term, $\sqrt{\mathbb{E}\left[\rho_{b_0}\right]\left|\mathcal{I}_{\epsilon_{b_0}}\right| MT} \leq \sqrt{\mathbb{E}\left[\rho_{b_0}\right] KMT}$. As $|\mathcal{I}_{2\epsilon}| + M\left|\mathcal{I}_{2\epsilon}^C\right| \geq K$, we have $\sqrt{\mathbb{E}\left[\rho_{b_0}\right]\left|\mathcal{I}_{\epsilon_{b_0}}\right| MT} \leq \sqrt{\mathbb{E}\left[\rho_{b_0}\right](|\mathcal{I}_{2\epsilon}| + M\left|\mathcal{I}_{2\epsilon}^C\right|)MT}$. Combining the above two, Eq. (41) is proved. □

*Proof sketch of Lemma 27.* Since the proof of Lemma 17 can be almost directly carried over here, we only sketch the proof by pointing out the major differences. We also refer the reader to (Arora et al., 2020, Appendix C.3) for a similar reasoning.

We first consider the concentration of $\zeta_i^p(j,t)$. We define a filtration $\{\mathcal{B}_t\}_{t=1}^T$, where

$$\mathcal{B}_t = \sigma(\{w_s, i_s^{p'}, \hat{r}_s^{p'} : s \in [t], p' \in [M]\} \cup \left\{i_{t+1}^{p'} : p' \in [M]\right\})$$

is the $\sigma$-algebra generated by the history (including that of $w_s$'s) up to round $t$ and the arm selection of all players at time step $t+1$

Let $X_t = \mathbb{1}\{i_t^p = i\}\left(\hat{r}_t^p - \mu_i^p\right)$. We have $\mathbb{E}\left[X_t \mid \mathcal{B}_{t-1}\right] = 0$. In addition,

$$\begin{aligned}
\mathbb{V}\left[X_t \mid \mathcal{B}_{t-1}\right] &= \mathbb{E}\left[(X_t - \mathbb{E}[X_t \mid \mathcal{B}_{t-1}])^2 \mid \mathcal{B}_{t-1}\right] \\
&= \mathbb{E}\left[X_t^2 \mid \mathcal{B}_{t-1}\right] \\
&\le \mathbb{1}\{i_t^p = i\}\mathbb{E}\left[w_t\left(\frac{r_t^p}{w_t} - \mu_i^p\right)^2 + (1-w_t)0 \mid \mathcal{B}_{t-1}\right] \\
&\le \mathbb{1}\{i_t^p = i\}\mathbb{E}\left[w_t\left(\frac{r_t^p}{w_t}\right)^2 \mid \mathcal{B}_{t-1}\right] \\
&\le \mathbb{1}\{i_t^p = i\}\rho.
\end{aligned}$$

Also, $|X_t| \le \rho$ with probability 1. Applying Freedman's inequality (Bartlett et al., 2008, Lemma 2) with $\sigma = \sqrt{\sum_{s=1}^{t-1}\mathbb{V}\left[X_s \mid \mathcal{B}_{s-1}\right]}$ and $b = \rho$, and using $\sigma \le \sqrt{\sum_{s=1}^{t-1}\mathbb{1}\{i_s^p = i\}\rho}$, we have that with probability at least $1 - 2T^{-5}$,

$$\left|\sum_{s=1}^{t-1} X_s\right| \le 4\sqrt{\sum_{s=1}^{t-1}\mathbb{1}\{i_s^p = i\}\rho \cdot \ln(T^5\log_2 T)} + 2\rho\ln(T^5\log_2 T). \tag{42}$$

We can then show that

$$\left|\frac{\sum_{s=1}^{t-1}\mathbb{1}\{i_s^p = i\}\hat{r}_s^p}{\overline{n_i^p}(t-1)} - \mu_i^p\right| \le 8\sqrt{\frac{3\rho\ln T}{\overline{n_i^p}(t-1)}}.$$

following the same strategy in the proof for Lemma 17.

Similarly, we show the concentration of $\eta_i^p(t)$. We define a filtration $\{\mathcal{G}_{t,q}\}_{t\in[T],q\in[M]}$, where

$$\mathcal{G}_{t,q} = \sigma(\left\{w_s, i_s^{p'}, \hat{r}_s^{p'} : s \in [t], p' \in [M], i \in [K]\right\} \cup \left\{i_{t+1}^{p'} : p' \in [M], p' \le q\right\})$$

is the $\sigma$-algebra generated by the history (including that of $w_s$'s) up to round $t$ and the arm selection of players $1, 2, \ldots, q$ at round $t+1$.

Let random variable $Y_{t,q} = \mathbb{1}\{q \ne p,\ i_t^q = i\}\left(\hat{r}_s^q - \mu_i^q\right)$. We have $\mathbb{E}\left[Y_{t,q} \mid \mathcal{G}_{t-1,q}\right] = 0$; in addition, $\mathbb{V}\left[Y_t \mid \mathcal{G}_{t-1}\right] = \mathbb{E}\left[Y_{t,q}^2 \mid \mathcal{G}_{t-1,q}\right] \le \mathbb{1}\{q \ne p,\ i_t^q = i\}\rho$ and $\left|Y_{t,q}\right| \le \rho$.

Again, applying Freedman's inequality (Bartlett et al., 2008, Lemma 2), we have that with probability at least $1 - 2T^{-5}$,

$$\left|\sum_{s=1}^{t-1}\sum_{q=1}^{M} Y_{s,q}\right| \le 4\sqrt{\sum_{s=1}^{t-1}\sum_{q=1}^{M}\mathbb{1}\{q \ne p,\ i_s^q = i\}\rho \cdot \ln(T^5\log_2(TM))} + 2\rho\ln(T^5\log_2(TM)). \tag{43}$$

Using the same strategy from the proof for Lemma 17, we can show that

$$\left|\eta_i^p(t-1) - \sum_{q\ne p}\frac{n_i^q(t-1)}{\overline{m_i^p}(t-1)}\mu_i^q\right| \le 4\sqrt{\frac{14\rho\ln T}{\overline{m_i^p}(t-1)}}. \qquad \square$$

The lemma then follows by applying the union bound.

*Proof sketch of Lemma 28.* Similar to the proof of Theorem 8, we define $\mathcal{E} = \cap_{t=1}^T \cap_{i=1}^K \mathcal{Q}_i(t)$, where

$$\mathcal{Q}_i(t) = \left\{\forall p, \left|\zeta_i^p(t) - \mu_i^p\right| \le 8\sqrt{\frac{3\rho\ln T}{\overline{n_i^p}(t-1)}},\quad \left|\eta_i^p(t) - \sum_{q\ne p}\frac{n_i^q(t-1)}{\overline{m_i^p}(t-1)}\mu_i^q\right| \le 4\sqrt{\frac{14\rho\ln T}{\overline{m_i^p}(t-1)}}\right\};$$

note that the new definition of $\mathcal{Q}_i(t)$ has a dependence on $\rho$.

Similar to the proof of Theorem 5, we have,

$$\mathbb{E}[\mathcal{R}(T)] \le \mathbb{E}[\mathcal{R}(T)|\mathcal{E}] + O(1),$$

and

$$\begin{aligned}\mathbb{E}[\mathcal{R}(T)|\mathcal{E}] &= \sum_{i\in[K]}\sum_{p\in[M]} \mathbb{E}[n_i^p(T)|\mathcal{E}]\cdot\Delta_i^p \\ &\le \sum_{i\in\mathcal{I}_\epsilon} \mathbb{E}[n_i(T)|\mathcal{E}]\cdot\Delta_i^{\max} + \sum_{i\in\mathcal{I}_\epsilon^C}\sum_{p\in[M]:\Delta_i^p>0} \mathbb{E}[n_i^p(T)|\mathcal{E}]\cdot\Delta_i^p\end{aligned}$$

We bound these two terms respectively, applying the technique from (Lattimore and Szepesvári, 2020, Theorem 7.2).

1. We can show the following analogue of Lemma 20: there exists some constant $C_1 > 0$ such that for each $i \in \mathcal{I}_\epsilon$,

$$\mathbb{E}[n_i(T)|\mathcal{E}] \le C_1\left(\frac{\rho\ln T}{(\Delta_i^{\min})^2} + M\right).$$

   Using the above fact, and from a similar calculation of Equation (19) in the proof of Theorem 8, we get

$$\sum_{i\in\mathcal{I}_\epsilon} \mathbb{E}[n_i(T)|\mathcal{E}]\cdot\Delta_i^{\max} \le 4\sqrt{C_1\rho|\mathcal{I}_\epsilon|MT\ln T} + 2C_1M|\mathcal{I}_\epsilon|.$$

2. We can show the following analogue of Lemma 21: there exists some constant $C_2 > 0$ such that for each $i \in \mathcal{I}_\epsilon^C$ and $p \in [M]$ with $\Delta_i^p > 0$,

$$\mathbb{E}[n_i^p(T)|\mathcal{E}] \le C_2\left(\frac{\ln T}{(\Delta_i^p)^2}\right).$$

   Using the above fact, and from a similar calculation of Equation (22) in the proof of Theorem 8, we get

$$\sum_{i\in\mathcal{I}_\epsilon^C}\sum_{p\in[M]:\Delta_i^p>0} \mathbb{E}[n_i^p(T)|\mathcal{E}]\cdot\Delta_i^p \le 2M\sqrt{C_2\rho|\mathcal{I}_\epsilon^C|T\ln T}.$$

   On the other hand, we trivially have that for all $i$ in $\mathcal{I}_\epsilon^C$, $\Delta_i^p \le 5\epsilon$; therefore,

$$\sum_{i\in\mathcal{I}_\epsilon^C}\sum_{p\in[M]:\Delta_i^p>0} \mathbb{E}[n_i^p(T)]\cdot\Delta_i^p \le 5\epsilon MT.$$

Therefore,

$$\begin{aligned}\mathbb{E}[\mathcal{R}(T)] &\le \left(4\sqrt{C_1\rho|\mathcal{I}_\epsilon|MT\ln T} + 2C_1M|\mathcal{I}_\epsilon|\right) + \min\left(2M\sqrt{C_2\rho|\mathcal{I}_\epsilon^C|T\ln T}, 5\epsilon MT\right) + O(1) \\ &\le \tilde{O}\left(\sqrt{\rho\,|\mathcal{I}_\epsilon|\,MT} + M\,|\mathcal{I}_\epsilon| + \min\left(M\sqrt{\rho\,|\mathcal{I}_\epsilon^C|\,T}, \epsilon MT\right)\right).\end{aligned}$$

□

*Proof of Lemma 30.* The proof closely follows (Agarwal et al., 2017, Theorem 15); we cannot directly repeat that proof here, because Lemma 28 is not precisely a weak stability statement (see footnote 5).

For base learner $b$, suppose that its $\rho$ gets doubled $n_b$ times throughout the process, where $n_b$ is a random number in $[\lceil\log T\rceil]$. For every $i \in [n_b]$, denote by random variable $t_i$ the $i$-th time step where the value of $\rho$ gets doubled.

In addition, denote by $t_0 = 0$ and $t_{n_b+1} = T$. In this notation, for all $t \in \{t_i + 1, t_i + 2, \ldots, t_{i+1}\}$, the value of $\rho$ is equal to $\rho^i = 2B \cdot 2^i$; in addition, $\rho_b = 2B \cdot 2^{n_b}$.

Therefore, we have:

$$\begin{aligned}
\mathbb{E}\left[\mathcal{R}(T) \mid n_b = n\right] &= \sum_{i=0}^{n} \mathbb{E}\left[\sum_{t=t_i+1}^{t_{i+1}} \left(\sum_{p=1}^{M} \mu_*^p - \sum_{p=1}^{M} \mu_{i_t^p(b)}^p\right) \mid n_b = n\right] \\
&= \sum_{i=0}^{n} \tilde{O}\left(\sqrt{\rho^i \left|\mathcal{I}_{\epsilon_b}\right| MT} + M\left|\mathcal{I}_{\epsilon_b}\right| + \min\left(M\sqrt{\rho^i \left|\mathcal{I}_{\epsilon_b}^C\right| T}, \epsilon_b MT\right)\right) \\
&= \tilde{O}\left(\sqrt{\rho^n \left|\mathcal{I}_{\epsilon_b}\right| MT} + M\left|\mathcal{I}_{\epsilon_b}\right| + \min\left(M\sqrt{\rho^n \left|\mathcal{I}_{\epsilon_b}^C\right| T}, \epsilon_b MT\right)\right),
\end{aligned}$$

where the first equality of by the definition of $\mathcal{R}(T)$, and $[T] = \cup_{i=1}^{n} \{t_i + 1, t_i + 2, \ldots, t_{i+1}\}$; the second equality is from Lemma 30's guarantee in each time interval $\{t_i + 1, t_i + 2, \ldots, t_{i+1}\}$ and $\epsilon_b \geq \epsilon$; and the third equality is by algebra.

As $n_b = n$ is equivalent to $\rho^n = \rho_b$, this implies that

$$\mathbb{E}\left[\mathcal{R}(T) \mid \rho_b\right] = \tilde{O}\left(\sqrt{\rho_b \left|\mathcal{I}_{\epsilon_b}\right| MT} + M\left|\mathcal{I}_{\epsilon_b}\right| + \min\left(M\sqrt{\rho_b \left|\mathcal{I}_{\epsilon_b}^C\right| T}, \epsilon_b MT\right)\right);$$

observe that the expression inside $\tilde{O}$ in the last line is a concave function of $\rho_b$.

Now, by the law of total expectation,

$$\begin{aligned}
\mathbb{E}\left[\mathcal{R}(T)\right] &= \mathbb{E}\left[\mathbb{E}\left[\mathcal{R}(T) \mid \rho_b\right]\right] \\
&= \mathbb{E}\left[\tilde{O}\left(\sqrt{\rho_b \left|\mathcal{I}_{\epsilon_b}\right| MT}\right) + M\left|\mathcal{I}_{\epsilon_b}\right| + \min\left(M\sqrt{\rho_b \left|\mathcal{I}_{\epsilon_b}^C\right| T}, \epsilon_b MT\right)\right] \\
&= \tilde{O}\left(\mathbb{E}\left[\sqrt{\rho_b \left|\mathcal{I}_{\epsilon_b}\right| MT} + M\left|\mathcal{I}_{\epsilon_b}\right| + \min\left(M\sqrt{\rho_b \left|\mathcal{I}_{\epsilon_b}^C\right| T}, \epsilon_b MT\right)\right]\right) \\
&= \tilde{O}\left(\sqrt{\mathbb{E}\left[\rho_b\right] \left|\mathcal{I}_{\epsilon_b}\right| MT} + M\left|\mathcal{I}_{\epsilon_b}\right| + \min\left(M\sqrt{\mathbb{E}\left[\rho_b\right] \left|\mathcal{I}_{\epsilon_b}^C\right| T}, \epsilon_b MT\right)\right),
\end{aligned}$$

where the third equality is by algebra, and the last equality uses Jensen's inequality. □

# G Experimental Details

In Appendix G.1, we present a pseudocode for RobustAgg-Adapted($\epsilon$) which was used in the experiments. Then, in Appendix G.2, we provide a proof of Fact 13 which is about the instance generation procedure. Finally, in Appendix G.3, we present comprehensive results from the simulations we performed.

## G.1 RobustAgg-Adapted($\epsilon$)

In our experiments, along with the two baselines—Ind-UCB and Naive-Agg—we evaluated a more practical algorithm adapted from RobustAgg($\epsilon$), which we call RobustAgg-Adapted($\epsilon$). Algorithm 2 provides a pseudocode for RobustAgg-Adapted($\epsilon$), in which we added an initialization phase (lines 2 to 5), and used a more aggressive upper confidence bound with length $\min_{\lambda \in [0,1]} \sqrt{2 \ln T [\frac{\lambda^2}{n_i^p} + \frac{(1-\lambda)^2}{m_i^p}]} + (1-\lambda)\epsilon$ (line 10).

**Algorithm 2:** RobustAgg-Adapted($\epsilon$)

**Input:** A parameter $\epsilon \in [0,1]$;

**Initialization:** Set $n_i^p = 0$ for all $p \in [M]$ and all $i \in [K]$.

**for** $t = 1, 2, \dots K$ **do**

  **for** $p \in [M]$ **do**

    Player $p$ pulls arm $i_t^p = t$ and observes reward $r_t^p$;

    Set $n_i^p = n_i^p + 1$.

**for** $t = K+1, K+2 \dots, T$ **do**

  **for** $p \in [M]$ **do**

    **for** $i \in [K]$ **do**

      Let $m_i^p = \sum_{q \in [M]: q \neq p} n_i^q$;

      Let $F(n_i^p, m_i^p, \lambda, \epsilon) = \sqrt{2 \ln T \left[ \frac{\lambda^2}{n_i^p} + \frac{(1-\lambda)^2}{m_i^p} \right]} + (1-\lambda)\epsilon$ ;

      Compute $\lambda^* = \operatorname{argmin}_{\lambda \in [0,1]} F(n_i^p, m_i^p, \lambda, \epsilon)$;

      Let

$$\zeta_i^p(t) = \frac{1}{n_i^p} \sum_{\substack{s<t \\ i_s^p = i}} r_s^p, \ \eta_i^p(t) = \frac{1}{m_i^p} \sum_{\substack{q \in [M] \\ q \neq p}} \sum_{\substack{s<t \\ i_s^q = i}} r_s^q, \text{and } \kappa_i^p(t, \lambda) = \lambda \zeta_i^p(t) + (1-\lambda)\eta_i^p(t);$$

      Compute the upper confidence bound of the reward of arm $i$ for player $p$:

$$\mathrm{UCB}_i^p(t) = \kappa_i^p(t, \lambda^*) + F(n_i^p, m_i^p, \lambda^*, \epsilon).$$

    Let $i_t^p = \operatorname{argmax}_{i \in [K]} \mathrm{UCB}_i^p(t)$;

    Player $p$ pulls arm $i_t^p$ and observes reward $r_t^p$;

  **for** $p \in [M]$ **do**

    Let $i = i_t^p$ and set $n_i^p = n_i^p + 1$.

### G.2 Proof of Fact 13

*Proof of Fact 13.* For every $i$, as $\mu_i^p \in [\mu_i^1 - \frac{\epsilon}{2}, \mu_i^1 + \frac{\epsilon}{2}]$ for all $p$ in $[M]$, we have that for all $p, q$ in $[M]$, $\left|\mu_i^p - \mu_i^q\right| \leq \epsilon$. This proves that $\mu = (\mu_i^p)_{i \in [K], p \in [M]}$ is indeed a Bernoulli $\epsilon$-MPMAB instance.

Recall that $d = \max_{i \in [c]} \mu_i^1 = \max_{i \in [K]} \mu_i^1$ is the optimal mean reward for player 1. We now show that $\mathcal{I}_\epsilon = \{c+1, \dots, K\}$ by a case analysis:

1. First, we show that for all $i$ in $\{c+1, \dots, K\}$, $i$ is in $\mathcal{I}_\epsilon$. This is because $\mu_i^1$ is chosen from $[0, d - 5\epsilon)$, which implies that $\Delta_i^1 > 5\epsilon$.

2. Second, for all $i$ in $\{1, \dots, c\}$, we claim that $i \notin \mathcal{I}_\epsilon$. To this end, we show that for all $p$, $\Delta_i^p \leq 5\epsilon$.

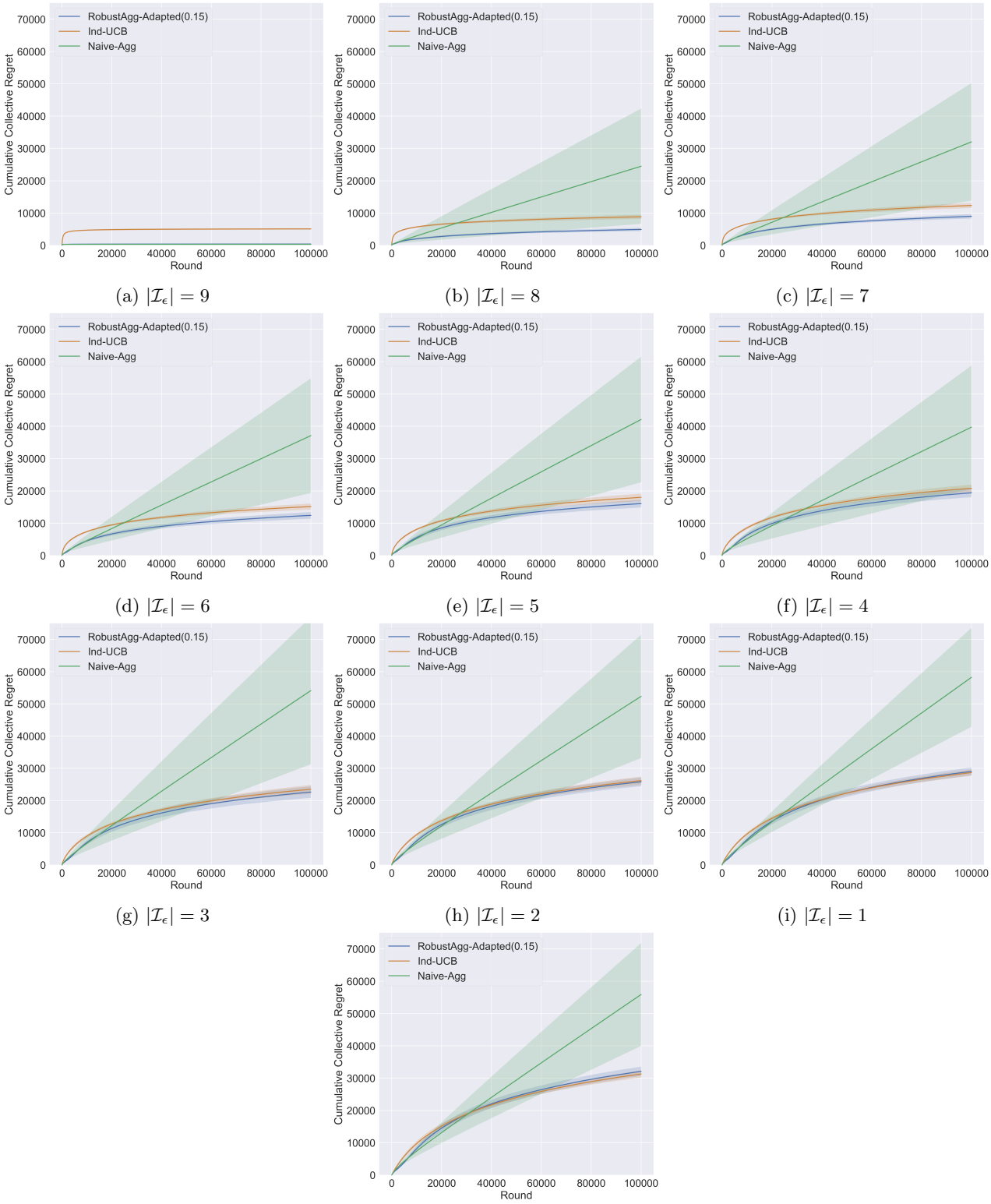
   We start with the observation that $\mu_i^1 \in (d - \epsilon, d]$, which implies that $\Delta_i^1 = d - \mu_i^1 \leq \epsilon$. Now, it follows from Fact 14 in Appendix C that for any $i \in [K]$ and $p \in [M]$, $|\Delta_i^p - \Delta_i^1| \leq 2\epsilon$. Therefore, we have $\Delta_i^p \leq 3\epsilon$ for all $p$, which implies that any $i \in [c]$ cannot be in $\mathcal{I}_\epsilon$.

□

### G.3 Extended results

Here, we present comprehensive results from the simulations we performed.

**Experiment 1.** Recall that for each $v \in \{0, 1, 2, \dots, 9\}$, we generated 30 Bernoulli 0.15-MPMAB problem instances such that $|\mathcal{I}_\epsilon| = v$. Figure 3 compares the average cumulative collective regrets of the three algorithms in a horizon of $100,000$ rounds over instances with different values of $|\mathcal{I}_\epsilon|$:



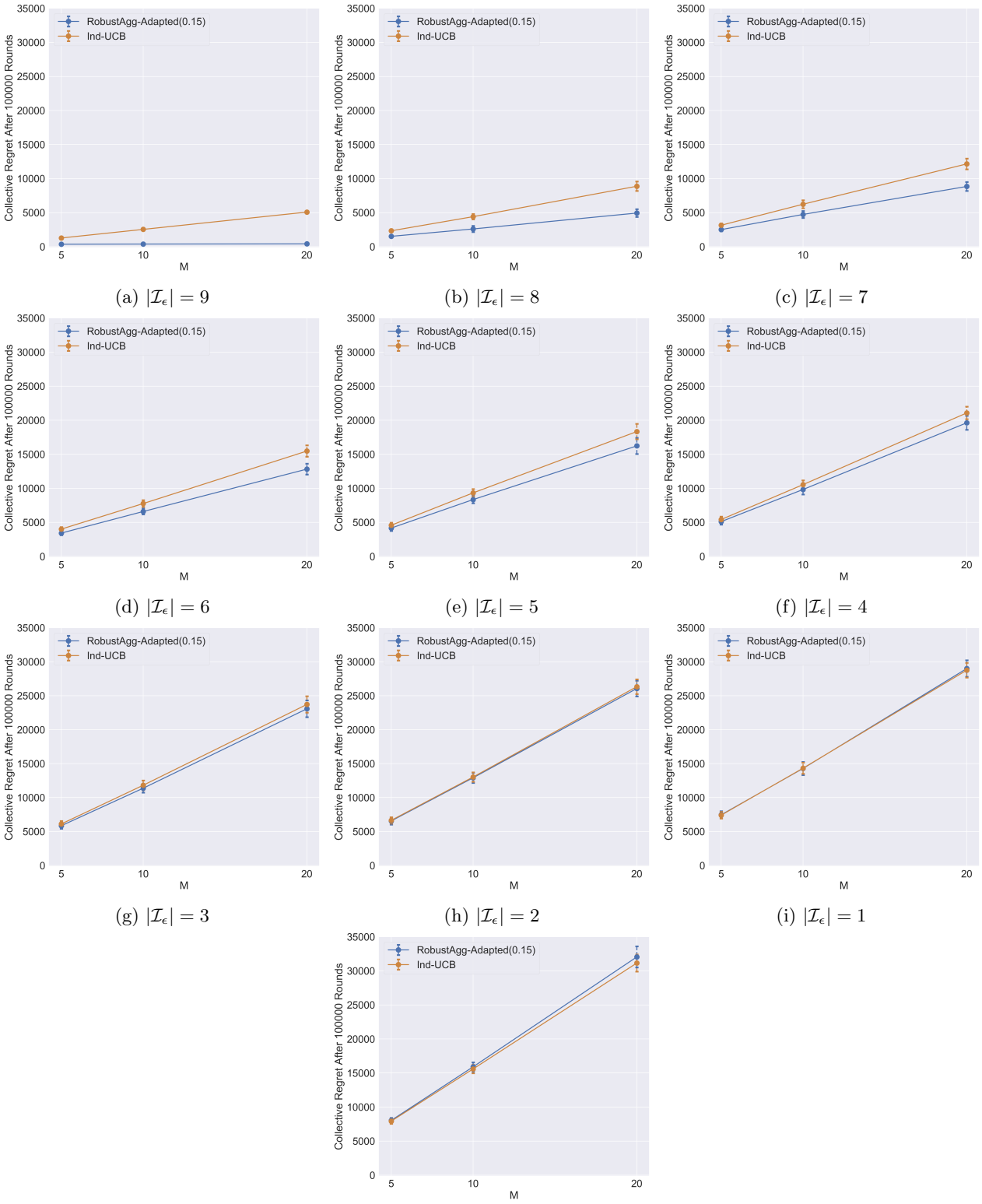

(a) $|\mathcal{I}_\epsilon| = 9$ (b) $|\mathcal{I}_\epsilon| = 8$ (c) $|\mathcal{I}_\epsilon| = 7$

(d) $|\mathcal{I}_\epsilon| = 6$ (e) $|\mathcal{I}_\epsilon| = 5$ (f) $|\mathcal{I}_\epsilon| = 4$

(g) $|\mathcal{I}_\epsilon| = 3$ (h) $|\mathcal{I}_\epsilon| = 2$ (i) $|\mathcal{I}_\epsilon| = 1$

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

Figure 3: Compares the average performance of RobustAgg-Adapted(0.15), Ind-UCB, and Naive-Agg over randomly generated Bernoulli 0.15-MPMAB problem instances with $K = 10$ and $M = 20$. The $x$-axis shows a horizon of $T = 100,000$ rounds, and the $y$-axis shows the cumulative collective regret of the players.




(a) $|\mathcal{I}_\epsilon| = 9$ (b) $|\mathcal{I}_\epsilon| = 8$ (c) $|\mathcal{I}_\epsilon| = 7$

(d) $|\mathcal{I}_\epsilon| = 6$ (e) $|\mathcal{I}_\epsilon| = 5$ (f) $|\mathcal{I}_\epsilon| = 4$

(g) $|\mathcal{I}_\epsilon| = 3$ (h) $|\mathcal{I}_\epsilon| = 2$ (i) $|\mathcal{I}_\epsilon| = 1$

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

Figure 4: Compares the average performance of RobustAgg-Adapted(0.15) and Ind-UCB over randomly generated Bernoulli 0.15-MPMAB problem instances with $K = 10$ and $M \in \{5, 10, 20\}$. The $x$-axis shows different values of $M$, and the $y$-axis shows the cumulative collective regret of the players after $100,000$ rounds.

- Notice that RobustAgg-Adapted(0.15) outperforms both baseline algorithms when $|\mathcal{I}_\epsilon| \in [2, 8]$, as shown in Figures 3b, 3c, ..., 3h, especially when $|\mathcal{I}_\epsilon|$ is large.

- Figure 3a shows that when $|\mathcal{I}_\epsilon| = 9$—i.e., when one arm is optimal for all players and the other arms are all subpar arms—Naive-Agg and RobustAgg-Adapted(0.15) perform much better than Ind-UCB with little difference between themselves. However, note that as long as there are more than one "competitive" arms—e.g., in Figure 3b when $|\mathcal{I}_\epsilon^C| = 2$—the collective regret of Naive-Agg can easily be nearly linear in the number of rounds.

- Figure 3i and Figure 3j demonstrate that when when there are very few arms or even no arm that is amenable to data aggregation, the performance of RobustAgg-Adapted(0.15) is still on par with that of Ind-UCB.

**Experiment 2.** Recall that for each $M \in \{5, 10, 20\}$ and $v \in \{0, 1, 2, \ldots, 9\}$, we generated 30 Bernoulli 0.15-MPMAB problem instances with $M$ players such that $|\mathcal{I}_\epsilon| = v$. Figure 4 shows and compares the average collective regrets of RobustAgg-Adapted(0.15) and Ind-UCB after 100,000 rounds in problem instances with $M = 5, 10$, and 20, and in each subfigure, $|\mathcal{I}_\epsilon|$ takes a different value.

Observe that when $|\mathcal{I}_\epsilon|$ is large (e.g., in Figures 4a, 4b,..., 4e), the collective regret of RobustAgg-Adapted(0.15) is less sensitive to the number of players $M$, in comparison with Ind-UCB. Especially, in the extreme case when $|\mathcal{I}_\epsilon| = 9$—i.e., all suboptimal arms are subpar arms—Figure 4a shows that the collective regret of RobustAgg-Adapted(0.15) has negligible dependence on $M$.

In conclusion, our empirical evaluation validate our theoretical results in Section 3.

## H Analytical Solution to $\lambda^*$

We first present the following proposition simliar to the results in (Ben-David et al., 2010, Section 6 thereof). The original solution in Ben-David et al. (2010) has a $\min(1, \cdot)$ operation in the second case; we slightly simplify that result by showing that this operation is unnecessary.[6]

**Proposition 32.** *Suppose $\beta \in (0, 1)$. Define function*

$$f(\alpha) = 2B\sqrt{\left(\frac{\alpha^2}{\beta} + \frac{(1-\alpha)^2}{1-\beta}\right)} + 2(1-\alpha)A,$$

*Then, $\alpha^* = \operatorname{argmin}_{\alpha\in[0,1]} f(\alpha)$ has the following form:*

$$\alpha^* = \begin{cases} 1 & \beta \geq \frac{B^2}{A^2}, \\ \beta\left(1 + \frac{1-\beta}{\sqrt{\frac{B^2}{A^2} - \beta(1-\beta)}}\right) & \beta < \frac{B^2}{A^2}. \end{cases}$$

Observe that when $\beta < \frac{B^2}{A^2}$, $\frac{B^2}{A^2} - \beta(1-\beta) > 0$, so the expression in the second case is well defined.

*Proof.* First, observe that $f$ is a strictly convex function, and therefore has at most one stationary point in $\mathbb{R}$; and if it exists, it must be $f$'s global minimum.

Second, we study the monotonicity property of $f$ in $\mathbb{R}$. To this end, we calculate $\alpha_0$, the stationary point of $f$. We have

$$f'(\alpha) = 2B\frac{\frac{\alpha}{\beta} - \frac{1-\alpha}{1-\beta}}{\sqrt{\frac{\alpha^2}{\beta} + \frac{(1-\alpha)^2}{1-\beta}}} - 2A$$

[6] In Ben-David et al. (2010)'s notation, this can also be seen directly by observing that when $m_T \geq D^2$, $v = \frac{m_T}{m_T+m_S} \cdot \left(1 + \frac{m_S}{\sqrt{D^2(m_S+m_T)-m_S m_T}}\right) \leq \frac{m_T}{m_T+m_S} \cdot (1 + \frac{m_S}{m_T}) = 1$.

By algebraic calculations, $f'(\alpha) = 0$ is equivalent to

$$\frac{\alpha - \beta}{\beta(1-\beta)} = \frac{A}{B}\sqrt{\frac{\alpha^2 - 2\beta\alpha + 1}{\beta(1-\beta)}}.$$

This yields the following quadratic equation:

$$\left(\frac{B^2}{A^2} - \beta(1-\beta)\right)\alpha^2 - 2\beta\left(\frac{B^2}{A^2} - \beta(1-\beta)\right)\alpha + \beta^2\left(\frac{B^2}{A^2} - (1-\beta)\right) = 0,$$

with the constraint that $\alpha > \beta$. The discriminant of the above quadratic equation is $\Delta = 4\beta^2(1-\beta)^2(\frac{B^2}{A^2} - \beta(1-\beta))$. If $\Delta \geq 0$, the stationary point is

$$\alpha_0 = \frac{2\beta(\frac{B^2}{A^2} - \beta(1-\beta)) + \sqrt{\Delta}}{2(\frac{B^2}{A^2} - \beta(1-\beta))} = \beta\left(1 + \frac{1-\beta}{\sqrt{\frac{B^2}{A^2} - \beta(1-\beta)}}\right)$$

We now consider two cases:

1. If $\beta(1-\beta) > \frac{B^2}{A^2}$, it can be checked that $\Delta < 0$, and consequently $f'(\alpha) < 0$ for all $\alpha \in \mathbb{R}$, i.e., $f$ is monotonically decreasing in $\mathbb{R}$.

2. $\beta(1-\beta) \leq \frac{B^2}{A^2}$, we have that $f$ is monotonically decreasing in $(-\infty, \alpha_0]$, and monotonically increasing in $[\alpha_0, +\infty)$.

We are now ready to calculate $\alpha^* = \operatorname{argmin}_{\alpha\in[0,1]} f(\alpha)$.

1. If $\beta(1-\beta) > \frac{B^2}{A^2}$, as $f$ is monotonically decreasing in $\mathbb{R}$, $\alpha^* = 1$.

2. If $\beta(1-\beta) \leq \frac{B^2}{A^2}$ and $\beta > \frac{B^2}{A^2}$, it can be checked that $\alpha_0 \geq 1$. As $f$ is monotonically decreasing in $(-\infty, \alpha_0] \supset [0, 1]$, we also have $\alpha^* = 1$.

3. If $\beta \leq \frac{B^2}{A^2}$, $\alpha_0 \in [0, 1]$. Therefore, $\alpha^* = \alpha_0 = \beta\left(1 + \frac{1-\beta}{\sqrt{\frac{B^2}{A^2} - \beta(1-\beta)}}\right)$.

In summary, we have the expression of $\alpha^*$ as desired. □

Algorithm 1's line 9 computes

$$\lambda^* = \operatorname*{argmin}_{\lambda\in[0,1]} \ 8\sqrt{13(\ln T)[\frac{\lambda^2}{\overline{n_i^p}(t-1)} + \frac{(1-\lambda)^2}{\overline{m_i^p}(t-1)}]} + (1-\lambda)\epsilon;$$

we now use Proposition 32 to give its analytical form. For notational simplicity, let $n = \overline{n_i^p}(t-1)$ and $m = \overline{m_i^p}(t-1)$. Applying Proposition 32 with $A = \frac{\epsilon}{2}, B = 4\sqrt{\frac{13(\ln T)}{n+m}}$, and $\beta = \frac{n}{n+m}$, we have

$$\lambda^* = \begin{cases} 1 & \epsilon > 0 \text{ and } n \geq \frac{832(\ln T)}{\epsilon^2}, \\ \frac{n}{n+m}\left(1 + \epsilon m\sqrt{\frac{1}{832(\ln T)(n+m) - \epsilon^2 nm}}\right) & \text{otherwise.} \end{cases}$$

# I On adaptive reward confidence interval construction under unknown $\epsilon$

Recall that RobustAgg($\epsilon$) carefully utilizes the dissimilarity parameter $\epsilon$ to construct high-probability reward confidence bounds for all arms and players by inter-player information sharing. In this section, we investigate the limitations of constructing reward confidence bounds by utilizing auxiliary data, in the setting when $\epsilon$ is unknown.

To this end, we consider a basic interval estimation problem: given source data $(z_1, \ldots, z_n)$ and target data $(x_1, \ldots, x_m)$ drawn iid from distributions $D_Z$ and $D_X$ supported on $[0,1]$ of unknown dissimilarity, how can one design an adaptive interval estimator for $\mu_X$, the mean of $D_X$? A naive baseline is to build the estimator by ignoring the source data: by Hoeffding's inequality, the interval centered at $\bar{x} = \frac{1}{m}\sum_{i=1}^m x_i$ of width $O(\sqrt{\frac{\ln \frac{1}{\delta}}{m}})$ is a $(1-\delta)$-confidence interval. This motivates the question: can one construct adaptive $(1-\delta)$-confidence intervals that become much narrower when $D_X$ and $D_Z$ are very close and $n$ is large?

We show in this section that the aforementioned naive interval estimation is about the best one can do: any "valid" confidence interval construction for $\mu_X$ must have width $\Omega(\sqrt{\frac{1}{m}})$ for a wide family of $(D_X, D_Z)$ where $D_X$ and $D_Z$ are identical to $D_Z$, regardless of the value of $n$. This is in sharp contrast to the results in the setting when a dissimiliarity parameter between $D_X$ and $D_Z$ is known: if we know that $D_X = D_Z$ *apriori*, it is easy to construct a confidence interval of $\mu_X$ of length $O(\sqrt{\frac{1}{n+m}})$ by setting its center at $\frac{1}{m+n}(\sum_{i=1}^m x_i + \sum_{i=1}^n z_i)$. Similar impossibility results of constructing adaptive and honest confidence intervals have appeared in (Low, 1997; Juditsky and Lambert-Lacroix, 2003) in nonparametric regression.

To formally present our results, we set up some useful notation. Denote by $S = (x_1, \ldots, x_m, z_1, \ldots, z_n)$ the sample we observe; in this notation, sample $S$ is drawn from $D_X^m \otimes D_Z^n$. The following notion formalizes the idea of valid confidence interval construction.

**Definition 33.** *$I : \mathbb{R}^{m+n} \times (0,1) \to \{[a,b] : 0 \le a \le b \le 1\}$ is said to be an honest confidence interval construction procedure for $\mu_X$, if under all distributions $D_X$ and $D_Z$ supported on $[0,1]$, and $\delta \in (0,1)$,*

$$\mathbb{P}_{S \sim D_X^m \otimes D_Z^n}\left(\mu_X \in I(S,\delta)\right) \ge 1 - \delta.$$

We have the following theorem.

**Theorem 34.** *Suppose $I$ is an honest confidence interval construction procedure for $\mu_X$. Then for any $m \ge 10$, $n$, $\mu \in [\frac{3}{8}, \frac{5}{8}]$, we have the following: under distributions $D_X = D_Z = \mathrm{Ber}(\mu)$,*

$$\mathbb{E}_{S \sim D_X^m \otimes D_Z^n}\left[\lambda\left(I(S,0.1)\right)\right] \ge \frac{1}{10\sqrt{m}},$$

*where $\lambda(J)$ denotes the length of interval $J$.*

*Proof.* We consider two hypotheses:

1. $H_1 : D_X = \mathrm{Ber}(\mu), D_Z = \mathrm{Ber}(\mu)$; under this hypothesis, $\mu_X = \mu$.
2. $H_2 : D_X = \mathrm{Ber}(\mu + \frac{1}{3\sqrt{m}}), D_Z = \mathrm{Ber}(\mu)$; under this hypothesis, $\mu_X = \mu + \frac{1}{3\sqrt{m}}$.

Denote by $\mathbb{P}_{H_1}$ and $\mathbb{P}_{H_2}$ the probability distributions of $S$ under hypotheses $H_1$ and $H_2$ respectively. As $I$ is an honest confidence interval procedure for $\mu_X$, we must have

$$\mathbb{P}_{H_1}\left(\mu \in I(S,0.1)\right) \ge 0.9, \tag{44}$$

and

$$\mathbb{P}_{H_2}\left(\mu + \frac{1}{3\sqrt{m}} \in I(S,0.1)\right) \ge 0.9, \tag{45}$$

holding simultaneously. We now show that $\mathbb{E}_{H_1}\left|I(S,0.1)\right| \ge \frac{1}{10\sqrt{m}}$.

We first establish a lower bound on $\mathbb{P}_{H_1}\left(\mu+\frac{1}{3\sqrt{m}}\in I(S,0.1)\right)$ using Eq. (45). The KL divergence between $\mathbb{P}_{H_1}$ and $\mathbb{P}_{H_2}$ can be bounded by:

$$\begin{aligned}\mathrm{KL}(\mathbb{P}_{H_1},\mathbb{P}_{H_2}) &= \sum_{i=1}^{m}\mathrm{KL}\left(\mathrm{Ber}(\mu),\mathrm{Ber}(\mu+\frac{1}{3\sqrt{m}})\right)+\sum_{i=1}^{n}\mathrm{KL}\left(\mathrm{Ber}(\mu),\mathrm{Ber}(\mu)\right)\\ &\le m\cdot 3(\frac{1}{3\sqrt{m}})^2+n\cdot 0=\frac{1}{3}.\end{aligned}$$

where the inequality uses Lemma 25 and $\mu,\mu+\frac{1}{3\sqrt{m}}\in[\frac{1}{4},\frac{3}{4}]$.

By Pinsker's inequality,

$$d_{\mathrm{TV}}(\mathbb{P}_{H_1},\mathbb{P}_{H_2})\le\sqrt{\frac{1}{2}\mathrm{KL}(\mathbb{P}_{H_1},\mathbb{P}_{H_2})}\le 0.5.$$

Therefore,

$$\mathbb{P}_{H_1}\left(\mu+\frac{1}{3\sqrt{m}}\in I(S,0.1)\right)\ge\mathbb{P}_{H_2}\left(\mu+\frac{1}{3\sqrt{m}}\in I(S,0.1)\right)-d_{\mathrm{TV}}(\mathbb{P}_{H_1},\mathbb{P}_{H_2})\ge 0.4.$$

Combining the above inequality with Eq. (44), using the fact that $\mathbb{P}(U\cap V)\ge\mathbb{P}(U)+\mathbb{P}(V)-1$, we have

$$\mathbb{P}_{H_1}\left(\mu\in I(S,0.1),\mu+\frac{1}{3\sqrt{m}}\in I(S,0.1)\right)\ge 0.9+0.4-1\ge 0.3.$$

Observe that if $\mu\in I(S,0.1)$ and $\mu+\frac{1}{3\sqrt{m}}\in I(S,0.1)$ both happens, $\lambda\left(I(S,0.1)\right)\ge\frac{1}{3\sqrt{m}}$. Therefore,

$$\mathbb{E}_{H_1}\left[\lambda\left(I(S,0.1)\right)\right]\ge\mathbb{P}_{H_1}\left(\mu\in I(S,0.1),\mu+\frac{1}{3\sqrt{m}}\in I(S,0.1)\right)\cdot\frac{1}{3\sqrt{m}}\ge\frac{1}{10\sqrt{m}}.$$ □

## References


A. Agarwal, H. Luo, B. Neyshabur, and R. E. Schapire. Corralling a band of bandit algorithms. In *Conference on Learning Theory*, pages 12–38. PMLR, 2017.

N. Alon, N. Cesa-Bianchi, C. Gentile, S. Mannor, Y. Mansour, and O. Shamir. Nonstochastic multi-armed bandits with graph-structured feedback. *SIAM Journal on Computing*, 46(6):1785–1826, 2017.

R. Arora, T. V. Marinov, and M. Mohri. Corralling stochastic bandit algorithms. *arXiv preprint arXiv:2006.09255*, 2020.

B. Awerbuch and R. D. Kleinberg. Competitive collaborative learning. In *International Conference on Computational Learning Theory*, pages 233–248. Springer, 2005.

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

Y. Bar-On and Y. Mansour. Individual regret in cooperative nonstochastic multi-armed bandits. In *Advances in Neural Information Processing Systems*, pages 3110–3120, 2019.

P. Bartlett, V. Dani, T. Hayes, S. Kakade, A. Rakhlin, and A. Tewari. High-probability regret bounds for bandit online linear optimization. In T. Zhang and R. A. Sevedio, editors, *Proceedings of the 21st Annual Conference on Learning Theory - COLT 2008*, pages 335–342, United States, 2008. Omnipress.

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

I. Bistritz, T. Baharav, A. Leshem, and N. Bambos. My fair bandit: Distributed learning of max-min fairness with multi-player bandits. In *Proceedings of the 37th International Conference on Machine Learning*, pages 11–21, 2020.

E. Boursier and V. Perchet. Selfish robustness and equilibria in multi-player bandits. volume 125 of *Proceedings of Machine Learning Research*, pages 530–581, 2020.

E. Boursier, E. Kaufmann, A. Mehrabian, and V. Perchet. A practical algorithm for multiplayer bandits when arm means vary among players. volume 108 of *Proceedings of Machine Learning Research*, pages 1211–1221, 2020.

G. Bresler, G. H. Chen, and D. Shah. A latent source model for online collaborative filtering. In *Advances in Neural Information Processing Systems*, pages 3347–3355, 2014.

S. Bubeck and T. Budzinski. Coordination without communication: optimal regret in two players multi-armed bandits. volume 125 of *Proceedings of Machine Learning Research*, pages 916–939, 2020.

S. Bubeck, Y. Li, Y. Peres, and M. Sellke. Non-stochastic multi-player multi-armed bandits: Optimal rate with collision information, sublinear without. In *Conference on Learning Theory*, pages 961–987, 2020.

S. Buccapatnam, A. Eryilmaz, and N. B. Shroff. Stochastic bandits with side observations on networks. In *The 2014 ACM international conference on Measurement and modeling of computer systems*, pages 289–300, 2014.

S. Caron, B. Kveton, M. Lelarge, and S. Bhagat. Leveraging side observations in stochastic bandits. In *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*, UAI'12, page 142–151, Arlington, Virginia, USA, 2012. AUAI Press. ISBN 9780974903989.

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

N. Cesa-Bianchi, C. Gentile, and Y. Mansour. Delay and cooperation in nonstochastic bandits. *The Journal of Machine Learning Research*, 20(1):613–650, 2019.

R. Chawla, A. Sankararaman, A. Ganesh, and S. Shakkottai. The gossiping insert-eliminate algorithm for multi-agent bandits. volume 108 of *Proceedings of Machine Learning Research*, pages 3471–3481, 2020.

K. Christakopoulou and A. Banerjee. Learning to interact with users: A collaborative-bandit approach. In *Proceedings of the 2018 SIAM International Conference on Data Mining*, pages 612–620. SIAM, 2018.

A. A. Deshmukh, U. Dogan, and C. Scott. Multi-task learning for contextual bandits. In *Advances in neural information processing systems*, pages 4848–4856, 2017.

A. Dubey and A. Pentland. Cooperative multi-agent bandits with heavy tails. In *Proceedings of the 37th International Conference on Machine Learning*, pages 451–460, 2020a.

A. Dubey and A. Pentland. Kernel methods for cooperative contextual bandits. In *Proceedings of the 37th International Conference on Machine Learning*, pages 428–450, 2020b.

R. Feraud, R. Alami, and R. Laroche. Decentralized exploration in multi-armed bandits. In *International Conference on Machine Learning*, pages 1901–1909, 2019.

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

A. Juditsky and S. Lambert-Lacroix. Nonparametric confidence set estimation. *Mathematical Methods of Statistics*, 12(4):410–428, 2003.

D. Kalathil, N. Nayyar, and R. Jain. Decentralized learning for multiplayer multiarmed bandits. *IEEE Transactions on Information Theory*, 60(4):2331–2345, 2014.

S. Kar, H. V. Poor, and S. Cui. Bandit problems in networks: Asymptotically efficient distributed allocation rules. In *2011 50th IEEE Conference on Decision and Control and European Control Conference*, pages 1771–1778. IEEE, 2011.

R. K. Kolla, K. Jagannathan, and A. Gopalan. Collaborative learning of stochastic bandits over a social network. *IEEE/ACM Transactions on Networking*, 26(4):1782–1795, 2018.

N. Korda, B. Szorenyi, and S. Li. Distributed clustering of linear bandits in peer to peer networks. In *International Conference on Machine Learning*, pages 1301–1309, 2016.

P. Landgren, V. Srivastava, and N. E. Leonard. Distributed cooperative decision-making in multiarmed bandits: Frequentist and bayesian algorithms. In *2016 IEEE 55th Conference on Decision and Control (CDC)*, pages 167–172. IEEE, 2016.

P. C. Landgren. *Distributed Multi-Agent Multi-Armed Bandits*. PhD thesis, Princeton University, 2019.

T. Lattimore and C. Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

S. Li, A. Karatzoglou, and C. Gentile. Collaborative filtering bandits. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 539–548, 2016.

S. Li, W. Chen, S. Li, and K.-S. Leung. Improved algorithm on online clustering of bandits. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 2923–2929. AAAI Press, 2019.

K. Liu and Q. Zhao. Distributed learning in multi-armed bandit with multiple players. *IEEE Transactions on Signal Processing*, 58(11):5667–5681, 2010.

M. G. Low. On nonparametric confidence intervals. *The Annals of Statistics*, 25(6):2547–2554, 1997.

S. Mannor and O. Shamir. From bandits to experts: On the value of side-observations. In *Advances in Neural Information Processing Systems*, pages 684–692, 2011.

T. T. Nguyen and H. W. Lauw. Dynamic clustering of contextual multi-armed bandits. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 1959–1962, 2014.

A. Sankararaman, A. Ganesh, and S. Shakkottai. Social learning in multi agent multi armed bandits. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 3(3):1–35, 2019.

S. Shahrampour, A. Rakhlin, and A. Jadbabaie. Multi-armed bandits in multi-agent networks. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2786–2790. IEEE, 2017.

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

C. Shi, W. Xiong, C. Shen, and J. Yang. Decentralized multi-player multi-armed bandits with no collision information. volume 108 of *Proceedings of Machine Learning Research*, pages 1519–1528, 2020.

P. Shivaswamy and T. Joachims. Multi-armed bandit problems with history. In *Artificial Intelligence and Statistics*, pages 1046–1054, 2012.

A. Slivkins. Contextual bandits with similarity information. *The Journal of Machine Learning Research*, 15(1): 2533–2568, 2014.

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

L. Song, C. Tekin, and M. Van Der Schaar. Online learning in large-scale contextual recommender systems. *IEEE Transactions on Services Computing*, 9(3):433–445, 2014.

B. Szörényi, R. Busa-Fekete, I. Hegedűs, R. Ormándi, M. Jelasity, and B. Kégl. Gossip-based distributed stochastic bandit algorithms. In *Journal of Machine Learning Research Workshop and Conference Proceedings*, volume 2, pages 1056–1064. International Machine Learning Society, 2013.

D. Vial, S. Shakkottai, and R. Srikant. Robust multi-agent multi-armed bandits. *arXiv preprint arXiv:2007.03812*, 2020.

H. Wang, Q. Wu, and H. Wang. Factorization bandits for interactive recommendation. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017a.

P.-A. Wang, A. Proutiere, K. Ariu, Y. Jedra, and A. Russo. Optimal algorithms for multiplayer multi-armed bandits. volume 108 of *Proceedings of Machine Learning Research*, pages 4120–4129, 2020.

Q. Wang, C. Zeng, W. Zhou, T. Li, S. S. Iyengar, L. Shwartz, and G. Y. Grabarnik. Online interactive collaborative filtering using multi-armed bandit with dependent arms. *IEEE Transactions on Knowledge and Data Engineering*, 31(8):1569–1580, 2018.

X. Wang, S. C. Hoi, C. Liu, and M. Ester. Interactive social recommendation. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 357–366, 2017b.

Y. Wang, J. Hu, X. Chen, and L. Wang. Distributed bandit learning: How much communication is needed to achieve (near) optimal regret. *arXiv preprint arXiv:1904.06309*, 2019.

C.-Y. Wei, H. Luo, and A. Agarwal. Taking a hint: How to leverage loss predictors in contextual bandits? volume 125 of *Proceedings of Machine Learning Research*, pages 3583–3634, 2020.

Q. Wu, H. Wang, Q. Gu, and H. Wang. Contextual bandits in a collaborative environment. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 529–538, 2016.

Y. Wu, A. György, and C. Szepesvári. Online learning with gaussian payoffs and side observations. In *Advances in Neural Information Processing Systems*, pages 1360–1368, 2015.

X. Xu, S. Vakili, Q. Zhao, and A. Swami. Online learning with side information. In *MILCOM 2017 - 2017 IEEE Military Communications Conference (MILCOM)*, pages 303–308, 2017.

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

Z. Zhu, L. Huang, and H. Xu. Collaborative thompson sampling. *Mobile Networks and Applications*, 2020. doi: 10.1007/s11036-019-01453-x.