The Learnability of the Multiplayer Adversarial Bandit Problem

Anonymous authors
Paper under double-blind review

Abstract

We examine a novel multiplayer extension of the latent multi-armed bandit problem as formulated in Maillard & Mannor (2014), with broad applications such as recommendation systems and cognitive radio. Following Chang et al. (2022), we examine three information asymmetric scenarios: Problem A, in which players receive identical rewards but cannot observe each other's actions; Problem B, players receive private i.i.d rewards but can observe others' actions; and Problem C, players receive private i.i.d rewards and cannot observe others' actions. For problems A and B, we provide nearly optimal gap-independent regret bounds. When reduced to the single agent setting, our results improve on Maillard & Mannor (2014) by allowing for adversarial nature's actions. For Problem C, we use the knowledge of the reward means to improve on the results in Chang et al. (2022).

1 Introduction

The single-player multi-armed bandit problem is a general reinforcement learning problem where a single player has access to multiple arms. Each arm is associated with an unknown reward distribution, varying potentially in rewards, and the agent attempts to maximize cumulative reward by observing the empirical rewards, which converge to the true mean of each arm. Maillard & Mannor (2014) proposes the latent multi-armed bandit problem. This has applications in recommendation systems (Li et al. (2010)), where the agent selects an advertisement to show each arriving user, using a context vector that captures information such as the user's browsing behavior or geographic location. In cognitive radio (Avner et al. (2012)), the agent must choose a communication channel based on its current location and network status, while avoiding interference with other sources like radar or WiFi. However, relying solely on observable context may not be enough to achieve optimal solutions. In recommendation systems, key details like a user's gender or income are often unavailable due to privacy concerns. Similarly, in cognitive radio settings, it is unclear whether other sources or users are nearby or distant. In both scenarios, crucial aspects of the reward structure remain unobserved, but the latent structures can be inferred. In this paper, we extend latent bandits to the multiplayer setting.

Multi-agent reinforcement learning (MARL) has been applied to various domains such as autonomous driving Shalev-Shwartz et al. (2016), strategic board games like Go Silver et al. (2016), real-time strategy games, robotic control Kober et al. (2013), and card games Brown et al. (2017); Brown & Sandholm (2019). These successes are often enabled by deep neural networks and frequently involve multiple agents, highlighting the importance of studying MARL—where autonomous agents interact within a shared environment to maximize long-term rewards Busoniu et al. (2008). Beyond gaming, MARL is increasingly relevant in domains like cyber-physical systems Wang et al. (2016), finance Lee et al. (2007), sensor networks Choi et al. (2009), and social sciences Leibo et al. (2017).

The main drawback in previous multi-agent works is that they usually allow for communication or do not consider joint actions. In the case of games where both of these assumptions are not necessary, they do not seek to find a global optimal action for the players but rather some type of equilibrium. Therefore we consider the line of work that addresses partial or no observability of other players' actions, rewards or both. To address the complexity of multi-agent systems, Chang et al. (2022); Chang & Lu (2023) proposes algorithms

where players face information asymmetry in rewards, actions or both; Chang & Lu (2025) and Chang & Karthik (2025) extend this framework to contextual and metric bandits, achieving sub-linear upper bounds on regret. We address the limited-communication setting by leveraging players' ability to coordinate before online learning begins; we also account for the fact that one player's actions may influence another's rewards through the analysis of *joint actions*.

We also examine three cases of information asymmetry, in which players do not observe others' rewards or actions but must coordinate to minimize their regret without utilizing an explicit communication channel. We investigate three types of information asymmetry: In Problem A, we are faced with the problem of asymmetry in actions, where the players cannot see each other's actions but receive identical rewards. Problem B is the opposite, as players can see each other's actions but receive private i.i.d rewards. Finally, in Problem C, players are unable to see others' actions and receive private i.i.d rewards.

Our contribution First, we utilize the framework proposed in Chang et al. (2022) to extend the multiple-cluster arrival problem in Maillard & Mannor (2014) to a novel multi-agent setting, accounting for limited communication and information asymmetry between agents. Second, we show that our results improve on the work of Maillard & Mannor (2014) in the single-agent case by allowing for adversarial, rather than stochastic, nature's actions. Finally, we provide a gap-independent regret bound which also holds for the single-agent case, improving on Maillard & Mannor (2014). In section 3, we show that the cumulative regret of `Multiple-K-Intervals-A` under problem A satisfies $\mathcal{R}_T \leq O(\sqrt{T \log T} BC)$. In section 3.2, we propose `Multiple-K-Intervals-B` for problem B, which also satisfies $\mathcal{R}_T \leq O(\sqrt{T \log T} BC)$. In section 3.3, we propose `Multiple-K-mDSEE` for problem C, with $\mathcal{R}_T \leq O(K^M B \log(T))$.

Related Works Past works have explored multi-armed bandits with latent clusters (MAB-LC) within various settings. Agrawal et al. (1989) provides an early formulation of the MAB problem in which reward distributions are parametrized by an unknown parameter from a known parameter space, achieving asymptotic efficiency while providing a lower bound significantly differing from the standard bound for multi-armed bandits.

We focus on latent bandits as introduced in Maillard & Mannor (2014), where nature's actions belong to clusters with known reward distributions (both spaces being discrete and finite). Maillard & Mannor (2014) considers the setting in which all states belong to a single cluster and propose `Single-K-UCB`, which improves on Agrawal et al. (1989). The authors also consider a setting in which nature's actions can belong to any of several known clusters (the multiple-cluster arrival case), achieving gap-dependent sublinear regret with `Multiple-K-UCB`; this is the problem we examine below.

Motivated by problems in which abundant data allows patterns to be learned offline, there exists a large body of work on regret minimization for latent bandits, such as Pal et al. (2023b), Pal et al. (2023a) and Hong et al. (2020a). Kinyanjui et al. (2023) addresses fixed-confidence pure-exploration for latent bandits. Another line of work (Hong et al. (2020b), Gentile et al. (2014), Zhou & Brunskill (2016), Bresler et al. (2014)) applies the latent bandit problem to recommendation systems, using a combination of offline learning to infer latent structures and online learning to provide personalized recommendations.

Kausik et al. (2024) and Gupta et al. (2020) investigate more structured variants of the MAB-LC problem; in particular, Gupta et al. (2020) assumes rewards are functions of a common latent variable and identifies competitive arms in O(1).

In cooperative multiplayer bandit problems, players aim to collectively identify the best arm from a shared set. Communication between players is often modeled by a graph structure, a framework initially proposed by Awerbuch & Kleinberg (2008). Over time, various strategies have been developed, including ϵ-greedy methods Szorenyi et al. (2013), gossip-based UCB variants Landgren et al. (2016); Martínez-Rubio et al. (2019), and leader-based coordination Wang et al. (2020). In adversarial contexts, Bar-On & Mansour (2019) proposed a setup where non-leader players follow the EXP3 algorithm. Other studies allow players to observe neighbors' rewards according to the graph topology Cesa-Bianchi et al. (2016), and some consider asynchronous environments where only certain players are active in each round Bonnefoi et al. (2017); Cesa-Bianchi et al. (2020). In collision models—where players receive no reward if they select the same

arm—joint actions are not considered. An extension to Lipschitz bandits was explored in Proutiere & Wang (2019), which introduced the DPE algorithm to reduce communication while maximizing collective rewards.

The competing bandits model, introduced by Liu et al. (2020), builds on collision settings by introducing player preferences: when multiple players choose the same arm, only the top-ranked one gains the reward. A centralized algorithm (CUB) was proposed for this setup, with players reporting their UCB indices to a central authority. Later work Cen & Shah (2022) showed that optimal logarithmic regret is attainable when the platform handles transfers between players and arms. Jagadeesan et al. (2021) further enhanced this with stronger equilibrium concepts involving negotiated transfers. An Explore-Then-Commit (ETC) algorithm from Liu et al. (2020) achieves similar regret bounds without requiring transfers, assuming known reward gaps. This assumption was later removed by Sankararaman et al. (2021). Additionally, Liu et al. (2021a) proposed a decentralized UCB method with built-in collision avoidance.

Competitive MARL, often modeled as zero-sum Markov games Littman (1994), captures adversarial dynamics where one agent's gain is another's loss—supporting robust policy development under uncertainty Zhang et al. (2020). General-sum games lie between these extremes, featuring agents with differing or conflicting objectives. In such settings, equilibrium concepts like Nash equilibrium guide the learning process Basar & Olsder (1999); Bistritz & Leshem (2018). Recent work by Ding et al. (2023) tackles the problem of last-iterate convergence in constrained Markov decision processes (MDPs), a setting previously dominated by approaches using average-policy or two-timescale updates. In contrast to earlier efforts Moskovitz et al. (2023); Liu et al. (2021b); Li et al. (2024), they provide the first global non-asymptotic guarantees for single-timescale methods in constrained RL.

2 Preliminary

2.1 Single-player latent bandits

First, we present the single-player latent bandit problem, following some notation from Maillard & Mannor (2014). For each round t from 1 to the horizon T, nature selects an action b from the set of all possible states $\mathcal{B}$, where $|\mathcal{B}| = B$, which is visible to the player. The player then takes an action a from $\mathscr{A}$ (with $|\mathscr{A}| = K$) and receives a real-valued reward sampled from the distribution $\nu_{a,b}$. We assume that each $\nu_{a,b}$ is 1-sub-Gaussian and $\mathbb{E}[\nu_{a,b}] = \mu_{a,b}$. The set $\mathcal{B}$ is further partitioned into C clusters $\{\mathcal{B}_c\}_{c=1,\dots,C}$, where the reward distributions $\nu_{a,b}$ for a given action a are identical for all b in a cluster $\mathcal{B}_c$.

2.2 Multiplayer latent bandits

We now extend this to the multiplayer problem as examined in Chang et al. (2022), Chang & Lu (2023). Suppose there are M players. The players take a *joint action* $\boldsymbol{a}$, where we express $\boldsymbol{a}$ as a tuple of individual player actions at time t, $\boldsymbol{a}_t = (a_1, a_2, \dots, a_M)$. Each player then receives a real-valued reward sampled from $\nu_{\boldsymbol{a},b}$.

We examine the scenario from Maillard & Mannor (2014) where $\{\nu_{\boldsymbol{a},c}\}_{\boldsymbol{a}\in\mathscr{A},c\in\mathcal{C}}$ is known to the players, i.e. b can come from any of several clusters with known reward distributions. Players are not given the true cluster of each state b; however, they are given $\star_c = \arg\max_{\boldsymbol{a}} \mu_{\boldsymbol{a},c}$, the optimal arm within each cluster. Thus the problem reduces to identifying the true cluster each state belongs to. Let $\star_b = \arg\max_{\boldsymbol{a}} \mu_{\boldsymbol{a},b}$, the optimal joint action for each state b. Then maximizing reward is equivalent to minimizing the regret, which we define as

$$\mathcal{R}_T = \sum_{t=1}^{T} \left(\mathbb{E}_{X_t \sim \nu_{\star_{b_t},b_t}}[X_{\star_{\boldsymbol{b_t}},b_t}] - \mathbb{E}_{X_t \sim \nu_{\boldsymbol{a_t},b_t}}[X_{\boldsymbol{a_t},b_t}] \right)$$

or equivalently, $\mathcal{R}_T = \sum_{t=1}^{T} \left(\mu_{\star_{c_{b_t}},c_{b_t}} - \mu_{\boldsymbol{a}_t,c_{b_t}} \right)$, where c_{b_t} is the cluster b_t belongs to.

Additionally, denote the number of observations for the pair $(\boldsymbol{a}, b)$ at time t by $N_{\boldsymbol{a},b}(t) = \sum_{n=1}^{t} \mathbb{I}\{\boldsymbol{a}_n = \boldsymbol{a}, b_n = b\}$ and let the empirical mean built from the same observations be $\hat{\mu}_{\boldsymbol{a},b}(t)$. Define $N_b(t) = \sum_{\boldsymbol{a}\in\mathscr{A}} N_{\boldsymbol{a},b}(t)$.

To estimate a confidence interval for each $\mu_{\boldsymbol{a},b}$ and identify the corresponding cluster for b, we construct a confidence interval based on $\hat{\mu}_{\boldsymbol{a},b}(t)$ (following Maillard & Mannor (2014)). Let $S_{\boldsymbol{a},b}(t) = (U_{\boldsymbol{a},b}(t), L_{\boldsymbol{a},b}(t))$, where

$$U_{\boldsymbol{a},b}(t) = \hat{\mu}_{\boldsymbol{a},b}(t) + \sqrt{4\log(T)/N_{\boldsymbol{a},b}(t)}$$

and

$$L_{\boldsymbol{a},b}(t) = \hat{\mu}_{\boldsymbol{a},b}(t) - \sqrt{4\log(T)/N_{\boldsymbol{a},b}(t)}.$$

Note that $\mu_{\boldsymbol{a},b} \in S_{\boldsymbol{a},b}(t)$ is true with high probability (see Lemma A). Additionally, if $\mu_{\boldsymbol{a},c} \in S_{\boldsymbol{a},b}(t)$ for all $\boldsymbol{a}$ then we consider c to be an admissible cluster for nature's action b.

We assume players cannot explicitly communicate with each other during the learning period, but can coordinate on a strategy beforehand. During this strategy meeting, they know the number of actions each player has access to, as well as the horizon T. We seek to minimize the cumulative regret across T rounds.

We consider several information asymmetric scenarios as in Chang et al. (2022); Chang & Lu (2023), as follows:

Problem A: Information Asymmetry in Actions. In Problem A, we assume that all players receive identical rewards but cannot explicitly observe the actions of other players. We propose a round-robin scheme that allows players to inductively infer each other's actions by maintaining identical admissible sets, enabling consistent updates and avoiding suboptimal joint actions.

Problem B: Information Asymmetry in Rewards. In Problem B, we assume that players receive independent and identically distributed (i.i.d.) rewards that are unknown to other players, but can observe the actions of others. We build on the algorithm presented for Problem A with the addition of *sabotage*, which allows players to implicitly signal belief changes regarding the admissibility of clusters.

Problem C: Asymmetry in Rewards and Actions. In Problem C, we consider the most information-constrained setting: players receive independent and identically distributed rewards, and observe neither the rewards nor the actions of other players. We propose an explore-then-commit-style scheme run in parallel for each nature's action b, where players follow a synchronized exploration schedule based on prior coordination and then independently commit to their estimated optimal arm once sufficient confidence is achieved.

3 Main Results

3.1 Problem A: Information Asymmetry in Actions

In Problem A, players receive identical rewards but cannot observe others' actions. Therefore, players may become miscoordinated in case of ties between arms. Consider a scenario where there are two players and two optimal actions. Player 1 intends to take $(1, 2)$ whereas Player 2 intends to take $(2, 1)$; since each player controls only their respective entry in the tuple, the resulting joint action taken is $(1, 1)$. Thus players must agree on an ordering scheme for arms and clusters before the start of learning to ensure that ties are broken consistently across all players. Any order for arms suffices if all players agree—one option is the lexicographical ordering proposed in Chang et al. (2021). Similarly, players should agree on an ordering for clusters.

As rewards from chosen actions are observed and empirical reward means change, the ranking of arms may change. However, the original ordering remains a constant reference point for resolving any new ties that occur.

Intuition: We propose Algorithm 1, which enables players to gradually eliminate non-admissible clusters utilizing confidence intervals $S_{\boldsymbol{a},b}(t)$. Note that we take an optimistic approach, where at round $t = 1$, all clusters are assumed to be admissible for all states. When multiple clusters are admissible for a nature's

action, players "test" clusters according to a round-robin scheme as follows, where clusters are ordered as above:

$$c_1 \to c_2 \to \cdots \to c_C \tag{1}$$

The confidence intervals will converge around the true means $\mu_{\boldsymbol{a},b}$, allowing players to identify the true cluster for each nature's action. Because players receive identical rewards and maintain identical confidence intervals, they also maintain identical admissible sets at all times. Thus players remain synchronized and can infer other players' actions, allowing them to update confidence intervals correctly.

We assume that the true mean $\mu_{\boldsymbol{a},b}$ is inside all players' confidence intervals $S_{a,b}(t)$ with high probability, enabling a gap-independent regret analysis based on confidence interval width. Further, Algorithm 1 serves as a baseline that can be modified for multiple types of information asymmetry, as in section 3.2.

Algorithm 1 The `Multiple-K-Intervals-A` Algorithm

Input: The cluster distributions $\{\nu_{\boldsymbol{a},c}\}$
for all $b \in \mathcal{B}$ **do**
    $\mathcal{C}_b \leftarrow \mathcal{C}$
end for
for $t = 1, \ldots, T$ **do**
    Receive $b \in \mathcal{B}$ and suppose this action was chosen at rounds $t_1, \ldots, t_n = t$
    Players construct $S_{\boldsymbol{a},b}(t)$ for all $\boldsymbol{a}$
    if $N_{\boldsymbol{a},b}(t) = 0$ **then**
        $S_{\boldsymbol{a},b}(t) \leftarrow (-\infty, \infty)$
    end if
    Select the next cluster $c(t_n)$ following $c(t_{n-1})$ in the ordered admissible set $\mathcal{C}_b$
    if $\mu_{\boldsymbol{a},c(t_n)} \notin S_{\boldsymbol{a},b}(t)$ for any $\boldsymbol{a}$ **then**
        $\mathcal{C}_b \leftarrow \mathcal{C}_b \setminus \{c(t_n)\}$
    end if
    Choose the next arm (break ties with ordering of arms)
    $\boldsymbol{a}_t \leftarrow \star_{c(t_n)}$
end for

The regret bound is as follows, with the proof deferred to the Appendix.

Theorem 1 *The regret of Multiple-K-Intervals-A satisfies*

$$\mathcal{R}_{\mathcal{T}} \leq 8\sqrt{4T(\log T)}BC + 2BC \tag{2}$$

Note a comparison with the regret bound of `Multiple-K-UCB` in Theorem 6 of Maillard & Mannor (2014). Their gap dependent bound contains a $\frac{\Delta_{\boldsymbol{a},c_b}}{(\Delta^+_{\boldsymbol{a},c_b})^2}$ term, where $\Delta^+_{\boldsymbol{a},\mathbf{c}} = \inf_{c' \in \mathcal{C}} \left\{\mu_{a,c'} - \mu_{\boldsymbol{a},c} : \star_{c'} = a \cap \mu_{\star_{c'},c'} \geqslant \mu_{\star_{\mathbf{c}},c}\right\}$. Since $\Delta^+_{\boldsymbol{a},c} \geq \Delta_{\boldsymbol{a},c}$, their regret bound is bounded by $\frac{1}{\Delta^+_{\boldsymbol{a},c_b}}$, which essentially results in a gap-independent bound of the same order as the one presented above. However, the proof technique we employ allows us to account for the adversarial nature's actions. Furthermore, a lower bound on the cumulative regret of any latent bandit algorithm, in the case $C > K^M$, is $\frac{1}{20}\sqrt{TK^MC}$, as shown in Maillard & Mannor (2014). Our bound is nearly optimal up to factor of logs.

3.1.1 Example

We now consider a setting to illustrate `Multiple-K-Intervals-A` in the case where there are three nature's actions and two clusters. Let nature's actions be denoted by b_1, b_2, b_3, with $b_1, b_3 \in \mathcal{B}_1$ (i.e., belonging to Cluster 1) and $b_2 \in \mathcal{B}_2$. Each cluster defines a reward distribution following Gaussian distributions $\mathcal{N}(\mu_{a,c}, 0.2^2)$ over joint actions $\boldsymbol{a} = (a_1, a_2) \in \{1, 2\} \times \{1, 2\}$.

Define $\mu_{\boldsymbol{a},c}$ and hence the cluster distributions as:

$$\underbrace{\begin{array}{cc} & \begin{array}{cc} 1 & 2 \end{array} \\ \begin{array}{c} 1 \\ 2 \end{array} & \begin{bmatrix} 0.2 & 0.6 \\ 0.8 & 0.5 \end{bmatrix} \end{array}}_{\text{Cluster 1}}, \underbrace{\begin{array}{cc} & \begin{array}{cc} 1 & 2 \end{array} \\ \begin{array}{c} 1 \\ 2 \end{array} & \begin{bmatrix} 0.8 & 0.1 \\ 0.7 & 0.3 \end{bmatrix} \end{array}}_{\text{Cluster 2}}$$

The rows represent Player 1's actions while the columns represent Player 2's actions.

Suppose the players agreed prior to learning to let the ordering of clusters be $(1, 2)$, and the order of joint arms be:

$$(1,1) \to (1,2) \to (2,1) \to (2,2) \tag{3}$$

Note that any order will suffice as long as the players use the same ordering. Initially, all clusters are admissible for all players under each state; as time progresses, players update their empirical means and confidence intervals for each joint arm under each state.

Suppose at round $t = 30$, the following occurs:

- Nature draws state $b_1 \in \mathcal{B}_1$ (true cluster is Cluster 1).
- The algorithm is currently testing Cluster 2 for b_1. Both players believe both clusters are admissible for b_1.
- For the optimal joint arm under Cluster 2, $(1, 1)$, the true mean under Cluster 2 is 0.8.
- Both players have pulled $(1, 1)$ many times under b_1 and observe (note that rewards are identical):

$$\text{Player 1: } \hat{\mu}^{(1)}_{(1,1),b_1} = 0.3, \quad S^{(1)} = [0.1, 0.5]$$

$$\text{Player 2: } \hat{\mu}^{(2)}_{(1,1),b_1} = 0.3, \quad S^{(2)} = [0.1, 0.5]$$

Since $\mu_{(1,1),2} = 0.8 \notin S^{(1)}$ and $\mu_{(1,1),2} \notin S^{(2)}$, both players simultaneously identify Cluster 2 as inadmissible under b_1. Both players update: $\mathcal{C}_{b_1} \leftarrow \mathcal{C}_{b_1} \setminus \{2\}$. If nature selects b_1 in the future, both players will aim for action $(2, 1)$, the optimal arm of the only admissible cluster for b_1. Later, nature selects b_3, which also belongs to $\mathcal{B}_1$. If both clusters are currently admissible for b_3, then players may still test Cluster 2 by pulling its optimal arm. Suppose nature selects b_3 again; then players will test the next admissible cluster, which is Cluster 1.

3.2 Problem B: Information Asymmetry in Rewards

The approach used in Problem A relies on all players observing identical rewards, allowing them to construct identical confidence intervals. This enables players to remain synchronized without needing to observe others' actions. In Problem B, however, players receive different rewards and maintain different confidence intervals. Consequently, players may disagree on which clusters are admissible, creating the potential for miscoordination.

We address this by building on Algorithm 1 and introducing *sabotage,* in which individual players use observability of actions to signal that a cluster is no longer admissible based on their confidence intervals. The pseudocode is given in Algorithm 2, with the sabotage step in step 13.

Intuition: `Multiple-K-Intervals-B` builds on `Multiple-K-Intervals-A`, with one key difference to account for private rewards. Since players no longer maintain the same confidence intervals and admissible sets, they must implicitly communicate any belief changes. In `Multiple-K-Intervals-B`, agents sabotage by purposely deviating from the agreed-upon order when they observe that the cluster currently being tested, $c(t_n)$, is no longer admissible. This signals to other players that they should also remove that cluster from their admissible set for the current nature's action b.

The regret bound is as follows; the proof is deferred to the Appendix.

Theorem 2 *The regret of Multiple-K-Intervals-B satisfies*

$$\mathcal{R}_T \leq 8\sqrt{4T(\log T)}BC + 2BC + N_s, \tag{4}$$

where N_s is the number of rounds in which any player selects a suboptimal action to sabotage.

Algorithm 2 The `Multiple-K-Intervals-B` Algorithm

Input: The cluster distributions $\{\nu_{\boldsymbol{a},c}\}$
for all $b \in \mathcal{B}$ **do**
    $\mathcal{C}_b \leftarrow \mathcal{C}$
end for
for $t = 1, \ldots, T$ **do**
    Receive $b \in \mathcal{B}$, and suppose this action was chosen at rounds $t_1, \ldots, t_n = t$
    Players construct $S^i_{\boldsymbol{a},b}(t)$ for all $\boldsymbol{a}$
    if $N_{\boldsymbol{a},b}(t) = 0$ **then**
        $S^i_{\boldsymbol{a},b}(t) \leftarrow (-\infty, \infty)$
    end if
    Select the next cluster $c(t_n)$ following $c(t_{n-1})$ in the ordered admissible set $\mathcal{C}_b$
    if player i observes $\mu_{\boldsymbol{a},c(t_n)} \notin S^i_{\boldsymbol{a},b}(t)$ for some $\boldsymbol{a}$ **then**
        Player i pulls a non-optimal arm to indicate to other players that $c(t_n)$ is no longer admissible for state b
        Each player observes the action and updates $\mathcal{C}_b \leftarrow \mathcal{C}_b \setminus \{c(t_n)\}$
    end if
    Choose the next arm:
    $\boldsymbol{a}_t \leftarrow \star_{c(t_n)}$
end for

3.2.1 Example

We now consider a setting to illustrate `Multiple-K-Intervals-B`, using the same setup as the example in section 3.1.1. Now suppose $b_3 \in \mathcal{B}_1$ and $b_1, b_2 \in \mathcal{B}_2$.

Define the cluster distributions as:

$$\underbrace{\begin{array}{cc} & \begin{array}{cc} 1 & 2 \end{array} \\ \begin{array}{c} 1 \\ 2 \end{array} & \begin{bmatrix} 0.2 & 0.6 \\ 0.8 & 0.5 \end{bmatrix} \end{array}}_{\text{Cluster 1}}, \underbrace{\begin{array}{cc} & \begin{array}{cc} 1 & 2 \end{array} \\ \begin{array}{c} 1 \\ 2 \end{array} & \begin{bmatrix} 0.8 & 0.1 \\ 0.7 & 0.3 \end{bmatrix} \end{array}}_{\text{Cluster 2}}$$

The rows represent Player 1's actions while the columns represent Player 2's actions. Again let the ordering of clusters be $(1, 2)$, and the order of joint arms be the same as above.

Suppose at round $t = 30$, the following occurs:

- Nature draws state $b_1 \in \mathcal{B}_1$ (true cluster is Cluster 1).
- The algorithm is currently testing Cluster 1 for b_1. Both players believe both clusters are admissible for b_1.
- For the joint arm $(2, 1)$, the true mean under Cluster 1 is 0.8.
- Both players have pulled $(2, 1)$ several times under b_1 and observe:

$$\text{Player 1: } \hat{\mu}^{(1)}_{(2,1),b_1} = 0.75, \quad S^{(1)} = [0.68,\ 0.82]$$

$$\text{Player 2: } \hat{\mu}^{(2)}_{(2,1),b_1} = 0.72, \quad S^{(2)} = [0.65,\ 0.79]$$

Since $\mu_{(2,1),1} = 0.8 \notin S^{(2)}$, Player 2 identifies Cluster 1 as inadmissible under b_1, while the interval for Player 1 still suggests both clusters are admissible. Player 2 deviates from the optimal arm of Cluster 1 to signal that it should be eliminated.

Player 1 observes this deviation, and both players update:

$$\mathcal{C}_{b_1} \leftarrow \mathcal{C}_{b_1} \setminus \{1\}$$

Later, suppose nature selects state b_2, which also belongs to $\mathcal{B}_2$. Since $\mathcal{C}_{b_2}$ has not yet been updated, players begin testing Cluster 1 again. Assume Player 1 observes Cluster 1 is no longer admissible according to their own confidence interval, but Player 2 does not. In that round: Player 1 deviates from the expected arm (the optimal arm of Cluster 1) to indicate Cluster 1 is inadmissible. Player 2 observes the deviation, and both players update $\mathcal{C}_{b_3} \leftarrow \mathcal{C}_{b_3} \setminus \{1\}$.

3.3 Problem C: Information Asymmetry in Both Rewards and Actions

In problem C, players can observe neither other players' rewards nor their actions, necessitating a different approach. Players maintain different confidence intervals, but cannot communicate changing beliefs to other players and may become miscoordinated.

To address this, we propose an explore-then-commit style algorithm, `Multiple-K-mDSEE`. In the basic `mDSEE` algorithm (Chang et al. (2022)), players alternate between exploration and commitment phases, spaced at increasing powers of 2. Players individually pick an action to commit to based on the empirical means observed during exploration phases. As the empirical means converge to the true means, commitment phases also increase in length; intuitively, this allows players to commit for longer periods once they are more confident and agree on the same optimal arm for each b.

Intuition: In our multiplayer latent bandit setting, we run `mDSEE` separately for each nature's action b, alternating between exploration and exploitation at predetermined intervals (with exploitation intervals increasing in length). We set a fixed exploration parameter E based on gaps within clusters to ensure sufficient concentration of the empirical means. This allows us to improve on the results of Chang et al. (2022) in several ways: First, the fact that K depends on the phase λ in Chang et al. (2022) is no longer necessary because we know the reward distributions. In this setting, we improve the order of regret in T by utilizing knowledge of the reward gaps. In addition, we allow for non-unique optimal joint actions.

Note that players do not update their empirical means based on rewards observed during commitment phases; this is because players may commit to different arms during these phases but cannot observe the actions taken by other players, and thus do not know what joint action is taken.

The regret bound is as follows.

Theorem 3 *The regret of Multiple-K-mDSEE satisfies*

$$\mathcal{R}_T \leq O(BK^M \log(T) + BK^M M \frac{\pi^2}{3}) \tag{5}$$

Hidden in this regret bound is the asymptotic dependence on the gaps within clusters, which makes our bound non-standard. However, we are still able to show that the regret grows logarithmically with respect to a fixed set of reward gaps. The proof is deferred to the Appendix.

4 Experiments

We empirically evaluate the three algorithms in the multi-agent multi-armed bandit setting with nature's actions. The algorithms differ in how players handle uncertainty in state-cluster mappings and how they coordinate under information asymmetry.

Setup: We ran each algorithm 10 times and computed the average and standard deviation of regret. Each nature's action $b \in \{0, 1, \ldots, B-1\}$ is deterministically assigned to a cluster $c_b \in \{0, 1, \ldots, C-1\}$ using the

Algorithm 3 `Multiple-K-mDSEE` Algorithm

Input: The cluster distributions $\{\nu_{\boldsymbol{a},c}\}$, exploration parameter $E > \dfrac{4}{\left(\frac{1}{2}\min_{\boldsymbol{a},b}\Delta_{\boldsymbol{a},b}\right)^2}$.
Initialize: $\lambda \leftarrow 1$.
for $t = 1, \ldots, T$ **do**
  Receive $b = b_t \sim \Upsilon$.
  if $\exists\, \boldsymbol{a}$ such that arm $\boldsymbol{a}$ has been pulled fewer than E times in state b for the λth phase **then**
    Pull arm $\boldsymbol{a}$ in state b.
    Each player i observes their own reward and updates $\widehat{\mu}^i_{\boldsymbol{a},b}$.
  else
    Each player pulls their optimal arm $\boldsymbol{a}_t \leftarrow \arg\max_{\boldsymbol{a}} \widehat{\mu}^i_{\boldsymbol{a},b}(t)$.
    Do not update $\widehat{\mu}^i_{\boldsymbol{a},b}$.
    if $t + 1 = 2^n$ for some $n \geq \lambda$ **then**
      $\lambda \leftarrow \lambda + 1$.
    end if
  end if
end for

rule:

$$c_b = \min\left(\left\lfloor \frac{b}{\lfloor B/C \rfloor} \right\rfloor, C - 1\right)$$

The reward for each arm-state pair is sampled from a Gaussian distribution $\mathcal{N}(\mu_{\boldsymbol{a},c}, 0.2^2)$, where the cluster reward tables $\{\mu_{\boldsymbol{a},c}\}_{\boldsymbol{a}\in[K]^M,\, c\in[C]}$ are initialized randomly from a uniform distribution at the start of each trial. Each table $\mu^{(c)}$ corresponds to a specific cluster c and defines the mean reward for every joint arm $\boldsymbol{a} = (a_1, \ldots, a_M)$ played by the M players.

Results: Figure 1 shows the cumulative regret averaged over 10 trials for each of the three algorithms evaluated. The red curve, `Multiple-K-UCB` from Maillard & Mannor (2014) adapted for the multiplayer joint-action problem, achieves sublinear regret performance across the horizon as expected.

The yellow curve represents a naive UCB approach, in which UCB is run "in parallel" for each nature's action b (where $\hat{\mu}(t)$ and $N(t)$ are tracked separately for each b.) Note also that this naive approach is only possible for Problem A, in which players can stay synchronized because they receive identical rewards and can infer each others' actions; this algorithm cannot be applied directly to Problem B or C. The naive UCB algorithm takes a notably longer time to find optimal arms, incurring higher regret across the horizon. This is because without cluster information, it is required to learn all the environments on its own for each choice of nature's actions. When nature has a large action space and feeds it to the learner adversarially, this results in a slower convergence in the regret plot as well as a higher variance.

The blue curve, our proposed algorithm `Multiple-K-Intervals-B`, demonstrates performance of a similar order to `Multiple-K-UCB` without requiring observability of both actions and rewards.

The green curve incurs significantly higher cumulative regret. The "staircase" pattern is a result of periodic resets in the explore-then-commit cycles of `Multiple-K-mDSEE`: players re-enter exploration phases upon reaching the next power of 2. Each jump corresponds to a phase shift, and the wide confidence band indicates high sensitivity to early misclassifications. Nevertheless, the curve eventually flattens, indicating eventual commitment to the right clusters, albeit at a much slower rate compared to the other two methods; this is to be expected due to the information-constrained nature of Problem C.

5 Conclusion

In this paper, we extended the latent multi-armed bandit framework to the novel multiplayer setting with information asymmetry and limited communication. We introduced three algorithms tailored to distinct asymmetry scenarios: `Multiple-K-Intervals-A` for shared rewards with unobservable actions,

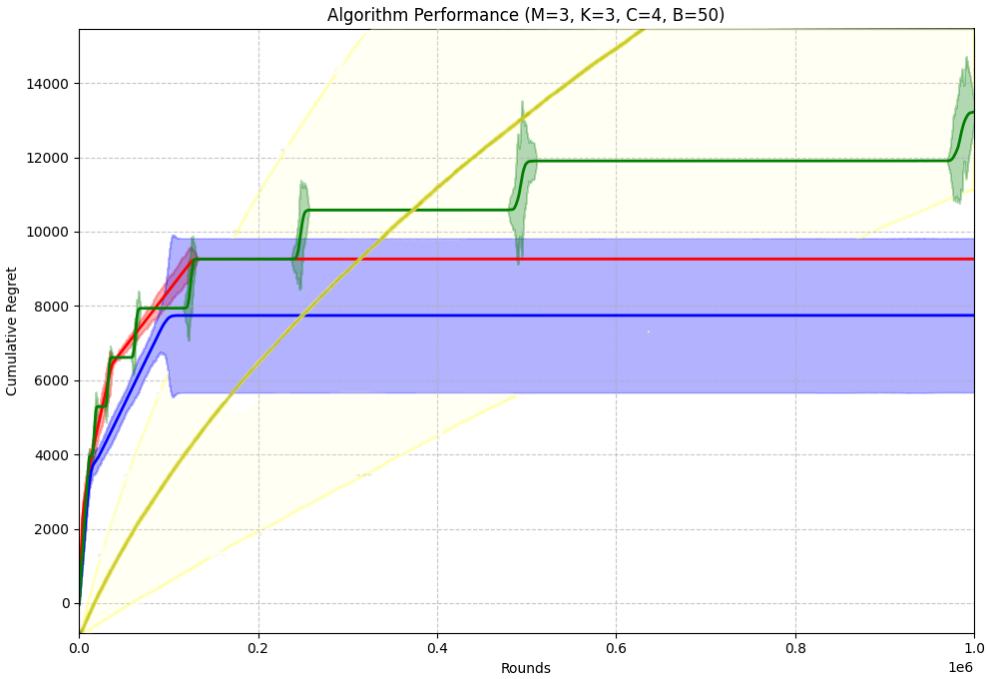

Figure 1: Cumulative regret over $T = 10^6$ rounds with $M = 3$ players, $K = 3$ arms, $C = 4$ clusters, the exploration parameter $E = 30$, and $B = 50$ nature's actions. Shaded regions indicate standard deviations between trials (scaled for visibility).

`Multiple-K-Intervals-B` for observable actions but private i.i.d. rewards, and `Multiple-K-mDSEE` for the case where neither rewards nor actions are shared. We verify our results through numerical experiments. Our results show that for Problems A and B, we achieve gap-independent regret bounds of order $O(\sqrt{T \log T} B)$ while still allowing for nature's actions to be adversarial, improving on the original single-agent analysis in Maillard & Mannor (2014). For Problem C, we utilize knowledge of cluster structures to design an explore-then-commit strategy that results in a regret bound of order $O(BK^M \log(T))$, improving on the results in Chang et al. (2022).

However, several directions for future work remain. The regret bound for problem C scales on the minimum gap within clusters, possibly leading to poor performance in certain cases. Future work could explore gap-free strategies for a tighter bound on regret. In addition, our algorithms rely on known reward distributions for each cluster, which may not be realistic in all applications. Investigating multiplayer algorithms for unknown cluster distributions (the agnostic case from Maillard & Mannor (2014)) remains an open challenge.

References

Rajeev Agrawal, Demosthenis Teneketzis, and Venkatachalam Anantharam. Asymptotically efficient adaptive allocation schemes for controlled i. i. d. processes: Finite parameter space. *IEEE Transactions on Automatic Control*, 34(3):258–267, 1989.

Orly Avner, Shie Mannor, and Ohad Shamir. Decoupling exploration and exploitation in multi-armed bandits. *arXiv preprint arXiv:1205.2874*, 2012.

Baruch Awerbuch and Robert Kleinberg. Competitive collaborative learning. *Journal of Computer and System Sciences*, 74(8):1271–1288, 2008.

Yogev Bar-On and Yishay Mansour. Individual regret in cooperative nonstochastic multi-armed bandits. *Advances in Neural Information Processing Systems*, 32, 2019.

T Basar and GJ Olsder. Dynamic noncooperative game theory, vol. 23 (siam, philadelphia). 1999.

Ilai Bistritz and Amir Leshem. Distributed multi-player bandits-a game of thrones approach. *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Rémi Bonnefoi, Lilian Besson, Christophe Moy, Emilie Kaufmann, and Jacques Palicot. Multi-armed bandit learning in iot networks: Learning helps even in non-stationary settings. In *International Conference on Cognitive Radio Oriented Wireless Networks*, pp. 173–185. Springer, 2017.

Guy Bresler, George H Chen, and Devavrat Shah. A latent source model for online collaborative filtering. *Advances in neural information processing systems*, 27, 2014.

Noam Brown and Tuomas Sandholm. Superhuman ai for multiplayer poker. *Science*, 365(6456):885–890, 2019.

Noam Brown, Tuomas Sandholm, and Strategic Machine. Libratus: The superhuman ai for no-limit poker. In *IJCAI*, pp. 5226–5228, 2017.

Lucian Busoniu, Robert Babuska, and Bart De Schutter. A comprehensive survey of multiagent reinforcement learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 38 (2):156–172, 2008.

Sarah H Cen and Devavrat Shah. Regret, stability & fairness in matching markets with bandit learners. In *International Conference on Artificial Intelligence and Statistics*, pp. 8938–8968. PMLR, 2022.

Nicolò Cesa-Bianchi, Tommaso Cesari, and Claire Monteleoni. Cooperative online learning: Keeping your neighbors updated. In *Algorithmic learning theory*, pp. 234–250. PMLR, 2020.

Nicol'o Cesa-Bianchi, Claudio Gentile, Yishay Mansour, and Alberto Minora. Delay and cooperation in nonstochastic bandits. In *Conference on Learning Theory*, pp. 605–622. PMLR, 2016.

William Chang and Aditi Karthik. Multiplayer information asymmetric bandits in metric spaces. *arXiv preprint arXiv:2503.08004*, 2025.

William Chang and Yuanhao Lu. Optimal cooperative multiplayer learning bandits with noisy rewards and no communication. *arXiv preprint arXiv:2311.06210*, 2023.

William Chang and Yuanhao Lu. Multiplayer information asymmetric contextual bandits. *arXiv preprint arXiv:2503.08961*, 2025.

William Chang, Mehdi Jafarnia-Jahromi, and Rahul Jain. Online learning for cooperative multi-player multi-armed bandits. *arXiv preprint arXiv:2109.03818*, 2021.

William Chang, Mehdi Jafarnia-Jahromi, and Rahul Jain. Online learning for cooperative multi-player multi-armed bandits. In *2022 IEEE 61st Conference on Decision and Control (CDC)*, pp. 7248–7253. IEEE, 2022.

Jongeun Choi, Songhwai Oh, and Roberto Horowitz. Distributed learning and cooperative control for multi-agent systems. *Automatica*, 45(12):2802–2814, 2009.

Dongsheng Ding, Chen-Yu Wei, Kaiqing Zhang, and Alejandro Ribeiro. Last-iterate convergent policy gradient primal-dual methods for constrained mdps. *Advances in Neural Information Processing Systems*, 36:66138–66200, 2023.

Claudio Gentile, Shuai Li, and Giovanni Zappella. Online clustering of bandits. In *International conference on machine learning*, pp. 757–765. PMLR, 2014.

Samarth Gupta, Gauri Joshi, and Osman Yağan. Correlated multi-armed bandits with a latent random source. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3572–3576. IEEE, 2020.

Joey Hong, Branislav Kveton, Manzil Zaheer, Yinlam Chow, Amr Ahmed, and Craig Boutilier. Latent bandits revisited. *Advances in Neural Information Processing Systems*, 33:13423–13433, 2020a.

Joey Hong, Branislav Kveton, Manzil Zaheer, Yinlam Chow, Amr Ahmed, Mohammad Ghavamzadeh, and Craig Boutilier. Non-stationary latent bandits. *arXiv preprint arXiv:2012.00386*, 2020b.

Meena Jagadeesan, Alexander Wei, Yixin Wang, Michael Jordan, and Jacob Steinhardt. Learning equilibria in matching markets from bandit feedback. *Advances in Neural Information Processing Systems*, 34:3323–3335, 2021.

Chinmaya Kausik, Kevin Tan, and Ambuj Tewari. Leveraging offline data in linear latent bandits. *arXiv preprint arXiv:2405.17324*, 2024.

Newton Mwai Kinyanjui, Emil Carlsson, and Fredrik D Johansson. Fast treatment personalization with latent bandits in fixed-confidence pure exploration. *Transactions on Machine Learning Research*, 2023.

Jens Kober, J Andrew Bagnell, and Jan Peters. Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research*, 32(11):1238–1274, 2013.

Peter Landgren, Vaibhav Srivastava, and Naomi Ehrich Leonard. On distributed cooperative decision-making in multiarmed bandits. In *2016 European Control Conference (ECC)*, pp. 243–248. IEEE, 2016.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Jae Won Lee, Jonghun Park, Jongwoo Lee, Euyseok Hong, et al. A multiagent approach to q-learning for daily stock trading. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 37(6):864–877, 2007.

Joel Z Leibo, Vinicius Zambaldi, Marc Lanctot, Janusz Marecki, and Thore Graepel. Multi-agent reinforcement learning in sequential social dilemmas. *arXiv preprint arXiv:1702.03037*, 2017.

Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pp. 661–670, 2010.

Tianjiao Li, Ziwei Guan, Shaofeng Zou, Tengyu Xu, Yingbin Liang, and Guanghui Lan. Faster algorithm and sharper analysis for constrained markov decision process. *Operations Research Letters*, 54:107107, 2024.

Michael L Littman. Markov games as a framework for multi-agent reinforcement learning. In *Machine learning proceedings 1994*, pp. 157–163. Elsevier, 1994.

Lydia T Liu, Horia Mania, and Michael Jordan. Competing bandits in matching markets. In *International Conference on Artificial Intelligence and Statistics*, pp. 1618–1628. PMLR, 2020.

Lydia T Liu, Feng Ruan, Horia Mania, and Michael I Jordan. Bandit learning in decentralized matching markets. *Journal of Machine Learning Research*, 22(211):1–34, 2021a.

Tao Liu, Ruida Zhou, Dileep Kalathil, PR Kumar, and Chao Tian. Policy optimization for constrained mdps with provable fast global convergence. *arXiv preprint arXiv:2111.00552*, 2021b.

Odalric-Ambrym Maillard and Shie Mannor. Latent bandits. In *International Conference on Machine Learning*, pp. 136–144. PMLR, 2014.

David Martínez-Rubio, Varun Kanade, and Patrick Rebeschini. Decentralized cooperative stochastic bandits. 2019.

Ted Moskovitz, Brendan O'Donoghue, Vivek Veeriah, Sebastian Flennerhag, Satinder Singh, and Tom Zahavy. Reload: Reinforcement learning with optimistic ascent-descent for last-iterate convergence in constrained mdps. In *International Conference on Machine Learning*, pp. 25303–25336. PMLR, 2023.

Soumyabrata Pal, Arun Suggala, Karthikeyan Shanmugam, and Prateek Jain. Blocked collaborative bandits: online collaborative filtering with per-item budget constraints. *Advances in Neural Information Processing Systems*, 36:25281–25324, 2023a.

Soumyabrata Pal, Arun Sai Suggala, Karthikeyan Shanmugam, and Prateek Jain. Optimal algorithms for latent bandits with cluster structure. In *International Conference on Artificial Intelligence and Statistics*, pp. 7540–7577. PMLR, 2023b.

Alexandre Proutiere and Po-An Wang. An optimal algorithm for multiplayer multi-armed bandits. *arXiv preprint arXiv:1909.13079*, 2019.

Abishek Sankararaman, Soumya Basu, and Karthik Abinav Sankararaman. Dominate or delete: Decentralized competing bandits in serial dictatorship. In *International Conference on Artificial Intelligence and Statistics*, pp. 1252–1260. PMLR, 2021.

Shai Shalev-Shwartz, Shaked Shammah, and Amnon Shashua. Safe, multi-agent, reinforcement learning for autonomous driving. *arXiv preprint arXiv:1610.03295*, 2016.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

Balazs Szorenyi, Róbert Busa-Fekete, István Hegedus, Róbert Ormándi, Márk Jelasity, and Balázs Kégl. Gossip-based distributed stochastic bandit algorithms. In *International Conference on Machine Learning*, pp. 19–27. PMLR, 2013.

Po-An Wang, Alexandre Proutiere, Kaito Ariu, Yassir Jedra, and Alessio Russo. Optimal algorithms for multiplayer multi-armed bandits. In *International Conference on Artificial Intelligence and Statistics*, pp. 4120–4129. PMLR, 2020.

Shiyong Wang, Jiafu Wan, Daqiang Zhang, Di Li, and Chunhua Zhang. Towards smart factory for industry 4.0: a self-organized multi-agent system with big data based feedback and coordination. *Computer networks*, 101:158–168, 2016.

Kaiqing Zhang, Bin Hu, and Tamer Basar. Policy optimization for h_2 linear control with h_∞ robustness guarantee: Implicit regularization and global convergence. In *Learning for Dynamics and Control*, pp. 179–190. PMLR, 2020.

Li Zhou and Emma Brunskill. Latent contextual bandits and their application to personalized recommendations for new users. *arXiv preprint arXiv:1604.06743*, 2016.

A Important Lemmas

Lemma 4 (corollary 5.5 of Lattimore & Szepesvári (2020)) *Assume that $X_i - \mu$ are independent, σ-subgaussian random variables. Then for any $\varepsilon \geq 0$,*

$$\mathbb{P}(\hat{\mu} \geq \mu + \varepsilon) \leq \exp\left(-\frac{n\varepsilon^2}{2\sigma^2}\right) \quad \textit{and} \quad \mathbb{P}(\hat{\mu} \leq \mu - \varepsilon) \leq \exp\left(-\frac{n\varepsilon^2}{2\sigma^2}\right),$$

where $\hat{\mu} = \frac{1}{n}\sum_{t=1}^{n} X_t$.

Lemma 5 *With the good event defined in equation (9), we have the following upper bound*

$$P(G^c) \leq \frac{2BC}{T} \tag{6}$$

Proof: By DeMorgan's rule, taking the complement of the good event we have

$$\begin{aligned} G^c &= \bigcup_{t=1}^{T}\bigcup_{b}\bigcup_{c}\{\mu_{\star_c,b} \leq L_{\star_c,b}(t) \text{ or } \mu_{\star_c,b} \geq U_{\star_c,b}(t)\} \\ &= \bigcup_{t=1}^{T}\bigcup_{b}\bigcup_{c}\{\mu_{\star_c,b} \leq L_{\star_c,b}(t)\} \cup \{\mu_{\star_c,b} \geq U_{\star_c,b}(t)\} \end{aligned}$$

By the probability union bound, we have

$$P(G^c) \leq \sum_{t=1}^{T}\sum_{b}\sum_{c} P(\mu_{\star_c,b} \leq L_{\star_c,b}(t)) + P(\mu_{\star_c,b} \geq U_{a,b}(t)) \tag{7}$$

By Lemma 4, we upper bound

$$\begin{aligned} P(\mu_{\star_c,b} \leq L_{\star_c,b}(t)) &= P\left(\widehat{\mu}_{a,b} \leq \mu_{a,b} - \sqrt{\frac{4\log(T)}{N_{\star_c,b}(t)}}\right) \\ &\leq e^{-\frac{N_{a,b}(t)\left(\sqrt{\frac{4\log(T)}{N_{\star_c,b}(t)}}\right)^2}{2}} \\ &= e^{-2\log(T)} \\ &= \frac{1}{T^2} \end{aligned}$$

Now we plug this into our upper bound for $P(G^c)$

$$P(G^c) \leq \sum_{t=1}^{T}\sum_{b}\sum_{c}\frac{2}{T^2} = \frac{2BC}{T} \tag{8}$$

□

B Regret bound of `Multiple-K-Intervals-A`

We present the proof of Theorem 1, the regret bound of `Multiple-K-Intervals-A` for Problem A, information asymmetry in actions.

Proof: Consider the good event

$$G = \bigcap_{t=1}^{T} \bigcap_{b} \bigcap_{c} \bigcap_{i} G^i_{\star_c,b}(t), \quad G^i_{\boldsymbol{a},b}(t) = \{\mu_{\boldsymbol{a},b} \in S^i_{\boldsymbol{a},b}(t)\} \tag{9}$$

where $S^i_{\boldsymbol{a},b}(t)$ is the confidence interval of player i at time t. Recall the definition of regret

$$\begin{aligned}
\mathcal{R}_T &= \sum_{t=1}^{T} \mathbb{E}[X_{\star_{\boldsymbol{b_t}},b_t} - X_{\boldsymbol{a_t},b_t}] && (10)\\
&= \sum_{t=1}^{T} \mathbb{E}[X_{\star_{\boldsymbol{b_t}},b_t} - X_{\boldsymbol{a_t},b_t}|G]P(G) + \mathbb{E}[X_{\star_{\boldsymbol{b_t}},b_t} - X_{\boldsymbol{a_t},b_t}|G^c]P(G^c) && (11)\\
&\leq \sum_{t=1}^{T} \mathbb{E}[X_{\star_{\boldsymbol{b_t}},b_t} - X_{\boldsymbol{a_t},b_t}|G] + TP(G^c) && (12)
\end{aligned}$$

Denote

$$\delta_{\boldsymbol{a},b}(t) = U_{\boldsymbol{a},b}(t) - L_{\boldsymbol{a},b}(t) = 2\sqrt{\frac{4\log T}{N_{\boldsymbol{a},b}(t)}}.$$

We aim to show that

$$\mathcal{R}_T \leq \sum_{t=1}^{T} (\delta_{\star_{\boldsymbol{c_t}},b_t}(t) + \delta_{\boldsymbol{a_t},b_t}(t)).$$

Suppose that at time t, nature selects an action b_t belonging to cluster c_t. At each round, the players select the best arm for a cluster that is admissible. More explicitly, $\boldsymbol{a_t} = \star_{\boldsymbol{c'}}$, where c' is admissible. Since c' is admissible, $\mu_{\star_{\boldsymbol{c_t}},c'} \in S_{\star_{\boldsymbol{c_t}},b_t}(t)$. We also assume that the true cluster c_t is admissible under the good event, thus $\mu_{\star_{c_t},c_t} \in S_{\star_{c_t},b_t}(t)$, and thus

$$\mu_{\star_{\boldsymbol{c_t}},c_t} - \mu_{\star_{\boldsymbol{c_t}},c'} \leq \delta_{\star_{\boldsymbol{c_t}},b_t}(t).$$

Similarly, since $\mu_{\star_{\boldsymbol{c'}},c'} \in S_{\star_{\boldsymbol{c'}},b_t}(t)$, we have

$$\mu_{\star_{\boldsymbol{c'}},c'} - \mu_{\star_{\boldsymbol{c'}},c_t} \leq \delta_{\star_{\boldsymbol{c'}},b_t}(t).$$

Further, $\mu_{\star_{\boldsymbol{c'}},c'} \geq \mu_{\star_{\boldsymbol{c_t}},c'}$ by definition. Thus

$$\sum_{t=1}^{T} \mathbb{E}[X_{\star_{b_t},b_t} - X_{a_t,b_t}|G] = \sum_{t=1}^{T} (\mu_{\star_{\boldsymbol{c_t}},c_t} - \mu_{\star_{\boldsymbol{c'}},c_t}) \tag{13}$$

$$= \sum_{t=1}^{T} \big((\mu_{\star_{\boldsymbol{c_t}},c_t} - \mu_{\star_{\boldsymbol{c_t}},c'}) + (\mu_{\star_{\boldsymbol{c_t}},c'} - \mu_{\star_{\boldsymbol{c'}},c'}) + (\mu_{\star_{\boldsymbol{c'}},c'} - \mu_{\star_{\boldsymbol{c'}},c_t})\big) \tag{14}$$

$$\leq \sum_{t=1}^{T} \big(\delta_{\star_{\boldsymbol{c_t}},b_t}(t) + \delta_{\star_{\boldsymbol{c'}},b_t}(t)\big) \tag{15}$$

$$= \sum_{t=1}^{T} \big(\delta_{\star_{\boldsymbol{c_t}},b_t}(t) + \delta_{\boldsymbol{a_t},b_t}(t)\big) \tag{16}$$

$$= 2\sum_{t=1}^{T} \left(\sqrt{\frac{4\log T}{N_{\star_{\boldsymbol{c_t}},b}(t)}} + \sqrt{\frac{4\log T}{N_{\boldsymbol{a_t},b}(t)}}\right) \tag{17}$$

$$\leq 2\sum_{b}\sum_{c}\sum_{s=1}^{N_{\boldsymbol{a},b}(T)} 2\sqrt{\frac{4\log T}{s}} \tag{18}$$

$$\leq 4\sqrt{4\log T}\sum_{b}\sum_{c}\sum_{s=1}^{T} \frac{1}{\sqrt{s}} \tag{19}$$

$$\leq 4\sqrt{4\log T}\sum_{b}\sum_{c}\int_{0}^{T} \frac{1}{\sqrt{s}} ds \tag{20}$$

$$= 8\sqrt{4\log T}\sum_{b}\sum_{c}\sqrt{T} \tag{21}$$

$$= 8\sqrt{4\log T}BC\sqrt{T} \tag{22}$$

Note that in equation (18) we take a summation across the clusters rather than the set of all arms, because for each nature's action b we try at most C arms (the optimal arms in each cluster). □

C Regret bound of Multiple-K-Intervals-B

We present the proof of Theorem 2, the regret bound of `Multiple-K-Intervals-B` for Problem B, information asymmetry in rewards.

Proof: The only deviation from Algorithm 1 is the addition of sabotage. The cumulative regret incurred by sabotage across all rounds is at most $N_s \leq BC$. Thus $\mathcal{R}_T \leq 8\sqrt{4T(\log T)}BC + 2BC + N_s$. □

D Regret bound of Multiple-K-mDSEE

We present the proof of Theorem 3, the regret bound of `Multiple-K-mDSEE` for Problem C, information asymmetry in actions and rewards.

Proof: Decompose $\mathcal{R}_T = \mathcal{R}_{T,E} + \mathcal{R}_{T,C}$, where $\mathcal{R}_{T,E}$ is the regret incurred from the exploration phases spaced at powers of 2, and $\mathcal{R}_{T,C}$ is the regret incurred from all commitment phases.

Thus, the regret incurred during exploration is at most:

$$\mathcal{R}_{T,E} \leq \sum_{b}\sum_{\boldsymbol{a}} E\lceil \log_2(T)\rceil \Delta_{\boldsymbol{a},b} \tag{23}$$

where $\Delta_{\boldsymbol{a},b} = \mu_{\star_b,b} - \mu_{\boldsymbol{a},b}$ is the suboptimality gap for arm $\boldsymbol{a}$ under state b.

We now analyze the **commitment phase** regret $\mathcal{R}_{T,C}$.

Let $\epsilon = \frac{1}{2}\min_{\boldsymbol{a},b} \Delta_{\boldsymbol{a},b}$.

In the latent multi-armed bandit setting with multiple-cluster arrivals, the sub-optimality gaps within clusters are known due to the known reward structure. This allows us to choose a fixed exploration constant E large enough to ensure sufficient concentration of the empirical means. If each arm is pulled E times during the previous exploration phases, then we have the following inequality when t is in the committing phase,

$$N_{\boldsymbol{a},b}(t) \geq E\log_2(t) \tag{24}$$

Define the good event for an individual player i:

$$G^i_{\boldsymbol{a},b}(t) = \left\{\left|\hat{\mu}^i_{\boldsymbol{a},b}(t) - \mu_{\boldsymbol{a},b}\right| < \epsilon\right\} \tag{25}$$

When $\bigcap_b \bigcap_{\boldsymbol{a}} \bigcap_i G^i_{\boldsymbol{a},b}(t)$ occurs, the optimal arm is selected at round t, and regret is 0. Thus,

$$\begin{aligned}
\mathcal{R}_{T,C} &\leq \sum_{b\in\mathcal{B}} \sum_{t:b_t=b} \mathbb{E}\left[\mathbb{I}\left[\left(\bigcap_{\boldsymbol{a}}\bigcap_i G^i_{\boldsymbol{a},b}(t)\right)^c\right]\right] && (26)\\
&= \sum_{b\in\mathcal{B}} \sum_{t:b_t=b} \mathbb{P}\left[\left(\bigcap_{\boldsymbol{a}}\bigcap_i G^i_{\boldsymbol{a},b}(t)\right)^c\right] && (27)\\
&\leq \sum_{b\in\mathcal{B}} \sum_{t=1}^{n_b(T)} \sum_{\boldsymbol{a}} \sum_{i=1}^{M} \mathbb{P}\left[G^i_{\boldsymbol{a},b}(t)^c\right] && (28)\\
&= \sum_{b\in\mathcal{B}} \sum_{t=1}^{n_b(T)} \sum_{\boldsymbol{a}} \sum_{i=1}^{M} \mathbb{P}\left[\left|\hat{\mu}^i_{\boldsymbol{a},b}(t) - \mu_{\boldsymbol{a},b}\right| > \epsilon\right] && (29)\\
&\leq \sum_{b\in\mathcal{B}} \sum_{t=1}^{n_b(T)} \sum_{\boldsymbol{a}} \sum_{i=1}^{M} 2e^{-\frac{n_{\boldsymbol{a},b}(t)\epsilon^2}{2}} && (30)\\
&\leq \sum_{b\in\mathcal{B}} \sum_{\boldsymbol{a}} \sum_{t=1}^{n_b(T)} M\cdot 2e^{-\frac{E\log_2(t)\epsilon^2}{2}} && (31)\\
&\leq \sum_{b\in\mathcal{B}} \sum_{\boldsymbol{a}} \sum_{t=1}^{n_b(T)} M\cdot 2t^{-2} && (32)\\
&\leq BK^M M\frac{\pi^2}{3} && (33)
\end{aligned}$$

where in the second inequality we use the probability union bound and in the third inequality we use the fact that the rewards are i.i.d., so the probability of the complement of the good event has the same upper bound for each player. Note that unlike in Chang et al. (2022), we can define ϵ explicitly, leveraging knowledge of the gaps, rather than fixing an exploration schedule to guarantee theoretical convergence.

Thus our total regret $\mathcal{R}_T$ satisfies $O(BK^M\log(T) + BK^M M\frac{\pi^2}{3})$.

□