Decentralized Cooperative Reinforcement Learning with Hierarchical Information Structure

Hsu Kao HSUKAO@UMICH.EDU
University of Michigan

Chen-Yu Wei CHENYU.WEI@USC.EDU
University of Southern California

Vijay Subramanian VGSUBRAM@UMICH.EDU
University of Michigan

Editors: Sanjoy Dasgupta and Nika Haghtalab

Abstract

Multi-agent reinforcement learning (MARL) problems are challenging due to information asymmetry. To overcome this challenge, existing methods often require high level of coordination or communication between the agents. We consider two-agent multi-armed bandits (MABs) and Markov decision processes (MDPs) with a hierarchical information structure arising in applications, which we exploit to propose simpler and more efficient algorithms that require no coordination or communication. In the structure, in each step the "leader" chooses her action first, and then the "follower" decides his action after observing the leader's action. The two agents observe the same reward (and the same state transition in the MDP setting) that depends on their joint action. For the bandit setting, we propose a hierarchical bandit algorithm that achieves a near-optimal gap-independent regret of $\widetilde{\mathcal{O}}(\sqrt{ABT})$ and a near-optimal gap-dependent regret of $\mathcal{O}(\log(T))$, where A and B are the numbers of actions of the leader and the follower, respectively, and T is the number of steps. We further extend to the case of multiple followers and the case with a deep hierarchy, where we both obtain near-optimal regret bounds. For the MDP setting, we obtain $\widetilde{\mathcal{O}}(\sqrt{H^7S^2ABT})$ regret, where H is the number of steps per episode, S is the number of states, T is the number of episodes. This matches the existing lower bound in terms of A, B, and T.

Keywords: hierarchical information structure, multi-agent online learning, multi-armed bandit, Markov decision process

1. Introduction

Multi-agent reinforcement learning (MARL) has received great attention due to its wide variety of applications and the tremendous advances in single-agent RL techniques (Zhang et al., 2020a). In a multi-agent environment, each agent has different observations and may have different sets of information. This is referred to as the *information asymmetry* property (Chang et al., 2021). One straightforward method used with information asymmetry is to let every agent concurrently learn based on its own information using single-agent algorithms. However, this creates the *non-stationarity issue* since the effective environment observed by each agent is time-varying, which sometimes causes non-convergence of the algorithms. Another line of solutions is to enforce coordination among agents, essentially transforming a multi-agent system back to (or making it more similar to) a single-agent one. One way to achieve this is through communication (Shahrampour et al., 2017; Zhang et al., 2021a), which introduces extra costs that may be intolerable in some cases. A more broadly applicable scheme is through the common information (CI) approach (Nayyar et al., 2013; Chang et al., 2021; Dibangoye and Buffet, 2018). The CI approach relies on a set of CI shared by all agents, and all agents need to agree on a protocol that specifies the joint policy updates of all agents upon receiving a certain piece of CI. With this protocol, an agent may be able to infer the actions taken by other agents without observing them. However, this approach has several shortcomings, making it hard to apply in practice: it has high computational complexity (since every agent has to perform policy updates for all the other agents), and it requires all agents to have very tight coordination (e.g., sharing randomization seeds in each round and knowing all details in the algorithms of other agents), which may be infeasible if synchronization among agents or agent privacy is an issue.

In this paper, we address the aforementioned issues in a special but widely applicable MARL setting. We consider MARL team problems[1] with a particular hierarchical information structure between agents under the settings of multi-agent multi-armed bandits (MAMABs) and multi-agent Markov decision processes (MDPs). In this structure, decisions are made sequentially, and a decision maker has all the decisions from decision makers that act before it in the sequence. In the two-agent case, one of the agents (the "leader") chooses her action first, while the other agent (the "follower") chooses his action after observing the leader's action. This setting is similar to the Stackelberg game but with the players cooperating to achieve the same objective. Such hierarchical information structure arises in many applications. For example, in a cognitive radio (CR) wireless network, the primary user (PU) first decides its resource allocation scheme; then based on this scheme the secondary user (SU) chooses its own resource allocation scheme that minimally interferes with the PU's transmission (Ning et al., 2020). While this problem can also be solved using the CI approach or other MAMAB algorithms (Kalathil et al., 2014; Chang et al., 2021), as discussed earlier, they are intensive either in computation or in communication. In this work, we exploit the hierarchical information structure, and propose simpler and more efficient MARL algorithms that require neither communication nor explicit coordination, while achieving near-optimal regret bounds. Such algorithms could be much easier to deploy in practice.

In more detail, we first consider the two-agent bandit setting, where both agents observe the same reward determined by their *joint action*, but only the follower observes the leader's action but not *vice versa*. For this setting, we propose a decentralized algorithm that achieves a near-optimal gap-independent regret bound of $\widetilde{\mathcal{O}}(\sqrt{ABT})$[2] and a gap-dependent bound of $\mathcal{O}(\log(T))$, where A and B are the numbers of actions of the leader and the follower, respectively, and T is the number of time-steps. In our method, the leader performs an Upper Confidence Bound (UCB)-based algorithm (Auer et al., 2002a) with a modified bonus term related to the follower's regret bound, while the follower can use any algorithm that achieves sub-linear regret. Without explicit coordination, the agents perform joint exploration over the action space and learn with low regret. Interestingly, our hierarchical bandit setting mathematically coincides with the bandit-over-bandit framework considered in model selection (Agarwal et al., 2017; Arora et al., 2021; Cutkosky et al., 2021), although the two problems are motivated by very different applications. Our algorithm can be readily used for model selection, and our gap-dependent bound (Theorem 2) answers the open question in (Arora et al., 2021) by improving their regret bound by a factor of $\log(T)$ (to our knowledge, (Arora et al.,

[1] We consider the agents share the same utility, with reward samples commonly observed. This reward structure is a special case of potential games (Leonardos et al., 2021), where there exists a global potential function that reflects the change in any agent's utility from any unilateral deviation.

[2] We use $\widetilde{\mathcal{O}}(\cdot)$ to hide poly-logarithmic factors.

2021) is the only work on model selection that achieves a logarithmic gap-dependent bound). We further extend our idea to two more complicated settings. The first is the case of multiple followers, where each follower only observes its own reward while the leader only observes the sum of the rewards of all followers. The other is the case with a deep hierarchy, where more than two agents make decisions sequentially based on the decisions made by prior agents, and all agents observe the same reward. In both extensions, our algorithms also achieve near-optimal regret bounds.

Next, we generalize the above idea to the two-agent MDP setting. In this setting, the state evolution and reward observable by both agents are sampled from distributions depending on the current state and the joint action of the agents; as before, the follower observes the leader's action but not vice versa. Similar to the bandit case, we propose a decentralized learning method that enables the agents to perform joint exploration without communication or explicit coordination. Our algorithm is based on an intriguing combination of two exploration strategies developed for single-agent reinforcement learning: UCB-H (Jin et al., 2018) and UCBVI (Azar et al., 2017). By letting the leader execute a UCB-H-styled algorithm and the follower use a UCBVI-styled one, the agents jointly achieve a regret upper bound of $\widetilde{\mathcal{O}}(\sqrt{H^7S^2ABT})$ (H is the horizon length, and T is the number of episodes), while the regret lower bound is $\Omega(\sqrt{H^2SABT})$, inherited from the single-agent MDP setting (Azar et al., 2017). Tightening our bound without sacrificing the benefit of decentralized learning is left as an open question.

Related work. Algorithms for various MAMAB settings have gained increasing interest recently, but there is only a limited literature that investigates the effects and challenges caused by information asymmetry in the setting where agents jointly interact with the environment as is common in MARL applications, with the corresponding MDP setting receiving even less attention. Chang et al. (2021) study the MAMAB setting where the reward is determined by the joint action with three types of information asymmetry: unobserved actions and common rewards, observed actions and independent rewards, and unobserved actions and independent rewards; the first two settings can be solved by the notion of the CI approach[3], while in the last setting they propose an "explore then commit" type algorithm that achieves an $\mathcal{O}\left(\log(T)\right)$ regret. Bai et al. (2021) consider sample-efficient learning in bandit games and bandit-RL games. Their bandit game corresponds to our hierarchical bandits, but under a general reward setup (i.e., in their setting, the rewards of the two agents have different means, unlike the common-reward setting we consider here). They consider centralized and offline learning assuming access to a sample generator, while we consider decentralized and online learning through interactions with the environment. While their results imply a worst case information-theoretic gap to the Stackelberg game value that cannot be closed, it is not the case in our team problem. On the other hand, our hierarchical MDP has a more general transition structure than their bandit-RL game (in their setting, only the follower is involved in an MDP). Arora et al. (2021) study meta-learning over bandit algorithms, which exhibits a similar mathematical structure to our hierarchical bandit problem, though from a very different perspective. Our hierarchical bandit algorithm can be readily used as an algorithm for their setting, and our gap-dependent bound improves their $\mathcal{O}(\log^2(T))$ bound by a factor of $\log(T)$, resolving their question on the tightness of their result.

Most papers on MAMAB consider a set of agents pulling *the same set of arms* simultaneously, and in most of them the agents coordinate through *real-time communications* to collaboratively find the optimal policy, with a few exceptions (Bistritz and Leshem, 2021; Bistritz et al., 2021; Bubeck

[3] Their mUCB algorithm for the first setting is equivalently the CI approach in combination with the UCB1 algorithm.

et al., 2021). In the former, the communication resource is either costly (Madhushani and Leonard, 2020), limited by budget (Lalitha and Goldsmith, 2021; Vial et al., 2021; Sankararaman et al., 2019; Chawla et al., 2020), or constrained through communication networks (Landgren et al., 2021; Shahrampour et al., 2017), so the main focus is on designing communication efficient schemes that achieve the same performance as if there were no information asymmetry. In another related thread, referred to as the matching bandits problem, agents choosing the same arm collide and obtain zero rewards (Kalathil et al., 2014; Bistritz and Leshem, 2021; Bistritz et al., 2021); here and in a few other works (Shahrampour et al., 2017), different agents get different distribution of rewards from the same arm, while in other referenced work they get independent and identically distributed samples from the same arm.

Regret minimization in MARL for Markov games is in general challenging due to the fact that every agent faces a non-stationary environment. It has been shown in (Abbasi-Yadkori et al., 2013; Radanovic et al., 2019; Tian et al., 2021) that for single-agent non-stationary MDP problems, to have a sub-linear-in-T regret bound against the best policy is both computationally and statistically hard. Therefore, to establish meaningful guarantees in MARL while keeping the algorithm efficient, special properties of the game have to be considered. Radanovic et al. (2019) consider the same two-agent collaborative setting as ours, but requires that the agents exchange their policies after each episode. Tian et al. (2021) study another two-agent setting where each agent is agnostic about the actions of the other; however, their algorithm is conservative (with the goal of guarding against an adversarial opponent) and does not exploit the cooperative setting of our problem. Leonardos et al. (2021); Zhang et al. (2021b); Fox et al. (2021); Ding et al. (2022); Zhang et al. (2022) study multi-agent Markov potential games (more general than the team problem) and establish finite convergence bounds; however, their algorithm does not handle the state-space exploration issue (which is a key element in our work) so their regret bound has an extra problem-dependent factor; besides, only convergence to local Nash equilibria is shown, and there is no guarantee about attaining global optima. Another line of research studies efficient learning in general-sum Markov games with state-space exploration (Song et al., 2021; Jin et al., 2021; Mao and Başar, 2022; Mao et al., 2021). However, they consider convergence to (coarse) correlated equilibrium, which is again a weaker equilibrium concept than the global optimum when specialized to the common-payoff setting we consider.

2. Preliminaries

We first define some notation. For a positive integer n, we denote $[n] = \{1, 2, \dots, n\}$. For an integer n, we define $n^+ = \max\{n, 1\}$.

Two-agent hierarchical bandits. Consider a two-agent MAB where the rewards are decided by the joint action of the two agents U1 (leader) and U2 (follower). Let A and B be the numbers of actions (which are arms in the context) of U1 and U2, respectively, and T be the number of time steps. Without loss of generality, we assume that $A, B \leq T$. Under the hierarchical information structure, in round $t \in [T]$, U1 first chooses an action $a_t \in [A]$; after observing U1's action a_t, U2 then chooses another action $b_t \in [B]$. However, U1 cannot observe U2's action b_t. These two actions jointly generate a noisy reward $r_t \in [0, 1]$ with expectation μ_{a_t,b_t}, and both agents observe[4]

[4] Our analysis can straightforwardly handle a more general case where U1 and U2 receive different (independent) noisy copies of the reward with the same mean. For simplicity, we assume that they receive the same copy.

r_t. For ease of presentation, we assume without loss of generality that the best action of U2 given any choice of U1 is indexed by 1, i.e., $\mu_{a,1} \geq \mu_{a,b}$ for all a, b; similarly, the best action of U1 is indexed by 1, i.e., $\mu_{1,1} \geq \mu_{a,1}$ for all a. Then the (common) goal of the agents is to minimize the pseudo-regret defined as follows:

$$\mathrm{Reg}(T) = \sum_{t=1}^{T} (\mu_{1,1} - \mu_{a_t,b_t}).$$

Two-agent hierarchical MDPs. Consider a two-agent H-step finite-horizon MDP where the rewards and state transitions depend on the joint action of the two agents U1 and U2, with the process run over T episodes. This generalizes the previous two-agent bandit setting. The state space is $\mathcal{S}$, with a number of $S = |\mathcal{S}|$ states. In each state, U1 and U2 choose actions from $[A]$ and $[B]$, respectively. We assume that S, A, B, H are all upper bounded by T. Every episode t starts with an initial state $s_{t,1} \in \mathcal{S}$. In the h-th step of the t-th episode, the agents first observe $s_{t,h} \in \mathcal{S}$. Under the hierarchical information structure, U1 chooses an action $a_{t,h} \in [A]$, followed by U2 choosing another action $b_{t,h} \in [B]$ upon seeing $a_{t,h}$. After the actions are chosen, both agents receive a reward $r_{t,h} \in [0, 1]$ with $\mathbb{E}[r_{t,h}] = R(s_{t,h}, a_{t,h}, b_{t,h})$, and then the state transitions to the next state $s_{t,h+1} \sim P(\cdot|s_{t,h}, a_{t,h}, b_{t,h})$. The episode ends right after the state transitions to $s_{t,H+1}$. In the RL setting we consider, rewards are commonly observed by both agents, but they do not know the reward function R or the transition probability P.

An H-step policy for U1 can be represented as $\pi^1 = \{\pi^1_1, \ldots, \pi^1_H\}$, where $\pi^1_h : \mathcal{S} \to [A]$ specifies the choice of her action on each state when she is at step h; a policy for U2 can be represented as $\pi^2 = \{\pi^2_1, \ldots, \pi^2_H\}$, where $\pi^2_h : \mathcal{S} \times [A] \to [B]$ specifies the choice of his action on each state and under each possible choice of U1, when he is at step h. We define the state value function at step h under a policy pair (π^1, π^2) as

$$V_h^{\pi^1,\pi^2}(s) = \mathbb{E}\left[\sum_{k=h}^{H} R(s_k, a_k, b_k) \,\middle|\, s_h = s, a_k = \pi^1_k(s_k), b_k = \pi^2_k(s_k, a_k), s_{k+1} \sim P(\cdot|s_k, a_k, b_k), \forall k \geq h\right].$$

with $V_{H+1}^{\pi^1,\pi^2}(\cdot) \triangleq 0$. Also, we define the state-action value function as

$$Q_h^{\pi^1,\pi^2}(s, a, b) = R(s, a, b) + \mathbb{E}\left[V_{h+1}^{\pi^1,\pi^2}(s_{h+1}) \middle| s_{h+1} \sim P(\cdot|s, a, b)\right].$$

The optimal value functions are then given by $V_{*,h}(s) = \max_{\pi^1,\pi^2} V_h^{\pi^1,\pi^2}(s)$ and $Q_{*,h}(s, a, b) = \max_{\pi^1,\pi^2} Q_h^{\pi^1,\pi^2}(s, a, b)$. By dynamic programming, we have the following for all h, s, a, b:

$$V_{*,h}(s) = \max_{a,b} Q_{*,h}(s, a, b) \quad \text{and} \quad Q_{*,h}(s, a, b) = R(s, a, b) + \mathbb{E}_{s' \sim P(\cdot|s,a,b)}\left[V_{*,h+1}(s')\right],$$

with $V_{*,H+1}(\cdot) \triangleq 0$. We further define $Q_{*,h}(s, a) = \max_b Q_{*,h}(s, a, b)$. With this notation, we can write the regret of the agents as

$$\mathrm{Reg}(T) = \sum_{t=1}^{T} \left(V_{*,1}(s_{t,1}) - V_1^{\pi^1_t,\pi^2_t}(s_{t,1})\right).$$

Benchmark. In two-agent cases, there are three obvious types of information structure in terms of *action information asymmetry*: the complete information setting, the no information setting, and the hierarchical setting considered in this paper. Learning under the complete information setting is equivalent to the case of single-agent with action space being the product space $[A] \times [B]$; single-agent results suggest lower regret bounds of $\Omega(\sqrt{ABT})$ for the bandit setting and $\Omega(\sqrt{H^2SABT})$ for the MDP setting, which are achievable by the state-of-the-art algorithms (Auer et al., 2002a; Azar et al., 2017). Since both the other two information structures, i.e. the no information setting and the hierarchical setting, also involve exploring among the AB pairs of actions but with less information, they inherit the lower bounds. Interestingly, with higher complexity and stronger assumptions, one may achieve the lower regret bounds in these two settings as well using the CI approach. For details see Appendix A.

3. Learning Hierarchical Bandits

Since U1 does not observe U2's actions, it is unclear how U1 can utilize or interpret the samples she receives. For example, if U1 receives a low reward, one possibility is that U1 has chosen a bad action, so whatever action U2 chooses, the reward is going to be low; but it is also possible that the action chosen by U1 is actually good (i.e., the reward would be high if U2 chose a good subsequent action), but U2 has chosen a bad subsequent action. If the identity of U2's action is not revealed, in general, U1 cannot distinguish these two cases. As mentioned in Section 1, this issue can be resolved by the CI approach, which enables U1 to infer the actions taken by U2, even if they are not directly observable.

We show that providing U2 executes a *no-regret* algorithm (i.e., an algorithm that always guarantees a sub-linear-in-T regret), we can actually make the agents converge to playing optimal actions while keeping U1 agnostic to U2's actions during the whole process. The key observation is that when U2 is a no-regret learning agent, his choices of action under a given action of U1 will converge to the best one, hence avoiding the second possibility mentioned above in the long run.

The following Assumption 1 specifies the condition that U2's algorithm should satisfy, for which using an MAB algorithm with near-optimal, high-probability regret guarantee for each action $a \in [A]$ is enough. Some existing algorithms satisfying this assumption are: UCB (Agrawal, 1995; Auer et al., 2002a), KL-UCB (Garivier and Cappé, 2011), EXP3.P (Auer et al., 2002b), EXP3-IX (Neu, 2015). [5]

Assumption 1 *U2 guarantees the following for some universal constant $\kappa \geq 1$ with probability at least $1 - \delta$:*

$$\forall t, a, \qquad \sum_{\tau=1}^{t} \mathbb{I}[a_\tau = a](\mu_{a,1} - \mu_{a,b_\tau}) \leq \sqrt{\kappa B \sum_{\tau=1}^{t} \mathbb{I}[a_\tau = a] \log(T/\delta)}.$$

With Assumption 1, our algorithm is presented in Algorithm 1. In Algorithm 1, U2 simply executes the specified algorithm. On the other hand, U1 follows a selection rule that is very similar

[5] While in Assumption 1 we assume that U2 uses a near-optimal algorithm for MAB for ease of presentation, our framework can also handle the case where he uses some sub-optimal algorithm (the final regret bound will also be sub-optimal though).

Algorithm 1 UCB for Hierarchical Bandits

```
1 define: $c > 0$ is a universal constant.
2 U2 starts running algorithms satisfying Assumption 1 (or Assumption 2) with some $\kappa \geq 1$ (we denote the instance of algorithm under action $a \in [A]$ as $\mathsf{ALG}(a)$).
3 for $t = 1, 2, \ldots, T$ do
4     U1 chooses $a_t \in \operatorname*{argmax}_{a \in [A]} \hat{\mu}_t(a) + \sqrt{\frac{\kappa B \log(T/\delta)}{n_t(a)^+}} + c\sqrt{\frac{\log(T/\delta)}{n_t(a)^+}}$. ($\hat{\mu}_t(a), n_t(a)$ are defined in (1))
5     After observing $a_t$, U2 calls $\mathsf{ALG}(a_t)$, which outputs an action $b_t$.
6     U2 chooses $b_t$.
7     U1 and U2 observe $r_t$, and U2 updates $\mathsf{ALG}(a_t)$ using $r_t$.
  end
```

to the UCB1 algorithm of Auer et al. (2002a). Specifically, in the beginning of each round t, U1 calculates the *empirical mean* of each of her actions:

$$\hat{\mu}_t(a) = \frac{1}{n_t(a)^+} \sum_{\tau=1}^{t-1} \mathbb{I}[a_\tau = a] r_\tau, \qquad \text{where } n_t(a) = \sum_{\tau=1}^{t-1} \mathbb{I}[a_\tau = a] \tag{1}$$

Also, U1 calculates the *bonus* for action a, which is simply the *average regret upper bound* of U2 on action a up to time $t-1$ (by Assumption 1, the regret on action a up to time $t-1$ is upper bounded by $\sqrt{\kappa B n_t(a) \log(T/\delta)}$), plus a term given by the Hoeffding bound (see Line 4 of Algorithm 1). Then U1 simply chooses the arm with the largest empirical mean plus bonus. While appearing similar, it is not the same as the standard UCB1 algorithm. First, the empirical mean $\hat{\mu}_t(a)$ is an average over samples that are not identically distributed since each r_t also depends on the action chosen by U2. Second, the construction of the bonus term involves the *regret upper bound* of U2, in addition to another term implied by concentration inequalities. However, our algorithm is indeed inspired by the UCB1 algorithm — in UCB1, the bonus term is designed to fulfill the following two properties: 1) for every arm a, the empirical mean of reward plus bonus should upper bound the true mean of reward with high probability; 2) the sum of bonus of the chosen arms in rounds $t = 1, 2, \ldots, T$ should be sub-linear in T. With both properties, the UCB1 algorithm is guaranteed to have sub-linear regret.

In our case, we shall identify $\mu_{a,1}$ as the "true mean of reward" of arm a. To satisfy the first property above, we would like to add a bonus term that upper bounds $\mu_{a,1} - \hat{\mu}_t(a)$. Note that $\hat{\mu}_t(a)$ is the mean of reward of U2 in the rounds when U1 chooses a, so $\mu_{a,1} - \hat{\mu}_t(a)$ is simply the *average regret* of U2 under U1's action a. Therefore, adding the regret bound of U2 as the bonus of U1 gives the first property above. The second property is also satisfied as long as U2 uses algorithms with sub-linear regret guarantees.

We remark that it is important for U1 to use a slightly larger bonus (i.e., $\widetilde{\mathcal{O}}\big(\sqrt{B/n_t(a)}\big)$) rather than the standard one she would use when she plays alone (i.e., $\widetilde{\mathcal{O}}\big(\sqrt{1/n_t(a)}\big)$). This extra $\sqrt{B}$ factor makes the bonus of U1 decrease at a slower rate, leading to more exploration on each of U1's actions, and allowing U2 to have enough time to find the best action under each of U1's actions.

The analysis of this algorithm is straightforward given the above intuition, which we show in the following theorem.

Theorem 1 *Suppose that U2 uses algorithms that satisfy Assumption 1. Then Algorithm 1 guarantees that with probability at least* $1 - \mathcal{O}(\delta)$, $\mathrm{Reg}(T) = \mathcal{O}\big(\sqrt{ABT\log(T/\delta)}\big)$.

Proof

$$\begin{aligned}
&\sum_{t=1}^{T}(\mu_{1,1} - \mu_{a_t,b_t}) \\
&\le \sum_{t=1}^{T}\left(\frac{1}{n_t(1)^+}\sum_{\tau=1}^{t-1}\mathbb{I}[a_\tau = 1]\mu_{1,b_\tau} + \sqrt{\frac{\kappa B\log(T/\delta)}{n_t(1)^+}} - \mu_{a_t,b_t}\right) && \text{(by Assumption 1 with } a = 1) \\
&\le \sum_{t=1}^{T}\left(\hat{\mu}_t(1) + c\sqrt{\frac{\log(T/\delta)}{n_t(1)^+}} + \sqrt{\frac{\kappa B\log(T/\delta)}{n_t(1)^+}} - \mu_{a_t,b_t}\right) && \text{(by Lemma 10)} \\
&\le \sum_{t=1}^{T}\left(\hat{\mu}_t(a_t) + c\sqrt{\frac{\log(T/\delta)}{n_t(a_t)^+}} + \sqrt{\frac{\kappa B\log(T/\delta)}{n_t(a_t)^+}} - \mu_{a_t,b_t}\right) && \text{(by the selection rule of } a_t) \\
&= \sum_{t=1}^{T}(\hat{\mu}_t(a_t) - \mu_{a_t,b_t}) + \mathcal{O}\left(\sqrt{ABT\log(T/\delta)}\right) \\
&= \underbrace{\sum_{a\in[A]}\sum_{t=1}^{T}\mathbb{I}[a_t = a]\left(\hat{\mu}_t(a) - \mu_{a,1}\right)}_{\textbf{term}_1} + \underbrace{\sum_{a\in[A]}\sum_{t=1}^{T}\mathbb{I}[a_t = a]\left(\mu_{a,1} - \mu_{a_t,b_t}\right)}_{\textbf{term}_2} + \mathcal{O}\left(\sqrt{ABT\log(T/\delta)}\right)
\end{aligned}$$

Notice that $\textbf{term}_2 = \mathcal{O}\left(\sum_{a\in[A]}\sqrt{n_{T+1}(a)\log(T/\delta)}\right) = \mathcal{O}\left(\sqrt{ABT\log(T/\delta)}\right)$ due to Assumption 1. Besides, by Azuma's inequality, with probability at least $1 - \mathcal{O}(\delta)$, for all t and a,

$$\hat{\mu}_t(a) = \frac{\sum_{\tau=1}^{t-1}\mathbb{I}[a_\tau = a]r_\tau}{n_t(a)^+} \le \frac{\sum_{\tau=1}^{t-1}\mathbb{I}[a_\tau = a]\mu_{a,b_\tau}}{n_t(a)^+} + \mathcal{O}\left(\sqrt{\frac{\log(T/\delta)}{n_t(a)^+}}\right) \le \mu_{a,1} + \mathcal{O}\left(\sqrt{\frac{\log(T/\delta)}{n_t(a)^+}}\right).$$

Therefore, $\textbf{term}_1 \le \mathcal{O}\left(\sum_{a\in[A]}\sum_{t=1}^{T}\mathbb{I}[a_t = a]\sqrt{\frac{\log(T/\delta)}{n_t(a)^+}}\right) = \mathcal{O}\left(\sqrt{AT\log(T/\delta)}\right)$. Combining everything finishes the proof. ■

For MAB, there are also algorithms with refined gap-dependent regret bounds with only $\mathcal{O}(\log(T))$ dependence on T (e.g., the UCB1 algorithm of Auer et al. (2002a)). Below we show that if U2 executes such algorithms, the overall regret can also be of order $\mathcal{O}(\log(T))$. Such algorithms satisfy the following assumption:

Assumption 2 *U2 guarantees the following for some universal constant* $\kappa \ge 1$ *with probability at least* $1 - \delta$:

$$\forall t, a, \qquad \sum_{\tau=1}^{t}\mathbb{I}[a_\tau = a](\mu_{a,1} - \mu_{a,b_\tau}) \le \min\left\{\kappa\sum_{b\in\mathcal{B}_a^\times}\frac{\log(T/\delta)}{\mu_{a,1} - \mu_{a,b}},\ \sqrt{\kappa B\sum_{\tau=1}^{t}\mathbb{I}[a_\tau = a]\log(T/\delta)}\right\},$$

where $\mathcal{B}_a^\times \triangleq \{b \in [B] : \mu_{a,1} - \mu_{a,b} > 0\}$ *is the set of sub-optimal arms of U2 under U1's action* a.

With Assumption 2, Algorithm 1 has the following guarantee:

Theorem 2 *Suppose that U2 uses algorithms that satisfy Assumption 2. Then Algorithm 1 guarantees that with probability at least* $1 - \mathcal{O}(\delta)$*,*

$$\mathrm{Reg}(T) = \mathcal{O}\left(\sum_{a\in\mathcal{A}^\times} \frac{B\log(T/\delta)}{\mu_{1,1}-\mu_{a,1}} + \sum_{a\in\mathcal{A}^\circ}\sum_{b\in\mathcal{B}_a^\times} \frac{\log(T/\delta)}{\mu_{a,1}-\mu_{a,b}}\right),$$

where $\mathcal{A}^\circ \triangleq \{a \in [A] : \mu_{1,1} = \mu_{a,1}\}$*,* $\mathcal{A}^\times \triangleq \{a \in [A] : \mu_{1,1} - \mu_{a,1} > 0\} = [A]\backslash\mathcal{A}^\circ$*, and* $\mathcal{B}_a^\times$ *is as defined in Assumption 2.*

The proof is deferred to Appendix B. In Theorem 2, the regret consists of two parts. For an action $a \in [A]$ that is sub-optimal (i.e., $\mu_{1,1} > \mu_{a,1}$), the regret scales with $\frac{B}{\mu_{1,1}-\mu_{a,1}}$; for optimal ones (i.e., $\mu_{1,1} = \mu_{a,1}$), the regret scales with $\sum_{b\in\mathcal{B}_a^\times} \frac{1}{\mu_{a,1}-\mu_{a,b}}$, i.e., the sum of inverse gaps of all U2's sub-optimal actions under a.

Our hierarchical bandit setting coincides with the model selection problem studied in Arora et al. (2021). Our result in Theorem 2 improves their results in two ways. First, when using UCB-based algorithms as the base algorithm, our regret bound scales with $\log(T)$, while theirs scales with $\log^2(T)$. This answers their open question regarding whether $\log^2(T)$ is tight (see the discussion in their Section 4.1). Second, while they assume that U1's optimal action is unique (i.e., for all $a \neq 1$, $\mu_{1,1} > \mu_{a,1}$), we do not make such an assumption.

3.1. Extensions

Multiple Followers Case Our framework can be easily extended to the case when there is a leader and multiple followers where each of the followers' rewards is only a function of the choice of the leader and the individual follower as well as independent across others, and the reward of the leader is an average/sum of that of all the followers. This is particularly useful in modeling networks with a star topology, e.g., federated learning systems (with leader being the server and followers being the clients), mobile networks (with leader being the base station/access point and followers being the mobile users), etc. In these networks, it is usually the case that the followers are heterogeneous and move into or out of the networks dynamically. Since our proposed method does not require per-round coordination and treats the algorithms of the followers as black boxes, the coordination overhead for the leader can be relatively low. Besides, our algorithm largely preserves privacy for the followers, which is also much more preferable than other schemes where the leader is required to know the algorithm of the followers. In Appendix C, we show that near-optimal regret bounds can also be obtained in this case with an idea similar to that presented in Section 3.

Deep Hierarchy Case While in Section 3 we consider a two-layer model which only involves U1 and U2, the idea can be generalized to the case where there are $D > 2$ layers. In other words, in each round, the decision is made jointly by D agents with a fixed ordering, where agents making decision earlier cannot observe the actions taken by agents making decisions later. Such a protocol may be useful in modeling networks with deeper hierarchy, e.g., mobile networks where macro-, micro-, pico-, and femto- base stations are overlaid to serve user equipments (Jain et al., 2011; Wei and Liao, 2017; Sigwele et al., 2020). In Appendix D, we design a multi-layer UCB algorithm for this setting, and show near-optimal regret bounds for it.

4. Learning Hierarchical MDPs

This setting is much more challenging than the hierarchical bandit setting. First, notice that in this setting, both agents are facing non-stationary transition and reward because of the dependence of these quantities on the policy of the other agent, which varies with time. Obtaining regret bounds in such time-varying MDPs is in general hard (Abbasi-Yadkori et al., 2013; Radanovic et al., 2019; Tian et al., 2021), except for problems with special structures or extra assumptions (Radanovic et al., 2019; Tian et al., 2021; Leonardos et al., 2021) like our case here. Second, notice that in the bandit case, given any choice of U1, U2 is essentially facing a stationary MAB problem, and thus we can directly apply existing theorems for standard MAB; however, in the MDP case, the world that U2 sees on a certain step is still affected by the non-stationarity of U1's policies in future steps. In this case, standard analysis for stationary MDPs cannot be directly applied.

An initial idea to deal with this setting is to let both agents run existing UCB-based algorithms (e.g., UCBVI (Azar et al., 2017), UCB-H (Jin et al., 2018)) with an increased bonus term for U1 to compensate for the regret of U2, imitating our hierarchical bandit solution. However, as we point out above, U2's world is also affected by the policies of U1 in future steps. Therefore, a natural solution is to do the following: besides letting U1 add extra bonus to compensate the regret of U2, we also let U2 add extra bonus to compensate the regret of U1 in future steps. Unfortunately, for this hypothetical algorithm, it is unclear to us how to obtain a regret bound that is polynomial in the number of steps H. This is because by recursively adding extra bonus in each layer, we end up with a factor of $(AB)^{H/2}$ in the regret bound, similar to the "Deep Hierarchy Case" discussed in Section 3.1 and Appendix D.

To address this issue, instead of trying to let U2 best respond to the non-stationary world created by U1, we exploit the fact that U2 has full knowledge about the joint action space, and let U2 find the best *joint policy* of U1 and U2. Then U2 will execute his part of this joint policy even though U1 may not follow it. Although this brings other issues (discussed later), it avoids the need of U2 to compensate for the regret of U1 in later steps, and prevents the exponential blowup in the regret bound.

Our algorithm for hierarchical MDPs is presented in Algorithm 2. To avoid cluttered notation, we drop the episode index t when presenting the algorithm. Unlike Algorithm 1 where we can plug in any algorithm for U2 with the desired regret bound, in Algorithm 2 we specify both agents' algorithms. We leave as an open problem how to design a black-box-reduction-styled algorithm for the MDP setting similar to Algorithm 1.

In Algorithm 2, U1 maintains optimistic value function estimators $V_h^1(s), Q_h^1(s, a)$, and U2 maintains $V_h^2(s), Q_h^2(s, a, b)$ for every $h = 1, \dots, H$. Their constructions are based on two standard UCB-based algorithms. Specifically, the constructions of $V_h^1(s)$ and $Q_h^1(s, a)$ (Line 17-Line 18) are similar to those of UCB Q-learning (Jin et al., 2018), with the bonus term bns_τ^1 enlarged by a factor of $\sqrt{SB}$. Like in the bandit setting from Section 3, U1 faces a non-stationary environment, and the extra $\sqrt{B}$ factor is used to compensate for the regret of U2 in future steps[6]. On the other hand, the constructions of $V_h^2(s)$ and $Q_h^2(s, a, b)$ (Line 26-Line 27) are similar to those of UCBVI (Azar et al., 2017). In particular, $V_h^2(s)$ is obtained by jointly optimizing over the actions of U1 and U2 (Line 27), conforming to our previous discussions.

Perhaps the most intriguing is why we use UCB-H for U1 but UCBVI for U2. From a high level, this is because UCB-H shrinks its confidence set of value functions at a slower rate, while

[6] The extra $\sqrt{S}$ factor arises from a technical difficulty, and we are unsure whether it is necessary.

Algorithm 2 UCB-H/UCBVI for Hierarchical MDP

```
1  define: $\alpha_\tau = \frac{H+1}{H+\tau}$, $\mathsf{bns}^1_\tau = c'\sqrt{\frac{H^3SB\log(T/\delta)}{\tau^+}}$, $\mathsf{bns}^2_\tau = c\sqrt{\frac{H^2S\log(T/\delta)}{\tau^+}}$ where $c, c' \geq 1$ are universal constants.
2  initialize: $Q^1_h(s,a), Q^2_h(s,a,b) \leftarrow H \ \ \forall h, s, a, b,$
3  $n_h(s,a), n_h(s,a,b), n_h(s,a,b,s'), \theta_h(s,a,b) \leftarrow 0 \ \ \forall h, s, a, b, s'.$
4  for $t = 1, \dots, T$ do
5      U1 and U2 observes $s_1$.
6      for $h = 1, \dots, H$ do
7          U1 chooses $a_h \in \mathrm{argmax}_a\, Q^1_h(s_h, a)$.
8          U2 observes $a_h$.
9          U2 chooses $b_h \in \mathrm{argmax}_b\, Q^2_h(s_h, a_h, b)$.
10         U1 and U2 observe $r_h$ and $s_{h+1}$.
11         U1 updates counts of visits: $n_h(s_h, a_h) \overset{+}{\leftarrow} 1$.   ("$\overset{+}{\leftarrow}$ 1" means to increase the number by 1.)
12         U2 updates counts of visits: $n_h(s_h, a_h, b_h) \overset{+}{\leftarrow} 1$, $n_h(s_h, a_h, b_h, s_{h+1}) \overset{+}{\leftarrow} 1$.
13         U2 updates cumulative reward: $\theta_h(s_h, a_h, b_h) \overset{+}{\leftarrow} r_h$.
       end

14     U1 updates $Q/V$ functions ($\approx$ UCB-H update rule):
15     $V^1_{H+1}(\cdot) \leftarrow 0$.
16     for $h = 1, \dots, H$ do
17         $Q^1_h(s_h, a_h) \leftarrow (1-\alpha_\tau)Q^1_h(s_h, a_h) + \alpha_\tau \left(r_h + V^1_{h+1}(s_{h+1}) + \mathsf{bns}^1_\tau\right)$
18         $V^1_h(s_h) \leftarrow \min\{\max_a Q^1_h(s_h, a),\ H\}$
19         where $\tau = n_h(s_h, a_h)$.
       end

20     U2 updates $Q/V$ functions ($\approx$ UCBVI update rule):
21     Let $\hat{P}_h(s'|s,a,b) = \frac{n_h(s,a,b,s')}{n_h(s,a,b)}$ and $\hat{R}_h(s,a,b) = \frac{\theta_h(s,a,b)}{n_h(s,a,b)}$  $\forall h, s, a, b, s'$.
22     (if $n_h(s,a,b) = 0$, set $\hat{P}_h(s'|s,a,b) = \frac{1}{|S|}$ and $\hat{R}_h(s,a,b) = 0$).
23     $V^2_{H+1}(\cdot) \leftarrow 0$.
24     for $h = H, \dots, 1$ do
25         for all $s, a, b$ do
26             $Q^2_h(s,a,b) \leftarrow \min\left\{\hat{R}_h(s,a,b) + \mathbb{E}_{s' \sim \hat{P}_h(\cdot|s,a,b)}\left[V^2_{h+1}(s')\right] + \mathsf{bns}^2_\tau,\ \ Q^2_h(s,a,b)\right\}$
27             $V^2_h(s) \leftarrow \max_{a,b} Q^2_h(s,a,b)$
28             where $\tau = n_h(s,a,b)$.
           end
       end
   end
```

UCBVI is faster, which fulfills our need that U1 has to explore more in the early stages, for U2 to have enough time to find his optimal policy. Recall that the value iteration performed by U2 is through $V^2_h(s) \leftarrow \max_{a,b} Q^2_h(s,a,b)$ (Line 27). By the optimism principle, ideally we would like the agents to take actions $(a_h, b_h) = \operatorname{argmax}_{a,b} Q_h^2(s_h, a, b)$ to facilitate exploration. However, since U2 cannot control the actions taken by U1, and there is no communication between U1 and U2, it is unclear whether the optimism principle on Q_h^2 can be successfully carried out (the best U2 can do is to take $b_h = \operatorname{argmax}_b Q_h^2(s_h, a_h, b)$ for some a_h taken by U1, as done in Line 9). Our key finding is that if $Q_h^1(s,a)$ always upper bounds $\max_b Q_h^2(s,a,b)$, then the agents can still perform adequate joint exploration without explicit coordination. This key property can be shown straightforwardly if we use UCB-H for U1 and UCBVI for U2 (see the proof of Lemma 5).

Below, we establish some lemmas to be used in the regret bound analysis. The detailed proofs are deferred to Appendix B. We first define new notation with the episode indices.

Definition 3 *Let $Q_{t,h}^1(\cdot,\cdot)$, $Q_{t,h}^2(\cdot,\cdot,\cdot)$ be the $Q_h^1(\cdot,\cdot)$, $Q_h^2(\cdot,\cdot,\cdot)$ at the beginning of episode t in Algorithm 2. Let $s_{t,h}, a_{t,h}, b_{t,h}, r_{t,h}$ be the s_h, a_h, b_h, r_h within episode t in Algorithm 2.*

Lemma 4 below shows the optimism of U2's Q-function estimator, and relates the cumulative sum of $Q_{t,h}^2(s_{t,h}, a_{t,h}, b_{t,h}) - Q_{*,h}(s_{t,h}, a_{t,h}, b_{t,h})$ to that of $V_{t,h+1}^2(s_{t,h+1}) - V_{*,h+1}(s_{t,h+1})$. The proof is standard and we provide it in Appendix B for completeness.

Lemma 4 *With probability at least $1 - \mathcal{O}(\delta)$, $Q_{t,h}^2(s,a,b) \ge Q_{*,h}(s,a,b)$ for all t, h, s, a, b, and*

$$\sum_{t=1}^{T} \left(Q_{t,h}^2(s_{t,h}, a_{t,h}, b_{t,h}) - Q_{*,h}(s_{t,h}, a_{t,h}, b_{t,h})\right)$$
$$\le \sum_{t=1}^{T} \left(V_{t,h+1}^2(s_{t,h+1}) - V_{*,h+1}(s_{t,h+1})\right) + \widetilde{\mathcal{O}}\left(\sqrt{H^2S^2ABT}\right), \quad \forall h.$$

We remark that it is possible to improve the bound in Lemma 4 by a factor of $\sqrt{S}$ by using the more refined analysis in Azar et al. (2017) and defining bns_τ^2 to be a $\sqrt{S}$-factor smaller. However, as we will see below, this improvement will not lead to a better final regret bound, so in Lemma 4 we opt to use a simpler analysis with a looser bound.

Next, we establish our key lemma, Lemma 5, which states that U1 has *more optimism* than U2. In this lemma we have to make bns_τ^1 a $\sqrt{SB}$-factor larger than that in Jin et al. (2018). While the $\sqrt{B}$ factor is necessary for the same reasons as in the bandit case, it is unclear whether the $\sqrt{S}$ factor is necessary. We leave the improvement of this factor as a future direction.

Lemma 5 *With probability at least $1 - \mathcal{O}(\delta)$, $Q_{t,h}^1(s,a) \ge \max_b Q_{t,h}^2(s,a,b)$ for all t, h, s, a, b.*

Finally, in Lemma 6, we relate the cumulative sum of $Q_{t,h}^1(s_{t,h}, a_{t,h}) - Q_{*,h}^1(s_{t,h}, a_{t,h})$ to that of $V_{t,h+1}^1(s_{t,h+1}) - V_{*,h+1}^1(s_{t,h+1})$. The proof is similar to that of Jin et al. (2018), but the bound is a $\sqrt{SB}$-factor larger than theirs due to the use of a larger bonus bns_τ^1.

Lemma 6 *With probability at least $1 - \mathcal{O}(\delta)$,*

$$\sum_{t=1}^{T} \left(Q_{t,h}^1(s_{t,h}, a_{t,h}) - Q_{*,h}(s_{t,h}, a_{t,h})\right)$$
$$\le \left(1 + \frac{1}{H}\right) \sum_{t=1}^{T} \left(V_{t,h+1}^1(s_{t,h+1}) - V_{*,h+1}(s_{t,h+1})\right) + \widetilde{\mathcal{O}}\left(\sqrt{H^3S^2ABT} + HSA\right), \quad \forall h.$$

Thanks to the fact that $V^1_{t,h}(s_{t,h}) = Q^1_{t,h}(s_{t,h}, a_{t,h})$, Lemma 6 leads to the following simple corollary. Note that we do not have a similar corollary for Lemma 4 because $V^2_{t,h}(s_{t,h}) \neq Q^2_{t,h}(s_{t,h}, a_{t,h}, b_{t,h})$ (as discussed earlier, $(a_{t,h}, b_{t,h})$ is not necessarily equal to $\operatorname{argmax}_{a,b} Q^2_{t,h}(s_{t,h}, a, b)$). The proof of the corollary is also in Appendix B.

Corollary 3 $\sum_{t=1}^{T} \left(Q^1_{t,h}(s_{t,h}, a_{t,h}) - Q_{*,h}(s_{t,h}, a_{t,h})\right) = \widetilde{\mathcal{O}}\left(\sqrt{H^5 S^2 ABT} + H^2 SA\right)$.

Finally, we are able to show our main theorem:

Theorem 7 *With probability* $1 - \mathcal{O}(\delta)$*, Algorithm 2 guarantees* $\text{Reg}(T) = \widetilde{\mathcal{O}}\left(H^{3.5} S\sqrt{ABT} + H^3 SA\right)$.

Proof We perform regret decomposition as follows:

$$\begin{aligned}
\text{Reg}(T) &= \sum_{t=1}^{T} \left(V_{*,1}(s_{t,1}) - V_1^{\pi_t^1, \pi_t^2}(s_{t,1})\right) \\
&= \sum_{t=1}^{T} \sum_{h=1}^{H} \sum_{s,a,b} \mathbb{P}\left[(s_{t,h}, a_{t,h}, b_{t,h}) = (s, a, b) \;\middle|\; s_{t,1}, \pi_t^1, \pi_t^2\right] \left(V_{*,h}(s) - Q_{*,h}(s, a, b)\right) \\
&\qquad \text{(by the performance difference lemma (Kakade and Langford, 2002))} \\
&= \sum_{t=1}^{T} \sum_{h=1}^{H} \left(V_{*,h}(s_{t,h}) - Q_{*,h}(s_{t,h}, a_{t,h}, b_{t,h})\right) + \widetilde{\mathcal{O}}\left(H\sqrt{HT}\right) \qquad \text{(by Lemma 10)} \\
&= \sum_{t=1}^{T} \sum_{h=1}^{H} \left(V_{*,h}(s_{t,h}) - Q_{*,h}(s_{t,h}, a_{t,h})\right) + \sum_{t=1}^{T} \sum_{h=1}^{H} \left(Q_{*,h}(s_{t,h}, a_{t,h}) - Q_{*,h}(s_{t,h}, a_{t,h}, b_{t,h})\right) + \widetilde{\mathcal{O}}\left(\sqrt{H^3 T}\right).
\end{aligned} \tag{2}$$

Note that

$$\begin{aligned}
\text{Reg}_h^1 &\triangleq \sum_{t=1}^{T} \left(V_{*,h}(s_{t,h}) - Q_{*,h}(s_{t,h}, a_{t,h})\right) \le \sum_{t=1}^{T} \left(Q^1_{t,h}(s_{t,h}, a_{t,h}) - Q_{*,h}(s_{t,h}, a_{t,h})\right) \\
&\le \widetilde{\mathcal{O}}\left(\sqrt{H^5 S^2 ABT} + H^2 SA\right), \qquad \text{(by Corollary 3)}
\end{aligned}$$

where the first inequality is because Lemma 4, Lemma 5, and the way U1 chooses $a_{t,h}$ yield $V_{*,h}(s_{t,h}) = \max_{a,b} Q_{*,h}(s_{t,h}, a, b) \le \max_{a,b} Q^2_{t,h}(s_{t,h}, a, b) \le \max_a Q^1_{t,h}(s_{t,h}, a) = Q^1_{t,h}(s_{t,h}, a_{t,h})$. On the other hand,

$$\begin{aligned}
\text{Reg}_h^2 &\triangleq \sum_{t=1}^{T} \left(Q_{*,h}(s_{t,h}, a_{t,h}) - Q_{*,h}(s_{t,h}, a_{t,h}, b_{t,h})\right) \\
&\le \sum_{t=1}^{T} \left(Q^2_{t,h}(s_{t,h}, a_{t,h}, b_{t,h}) - Q_{*,h}(s_{t,h}, a_{t,h}, b_{t,h})\right) \\
&\le \sum_{t=1}^{T} \left(V^2_{t,h+1}(s_{t,h+1}) - V_{*,h+1}(s_{t,h+1})\right) + \widetilde{\mathcal{O}}\left(\sqrt{H^2 S^2 ABT} + H^2 SA\right) \qquad \text{(by Lemma 4)}
\end{aligned}$$

$$\leq \sum_{t=1}^{T} \left(V^1_{t,h+1}(s_{t,h+1}) - V_{*,h+1}(s_{t,h+1})\right) + \widetilde{\mathcal{O}}\left(\sqrt{H^5S^2ABT} + H^2SA\right)$$

(by Lemma 5 and the definitions of $V^1_{t,h}(s)$ and $V^2_{t,h}(s)$)

$$\leq \sum_{t=1}^{T} \left(Q^1_{t,h+1}(s_{t,h+1}, a_{t,h+1}) - Q_{*,h+1}(s_{t,h+1}, a_{t,h+1})\right) + \widetilde{\mathcal{O}}\left(\sqrt{H^5S^2ABT} + H^2SA\right)$$

(by the way U1 chooses $a_{t,h+1}$)

$$\leq \widetilde{\mathcal{O}}\left(\sqrt{H^5S^2ABT} + H^2SA\right), \qquad \text{(by Corollary 3)}$$

where the first inequality follows from Lemma 4 and the way U2 chooses actions, resulting in $Q_{*,h}(s_{t,h}, a_{t,h}) = \max_b Q_{*,h}(s_{t,h}, a_{t,h}, b) \leq \max_b Q^2_{t,h}(s_{t,h}, a_{t,h}, b) = Q^2_{t,h}(s_{t,h}, a_{t,h}, b_{t,h})$.

Combining the bounds on Reg^1_h and Reg^2_h with (2) proves the theorem. ■

5. Conclusion and Future Directions

In this work, we exploit the hierarchical information structure in hierarchical bandits/MDPs, and propose efficient and near-optimal algorithms that require no coordination or communication between the agents. A key feature of our algorithms is that the leader, i.e., the less-informed upper level agent, performs single-agent-like near-optimal subroutines with specially designed bonuses without the need of knowing or tracking the learning procedure of the lower level agent(s). One future direction is to explore other information structures that may allow simplification when more than two agents and reward information asymmetry come in. Extending the results to bandit problems with structural relationships between the payoffs of the arms and to general-sum games are also interesting directions.

Acknowledgments

Hsu Kao and Vijay Subramanian would like to acknowledge support from NSF via grant ECCS 2038416, CNS 1955777, CCF 2008130, and support via a MIDAS grant sponsored by General Dynamics. Chen-Yu Wei would like to acknowledge support from NSF via grant IIS 1943607.

References

Yasin Abbasi-Yadkori, Peter L. Bartlett, and Csaba Szepesvári. Online learning in Markov decision processes with adversarially chosen transition probability distributions. In *Advances in Neural Information Processing Systems*, 2013.

Alekh Agarwal, Haipeng Luo, Behnam Neyshabur, and Robert E. Schapire. Corralling a band of bandit algorithms. In *Conference on Learning Theory*, 2017.

Rajeev Agrawal. Sample mean based index policies by O(log n) regret for the multi-armed bandit problem. *Advances in Applied Probability*, 27(4):1054–1078, 1995.

Raman Arora, Teodor Vanislavov Marinov, and Mehryar Mohri. Corralling stochastic bandit algorithms. In *International Conference on Artificial Intelligence and Statistics*, 2021.

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

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002b.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, 2017.

Yu Bai, Chi Jin, Huan Wang, and Caiming Xiong. Sample-efficient learning of Stackelberg equilibria in general-sum games. In *Advances in Neural Information Processing Systems*, 2021.

Ilai Bistritz and Amir Leshem. Game of thrones: Fully distributed learning for multiplayer bandits. *Mathematics of Operations Research*, 46(1):159–178, 2021.

Ilai Bistritz, Tavor Z. Baharav, Amir Leshem, and Nicholas Bambos. One for all and all for one: Distributed learning of fair allocations with multi-player bandits. *IEEE Journal on Selected Areas in Information Theory*, 2(2):584–598, 2021.

Sébastien Bubeck, Thomas Budzinski, and Mark Sellke. Cooperative and stochastic multi-player multi-armed bandit: Optimal regret with neither communication nor collisions. In *Conference on Learning Theory*, 2021.

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

Ronshee Chawla, Abishek Sankararaman, Ayalvadi Ganesh, and Sanjay Shakkottai. The gossiping insert-eliminate algorithm for multi-agent bandits. In *International Conference on Artificial Intelligence and Statistic*, 2020.

Ashok Cutkosky, Christoph Dann, Abhimanyu Das, Claudio Gentile, Aldo Pacchiano, and Manish Purohit. Dynamic balancing for model selection in bandits and RL. In *International Conference on Machine Learning*, 2021.

Jilles Dibangoye and Olivier Buffet. Learning to act in decentralized partially observable MDPs. In *International Conference on Machine Learning*, 2018.

Dongsheng Ding, Chen-Yu Wei, Kaiqing Zhang, and Mihailo R. Jovanović. Independent policy gradient for large-scale Markov potential games: Sharper rates, function approximation, and game-agnostic convergence. *arXiv preprint arXiv:2202.04129*, 2022.

Roy Fox, Stephen McAleer, Will Overman, and Ioannis Panageas. Independent natural policy gradient always converges in Markov potential games. *arXiv preprint arXiv:2110.10614*, 2021.

Aurélien Garivier and Olivier Cappé. The KL-UCB algorithm for bounded stochastic bandits and beyond. In *Conference on Learning Theory*, 2011.

R. K. Jain, Sumit Katiyar, and N. K. Agrawal. Hierarchical cellular structures in high-capacity cellular communication systems. *International Journal of Advanced Computer Science and Applications*, 2(9):51–57, 2011.

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11:1563–1600, 2010.

Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I. Jordan. Is Q-learning provably efficient. In *Advances in Neural Information Processing Systems*, 2018.

Chi Jin, Qinghua Liu, Yuanhao Wang, and Tiancheng Yu. V-learning – a simple, efficient, decentralized algorithm for multiagent RL. *arXiv preprint arXiv:2110.14555*, 2021.

Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *International Conference on Machine Learning*, 2002.

Dileep Kalathil, Naumaan Nayyar, and Rahul Jain. Decentralized learning for multiplayer multi-armed bandits. *IEEE Transactions on Information Theory*, 60(4):2331–2345, 2014.

Anusha Lalitha and Andrea Goldsmith. Bayesian algorithms for decentralized stochastic bandits. *IEEE Journal on Selected Areas in Information Theory*, 2(2):564–583, 2021.

Peter Landgren, Vaibhav Srivastava, and Naomi Ehrich Leonard. Distributed cooperative decision making in multi-agent multi-armed bandits. *Automatica*, 2021.

Stefanos Leonardos, Will Overman, Ioannis Panageas, and Georgios Piliouras. Global convergence of multi-agent policy gradient in Markov potential games. *arXiv preprint arXiv:2106.01969*, 2021.

Udari Madhushani and Naomi Ehrich Leonard. A dynamic observation strategy for multi-agent multi-armed bandit problem. In *European Control Conference*, 2020.

Weichao Mao and Tamer Başar. Provably efficient reinforcement learning in decentralized general-sum Markov games. *Dynamic Games and Applications*, pages 1–22, 2022.

Weichao Mao, Lin F. Yang, Kaiqing Zhang, and Tamer Başar. On improving model-free algorithms for decentralized multi-agent reinforcement learning. *arXiv preprint arXiv:2110.05707*, 2021.

Ashutosh Nayyar, Aditya Mahajan, and Demosthenis Teneketzis. Decentralized stochastic control with partial history sharing: A common information approach. *IEEE Transactions on Automatic Control*, 58(7):1644–1658, July 2013.

Gergely Neu. Explore no more: Improved high-probability regret bounds for non-stochastic bandits. In *Advances in Neural Information Processing Systems*, 2015.

Bing Ning, Gangcan Sun, Jianjun Li, Aihua Zhang, Wanming Hao, and Shouyi Yang. Resource allocation in multi-user cognitive radio network with Stackelberg game. *IEEE Access*, 2020.

Goran Radanovic, Rati Devidze, David Parkes, and Adish Singla. Learning to collaborate in Markov decision processes. In *International Conference on Machine Learning*, 2019.

Abishek Sankararaman, Ayalvadi Ganesh, and Sanjay 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.

Shahin Shahrampour, Alexander Rakhlin, and Ali Jadbabaie. Multi-armed bandits in multi-agent networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2017.

Tshiamo Sigwele, Yim Fun Hu, and Misfa Susanto. Energy-efficient 5G cloud RAN with virtual BBU server consolidation and base station sleeping. *Computer Networks*, 177:107302, 2020.

Ziang Song, Song Mei, and Yu Bai. When can we learn general-sum Markov games with a large number of players sample-efficiently? *arXiv preprint arXiv:2110.04184*, 2021.

Yi Tian, Yuanhao Wang, Tiancheng Yu, and Suvrit Sra. Online learning in unknown Markov games. In *International Conference on Machine Learning*, 2021.

Daniel Vial, Sanjay Shakkottai, and R. Srikant. Robust multi-agent multi-armed bandits. In *Proceedings of ACM Mobihoc*, 2021.

Chen-Yu Wei and Wanjiun Liao. Multi-cell cooperative scheduling for network utility maximization with user equipment side interference cancellation. *IEEE Transactions on Wireless Communications*, 17(1):619–635, 2017.

Tsachy Weissman, Erik Ordentlich, Gadiel Seroussi, Sergio Verdu, and Marcelo J. Weinberger. Inequalities for the L1 deviation of the empirical distribution. *Hewlett-Packard Labs, Tech. Rep*, 2003.

Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. Multi-agent reinforcement learning: A selective overview of theories and algorithms. *Springer Studies in Systems, Decision and Control, Handbook on RL and Control*, 2020a.

Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. Decentralized multi-agent reinforcement learning with networked agents: Recent advances. *IEEE Transactions on Automatic Control*, 2021a.

Runyu Zhang, Zhaolin Ren, and Na Li. Gradient play in stochastic games: stationary points, convergence, and sample complexity. *arXiv preprint arXiv:2106.00198*, 2021b.

Runyu Zhang, Jincheng Mei, Bo Dai, Dale Schuurmans, and Na Li. On the effect of log-barrier regularization in decentralized softmax gradient play in multiagent systems. *arXiv preprint arXiv:2202.00872*, 2022.

Zihan Zhang, Yuan Zhou, and Xiangyang Ji. Almost optimal model-free reinforcement learning via reference-advantage decomposition. In *Advances in Neural Information Processing Systems*, 2020b.

Appendix A. The CI Approach

In two-agent cases, there are three obvious types of information structure in terms of *action information asymmetry*.

- Sequential decision making: U1 first chooses $a_t \in [A]$; after observing a_t, U2 then chooses $b_t \in [B]$. This is the hierarchical information structure considered in this paper.

- Simultaneous decision making: U1 and U2 choose $a_t \in [A]$ and $b_t \in [B]$ simultaneously, respectively. Depending on their respective feedback afterwards, the setting can be further divided as follows:
  - Complete information sharing: both agents observe (a_t, b_t) directly after they make their choices.
  - No information sharing: both agents do not observe (a_t, b_t). This setting is considered in Chang et al. (2021).

In the setting of sequential decision making, U1 also does not observe b_t. Otherwise, it will be identical to the setting of complete information sharing in the learning context since after the time-step ends both agents will know (a_t, b_t).

In the complete information sharing setting, it is evident that the agents may treat the joint action space $[A] \times [B]$ as the new action space and learns as if a single agent (the fictitious coordinator) is learning the policy of choosing the joint actions. The learning is centralized as there is no information asymmetry and the non-stationarity issue will not happen. Interestingly, the same approach can be carried through in the other two information structures as well. Suppose U1 learns with algorithm ALG^1 (which should also include any possible tie-breaking rule) and randomization seed $\mathfrak{R}^1$, and U2 learns with ALG^2 and randomization seed $\mathfrak{R}^2$. Suppose both agents have the information of $(\mathsf{ALG}^1, \mathfrak{R}^1, \mathsf{ALG}^2, \mathfrak{R}^2)$. In step 1, U2 can generate a_1 from $(\mathsf{ALG}^1, \mathfrak{R}^1)$, and U1 can generate b_1, so that (a_1, b_1) becomes CI. Going forward, in step t, since $\mathcal{I}_{t-1} = (a_{1:t-1}, b_{1:t-1}, r_{1:t-1})$ (where $a_{1:t-1} = (a_1, \ldots, a_{t-1})$, etc.) is CI, U2 can reproduce a_t from $\mathsf{ALG}^1(\mathcal{I}_{t-1}, \mathfrak{R}^1)$, and U1 can reproduce b_t from $\mathsf{ALG}^2(\mathcal{I}_{t-1}, \mathfrak{R}^2)$, so that (a_t, b_t) is again CI. We can see that with this approach there will be no information asymmetry. Clearly, if U1 and U2 treat $[A] \times [B]$ from the coordinator's perspective and adopt the same single-agent algorithm $\mathsf{ALG}^1 = \mathsf{ALG}^2$ with near-optimal regret guarantee and the same randomization device $\mathfrak{R}^1 = \mathfrak{R}^2$, the problem is equivalent to learning in the standard single-agent AB-armed bandit or standard single-agent MDP with action space being $[A] \times [B]$. Using the state of the art algorithms, i.e., UCB1 (Auer et al., 2002a) for the bandit setting, and UCBVI algorithm (with a Bernstein bonus design) (Azar et al., 2017) (model-based) or the UCB-Advantage algorithm (Zhang et al., 2020b) (model-free), one may achieve the lower regret bounds of $\widetilde{\mathcal{O}}(\sqrt{ABT})$ for the bandit setting and $\widetilde{\mathcal{O}}(\sqrt{H^2SABT})$ for the MDP setting.

Both agents knowing $(\mathsf{ALG}^1, \mathfrak{R}^1, \mathsf{ALG}^2, \mathfrak{R}^2)$ and being able to reproduce each other's computation is a strong assumption. In the case of hierarchical information structure, simpler and more efficient alternatives presented in this paper are possible.

Appendix B. Omitted Proofs

Proof of Theorem 2 We first bound the number of times a sub-optimal arm $a \in \mathcal{A}^\times$ can be drawn by U1. Notice that with probability at least $1 - \mathcal{O}(\delta)$, for any $t, a \in \mathcal{A}^\times$,

$$
\begin{aligned}
&\hat{\mu}_t(a) + c\sqrt{\frac{\log(T/\delta)}{n_t(a)^+}} + \sqrt{\frac{\kappa B \log(T/\delta)}{n_t(a)^+}} \\
&\leq \frac{1}{n_t(a)^+}\sum_{\tau=1}^{t-1} \mu_{a,b_\tau} + 2c\sqrt{\frac{\log(T/\delta)}{n_t(a)^+}} + \sqrt{\frac{\kappa B \log(T/\delta)}{n_t(a)^+}} && \text{(by Lemma 10)}
\end{aligned}
$$

$$\leq \mu_{a,1} + 2c\sqrt{\frac{\log(T/\delta)}{n_t(a)^+}} + \sqrt{\frac{\kappa B\log(T/\delta)}{n_t(a)^+}}. \tag{3}$$

If $n_t(a) > \frac{(16\kappa B + 64c^2)\log(T/\delta)}{(\mu_{1,1}-\mu_{a,1})^2}$, then the last expression can further be upper bounded by $\mu_{a,1} + \frac{(\mu_{1,1}-\mu_{a,1})}{4} + \frac{(\mu_{1,1}-\mu_{a,1})}{4} < \mu_{1,1}$.

On the other hand, by Assumption 2 (which implies Assumption 1), with high probability we have

$$\begin{aligned}
&\hat{\mu}_t(1) + c\sqrt{\frac{\log(T/\delta)}{n_t(1)^+}} + \sqrt{\frac{\kappa B\log(T/\delta)}{n_t(1)^+}} \\
&\geq \frac{1}{n_t(1)^+}\sum_{\tau=1}^{t-1}\mu_{1,b_\tau} + \sqrt{\frac{\kappa B\log(T/\delta)}{n_t(1)^+}} && \text{(by Lemma 10)} \\
&\geq \mu_{1,1}.
\end{aligned}$$

Combining them, we see that if $n_t(a) > \frac{(16\kappa B + 64c^2)\log(T/\delta)}{(\mu_{1,1}-\mu_{a,1})^2}$, then

$$\hat{\mu}_t(1) + c\sqrt{\frac{\log(T/\delta)}{n_t(1)^+}} + \sqrt{\frac{\kappa B\log(T/\delta)}{n_t(1)^+}} > \hat{\mu}_t(a) + c\sqrt{\frac{\log(T/\delta)}{n_t(a)^+}} + \sqrt{\frac{\kappa B\log(T/\delta)}{n_t(a)^+}}.$$

By the way U1 selects arms, with high probability, she will not draw arm a at round t. In other words, the number of draws for any $a \in \mathcal{A}^\times$ is upper bounded by $\mathcal{O}\left(\frac{B\log(T/\delta)}{(\mu_{1,1}-\mu_{a,1})^2}\right)$.

Then we bound the overall regret. Define $n_t(a,b) \triangleq \sum_{\tau=1}^{t-1}\mathbf{1}[(a_\tau, b_\tau) = (a,b)]$. We have

$$\begin{aligned}
\mathrm{Reg}(T) &= \sum_{(a,b)\in[A]\times[B]} (\mu_{1,1} - \mu_{a,b}) \cdot n_{T+1}(a,b) \\
&= \sum_{(a,b)\in[A]\times[B]} (\mu_{1,1} - \mu_{a,1} + \mu_{a,1} - \mu_{a,b}) \cdot n_{T+1}(a,b) \\
&= \sum_{a\in[A]} (\mu_{1,1} - \mu_{a,1}) \cdot \sum_{b\in[B]} n_{T+1}(a,b) + \sum_{a\in[A]}\sum_{b\in[B]} (\mu_{a,1} - \mu_{a,b}) \cdot n_{T+1}(a,b) \\
&\leq \sum_{a\in[A]} (\mu_{1,1} - \mu_{a,1}) \cdot n_{T+1}(a) + \sum_{a\in\mathcal{A}^\circ}\sum_{b\in\mathcal{B}_a^\times} \mathcal{O}\left(\frac{\log(T/\delta)}{\mu_{a,1} - \mu_{a,b}}\right) \\
&\qquad + \sum_{a\in\mathcal{A}^\times} \mathcal{O}\left(\sqrt{Bn_{T+1}(a)\log(T/\delta)}\right), && \text{(by Assumption 2)} \\
&\leq \sum_{a\in\mathcal{A}^\times} \mathcal{O}\left(\frac{B\log(T/\delta)}{\mu_{1,1} - \mu_{a,1}}\right) + \sum_{a\in\mathcal{A}^\circ}\sum_{b\in\mathcal{B}_a^\times} \mathcal{O}\left(\frac{\log(T/\delta)}{\mu_{a,1} - \mu_{a,b}}\right). \\
& && \text{(because } n_{T+1}(a) = \mathcal{O}\left(\tfrac{B\log(T/\delta)}{(\mu_{1,1}-\mu_{a,1})^2}\right) \text{ for } a \in \mathcal{A}^\times\text{)}
\end{aligned}$$

■

Proof of Lemma 4 First, we use induction to prove the following inequalities:

$$0 \leq Q^2_{t,h}(s,a,b) - Q_{*,h}(s,a,b) \leq \mathbb{E}_{s'\sim P(\cdot|s,a,b)}\left[V^2_{t,h+1}(s') - V_{*,h+1}(s')\right] + 2\mathsf{bns}^2_\tau, \tag{4}$$

where $\tau = n_{t,h}(s,a,b)$, for all s,a,b. The order of induction is from $t=1$ to $t=T$, and (within each t) from $h=H$ to $h=1$.

For $t=1$, we have $Q^2_{1,h}(s,a,b) - Q_{*,h}(s,a,b) = H - Q_{*,h}(s,a,b) \geq 0$ and that $Q^2_{1,h}(s,a,b) - Q_{*,h}(s,a,b) \leq H \leq 2\mathsf{bns}^2_0$. Suppose that the inequality holds for all (t', h') with either $t' < t$, or $t' = t$ and $h' > h$. Fix a (s,a,b) and let $\tau = n_{t,h}(s,a,b)$. By the update rule of $Q^2_{t,h}(s,a,b)$, we have $Q^2_{t,h}(s,a,b) = \min\left\{\hat{Q}^2_{t,h}(s,a,b), Q^2_{t-1,h}(s,a,b)\right\}$ where

$$\hat{Q}^2_{t,h}(s,a,b) = \hat{R}_{t,h}(s,a,b) + \mathbb{E}_{s'\sim\hat{P}_{t,h}(\cdot|s,a,b)}\left[V^2_{t,h+1}(s')\right] + \mathsf{bns}^2_\tau. \qquad (\tau = n_{t,h}(s,a,b))$$

Besides,

$$Q_{*,h}(s,a,b) = R(s,a,b) + \mathbb{E}_{s'\sim P(\cdot|s,a,b)}\left[V_{*,h}(s')\right].$$

Taking their difference, we get

$$\begin{aligned}
&\hat{Q}^2_{t,h}(s,a,b) - Q_{*,h}(s,a,b) \\
&= \left(\hat{R}_{t,h}(s,a,b) - R(s,a,b)\right) + \underbrace{\mathbb{E}_{s'\sim P(\cdot|s,a,b)}\left[V^2_{t,h+1}(s') - V_{*,h+1}(s')\right]}_{\textbf{term}_1} \\
&\quad + \underbrace{\left(\mathbb{E}_{s'\sim\hat{P}_{t,h}(\cdot|s,a,b)}\left[V^2_{t,h+1}(s')\right] - \mathbb{E}_{s'\sim P(\cdot|s,a,b)}\left[V^2_{t,h+1}(s')\right]\right)}_{\textbf{term}_2} + \mathsf{bns}^2_\tau. \qquad (5)
\end{aligned}$$

By Lemma 10 and Lemma 11, for some universal constant $c > 0$,

$$\left|\hat{R}_{t,h}(s,a,b) - R(s,a,b)\right| \leq \frac{1}{2}c\sqrt{\frac{\log(T/\delta)}{\tau^+}}, \qquad (6)$$

$$|\textbf{term}_2| \leq \left\|\hat{P}_{t,h}(\cdot|s,a,b) - P(\cdot|s,a,b)\right\|_1 \|V^2_{t,h+1}\|_\infty \leq \frac{1}{2}cH\sqrt{\frac{S\log(T/\delta)}{\tau^+}}, \qquad (7)$$

and therefore $\left|\hat{R}_{t,h}(s,a,b) - R(s,a,b)\right| + |\textbf{term}_2| \leq \mathsf{bns}^2_\tau$. Combining this with (5), we get

$$\textbf{term}_1 \leq \hat{Q}^2_{t,h}(s,a,b) - Q_{*,h}(s,a,b) \leq \textbf{term}_1 + 2\mathsf{bns}^2_\tau. \qquad (8)$$

Using $Q^2_{t,h}(s,a,b) \leq \hat{Q}^2_{t,h}(s,a,b)$, (8) implies the right inequality in (4).

To prove the left inequality in (4), notice that if $h = H$, then $\textbf{term}_1 = 0$; if $h < H$,

$$\begin{aligned}
\textbf{term}_1 &\geq \min_{s'} V^2_{t,h+1}(s') - V_{*,h+1}(s') \\
&= \min_{s'}\left(\min\left\{\max_{a',b'} Q^2_{t,h+1}(s',a',b'), H\right\} - \max_{a',b'} Q_{*,h+1}(s',a',b')\right) \\
&= \min_{s'}\left(\min\left\{\max_{a',b'} Q^2_{t,h+1}(s',a',b') - \max_{a',b'} Q_{*,h+1}(s',a',b'), H - \max_{a',b'} Q_{*,h+1}(s',a',b')\right\}\right) \\
&\geq 0.
\end{aligned}$$

where the last inequality is by the induction hypothesis.

Thus, $\mathbf{term}_1 \geq 0$. Together with (8), we get $\hat{Q}^2_{t,h}(s,a,b) - Q_{*,h}(s,a,b) \geq 0$. By the induction hypothesis, we also have $Q^2_{t-1,h}(s,a,b) \geq Q_{*,h}(s,a,b)$. Therefore, $Q^2_{t,h}(s,a,b) = \min\left\{\hat{Q}^2_{t,h}(s,a,b), Q^2_{t-1,h}(s,a,b)\right\} \geq Q_{*,h}(s,a,b)$, proving the left inequality in (4).

Based on (4), we can write

$$\begin{aligned}
&\sum_{t=1}^{T}\left(Q^2_{t,h}(s_{t,h},a_{t,h},b_{t,h}) - Q_{*,h}(s_{t,h},a_{t,h},b_{t,h})\right) \\
&\leq \sum_{t=1}^{T}\mathbb{E}_{s'\sim P(\cdot|s_{t,h},a_{t,h},b_{t,h})}\left[V^2_{t,h+1}(s') - V_{*,h+1}(s')\right] + \sum_{t=1}^{T} 2\mathsf{bns}^2_{\tau_{t,h}} \\
&\hspace{20em}\text{(define } \tau_{t,h} = n_{t,h}(s_{t,h},a_{t,h},b_{t,h})\text{)} \\
&\triangleq \sum_{t=1}^{T}\left(V^2_{t,h+1}(s_{t,h+1}) - V_{*,h+1}(s_{t,h+1}) + \varepsilon_{t,h}\right) + \sum_{t=1}^{T} 2\mathsf{bns}^2_{\tau_{t,h}} \quad \text{(define } \varepsilon_{t,h} \text{ to be the difference)}
\end{aligned}$$

Since $\varepsilon_{t,h}$ is zero-mean, by Lemma 10,

$$\sum_{t=1}^{T}\varepsilon_{t,h} = \mathcal{O}\left(H\sqrt{T\log(T/\delta)}\right).$$

Besides,

$$\sum_{t=1}^{T}\mathsf{bns}^2_{\tau_{t,h}} = \mathcal{O}\left(\sum_{t=1}^{T} H\sqrt{\frac{S\log(T/\delta)}{n_{t,h}(s_{t,h},a_{t,h},b_{t,h})^+}}\right) = \mathcal{O}\left(HS\sqrt{ABT\log(T/\delta)}\right).$$

Combining the three bounds above proves the second conclusion in the lemma. ■

Proof of Lemma 5 We use induction to show the desired inequality. Again, the order of induction is from $t=1$ to $t=T$, and (within each t) from $h=H$ to $h=1$. When $t=1$, $Q^1_{1,h}(s,a) = H = \max_b Q^2_{1,h}(s,a,b)$.

Suppose that the inequality holds for all (t',h') with $t' < t$, or $t' = t$ and $h' > h$. Let $\tau = n_{t,h}(s,a)$, and let $1 \leq t_1 < t_2 < \cdots < t_\tau < t$ be the episodes in which (s,a) is visited at step h. By the update rule of $Q^1_{t,h}(\cdot,\cdot)$, we have

$$\begin{aligned}
&Q^1_{t,h}(s,a) \\
&= \alpha^0_\tau H + \sum_{i=1}^{\tau}\alpha^i_\tau\left(r_{t_i,h} + V^1_{t_i,h+1}(s_{t_i,h+1}) + \mathsf{bns}^1_i\right) \\
&\hspace{8em}\text{(define } \alpha^i_\tau = \alpha_i\Pi_{j=i+1}^{\tau}(1-\alpha_j) \text{ for } 1\leq i\leq\tau \text{ and } \alpha^0_\tau = \Pi_{j=1}^{\tau}(1-\alpha_j)\text{)} \\
&\geq \alpha^0_\tau H + \sum_{i=1}^{\tau}\alpha^i_\tau\left(R(s,a,b_{t_i,h}) + \sum_{s'}P(\cdot|s,a,b_{t_i,h})V^1_{t_i,h+1}(s')\right) + \frac{1}{2}\mathsf{bns}^1_\tau \\
&\hspace{20em}\text{(see the explanation below indexed } \star\text{)} \\
&\geq \alpha^0_\tau H + \sum_{i=1}^{\tau}\alpha^i_\tau\left(R(s,a,b_{t_i,h}) + \sum_{s'}P(\cdot|s,a,b_{t_i,h})V^2_{t_i,h+1}(s')\right) + \frac{1}{2}\mathsf{bns}^1_\tau \\
&\hspace{22em}\text{(by the induction hypothesis)}
\end{aligned}$$

$$\geq \alpha_\tau^0 H + \sum_{i=1}^{\tau} \alpha_\tau^i \left(\hat{R}_{t_i,h}(s,a,b_{t_i,h}) + \sum_{s'} \hat{P}_{t_i,h}(\cdot|s,a,b_{t_i,h}) V_{t_i,h+1}^2(s') - \mathsf{bns}_{\xi_i}^2 \right) + \frac{1}{2}\mathsf{bns}_\tau^1$$

(define $\xi_i = n_{t_i,h}(s,a,b_{t_i,h})$ and use (6) and (7))

$$\geq \alpha_\tau^0 H + \sum_{i=1}^{\tau} \alpha_\tau^i Q_{t_i,h}^2(s,a,b_{t_i,h}) - 2\sum_{i=1}^{\tau} \alpha_\tau^i \mathsf{bns}_{\xi_i}^2 + \frac{1}{2}\mathsf{bns}_\tau^1 \quad \text{(by the definition of } Q_{t,h}^2(s,a,b))$$

$$\geq \alpha_\tau^0 H + \sum_{i=1}^{\tau} \alpha_\tau^i \max_b Q_{t_i,h}^2(s,a,b) \qquad \text{(see the explanation below indexed } \star\star)$$

$$\geq \max_b Q_{t,h}^2(s,a,b) \qquad \text{(because } Q_{t,h}^2 \text{ is non-increasing in } t \text{ and } \textstyle\sum_{i=0}^{\tau} \alpha_\tau^i = 1)$$

In the first inequality ($\star$), we use the fact that $\sum_{i=1}^{\tau} \alpha_\tau^i \mathsf{bns}_i^1 \geq \mathsf{bns}_\tau^1$ by the first item in Lemma 12, and that

$$\left| \sum_{i=1}^{\tau} \alpha_\tau^i \left(R(s,a,b_{t_i,h}) + \sum_{s'} P(s'|s,a,b_{t_i,h}) V_{t_i,h+1}^1(s') - r_{t_i,h} - V_{t_i,h+1}^1(s_{t_i,h+1}) \right) \right|$$

$$\leq \frac{1}{2} c' H \sqrt{\frac{HS\log(T/\delta)}{\tau^+}} \leq \frac{1}{2}\mathsf{bns}_\tau^1 \tag{9}$$

for some universal constant $c' > 0$ by Lemma 13.

In the penultimate inequality ($\star\star$), we first use the selection rule of $b_{t,h} = \operatorname{argmax}_b Q_{t,h}^2(s_{t,h}, a_{t,h}, b)$, and then use the following Lemma 8 to bound

$$\begin{aligned} 2\sum_{i=1}^{\tau} \alpha_\tau^i \mathsf{bns}_{\xi_i}^2 &= 2cH\sqrt{S\log(T/\delta)} \sum_{i=1}^{\tau} \alpha_\tau^i \frac{1}{\sqrt{n_{t_i,h}(s,a,b_{t_i,h})^+}} \\ &\leq 2cH\sqrt{S\log(T/\delta)} \times 4\sqrt{\frac{BH}{n_{t,h}(s,a)^+}} \\ &\leq \frac{1}{2}c'H\sqrt{\frac{HSB\log(T/\delta)}{n_{t,h}(s,a)^+}} = \frac{1}{2}\mathsf{bns}_\tau^1. \end{aligned}$$

■

Lemma 8 *Let $\{\tau_1, \ldots, \tau_B\}$ be non-negative integers such that $\sum_{b=1}^{B} \tau_b = \tau$. Define for all $b = 1, \ldots, B$:*

$$Y_{bi} = \frac{1}{\sqrt{(i-1)^+}}, \qquad \textit{for } i = 1, 2, \ldots, \tau_b.$$

Let $\{Z_1, Z_2, \ldots, Z_\tau\}$ be any permutation of

$$\{Y_{11}, Y_{12}, \ldots, Y_{1\tau_1}, Y_{21}, Y_{22}, \ldots, Y_{2\tau_2}, \ldots\ldots Y_{B1}, Y_{B2}, \ldots, Y_{B\tau_B}\}$$

Then

$$\sum_{i=1}^{\tau} \alpha_\tau^i Z_i \leq 4\sqrt{\frac{BH}{\tau^+}}.$$

Proof We write $i = \phi(b, j)$ if Y_{bj} is mapped to Z_i. Also, define $\Phi(b) = \{\phi(b, j) : j \in [\tau_b]\}$ as the set of indices in $\{Z_i\}$ that are mapped from $\{Y_{b1}, \dots, Y_{b\tau_b}\}$.

We first show the following claim: for all b,

$$\sum_{i\in\Phi(b)} \alpha_\tau^i Z_i \le 2\sqrt{2\alpha_\tau \sum_{i\in\Phi(b)} \alpha_\tau^i}. \tag{10}$$

To show (10), observe that the left-hand side is equal to

$$\sum_{i\in\Phi(b)} \alpha_\tau^i Z_i = \sum_{j=1}^{\tau_b} \alpha_\tau^{\phi(b,j)} Y_{bj} = \sum_{j=1}^{\tau_b} \alpha_\tau^{\phi(b,j)} \frac{1}{\sqrt{(j-1)^+}} \le \sum_{j=1}^{\tau_b} \alpha_\tau^{\phi(b,j)} \sqrt{\frac{2}{j}}. \tag{11}$$

By the definition of α_τ^i, we have $\alpha_\tau^i \le \alpha_\tau$ for any i. We see that the last expression in (11) is upper bounded by the optimal solution of the following programming:

$$\begin{aligned} \max_{\beta_j} & \sum_{j=1}^{\tau_b} \beta_j \sqrt{\frac{2}{j}} \\ \text{s.t.} & \sum_{j=1}^{\tau_b} \beta_j \le \sum_{i\in\Phi(b)} \alpha_\tau^i \\ & 0 \le \beta_j \le \alpha_\tau \ \ \forall j \end{aligned}$$

This programming exhibits a greedy solution that sets $\beta_j = \alpha_\tau$ for $j \le j^\star \triangleq \left\lfloor \frac{1}{\alpha_\tau} \sum_{i\in\Phi(b)} \alpha_\tau^i \right\rfloor$, $\beta_j = \sum_{i\in\Phi(b)} \alpha_\tau^i - \alpha_\tau j^\star$ for $j = j^\star + 1$, and $\beta_j = 0$ otherwise. The optimal value of this solution is upper bounded by

$$\alpha_\tau \sum_{j=1}^{j^\star} \sqrt{\frac{2}{j}} + \left(\sum_{i\in\Phi(b)} \alpha_\tau^i - \alpha_\tau j^\star \right) \sqrt{\frac{2}{j^\star + 1}} \le \alpha_\tau \int_0^{\frac{1}{\alpha_\tau}\sum_{i\in\Phi(b)}\alpha_\tau^i} \sqrt{\frac{2}{x}} \mathrm{d}x = 2\sqrt{2\alpha_\tau \sum_{i\in\Phi(b)} \alpha_\tau^i},$$

showing (10). To get the final bound, we sum this bound over b and use the definition of α_τ:

$$\begin{aligned} \sum_{i=1}^{\tau} \alpha_\tau^i Z_i &= \sum_{b=1}^{B} \sum_{i\in\Phi(b)} \alpha_\tau^i Z_i \le \sum_{b=1}^{B} 2\sqrt{2\alpha_\tau \sum_{i\in\Phi(b)} \alpha_\tau^i} \\ &\le 2\sqrt{2B\alpha_\tau \sum_{i=1}^{\tau} \alpha_\tau^i} = 2\sqrt{\frac{2B(H+1)}{H+\tau}} \le 4\sqrt{\frac{BH}{\tau^+}}, \end{aligned}$$

where in the second inequality we use the AM-GM inequality and in the last equality we use $\sum_{i=1}^{\tau} \alpha_\tau^i = 1$ for $\tau \ge 1$. ■

Proof of Lemma 6

Fix t, h, s, a. Let $\tau = n_{t,h}(s, a)$, and let $1 \le t_1 < t_2 < \cdots < t_\tau < t$ be the episodes in which (s, a) is visited at layer h. By the update rule of $Q^1_{t,h}(\cdot, \cdot)$, we have

$$
\begin{aligned}
&Q^1_{t,h}(s,a)\\
&= \alpha^0_\tau H + \sum_{i=1}^{\tau} \alpha^i_\tau \left(r_{t_i,h} + V^1_{t_i,h+1}(s_{t_i,h+1}) + \mathsf{bns}^1_i \right)\\
&\le \alpha^0_\tau H + \sum_{i=1}^{\tau} \alpha^i_\tau \left(R(s,a,b_{t_i,h}) + \mathbb{E}_{s'\sim P(\cdot|s,a,b_{t_i,h})}\left[V^1_{t_i,h+1}(s')\right] \right) + \mathcal{O}(\mathsf{bns}^1_\tau)\\
&\qquad\qquad \text{(by Lemma 13 and that } \textstyle\sum_{i=1}^{\tau} \alpha^i_\tau \mathsf{bns}^1_i \le 2\mathsf{bns}^1_\tau \text{ by the first item in Lemma 12)}\\
&= \alpha^0_\tau H + \sum_{i=1}^{\tau} \alpha^i_\tau \left(R(s,a,b_{t_i,h}) + \mathbb{E}_{s'\sim P(\cdot|s,a,b_{t_i,h})}\left[V_{*,h+1}(s')\right] \right)\\
&\quad + \sum_{i=1}^{\tau} \alpha^i_\tau \left(\mathbb{E}_{s'\sim P(\cdot|s,a,b_{t_i,h})}\left[V^1_{t_i,h+1}(s') - V_{*,h+1}(s')\right] \right) + \mathcal{O}(\mathsf{bns}^1_\tau)\\
&= \alpha^0_\tau H + \sum_{i=1}^{\tau} \alpha^i_\tau Q_{*,h}(s,a,b_{t_i,h}) + \sum_{i=1}^{\tau} \alpha^i_\tau \left(\mathbb{E}_{s'\sim P(\cdot|s,a,b_{t_i,h})}\left[V^1_{t_i,h+1}(s') - V_{*,h+1}(s')\right] \right) + \mathcal{O}(\mathsf{bns}^1_\tau)\\
&\le \alpha^0_\tau H + \sum_{i=1}^{\tau} \alpha^i_\tau Q_{*,h}(s,a,b_{t_i,h}) + \sum_{i=1}^{\tau} \alpha^i_\tau \left(V^1_{t_i,h+1}(s_{t_i,h+1}) - V_{*,h+1}(s_{t_i,h+1}) \right) + \mathcal{O}(\mathsf{bns}^1_\tau)\\
&\qquad\qquad \text{(by Lemma 13)}
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
&Q^1_{t,h}(s,a) - Q_{*,h}(s,a)\\
&= \alpha^0_\tau (H - Q_{*,h}(s,a)) + \sum_{i=1}^{\tau} \alpha^i_\tau \left(Q_{*,h}(s,a,b_{t_i,h}) - Q_{*,h}(s,a) \right)\\
&\quad + \sum_{i=1}^{\tau} \alpha^i_\tau \left(V^1_{t_i,h+1}(s_{t_i,h+1}) - V_{*,h+1}(s_{t_i,h+1}) \right) + \mathcal{O}(\mathsf{bns}^1_\tau)\\
&\le \alpha^0_\tau H + \sum_{i=1}^{\tau} \alpha^i_\tau \left(V^1_{t_i,h+1}(s_{t_i,h+1}) - V_{*,h+1}(s_{t_i,h+1}) \right) + \mathcal{O}(\mathsf{bns}^1_\tau) \qquad (12)
\end{aligned}
$$

Now consider the cumulative sum, and define $t_i(s,a)$ to be the index of the episode when it is the i-th time (s,a) is visited at layer h.

$$
\begin{aligned}
&\sum_{t=1}^{T} \left(Q^1_{t,h}(s_{t,h},a_{t,h}) - Q_{*,h}(s_{t,h},a_{t,h}) \right)\\
&= \sum_{s,a} \sum_{i=1}^{n_{T+1}(s,a)} \left(Q^1_{t_i(s,a),h}(s_{t_i(s,a),h}, a_{t_i(s,a),h}) - Q_{*,h}(s_{t_i(s,a),h}, a_{t_i(s,a),h}) \right)\\
&\le \sum_{s,a} \sum_{i=1}^{n_{T+1}(s,a)} \left(\alpha^0_i H + \sum_{j=1}^{i-1} \alpha^j_i \left(V^1_{t_j(s,a),h+1}(s_{t_j(s,a),h+1}) - V_{*,h+1}(s_{t_j(s,a),h+1}) \right) + \mathsf{bns}^1_{i-1} \right)\\
&\qquad\qquad \text{(by (12))}
\end{aligned}
$$

$$
\begin{aligned}
&= \sum_{s,a} \sum_{i=1}^{n_{T+1}(s,a)} \alpha_i^0 H + \sum_{s,a} \sum_{j=1}^{n_{T+1}(s,a)-1} \sum_{i=j+1}^{n_{T+1}(s,a)} \alpha_i^j \left(V^1_{t_j(s,a),h+1}(s_{t_j(s,a),h+1}) - V_{*,h+1}(s_{t_j(s,a),h+1}) \right) \\
&\qquad + \sum_{s,a} \sum_{i=1}^{n_{T+1}(s,a)} \mathcal{O}\left(\sqrt{\frac{H^3 SB \log(T/\delta)}{\max\{i-1,1\}}} \right) \\
&\le HSA + \sum_{s,a} \sum_{j=1}^{n_{T+1}(s,a)-1} \left(1 + \frac{1}{H}\right) \left(V^1_{t_j(s,a),h+1}(s_{t_j(s,a),h+1}) - V_{*,h+1}(s_{t_j(s,a),h+1}) \right) \\
&\qquad + \mathcal{O}\left(\sqrt{H^3 S^2 ABT} \right) \qquad \text{(by the third item of Lemma 12)} \\
&\le \left(1 + \frac{1}{H}\right) \sum_{t=1}^{T} \left(V^1_{t,h+1}(s_{t,h+1}) - V^1_{*,h+1}(s_{t,h+1}) \right) + \mathcal{O}\left(\sqrt{H^3 S^2 ABT} + HSA \right).
\end{aligned}
$$

■

Proof of Corollary 3 By Lemma 6 and the fact that $V^1_{t,h}(s_{t,h}) = Q^1_{t,h}(s_{t,h}, a_{t,h})$, we have

$$
\begin{aligned}
&\sum_{t=1}^{T} \left(Q^1_{t,h}(s_{t,h}, a_{t,h}) - Q_{*,h}(s_{t,h}, a_{t,h}) \right) \\
&\le \left(1 + \frac{1}{H}\right) \sum_{t=1}^{T} \left(Q^1_{t,h+1}(s_{t,h+1}, a_{t,h+1}) - Q_{*,h+1}(s_{t,h+1}, a_{t,h+1}) \right) + \mathcal{O}\left(\sqrt{H^3 S^2 ABT} + HSA \right),
\end{aligned}
$$

which gives

$$
\begin{aligned}
\sum_{t=1}^{T} \left(Q^1_{t,h}(s_{t,h}, a_{t,h}) - Q_{*,h}(s_{t,h}, a_{t,h}) \right) &\le H \times \left(1 + \frac{1}{H}\right)^H \times \mathcal{O}\left(\sqrt{H^3 S^2 ABT} + HSA \right) \\
&= \mathcal{O}\left(\sqrt{H^5 S^2 ABT} + H^2 SA \right)
\end{aligned}
$$

by expanding the recursion. ■

Appendix C. Multiple Follower Extension

In this section, we consider the multiple follower case. Let A be the number of actions of the leader, and let B^i be the number of actions of the i-th follower, for $i = 1, 2, \ldots, N$. We define $B = \frac{1}{N} \sum_{i=1}^{N} B^i$ to be the average number of actions of all followers. In each round t, the leader first selects an action $a_t \in [A]$. Then based on the information of a_t, each follower i selects an action $b_t^i \in [B^i]$. The reward that the i-th follower receives is r_t^i, whose mean is $\mu^i_{a_t, b_t^i}$; the reward the leader receives is $r_t = \frac{1}{N} \sum_{i=1}^{N} r_t^i$, the average-reward over i. [7]

Our algorithm is presented in Algorithm 3. On the follower side, the algorithm is identical to the single-agent case (Algorithm 1) – for each arm of the leader, a simple empirical mean is maintained

[7] Similar to the single-agent case, our framework also handles the case where the leader observes another fresh sample with mean $\frac{1}{N} \sum_{i=1}^{N} \mu^i_{a_t, b_t^i}$.

Algorithm 3 UCB for Hierarchical Bandits with Multiple Followers

```
define: $c > 0$ is a universal constant.
Followers start running algorithms satisfying Assumption 4 with some $\kappa \geq 1$ (we denote the instance of the algorithm by Follower $i$ under action $a \in [A]$ as $\mathsf{ALG}^i(a)$).
for $t = 1, 2, \ldots, T$ do
    Leader chooses $a_t \in \operatorname{argmax}_{a\in[A]} \hat{\mu}_t(a) + c\sqrt{\frac{\log(T/\delta)}{n_t(a)^+}} + \sqrt{\frac{\kappa B \log(T/\delta)}{n_t(a)^+}}$. ($\hat{\mu}_t(a)$, $n_t(a)$ defined in (1))
    for $i = 1, \ldots, N$ (in parallel) do
        Follower $i$ observes $a_t$ and calls $\mathsf{ALG}^i(a_t)$, which outputs an action $b_t^i$.
        Follower $i$ chooses $b_t^i$.
        Follower $i$ observes $r_t^i$ with $\mathbb{E}[r_t^i] = \mu^i_{a_t,b_t^i}$.
    end
    Leader observes $r_t = \frac{1}{N}\sum_{i=1}^N r_t^i$.
end
```

as in (1). For the followers, similarly, we assume that all of them use a no-regret algorithm that satisfies the following assumption:

Assumption 4 *Every Follower i guarantees the following for some universal constant $\kappa \geq 1$ with probability at least $1-\delta$:*

$$\forall t, a, \qquad \sum_{\tau=1}^{t} \mathbb{I}[a_\tau = a]\left(\max_{b^i} \mu^i_{a,b^i} - \mu^i_{a,b^i_\tau}\right) \leq \sqrt{\kappa B^i \sum_{\tau=1}^{t} \mathbb{I}[a_\tau = a]\log(T/\delta)}.$$

With Assumption 4, the regret bound of Algorithm 3 can be shown as in the following theorem:

Theorem 9 *With probability at least $1 - \mathcal{O}(\delta)$, for all a and $\{b^i\}_{i=1}^N$,*

$$\sum_{t=1}^{T}\left(\frac{1}{N}\sum_{i=1}^{N}\mu^i_{a,b^i} - \frac{1}{N}\sum_{i=1}^{N}\mu^i_{a_t,b^i_t}\right) = \mathcal{O}\left(\sqrt{ABT}\log(T/\delta)\right).$$

Proof

$$\begin{aligned}
&\sum_{t=1}^{T}\left(\frac{1}{N}\sum_{i=1}^{N}\mu^i_{a,b^i} - \frac{1}{N}\sum_{i=1}^{N}\mu^i_{a_t,b^i_t}\right)\\
&\leq \sum_{t=1}^{T}\left(\frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{n_t(a)^+}\times\sum_{\tau=1}^{t-1}\mu^i_{a,b^i_\tau} + \sqrt{\frac{\kappa B^i\log(T/\delta)}{n_t(a)^+}}\right) - \frac{1}{N}\sum_{i=1}^{N}\mu^i_{a_t,b^i_t}\right) && \text{(Assumption 4)}\\
&\leq \sum_{t=1}^{T}\left(\frac{1}{n_t(a)^+}\left(\frac{1}{N}\sum_{i=1}^{N}\sum_{\tau=1}^{t-1}\mu^i_{a,b^i_\tau}\right) + \sqrt{\frac{\kappa B\log(T/\delta)}{n_t(a)^+}} - \frac{1}{N}\sum_{i=1}^{N}\mu^i_{a_t,b^i_t}\right)\\
& && \text{(Cauchy-Schwarz inequality)}\\
&\leq \sum_{t=1}^{T}\left(\hat{\mu}_t(a) + c\sqrt{\frac{\log(T/\delta)}{n_t(a)^+}} + \sqrt{\frac{\kappa B\log(T/\delta)}{n_t(a)^+}} - \frac{1}{N}\sum_{i=1}^{N}\mu^i_{a_t,b^i_t}\right) && \text{(by Lemma 10)}
\end{aligned}$$

$$
\begin{aligned}
&\leq \sum_{t=1}^{T}\left(\hat{\mu}_t(a_t)+c\sqrt{\frac{\log(T/\delta)}{n_t(a_t)^+}}+\sqrt{\frac{\kappa B\log(T/\delta)}{n_t(a_t)^+}}-\frac{1}{N}\sum_{i=1}^{N}\mu^i_{a_t,b^i_t}\right) \qquad \text{(by the algorithm)}\\
&=\sum_{t=1}^{T}\left(\hat{\mu}_t(a_t)-\frac{1}{N}\sum_{i=1}^{N}\mu^i_{a_t,b^i_t}\right)+\mathcal{O}\left(\sqrt{ABT\log(T/\delta)}\right)\\
&=\underbrace{\sum_{a\in[A]}\sum_{t=1}^{T}\mathbb{I}[a_t=a]\left(\hat{\mu}_t(a)-\max_{\{b^i\}}\frac{1}{N}\sum_{i=1}^{N}\mu^i_{a,b^i}\right)}_{\textbf{term}_1}\\
&\qquad+\underbrace{\sum_{a\in[A]}\sum_{t=1}^{T}\mathbb{I}[a_t=a]\left(\max_{\{b^i\}}\frac{1}{N}\sum_{i=1}^{N}\left(\mu^i_{a,b^i}-\mu^i_{a,b^i_t}\right)\right)}_{\textbf{term}_2}+\mathcal{O}\left(\sqrt{ABT\log(T/\delta)}\right)
\end{aligned}
$$

Under Assumption 4, we can upper bound **term**$_2$ by

$$
\begin{aligned}
\mathcal{O}\left(\sum_{a\in[A]}\frac{1}{N}\sum_{i=1}^{N}\sqrt{B^i n_{T+1}(a)\log(T/\delta)}\right)&=\mathcal{O}\left(\sum_{a\in[A]}\sqrt{Bn_{T+1}(a)\log(T/\delta)}\right)\\
&\qquad\qquad\qquad\qquad\text{(Cauchy-Schwarz)}\\
&=\mathcal{O}\left(\sqrt{ABT\log(T/\delta)}\right).
\end{aligned}
$$

Besides, for all t and a, we have with probability at least $1-\mathcal{O}(\delta)$,

$$
\begin{aligned}
\hat{\mu}_t(a)&=\frac{1}{n_t(a)^+}\sum_{\tau=1}^{t-1}\mathbb{I}[a_\tau=a]r_\tau\\
&\leq\frac{1}{n_t(a)^+}\sum_{\tau=1}^{t-1}\mathbb{I}[a_\tau=a]\left(\frac{1}{N}\sum_{i=1}^{N}\mu_{a,b^i_\tau}\right)+\mathcal{O}\left(\sqrt{\frac{\log(T/\delta)}{n_t(a)^+}}\right) \qquad \text{(Azuma's inequality)}\\
&\leq\frac{1}{N}\max_{\{b^i\}}\sum_{i=1}^{N}\mu_{a,b^i}+\mathcal{O}\left(\sqrt{\frac{\log(T/\delta)}{n_t(a)^+}}\right).
\end{aligned}
$$

Therefore,

$$
\textbf{term}_1\leq\mathcal{O}\left(\sum_{a\in[A]}\sum_{t=1}^{T}\mathbb{I}[a_t=a]\sqrt{\frac{\log(T/\delta)}{n_t(a)^+}}\right)=\mathcal{O}\left(\sqrt{AT\log(T/\delta)}\right).
$$

Combining everything finishes the proof. ■

Appendix D. Deep Hierarchical Bandits Extension

In this subsection, we consider the deep hierarchical bandit setting. In this setting, there are D agents making decisions in a fixed order: in each round t, Agent 1 first chooses an action $a^1_t\in[A]$.

Algorithm 4 UCB for Deep Hierarchical Bandits

```
define: $C_D \geq 2$ and $C_d = 6C_{d+1} + 8$ for $d = D-1, \ldots, 1$.
for $t = 1, \ldots, T$ do
    for $d = 1, \ldots, D$ do
        Agent $d$ chooses $a_t^d \in \arg\max_{a \in [A]} \hat{\mu}_t^d(a_t^1, \ldots, a_t^{d-1}, a) + C_d \sqrt{\frac{A^{D-d}\log(A^D T/\delta)}{n_t^d(a_t^1,\ldots,a_t^{d-1},a)^+}}$,
    end
end
```

For $d = 2, \ldots, D$, after receiving $(a_t^1, a_t^2, \ldots, a_t^{d-1})$, Agent d chooses an action $a_t^d \in [A]$.[8] After all agents all choose an action, a reward $r_t \in [0,1]$ is generated based on the joint action $\mathbf{a}_t \triangleq (a_t^1, \ldots, a_t^D)$ with its mean equal to $\mu_{\mathbf{a}_t}$. As before, we assume that the first action is the best action in all layers, and therefore the goal of the agents is to have sub-linear regret with respect to the joint action $\mathbf{1} = (1, \ldots, 1)$.

We propose Algorithm 4 to solve this problem, which is based on the similar idea as Algorithm 3. At time t, Agent d maintains the number of times an arm $\mathbf{a}^{1:d} = (a^1, \ldots, a^d) \in [A]^d$ has been visited:

$$n_t^d(\mathbf{a}^{1:d}) = \sum_{s=1}^{t-1} \mathbb{I}\{\mathbf{a}_s^{1:d} = \mathbf{a}^{1:d}\} \tag{13}$$

and the empirical mean of the same arm

$$\hat{\mu}_t^d(\mathbf{a}^{1:d}) = \frac{1}{n_t^d(\mathbf{a}^{1:d})^+} \sum_{s=1}^{t-1} r_s \mathbb{I}\{\mathbf{a}_s^{1:d} = \mathbf{a}^{1:d}\}. \tag{14}$$

The bonus term of Agent d for the arm $\mathbf{a}^{1:d}$ is of order $\sqrt{\frac{A^{D-d}\log(A^D T/\delta)}{n_t^d(\mathbf{a}^{1:d})^+}}$, which is again the average regret upper bound of its direct subordinate. As in Section 3, such a design is to ensure that agent d's optimistic value is not smaller than the value of the optimal arm (with high probability).

In the following, we show the regret bounds of Algorithm 4. Lemma 5 and Theorem 6 are the gap-independent results, where Lemma 5 is a generalization of Assumption 1 and Theorem 6 is a generalization of Theorem 1. Lemma 7 and Theorem 8 are the gap-dependent results, where Theorem 8 is a generalization of Theorem 2. In the bounds, the regret grows exponentially in the depth D; however, this comes from the high model complexity and is unavoidable.

Lemma 5 *For any $t \in [T]$, $d \in [D-1]$, $\mathbf{a}^{1:d} = (a^1, \ldots, a^d) \in [A]^d$,*

$$\sum_{s=1}^{t} \mathbb{I}\{\mathbf{a}_s^{1:d} = \mathbf{a}^{1:d}\} \left[\mu_{(\mathbf{a}^{1:d}, \mathbf{1}_{D-d})} - r_s\right] \leq C_d \sqrt{A^{D-d} n_{t+1}^d(\mathbf{a}^{1:d}) \log(A^D T/\delta)} \tag{15}$$

Furthermore,

$$\sum_{s=1}^{t} (\mu_{\mathbf{1}} - r_s) \leq (6C_1 + 8)\sqrt{A^D t \log(A^D T/\delta)}. \tag{16}$$

[8] For simplicity, we assume that the number of actions on all layers are the same and equal to A.

Proof of Lemma 5

Notice that (16) can be viewed as a case of $d = 0$ in (15) by defining $\mathbb{I}\{\mathbf{a}_s^{1:0} = \mathbf{a}^{1:0}\} = 1$ and $n_{t+1}(\mathbf{a}^{1:d}) = t$. Therefore, below we prove by induction from $d = D - 1$ to $d = 0$.

Base $d = D - 1$:

$$\begin{aligned}&\sum_{s=1}^{t} \mathbb{I}\{\mathbf{a}_s^{1:D-1} = a^{1:D-1}\} \left[\mu_{(\mathbf{a}^{1:D-1},1)} - r_s\right] \\ &\leq C_{D-1}\sqrt{An_{t+1}^{D-1}(\mathbf{a}^{1:D-1}) \log(An_{t+1}^{D-1}(\mathbf{a}^{1:D-1})/\delta)} \leq C_{D-1}\sqrt{An_{t+1}^{D-1}(\mathbf{a}^{1:D-1}) \log(AT/\delta)}\end{aligned}$$

by the standard analysis of the UCB algorithm.

Induction step: assume the statement is true for $d + 1$, we show that it holds for d as well. Dividing both sides of the inequality in the induction hypothesis by $n_{t+1}^{d+1}(\mathbf{a}^{1:d+1})^+$ to get

$$\mu_{(\mathbf{a}^{1:d+1},\mathbf{1}_{D-d-1})} \leq \hat{\mu}_{t+1}^{d+1}(\mathbf{a}^{1:d+1}) + C_{d+1}\sqrt{\frac{A^{D-d-1}\log(A^D T/\delta)}{n_{t+1}^{d+1}(\mathbf{a}^{1:d+1})^+}}. \tag{17}$$

Now for the left hand side of the inequality for d, we have

$$\begin{aligned}&\sum_{s=1}^{t} \mathbb{I}\{\mathbf{a}_s^{1:d} = \mathbf{a}^{1:d}\} \left[\mu_{(\mathbf{a}^{1:d},\mathbf{1}_{D-d})} - r_s\right] \\ &\leq \sum_{s=1}^{t} \mathbb{I}\{\mathbf{a}_s^{1:d} = \mathbf{a}^{1:d}\} \left\{\max_{a^{d+1}} \left[\hat{\mu}_{t+1}^{d+1}(\mathbf{a}_s^{1:d}, a^{d+1}) + C_{d+1}\sqrt{\frac{A^{D-d-1}\log(A^D T/\delta)}{n_{t+1}^{d+1}(\mathbf{a}_s^{1:d}, a^{d+1})^+}}\right] - r_s\right\} \\ &\leq \sum_{s=1}^{t} \mathbb{I}\{\mathbf{a}_s^{1:d} = \mathbf{a}^{1:d}\} \Bigg[\hat{\mu}_{t+1}^{d+1}(\mathbf{a}_s^{1:d+1}) + C_{d+1}\sqrt{\frac{A^{D-d-1}\log(A^D T/\delta)}{n_{t+1}^{d+1}(\mathbf{a}_s^{1:d+1})^+}} - \mu_{(\mathbf{a}_s^{1:d+1},\mathbf{1}_{D-d-1})} \\ &\qquad + \mu_{(\mathbf{a}_s^{1:d+1},\mathbf{1}_{D-d-1})} - r_s\Bigg],\end{aligned} \tag{18}$$

where the first inequality follows from (17) and taking max, and the second inequality follows from the way a_s^{d+1} is selected.

For any $s \in [T]$, $\mathbf{a}^{1:d+1} \in [A]^{d+1}$, if $n_s^{d+1}(\mathbf{a}^{1:d+1}) \geq 1$, then we have

$$\begin{aligned}&\hat{\mu}_s^{d+1}(\mathbf{a}^{1:d+1}) - \mu_{(\mathbf{a}^{1:d+1},\mathbf{1}_{D-d-1})} \\ &\leq \frac{1}{n_s^{d+1}(\mathbf{a}^{1:d+1})^+} \sum_{u=1}^{s-1} \mathbb{I}\{\mathbf{a}_u^{1:d+1} = \mathbf{a}^{1:d+1}\} \left[r_u - \mu_{(\mathbf{a}^{1:d+1},\mathbf{1}_{D-d-1})}\right] \\ &\leq \frac{1}{n_s^{d+1}(\mathbf{a}^{1:d+1})^+} \sum_{u=1}^{s-1} \mathbb{I}\{\mathbf{a}_u^{1:d+1} = \mathbf{a}^{1:d+1}\} \left[r_u - \mu_{\mathbf{a}_u}\right] \\ &\leq \frac{2}{n_s^{d+1}(\mathbf{a}^{1:d+1})^+} \sqrt{\sum_{u=1}^{s-1} \mathbb{I}\{\mathbf{a}_u^{1:d+1} = \mathbf{a}^{1:d+1}\} \log(1/\delta)} \leq 2\sqrt{\frac{\log(1/\delta)}{n_s^{d+1}(\mathbf{a}^{1:d+1})^+}}\end{aligned}$$

with probability $1-\delta$. This also trivially holds if $n_s^{d+1}(\mathbf{a}^{1:d+1})=0$. Then, the inequality

$$\hat{\mu}_s^{d+1}(\mathbf{a}^{1:d+1})-\mu_{(\mathbf{a}^{1:d+1},\mathbf{1}_{D-d-1})}\le 2\sqrt{\frac{\log(A^{d+1}T/\delta)}{n_s^{d+1}(\mathbf{a}^{1:d+1})^+}} \tag{19}$$

holds for all $s\in[T]$ and all $\mathbf{a}^{1:d+1}\in[A]^{d+1}$ simultaneously by the union bound. Hence, the sum of the first three terms of (18) can be bounded by

$$\begin{aligned}
&\sum_{s=1}^{t}\mathbb{I}\{\mathbf{a}_s^{1:d}=\mathbf{a}^{1:d}\}\left[\hat{\mu}_t^{d+1}(\mathbf{a}_s^{1:d+1})+C_{d+1}\sqrt{\frac{A^{D-d-1}\log(A^DT/\delta)}{n_{t+1}^{d+1}(\mathbf{a}_s^{1:d+1})^+}}-\mu_{(\mathbf{a}_s^{1:d+1},\mathbf{1}_{D-d-1})}\right]\\
&\le\sum_{s=1}^{t}\mathbb{I}\{\mathbf{a}_s^{1:d}=\mathbf{a}^{1:d}\}\left[2\sqrt{\frac{\log(A^{d+1}T/\delta)}{n_{t+1}^{d+1}(\mathbf{a}_s^{1:d+1})^+}}+C_{d+1}\sqrt{\frac{A^{D-d-1}\log(A^DT/\delta)}{n_{t+1}^{d+1}(\mathbf{a}_s^{1:d+1})^+}}\right]\\
&\le(C_{d+1}+2)\sqrt{A^{D-d-1}\log(A^DT/\delta)}\cdot\sum_{s=1}^{t}\mathbb{I}\{\mathbf{a}_s^{1:d}=\mathbf{a}^{1:d}\}\sqrt{\frac{1}{n_s^{d+1}(\mathbf{a}_s^{1:d+1})^+}}\\
&\le 2(C_{d+1}+2)\sqrt{A^{D-d}n_{t+1}^{d}(\mathbf{a}^{1:d})\log(A^DT/\delta)}.
\end{aligned}$$

On the other hand, the last two terms sum up to

$$\begin{aligned}
&\sum_{s=1}^{t}\mathbb{I}\{\mathbf{a}_s^{1:d}=\mathbf{a}^{1:d}\}\left[\mu_{(\mathbf{a}_s^{1:d+1},\mathbf{1}_{D-d-1})}-r_s\right]\\
&=\sum_{a^{d+1}\in[A]}\sum_{s=1}^{t}\mathbb{I}\{\mathbf{a}_s^{1:d+1}=\mathbf{a}^{1:d+1}\}\left[\mu_{(\mathbf{a}_s^{1:d+1},\mathbf{1}_{D-d-1})}-r_s\right]\\
&\le\sum_{a^{d+1}\in[A]}C_{d+1}\sqrt{A^{D-d-1}n_{t+1}^{d+1}(\mathbf{a}^{1:d})\log(A^DT/\delta)}\\
&\le C_{d+1}\sqrt{A^{D-d-1}\cdot\left[\sum_{a^{d+1}\in[A]}n_{t+1}^{d+1}(\mathbf{a}^{1:d})\right]\log(A^DT/\delta)}\\
&\le C_{d+1}\sqrt{A^{D-d}n_{t+1}^{d}(\mathbf{a}^{1:d})\log(A^DT/\delta)}.
\end{aligned}$$

Combining both parts, with probability $1-3\delta:=1-\delta'$, we have

$$\begin{aligned}
&\sum_{s=1}^{t}\mathbb{I}\{\mathbf{a}_s^{1:d}=\mathbf{a}^{1:d}\}\left[\mu_{(\mathbf{a}^{1:d},\mathbf{1}_{D-d})}-r_s\right]\\
&\le[2(C_{d+1}+2)+C_{d+1}]\sqrt{A^{D-d}n_{t+1}^{d}(\mathbf{a}^{1:d})\log(A^DT/\delta)}\\
&\le\sqrt{3}(3C_{d+1}+4)\sqrt{A^{D-d}n_{t+1}^{d}(\mathbf{a}^{1:d})\log(A^DT/\delta')}\\
&\le(6C_{d+1}+8)\sqrt{A^{D-d}n_{t+1}^{d}(\mathbf{a}^{1:d})\log(A^DT/\delta')}\\
&\le C_d\sqrt{A^{D-d}n_{t+1}^{d}(\mathbf{a}^{1:d})\log(A^DT/\delta')}
\end{aligned}$$

whenever $C_d \geq 6C_{d+1} + 8$.

Notice that when $d = 0$, the same arguments follow except that $\mathbb{I}\{\mathbf{a}_s^{1:d} = \mathbf{a}^{1:d}\}$ degenerates to 1, $n_{t+1}^d(\mathbf{a}^{1:d})$ degenerates to t, and the inequality will end with $(6C_1 + 8)\sqrt{A^D t \log(A^D T/\delta')}$. ■

Theorem 6 *For the deep hierarchical bandit problem, with probability of at least* $1-\delta$, *Algorithm 4 achieves the regret bound of*

$$\sum_{t=1}^{T} (\mu_{\mathbf{1}_D} - \mu_{\mathbf{a}_t}) \leq O\left(\sqrt{A^D DT \log(AT/\delta)}\right). \tag{20}$$

Proof By Lemma 5 we have $\sum_{t=1}^{T} (\mu_{\mathbf{1}_D} - r_t) = \mathcal{O}\big(\sqrt{A^D DT \log(AT/\delta)}\big)$. Further using the fact that $\sum_{t=1}^{T} (r_t - \mu_{\mathbf{a}_t}) = \mathcal{O}\big(\sqrt{T \log(T/\delta)}\big)$ finishes the proof. ■

Lemma 7 *For any* $d \in [D]$, *let* $a^d \in [A]$ *be a sub-optimal arm of agent* d *given that* $\mathbf{a}^{1:d-1}$ *is chosen by the first* $d-1$ *agents, then with probability at least* $1-2\delta$,

$$\sum_{t=1}^{T} \mathbb{I}\{\mathbf{a}_t^{1:d} = \mathbf{a}^{1:d}\} \leq \frac{4(C_d+2)^2 A^{D-d} \log(A^D T/\delta)}{\left[\mu_{(\mathbf{a}^{1:d-1}, \mathbf{1}_{D-d+1})} - \mu_{(\mathbf{a}^{1:d}, \mathbf{1}_{D-d})}\right]^2}. \tag{21}$$

Proof of Lemma 7 Suppose at time $t+1$,

$$n_{t+1}^d(\mathbf{a}^{1:d}) \geq \frac{4(C_d+2)^2 A^{D-d} \log(A^D T/\delta)}{\left[\mu_{(\mathbf{a}^{1:d-1}, \mathbf{1}_{D-d+1})} - \mu_{(\mathbf{a}^{1:d}, \mathbf{1}_{D-d})}\right]^2}. \tag{22}$$

Then with probability at least $1-\delta$,

$$\begin{aligned}
&\hat{\mu}_{t+1}^d(\mathbf{a}^{1:d}) + C_d \sqrt{\frac{A^{D-d} \log(A^D T/\delta)}{n_{t+1}^d(\mathbf{a}^{1:d})^+}} \\
&\leq \mu_{(\mathbf{a}^{1:d}, \mathbf{1}_{D-d})} + 2\sqrt{\frac{\log(A^d T/\delta)}{n_{t+1}^d(\mathbf{a}^{1:d})^+}} + C_d \sqrt{\frac{A^{D-d} \log(A^D T/\delta)}{n_{t+1}^d(\mathbf{a}^{1:d})^+}} && \text{(by (19))} \\
&\leq \mu_{(\mathbf{a}^{1:d}, \mathbf{1}_{D-d})} + (C_d+2) \sqrt{\frac{A^{D-d} \log(A^D T/\delta)}{n_{t+1}^d(\mathbf{a}^{1:d})^+}} \\
&\leq \mu_{(\mathbf{a}^{1:d}, \mathbf{1}_{D-d})} + \frac{\mu_{(\mathbf{a}^{1:d-1}, \mathbf{1}_{D-d+1})} - \mu_{(\mathbf{a}^{1:d}, \mathbf{1}_{D-d})}}{2} < \mu_{(\mathbf{a}^{1:d-1}, \mathbf{1}_{D-d+1})}
\end{aligned}$$

On the other hand, by Lemma 5 with $(t, d, (\mathbf{a}^{1:d-1}, 1))$, with probability at least $1-\delta$,

$$\mu_{(\mathbf{a}^{1:d-1}, \mathbf{1}_{D-d+1})} \leq \hat{\mu}_t^d(\mathbf{a}^{1:d-1}, 1) + C_d \sqrt{\frac{A^{D-d} \log(A^D T/\delta)}{n_{t+1}^d(\mathbf{a}^{1:d-1}, 1)^+}}.$$

Hence, by the design of Algorithm 4, given that the first $d-1$ agents choose $\mathbf{a}^{1:d-1}$, agent d will not choose the sub-optimal arm $a^d \neq 1$ over the first arm before time T when the condition (22) is satisfied. ■

Theorem 8 *For the deep hierarchical bandit problem, with probability of at least* $1-\delta$*, Algorithm 4 achieves the regret bound of*

$$\sum_{t=1}^{T}[\mu_{\mathbf{1}_D} - \mu_{\mathbf{a}_t}] \leq \sum_{d=1}^{D} \sum_{\mathbf{a}^{1:d} \neq \mathbf{1}_d} O\left(\frac{A^{D-d} D \log(ADT/\delta)}{\mu_{(\mathbf{a}^{1:d-1}, \mathbf{1}_{D-d+1})} - \mu_{(\mathbf{a}^{1:d}, \mathbf{1}_{D-d})}}\right). \tag{23}$$

Proof of Theorem 8 From Lemma 7 and the union bound, the inequality (23) will hold simultaneously for all $d \in [D]$ and $\mathbf{a}^{1:d} \in [A]^d$ with probability at least $1 - 2A^D D\delta$. Hence, with probability at least $1 - 2A^D D\delta$,

$$\begin{aligned}
&\sum_{t=1}^{T}[\mu_{\mathbf{1}_D} - \mu_{\mathbf{a}_t}] = \sum_{\mathbf{a} \neq \mathbf{1}_D} (\mu_{\mathbf{1}_D} - \mu_{\mathbf{a}}) \cdot n_{T+1}^D(\mathbf{a}) \\
&= \sum_{\mathbf{a} \neq \mathbf{1}_D} \sum_{d=1}^{D} \left\{ \left[\mu_{(\mathbf{a}^{1:d-1}, \mathbf{1}_{D-d+1})} - \mu_{(\mathbf{a}^{1:d}, \mathbf{1}_{D-d})} \right] \cdot n_{T+1}^D(\mathbf{a}) \right\} \\
&= \sum_{d=1}^{D} \sum_{\mathbf{a}^{1:d} \neq \mathbf{1}_d} \left\{ \left[\mu_{(\mathbf{a}^{1:d-1}, \mathbf{1}_{D-d+1})} - \mu_{(\mathbf{a}^{1:d}, \mathbf{1}_{D-d})} \right] \cdot \left[\sum_{\mathbf{a}^{d+1:D}} n_{T+1}^D(\mathbf{a}) \right] \right\} \\
&= \sum_{d=1}^{D} \sum_{\mathbf{a}^{1:d} \neq \mathbf{1}_d} \left\{ \left[\mu_{(\mathbf{a}^{1:d-1}, \mathbf{1}_{D-d+1})} - \mu_{(\mathbf{a}^{1:d}, \mathbf{1}_{D-d})} \right] \cdot \left[\sum_{\mathbf{a}^{d+1:D}} n_{T+1}^D(\mathbf{a}) \right] \right\} \\
&= \sum_{d=1}^{D} \sum_{\mathbf{a}^{1:d} \neq \mathbf{1}_d} \left\{ \left[\mu_{(\mathbf{a}^{1:d-1}, \mathbf{1}_{D-d+1})} - \mu_{(\mathbf{a}^{1:d}, \mathbf{1}_{D-d})} \right] \cdot \left[\sum_{t=1}^{T} \mathbb{I}\{\mathbf{a}_t^{1:d} = \mathbf{a}^{1:d}\} \right] \right\} \\
&\leq \sum_{d=1}^{D} \sum_{\mathbf{a}^{1:d} \neq \mathbf{1}_d} \left\{ \left[\mu_{(\mathbf{a}^{1:d-1}, \mathbf{1}_{D-d+1})} - \mu_{(\mathbf{a}^{1:d}, \mathbf{1}_{D-d})} \right] \cdot \frac{4(C_d+2)^2 A^{D-d} \log(A^D T/\delta)}{\left[\mu_{(\mathbf{a}^{1:d-1}, \mathbf{1}_{D-d+1})} - \mu_{(\mathbf{a}^{1:d}, \mathbf{1}_{D-d})} \right]^2} \right\} \\
&= \sum_{d=1}^{D} \sum_{\mathbf{a}^{1:d} \neq \mathbf{1}_d} \frac{4(C_d+2)^2 A^{D-d} \log(A^D T/\delta)}{\mu_{(\mathbf{a}^{1:d-1}, \mathbf{1}_{D-d+1})} - \mu_{(\mathbf{a}^{1:d}, \mathbf{1}_{D-d})}}.
\end{aligned}$$

Letting $\delta' = 2A^D D\delta$ gives the claim. ■

Appendix E. Auxiliary Lemmas

Lemma 10 (Hoeffding-Azuma inequality) *Let* $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \cdots \subset \mathcal{F}_n$ *be a filtration, and* $X_1, \ldots, X_n$ *be real random variables such that* X_i *is* $\mathcal{F}_i$*-measurable,* $\mathbb{E}[X_i | \mathcal{F}_{i-1}] = 0$*,* $|X_i| \leq b$ *for some fixed* $b \geq 0$*. Furthermore, let* $\{y_i\}_{i=1}^n$ *be a fixed sequence. Then with probability at least* $1 - \delta$*,*

$$\sum_{i=1}^{n} y_i X_i \leq b \sqrt{2 \left(\sum_{i=1}^{n} y_i^2 \right) \log(1/\delta)}.$$

Lemma 11 (Weissman et al. (2003); Jaksch et al. (2010)) *Let $\hat{p}(\cdot) \in \mathbb{R}_+^d$ be the empirical over d distinct events from n samples, and let $p(\cdot)$ be the true underlying distribution. Then with probability at least $1-\delta$,*

$$\|\hat{p}(\cdot) - p(\cdot)\|_1 \leq \sqrt{\frac{2d \log(T/\delta)}{n}}.$$

Lemma 12 (Lemma 4.1 of (Jin et al., 2018)) *For a positive integer τ, define $\alpha_\tau^i = \alpha_i \Pi_{j=i+1}^{\tau}(1-\alpha_j)$ for $1 \leq i \leq \tau$ and $\alpha_\tau^0 = \Pi_{j=1}^{\tau}(1-\alpha_j)$ where $\alpha_\tau = \frac{H+1}{H+\tau}$. Then the following hold:*

1. $\frac{1}{\sqrt{\tau}} \leq \sum_{i=1}^{\tau} \frac{\alpha_\tau^i}{\sqrt{i}} \leq \frac{2}{\sqrt{\tau}}$ *for all $\tau \geq 1$.*
2. $\max_{i\in[\tau]} \alpha_\tau^i \leq \frac{2H}{\tau}$ *and* $\sum_{i=1}^{\tau}(\alpha_\tau^i)^2 \leq \frac{2H}{\tau}$ *for all $\tau \geq 1$.*
3. $\sum_{\tau=i}^{\infty} \alpha_\tau^i = 1 + \frac{1}{H}$ *for all $i \geq 1$.*

Lemma 13 *Let α_n^i be defined as in Lemma 12. Let $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \cdots \subset \mathcal{F}_n$ be a filtration, and $p_1, \ldots, p_n$ be distribution over $\mathcal{S}$ where p_i is deterministic given $\mathcal{F}_{i-1}$. Suppose that $s_i \in \mathcal{S}$ is drawn from p_i. Then with probability at least $1-\delta$, for all $V : \mathcal{S} \to [0,1]$,*

$$\left|\sum_{i=1}^{n} \alpha_n^i V(s_i) - \sum_{i=1}^{n} \alpha_n^i \sum_s p_i(s)V(s)\right| \leq 3\sqrt{\frac{SH}{n}\log(4n/\delta)}.$$

Proof Consider the following ϵ-cover for the space of V:

$$\mathcal{V} = \{V : \mathcal{S} \to \{0, \epsilon, 2\epsilon, \ldots, 1\}\}$$

For every fixed $V \in \mathcal{V}$, we have with probability at least $1-\delta'$,

$$\left|\sum_{i=1}^{n} \alpha_n^i V(s_i) - \sum_{i=1}^{n} \alpha_n^i \sum_s p_i(s)V(s)\right| \leq \sqrt{2\left(\sum_{i=1}^{n}(\alpha_n^i)^2\right)\log(2/\delta')} \leq 2\sqrt{\frac{H}{n}\log(2/\delta')}.$$

By an union bound, the above holds for all $V \in \mathcal{V}$ with probability at least $1 - |\mathcal{V}|\delta'$.

For any $V : \mathcal{S} \to [0,1]$, there is a $\tilde{V} \in \mathcal{V}$ such that $|V(s) - \tilde{V}(s)| \leq \frac{\epsilon}{2}$ for all s. Thus, with probability $1 - |\mathcal{V}|\delta'$, we have

$$\left|\sum_{i=1}^{n} \alpha_n^i V(s_i) - \sum_{i=1}^{n} \alpha_n^i \sum_s p_i(s)V(s)\right| \leq 2\sqrt{\frac{H}{n}\log(2/\delta')} + \epsilon.$$

Picking $\epsilon = \frac{1}{n}$ (which implies $|\mathcal{V}| = \left(\frac{1}{\epsilon}+1\right)^S \leq \left(\frac{2}{\epsilon}\right)^S = (2n)^S$), and $\delta' = \frac{\delta}{|\mathcal{V}|}$, the right-hand side above can be upper bounded by

$$2\sqrt{\frac{H}{n}\log(2(2n)^S/\delta)} + \frac{1}{n} \leq 3\sqrt{\frac{SH}{n}\log(4n/\delta)}.$$

■