Bounded regret in multi-agent bandit settings with linear contextual rewards

Hyunwook Kang HWKANG@TAMU.EDU
Department of Electrical & Computer Engineering
Texas A&M University
College Station, TX 77843-3128, USA

P. R. Kumar PRK@TAMU.EDU
Department of Electrical & Computer Engineering
Texas A&M University
College Station, TX 77843-3128, USA

Abstract

Asymptotically unbounded regret bounds, of the form $O(\log T)$, $O(\sqrt{T})$ or $O(\sqrt{T}\log T))$, have been proven to be the lowest possible regret upper bounds that can be achieved when an agent explores multiple arms in various bandit settings. What if there are multiple heterogeneous agents, i.e., agents who experience different rewards for the same arm according to the latent reward structure? In this paper, we show that a bounded, i.e., $O(1)$, expected regret can be achieved when there is a large number of agents who play a relatively small number of arms according to Counterfactual UCB (UFUCB) under the typical assumptions made in contextual bandits: 1) linear latent reward structure and 2) knowledge of contextual feature vector for each agent.

Keywords: Multi-agent bandit, Linear contextual rewards, Bounded regret

1. Introduction

The contextual bandits framework has been a popular theoretical tool for the analysis of the exploration-exploitation trade-off under the assumption that each agent's characteristics, usually modeled as a feature vector called context, are known a priori [Auer (2002); Chu et al. (2011); Abbasi-Yadkori et al. (2011); Krause and Ong (2011)]. This framework's potential usefulness for adaptive experimentation has been increasingly highlighted [Bouneffouf and Rish (2019); Dimakopoulou et al. (2021); Jourdan et al. (2021); Srinivas et al. (2009)]. However, when an expert in contextual bandits tries to claim the usefulness of her algorithm to the practitioners of real-world experiments, it becomes an issue of concern that the regret bounds of the contextual bandit framework increase unboundedly over time. To understand this, think of marketing campaigns shown to users in the advertisement platforms. Many of marketing campaign's needs, such as the ones for new movies or election candidates, are only ephemeral; they quickly lose their timeliness before they get old. In such cases, the theoretical claims based on asymptotic analysis may be limitedly persuasive.

In this paper, we suggest another bandit framework for which we can show bounded expected regret in the large. While similar in most parts, its problem setting differs from that of typical linear contextual bandit problems in that 1) we specify the set of agents

A and keep track of each individual's repeated arrivals, and 2) the result becomes highly probable as the number of agents gets relatively larger than the number of arms[1].

The main idea that enables the bounded expected regret result is as follows. An agent searching for the optimal arm alone under a typical UCB (Upper confidence bound) policy [Lai et al. (1985)] cannot utilize the linear latent structure and therefore suffers $O(\log T)$ regret due to the exploration requirements for all non-optimal arms. However, if there are enough number of agents, the linear latent structure can be utilized; this is done by additionally employing the *counterfactual UCBs* that are computed mainly from other agents' exploitation. Each agent is then fully exempt from the exploration requirements after finite time, achieving bounded expected regret.

2. The problem setting

Denote the set of agents' indices as A and the set of arms' indices as M. The feature of agent j is represented by a d-dimensional vector $\alpha^{(j)} \in \mathbb{R}^d$ and the feature of arm m is represented by a d-dimensional vector $\beta_m \in \mathbb{R}^d$.[2] For the feature vectors associated with the agents, we further assume that any d-sized subset of A is linearly independent (note that this is justified by the fact that fullness of rank is generic). We assume that each agent's feature vector is known a priori. On the other hand, there is no prior information on arm feature vectors.

An agent receives a random reward when it pulls an arm. Denote the kth reward of agent j from arm m as $X_{m,k}^{(j)}$. We assume that the reward is noisily determined by the inner product of agent feature vector and arm feature vector. That is, $X_{m,k}^{(j)} = \alpha^{(j)}\beta_m + \epsilon_m^{(j)}(k)$ where $\epsilon_m^{(j)}(k)$ follows a sub-Gaussian distribution with $E[\epsilon_m^{(j)}(k)] = 0$ and proxy variance σ^2 [Rivasplata (2012)]. We denote $E[X_{m,k}^{(j)}] = \alpha^{(j)}\beta_m$ by $\mu_m^{(j)}$.

We assume that $\mu_m^{(j)} \neq \mu_n^{(j)}\ \forall j \in A,\ \forall m, n \in M$. This implies that each agent will have a unique optimal arm. We denote the set of agents with m as their optimal arm by $A_m := \{j \in A : \mu_m^{(j)} > \mu_n^{(j)}\ \forall n \in M, n \neq m\}$. Note that $\{A_m\}_{m\in M}$ partitions A.

We suppose that agents independently arrive according to identical renewal processes, i.e., agent inter-arrival times are i.i.d. (independent and identically distributed). Specifically, two cases are considered: when all inter-arrival times are 1) i.i.d. subgaussian with a density on the real line, and 2) i.i.d. exponential. An agent gets to pull an arm when it arrives. Denote the time of the nth arrival of agent j by $S_n^{(j)}$, and the associated inter-arrival time by $Y_n^{(j)} := S_n^{(j)} - S_{n-1}^{j}$. We denote the associated counting process of agent j's arrival process by $N^{(j)}(t)$. That is, $\{S_n^{(j)} \leq t\} = \{N^{(j)}(t) \geq n\}$. Note that $N^{(j)}(t)$ is also the total number of pulls over all arms of agent j until time t. We further denote $N_m^{(j)}(t)$ as the number of agent j's pulls of arm m until time t.

1. This condition fits particularly well for the experimentation design problems as the size of the population is much larger than the number of treatments in typical experimentation settings.
2. Note that the existence of arm feature vectors is not assumed in the typical linear contextual bandit problems. In such problems, a prior knowledge on the feature mapping that takes an agent (=context) and an arm as the input elements and outputs a vector is assumed instead. The inner product of this output vector and an unknown common fixed vector forms the reward. Comparison between this setting and ours on the usefulness and the applicability of representation learning is left as an open problem.

We observe which agents have pulled a specific arm the most until time T. Specifically, let $A_m(n,T) := \{j \in A : |\{i \in A : N_m^{(i)}(T) > N_m^{(j)}(T)\}| < n\}$. This set includes the n agents most pulling arm m with all ties at the bottom being included. That is, $|A_m(n,T)| \geq n$ unless $|A| < n$. For each agent j, we arbitrarily choose one d-size subset of $A_m(d+1,T) \setminus j$ (or more precisely $A_m(d+1,T) \setminus \{j\}$), and fix it as $E_m^{(j)}(T)$.

Denote agent j's optimal arm by m_j^*. and the arm pulled at agent j's nth arrival as $m_j(n)$. Then we consider the finite time regret of agent j until time T defined as $Regret^{(j)}(T) := \sum_{n=1}^{N^{(j)}(T)} (\mu_{m_j^*}^{(j)} - \mu_{m_j(n)}^{(j)})$. Our objective is to upper bound $E[Regret^{(j)}(T)]$ by a constant for all $j \in A$.

In the following sections, we will show that $E[Regret^{(j)}(T)]$ is upper bounded by a constant under the following condition that intuitively holds when $|A|$ is large enough:

$$|A_m| \geq d+1, \quad \forall m \in M. \tag{1}$$

If this condition does not hold, then there will be some agents who suffer $O(\log T)$ expected regret instead of enjoying bounded expected regret. For the rest of the paper, we will assume that the condition (1) holds.

How large the number of agents should be to make this condition as probable as we want? This is an unproved different version of Double Dixie cup problem [Newman (1960)]. Theorem 1 answers this question.

Theorem 1. *Suppose that the optimal arms associated with agents $\{m_j^* : j \in A\}$ are independently and uniformly distributed over A. If*

$$|A| \geq |M|d + \max\{\eta|M|d, \frac{2(1+\eta)}{\eta}\left(|M|\ln|M| + |M|\ln\frac{1}{\epsilon} + d\right)\}, \tag{2}$$

then

$$P(|A_m| \geq d+1 \ \forall m \in M) \geq 1-\epsilon.$$

The proof of Theorem 1 is provided in Appendix A. It shows that at least a multiple of $(|M|\ln|M| + |M|d)$ number of agents is required, and an additional multiple of $|M|\ln\frac{1}{\epsilon}$ agents is needed if we want $(1-\epsilon)$ probability assurance.

3. Construction of the Upper Confidence Bounds (UCBs)

Assuming that the condition (1) holds, we can form the counterfactual means and counterfactual confidence bounds, by which mean those obtained from outside an individual agent's experience. Recall that we defined $N_m^{(j)}(t)$ as the number of agent j's pulls of arm m until time t. We first denote by $\overline{X}_m^{(j)}(t) = \frac{\sum_{k=1}^{N_m^{(j)}(t)} X_{m,k}^{(j)}}{N_m^{(j)}(t)}$ the empirical mean reward of agent j on arm m. Before defining the counterfactual mean of agent j for arm m, recall that $A_m(d,t) := \{j \in A : |\{i \in A : N_m^{(i)}(t) > N_m^{(j)}(t)\}| < d\}$. This set includes top d agents for arm m with all ties at the bottom being included. Taking into account Theorem 1, suppose that $|A| \geq d+1$. Note that this implies $|A_m(d+1,t)| \geq d+1$. Recall that we arbitrarily chose a d-size subset of $A_m(d+1,t) \setminus j$ and fixed it as $E_m^{(j)}(t)$. Since the

feature vectors of the d-size subset of A are linearly independent as we assumed earlier, we can express $\alpha^{(j)} = \sum_{i \in E_m^{(j)}(t)} a_i^{(j)} \alpha^{(i)}$ for some coefficients $\{a_i^{(j)}\}$, and subsequently $\mu_m^{(j)}$ as $\sum_{i \in E_m^{(j)}(t)} a_i^{(j)} \mu_m^{(i)}$. We define $\widehat{X}_m^{(j)}(t) := \sum_{i \in E_m^{(j)}(t)} a_i^{(j)} \overline{X}_m^{(i)}(t)$ and call it the counterfactual mean of agent j for arm m.

In choosing the confidence intervals, we follow the spirit of [Auer (2002)] - that is, we bound the violation probability by the inverse square of the total number of pulls at time t. Lemmas 2 and 3 describe this confidence interval choice.

Lemma 2 (Auer (2002))**.** *For* $\epsilon \geq \sqrt{\frac{\log N^{(j)}(t)}{N_m^{(j)}(t)}}$, $P(|\overline{X}_m^{(j)}(t) - \mu_m^{(j)}| > \epsilon) \leq N^{(j)}(t)^{-2}$.

Proof This follows from Hoeffding's inequality, $P(|\overline{X}_m^{(j)}(t) - \mu_m^{(j)}| > \epsilon) \leq exp(-2N_m^{(i)}(t)\epsilon^2)$. Since we want to upper bound $P(|\overline{X}_m^{(j)}(t) - \mu_m^{(j)}| > \epsilon) \leq$ by $N^{(j)}(t)^{-2}$, The value of ϵ that makes $exp(-2N_m^{(i)}(t)\epsilon^2) \leq N^{(j)}(t)^{-2}$ will do it. This gives us $\epsilon \geq \sqrt{\frac{\log N^{(j)}(t)}{N_m^{(j)}(t)}}$. ■

Lemma 3. *Denote* $c_{m,t} := \sum_{i \in E_m^{(j)}(t)} |a_i^{(j)}|$*, and define* $N_m^{(\min)}(d,t,j) := \min_{i \in E_m^{(j)}(t)} N_m^{(i)}(t)$*. Then, for* $\epsilon \geq \sqrt{\frac{\log(N^{(j)}(t)/d)}{N_m^{(\min)}(d,t,j)/c_{m,t}^2}}$, $P(|\widehat{X}_m^{(j)}(t) - \mu_m^{(j)}| > \epsilon) \leq N^{(j)}(t)^{-2}$.

Proof. $P(|\widehat{X}_m^{(j)}(t) - \mu_m^{(j)}| > \epsilon) = 1 - P(|\widehat{X}_m^{(j)}(t) - \mu_m^{(j)}| \leq \epsilon) \leq 1 - \Pi_{i \in E_m^{(j)}(t)} P(|a_i^{(j)}||\overline{X}_m^{(i)}(t) - \mu_m^{(i)}| \leq |a_i^{(j)}|\frac{\epsilon}{c_{m,t}}) = 1 - \Pi_{i \in E_m^{(j)}(t)}(1 - P(|\overline{X}_m^{(i)}(t) - \mu_m^{(i)}| > \frac{\epsilon}{c_{m,t}})) \leq 1 - \Pi_{i \in E_m^{(j)}(t)}((1 - \exp(\frac{-2N_m^{(i)}(t)\epsilon^2}{c_{m,t}^2}))) \leq 1 - \Pi_{i \in E_m^{(j)}(t)}(1 - \exp(\frac{-2N_m^{(\min)}(d,t,j)\epsilon^2}{c_{m,t}^2})) = 1 - (1 - \exp(\frac{-2N_m^{(\min)}(d,t,j)\epsilon^2}{c_{m,t}^2}))^d \leq d\exp(\frac{-2N_m^{(\min)}(d,t,j)\epsilon^2}{c_{m,t}^2})$. Therefore, $d\exp(\frac{-2N_m^{(\min)}(d,t,j)\epsilon^2}{c_{m,t}^2}) \leq N^{(j)}(t)^{-2}$, i.e., $\epsilon \geq \sqrt{\frac{\log(N^{(j)}(t)/d)}{N_m^{(\min)}(d,t,j)/c_{m,t}^2}}$ implies $P(|\widehat{X}_m^{(j)}(t) - \mu_m^{(j)}| > \epsilon) \leq N^{(j)}(t)^{-2}$. □

The original confidence interval's width for agent j's arm m reward, denoted by $w_m^{(j)}(t)$, is chosen as the minimum ϵ that suffices, i.e., $\sqrt{\frac{\log N^{(j)}(t)}{N_m^{(j)}(t)}}$ from Lemma 2. The width of counterfactual confidence interval, denoted by $\widehat{w}_m^{(j)}(t)$, is chosen as $\sqrt{\frac{\log(N^{(j)}(t)/d)}{N_m^{(\min)}(d,t,j)/c_{m,t}^2}}$ in the same manner. We define

$$ucb_m^{(j)}(t) := \overline{X}_m^{(j)}(t) + w_m^{(j)}(t),\ \widehat{ucb}_m^{(j)}(t) := \widehat{X}_m^{(j)}(t) + \widehat{w}_m^{(j)}(t). \tag{3}$$

4. Counterfactual UCB Algorithm (CFUCB)

Before describing the main algorithm we call Counterfactual UCB (CFUCB), we introduce a concept of *epoch* and some related notations. Define $S := \cup_{i \in A}\{S_n^{(i)}\}_{n \in \mathbb{N}}$, the set of all arrival times of all agents. Note that S is a totally ordered set with order structure $\leq$. We can sequentially order the elements of S as a monotone increasing sequence $\{s_k\}_{k \in \mathbb{N}}$, with s_k denoting the time of the kth arrival, irrespective of agent identity. From now on, we call

s_k as the time of the kth *arrival epoch*, or simply the kth *epoch*. We define a sequence of agent indices $\{a_k\}_{k\in\mathbb{N}}$ such that $a_k = i \in A$ if $s_k = S_n^{(i)}$ for some $n \in \mathbb{N}$. That is, $\{a_k\}_{k\in\mathbb{N}}$ indicates which agent arrives at each epoch. Ties between agents arriving at the same time is broken arbitrarily.Given $\{a_k\}_{k\in\mathbb{N}}$, we denote the the index of the arm pulled by agent a_k at epoch k by m_k, and the corresponding reward accrued by r_k, where $m_k \in M$ and $r_k \in \mathbb{R}^+$ ($r_k = \alpha^{(a_k)}\beta_{m_k} + \epsilon_k$, where ϵ_k is "noise"). Recall that $X_m^{(j)}(n)$ denotes the nth reward of agent j from arm m. Algorithm 1 describes the pseudocode for the CFUCB Algorithm.

Algorithm 1: CFUCB Algorithm

Input: $\{\alpha^{(j)}\}_{j\in A}$ where $\alpha^{(j)}$ denotes the feature vector of agent j

1 **for** $k = 1, 2, \ldots$ **do**
2 Observe s_k and a_k
3 **for** $m = 1, 2, \ldots, |M|$ **do**
4 Compute $ucb_m^{(a_k)}(s_k)$ (the original UCB) according to the Equation (3)
5 Compute $\widehat{ucb}_m^{(a_k)}(s_k)$ (the counterfactual UCB) according to the Equation (3)
6 $\widetilde{ucb}_m^{(a_k)}(s_k) = \min(ucb_m^{(a_k)}(s_k), \widehat{ucb}_m^{(a_k)}(s_k))$
7 Set $m_k = \arg\min_{m\in M}\{\widetilde{ucb}_m^{(a_k)}(s_k)\}$
8 Let agent a_k pull the arm m_k and obtain r_k
9 Store $X_{m_k}^{(a_k)}(N_{m_k}^{(a_k)}(s_k)) = r_k$ for the future use in later loop's line **4** and line **5**

4.1 Analysis of CFUCB

We start the analysis of Algorithm 1 from Lemma 4, which suggests the condition for the agent j to pull a non-optimal arm m. Lemma 4 is the key result of this paper in that it gives the intuition about why bounded regret will be achieved if the arrival rates of the agents are not too much different.

Note that as a consequence of Lemmas 2 and 3, at time t, for every arm n and every agent i, the original confidence interval $CI_n^{(i)}(t) := (\overline{X}_n^{(i)}(t) - w_n^{(i)}(t), \overline{X}_n^{(i)}(t) + w_n^{(i)}(t))$ and the counterfactual confidence interval $\widehat{CI}_n^{(i)}(t) := (\widehat{X}_n^{(i)}(t) - \widehat{w}_n^{(i)}(t), \widehat{X}_n^{(i)}(t) + \widehat{w}_n^{(i)}(t))$ both include the true mean $\mu_n^{(i)}$ with high probability.

Lemma 4. *If $CI_n^{(i)}(t)$ and $\widehat{CI}_n^{(i)}(t)$ both include the true mean $\mu_n^{(i)}$, then an agent j who arrives at time t pulls a non-optimal arm m, i.e., one with $\Delta_m^{(j)} > 0$, only if*

$$\min_{i\in A_m}\{N^{(i)}(t) - (\sum_{n\neq m}\frac{4}{{\Delta_n^{(i)}}^2})\log N^{(i)}(t)\} \leq \frac{4c_{m,t}^2\log(N^{(j)}(t)/d)}{{\Delta_m^{(j)}}^2}. \tag{4}$$

The proof of Lemma 4 is deferred to Appendix A. One may note that the LHS of (4) will increase far faster than the RHS of (4) unless some agent $i \in A_m$ arrives far slower than agent j. Soon, therefore, the inequality will cease to hold for all non-optimal arms, and only the optimal arm will be pulled afterwards.

Now we are ready to discuss how the bounded regret is achieved. Lemma 5 draws a connection between the expected regret and the probability of agent j arriving at time t pulling a non-optimal arm m. The proof is deferred to Appendix A.

Lemma 5. *Denote the event* $\{$*Agent* j *arrives at time* t *and pulls a non-optimal arm* $m\}$ *by* $G_m^{(j)}(t)$*, and the event* $\{\mu_m^{(j)} \in CI_m^{(j)}(t) \cap \widehat{CI}_m^{(j)}(t)\}$ *as* $V_m^{(j)}(t)$*. Suppose that there is a function* $g_m^{(j)}(t)$ *such that* $P(G_m^{(j)}(t)|V_m^{(j)}(t)) \leq g_m^{(j)}(t)$*. Then* $E[Regret^{(j)}(T)] \leq \sum_{m \in M \setminus m_j^*} \Delta_m \left(\frac{\pi^2}{6} + \sum_{n=1}^{\infty} \int_0^{+\infty} g_m^{(j)}(t) dF_n^{(j)}(t)\right)$ *holds, where* $F_n^{(j)}(t) := P(S_n^{(j)} \leq t)$*.*

Finding $g_m^{(j)}(t)$ such that $\sum_{n=1}^{\infty} \int_0^{+\infty} g_m^{(j)}(t) dF_n^{(j)}(t) < \infty$ will bring us the bounded expected regret result. We show this for the two most representative settings: agents arriving according to 1) sub-Gaussian inter-arrival times (Theorem 6) and 2) exponential inter-arrival times (Theorem 7). The proofs are deferred to the Appendix A.

Theorem 6. *Suppose that each agent of* $i \in A$ *arrives independently with i.i.d. 1-subgaussian inter-arrival times with mean* θ_i*. Then we can find* $g_m^{(j)}(t)$ *of Lemma 5 such that* $\int_0^{+\infty} g_m^{(j)}(t) dF_n^{(j)}(t) = O(\frac{1}{n^2})$ *holds and thus* $E[Regret^{(j)}(T)] < \infty$ *holds under CFUCB.*

Theorem 7. *Suppose that each agent* i *of* A *arrive independently with iid exponentially distributed inter-arrival times with* λ_i*. Then we can find* $g_m^{(j)}(t)$ *of Lemma 5 such that* $\int_0^{+\infty} g_m^{(j)}(t) dF_n^{(j)}(t) = O(\frac{1}{n^2})$ *holds and thus* $E[Regret^{(j)}(T)] < \infty$ *under CFUCB.*

5. Concluding remarks

This paper suggests a multi-armed bandit framework that shares the assumptions of typical contextual bandit problems such as 1) linear latent reward structure, and 2) knowledge of contextual feature vector for each agent. For this framework, which works in the large, the UCB policy we call Counterfactual-UCB (CFUCB) guarantees bounded expected regret. The key idea enabling this result is to gather together exploitation results of a set of agents to give another agent an exemption from exploration requirement.

The closest literature in terms of problem settings is on the linear contextual bandit problems [Auer (2002); Chu et al. (2011); Abbasi-Yadkori et al. (2011); Bogunovic et al. (2021)]. There has been progress in relaxing linear latent reward structure assumption Simchi-Levi and Xu (2021); Krishnamurthy et al. (2021); Foster et al. (2020); Foster and Rakhlin (2020), and it would be an interesting direction to see whether such progress can be extended to the framework suggested in this paper. Interestingly, Xu and Zeevi (2020) explores how the concept of counterfactuals can be used to address the UCB algorithm's weakness for non-linear contextual bandits with large context space.

On the idea of gathering information from multiple agents, there has been an enormous amount of previous works on cooperative multi-armed bandit problems [Shih et al. (2022); Sankararaman et al. (2019); Landgren et al. (2016); Martínez-Rubio et al. (2019); Wang et al. (2020); Chawla et al. (2020), as well as heterogeneous but non-cooperative multi-agent multi-armed bandit settings Immorlica et al. (2019); Chen et al. (2018)].

For the real-world applications, future works may consider relaxing the problem objective (e.g., finding a near-optimal arm instead of finding the true optimal arm) to address uncertainty in the knowledge of each agent's contextual feature vector.

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. Advances in neural information processing systems, 24, 2011.

Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. Journal of Machine Learning Research, 3(Nov):397–422, 2002.

Ilija Bogunovic, Arpan Losalka, Andreas Krause, and Jonathan Scarlett. Stochastic linear bandits robust to adversarial attacks. In International Conference on Artificial Intelligence and Statistics, pages 991–999. PMLR, 2021.

Djallel Bouneffouf and Irina Rish. A survey on practical applications of multi-armed and contextual bandits. arXiv preprint arXiv:1904.10040, 2019.

Ronshee Chawla, Abishek Sankararaman, and Sanjay Shakkottai. Multi-agent low-dimensional linear bandits. arXiv preprint arXiv:2007.01442, 2020.

Bangrui Chen, Peter Frazier, and David Kempe. Incentivizing exploration by heterogeneous users. In Conference On Learning Theory, pages 798–818. PMLR, 2018.

Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandits with linear payoff functions. In Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, pages 208–214. JMLR Workshop and Conference Proceedings, 2011.

Maria Dimakopoulou, Zhimei Ren, and Zhengyuan Zhou. Online multi-armed bandits with adaptive inference. Advances in Neural Information Processing Systems, 34, 2021.

Dylan Foster and Alexander Rakhlin. Beyond ucb: Optimal and efficient contextual bandits with regression oracles. In International Conference on Machine Learning, pages 3199–3210. PMLR, 2020.

Dylan J Foster, Claudio Gentile, Mehryar Mohri, and Julian Zimmert. Adapting to misspecification in contextual bandits. Advances in Neural Information Processing Systems, 33: 11478–11489, 2020.

Nicole Immorlica, Jieming Mao, Aleksandrs Slivkins, and Zhiwei Steven Wu. Bayesian exploration with heterogeneous agents. In The World Wide Web Conference, pages 751–761, 2019.

Marc Jourdan, Mojmír Mutnỳ, Johannes Kirschner, and Andreas Krause. Efficient pure exploration for combinatorial bandits with semi-bandit feedback. In Algorithmic Learning Theory, pages 805–849. PMLR, 2021.

Andreas Krause and Cheng Ong. Contextual gaussian process bandit optimization. Advances in neural information processing systems, 24, 2011.

Sanath Kumar Krishnamurthy, Vitor Hadad, and Susan Athey. Tractable contextual bandits beyond realizability. In International Conference on Artificial Intelligence and Statistics, pages 1423–1431. PMLR, 2021.

Tze Leung Lai, Herbert Robbins, et al. Asymptotically efficient adaptive allocation rules. Advances in applied mathematics, 6(1):4–22, 1985.

Peter Landgren, Vaibhav Srivastava, and Naomi Ehrich Leonard. Distributed cooperative decision-making in multiarmed bandits: Frequentist and bayesian algorithms. In 2016 IEEE 55th Conference on Decision and Control (CDC), pages 167–172. IEEE, 2016.

David Martínez-Rubio, Varun Kanade, and Patrick Rebeschini. Decentralized cooperative stochastic bandits. Advances in Neural Information Processing Systems, 32, 2019.

Donald J Newman. The double dixie cup problem. The American Mathematical Monthly, 67(1):58–61, 1960.

David Pollard. Miniempirical. Lecture note(Last accessed:05/10/2022):16, 2015. http://www.stat.yale.edu/ pollard/Courses/600.spring2017/Handouts/Basic.pdf.

Omar Rivasplata. Subgaussian random variables: An expository note. Internet publication, PDF, 5, 2012.

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.

Andy Shih, Stefano Ermon, and Dorsa Sadigh. Conditional imitation learning for multi-agent games. arXiv preprint arXiv:2201.01448, 2022.

David Simchi-Levi and Yunzong Xu. Bypassing the monster: A faster and simpler optimal algorithm for contextual bandits under realizability. Mathematics of Operations Research, 2021.

Niranjan Srinivas, Andreas Krause, Sham M Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. arXiv preprint arXiv:0912.3995, 2009.

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

Yunbei Xu and Assaf Zeevi. Upper counterfactual confidence bounds: a new optimism principle for contextual bandits. arXiv preprint arXiv:2007.07876, 2020.

Appendix A. Proof of Lemmas and Theorems

Proof [Proof of Theorem 2.] For simplicity, we denote $|A| = a$ and $|M| = b$. Let I_m be the indicator random variable for the event $\{|A_m| < d+1\}$, and $I := \sum_{m \in M} I_m$. What we want is to upper bound $P(I > 0)$ by ϵ. Note that

$$
\begin{aligned}
P(I>0) &= P(I \geq 1) \\
&\leq E[I] \quad \text{(because of Markov's inequality)} \\
&= bE[I_1] \\
&= bP(I_1 = 1) \\
&= b\sum_{k=0}^{d} \binom{a}{k} \left(1-\frac{1}{b}\right)^{a-k} \left(\frac{1}{b}\right)^{k} \\
&\leq b\sum_{k=0}^{d} \binom{a}{k} \left(1-\frac{1}{b}\right)^{a-d} \left(\frac{1}{b}\right)^{k} \\
&\leq b\sum_{k=0}^{d} \frac{a^k}{k!} \exp(-\frac{a-d}{b}) \left(\frac{1}{b}\right)^{k} \text{ (because } \binom{a}{k} \leq \frac{a^k}{k!}, \text{ and } 1+x \leq e^x) \\
&= \exp\left(\frac{d}{b}\right) \sum_{k=0}^{d} \frac{1}{k!} \left(\frac{a}{b}\right)^{k} \exp\left(-\frac{a}{b}\right) \\
&= b\exp\left(\frac{d}{b}\right) P\left(Z \leq d\right), \text{ where } Z \sim Poi(\frac{a}{b}) \\
&\overset{(a)}{\leq} b\exp\left(\frac{d}{b}\right) \exp\left(-\frac{1}{2}\frac{b}{a}\frac{(a-bd)^2}{b^2}\right) \qquad (5) \\
&= b\exp\left(\frac{1}{b}\left(d - \frac{(a-bd)^2}{2a}\right)\right) \\
&= \exp\left(\ln b - \frac{1}{b}\left(\frac{(a-bd)^2}{2a} - d\right)\right). \qquad (6)
\end{aligned}
$$

Above, the inequality (a) of (5) holds because $Z \sim Poi(\lambda)$, $\Pr[Z \leq \lambda - x] \leq e^{-\frac{x^2}{2\lambda}}$ for $0 \leq x \leq \lambda$ *Pollard* (2015), where in our case $\frac{a}{b} \geq d$ as assumed, $\lambda = \frac{a}{b}$, $\lambda - x = d$ and $x = \frac{a}{b} - d = \frac{a-bd}{b}$).

Let us further assume that $a \geq (1+\eta)bd$. Then

$$
\begin{aligned}
a &\geq (1+\eta)bd \\
&(\Leftrightarrow) \quad (1+\eta)(a-bd) \geq (1+\eta)a - a = \eta a \\
&(\Leftrightarrow) \quad a \leq (a-bd)\frac{(1+\eta)}{\eta}. \qquad (7)
\end{aligned}
$$

Then,

$$
\begin{aligned}
&P(I>0) \le \epsilon \\
&\quad (\Leftarrow) \quad \exp\left(\ln b - \frac{1}{b}\left(\frac{(a-bd)^2}{2a} - d\right)\right) \le \epsilon \quad \text{(because of (6))} \\
&\quad (\Leftrightarrow) \quad \exp\left(-\frac{(\frac{(a-bd)^2}{2a} - d) - b\ln b}{b}\right) \le \epsilon \\
&\quad (\Leftrightarrow) \quad \frac{(a-bd)^2}{2a} \ge b\ln b + b\ln\frac{1}{\epsilon} + d \\
&\quad (\Leftarrow) \quad a - bd \ge \frac{2(1+\eta)}{\eta}\left(b\ln b + b\ln\frac{1}{\epsilon} + d\right) \quad \text{(because of (7))} \qquad (8)
\end{aligned}
$$

■

Proof [Proof of Lemma 4.] Lemma 4 is based on the following Lemma 8:

Lemma 8. *Under the same conditions as in Lemma 4, agent j pulls arm m only if* $\min\left(2\sqrt{\frac{\log N^{(j)}(t)}{N_m^{(j)}(t)}}, 2\sqrt{\frac{\log(N^{(j)}(t)/d)}{N_m^{(\min)}(d,t,j)/c_{m,t}^2}}\right) \ge \Delta_m^{(j)}$. *That is, both* $N_m^{(j)}(t) \le \frac{4\log N^{(j)}(t)}{{\Delta_m^{(j)}}^2}$ *and* $N_m^{(\min)}(d,t,j) \le \frac{4c_{m,t}^2\log(N^{(j)}(t)/d)}{{\Delta_m^{(j)}}^2}$ *must hold for agent j to pull arm m.*

Proof [Proof of Lemma 8] Denote the optimal arm for agent j as arm m_j^*. Agent j pulls arm m only when the event $\{\arg\max_{q\in M} \mathrm{ucb}_q^{(j)}(t) = m\}$ happens. This means that $\mathrm{ucb}_m^{(j)}(t) \ge \mathrm{ucb}_{m_j^*}^{(i)}(t)$. Recall that we assume as in Lemma 4 that $\mu_m^{(j)} \in CI_m^{(j)}(t) \cap \widehat{CI}_m^{(j)}(t)$ and $\mu_{m_j^*}^{(j)} \in CI_{m_j^*}^{(j)}(t) \cap \widehat{CI}_{m_j^*}^{(j)}(t)$. Now we make an observation that $\mathrm{ucb}_m^{(j)}(t) \ge \mathrm{ucb}_{m_j^*}^{(j)}(t), \mu_m^{(j)} \le \mu_{m_j^*}^{(j)}, \mu_m^{(j)} \in CI_m^{(j)}(t) \cap \widehat{CI}_m^{(j)}(t)$ and $\mu_{m_j^*}^{(j)} \in CI_{m_j^*}^{(j)}(t) \cap \widehat{CI}_{m_j^*}^{(j)}(t)$ jointly imply $\mu_m^{(j)}, \mu_{m_j^*}^{(j)} \in CI_m^{(j)}(t) \cap \widehat{CI}_m^{(j)}(t)$. This means that $\min(2w_m^{(j)}(t), 2\widehat{w}_m^{(j)}(t)) \ge \Delta_m^{(j)}$ holds. Combining this with Lemma 2 and 3, yields the result. ■

We can now prove **Lemma 4**, starting as follows:

Fix an arm j. Now note that for any arm i, $N_m^{(i)}(t) = N^{(i)}(t) - \sum_{n\in M\setminus m} N_n^{(i)}(t)$. Let t^n be the last time prior to t at which a non-optimal arm n is played by agent i. Then $N_n^{(i)}(t) = N_n^{(i)}(t^n) \le \frac{4\log N^{(i)}(t^n)}{{\Delta_n^{(i)}}^2} \le \frac{4\log N^{(i)}(t)}{{\Delta_n^{(i)}}^2}$ holds by Lemma 8. Therefore, for agent $i \in A_m$, $N_m^{(i)}(t) \ge N^{(i)}(t) - (\sum_{n\ne m}\frac{4}{{\Delta_n^{(i)}}^2})\log N^{(i)}(t)$. By the assumption (1), $|A_m| \ge d+1$, and $N_m^{\min}(d,t,j) \ge N_m^{(i)}(t)$ for some $i \in A_m$. Therefore, $N_m^{\min}(d,t,j) \ge N_m^{(i)}(t) \ge N^{(i)}(t) - (\sum_{n\ne m}\frac{4}{{\Delta_n^{(i)}}^2})\log N^{(i)}(t)$ for some $i \in A_m$. That is, $N_m^{\min}(d,t,j) \ge \min_{i\in A_m}\{N^{(i)}(t) - (\sum_{n\ne m}\frac{4}{{\Delta_n^{(i)}}^2})\log N^{(i)}(t)\}$. Substituting this into $N_m^{(\min)}(d,t,j) \le \frac{4c_{m,t}^2\log(N^{(j)}(t)/d)}{{\Delta_m^{(j)}}^2}$ from Lemma 8, we see that arm m is pulled by agent j only when $\min_{i\in A_m}\{N^{(i)}(t) - (\sum_{n\ne m}\frac{4}{{\Delta_n^{(i)}}^2})\log N^{(i)}(t)\} \le$

$\frac{4c_{m,t}^2 \log(N^{(j)}(t)/d)}{{\Delta_m^{(j)}}^2}$. ■

Proof [Proof of Lemma 5.]
$E[Regret^{(j)}(T)] = \sum_{m\in M\setminus m_j^*} \Delta_m E[\#$ of agent j's non-optimal arm m pulls before $T]$
$= \sum_{m\in M\setminus m_j^*} \Delta_m \sum_{n=1}^{\infty} E[1_{G_m^{(j)}(S_n^{(j)})} 1_{S_n^{(j)}\le T}]$
$= \sum_{m\in M\setminus m_j^*} \Delta_m \sum_{n=1}^{\infty} E[E[1_{G_m^{(j)}(S_n^{(j)})} 1_{S_n^{(j)}\le T} | S_n^{(j)}]]$
$= \sum_{m\in M\setminus m_j^*} \Delta_m (\sum_{n=1}^{\infty} E[E[1_{G_m^{(j)}(S_n^{(j)})} 1_{S_n^{(j)}\le T} | V_m^{(j)}(S_n^{(j)})^c, S_n^{(j)}] P(V_m^{(j)}(S_n^{(j)})^c | S_n^{(j)}) +$
$E[1_{G_m^{(j)}(S_n^{(j)})} 1_{S_n^{(j)}\le T} | V_m^{(j)}(S_n^{(j)}), S_n^{(j)}] P(V_m^{(j)}(S_n^{(j)}) | S_n^{(j)})])$
$\le \sum_{m\in M\setminus m_j^*} \Delta_m \left(\sum_{n=1}^{\infty} E[P(V_m^{(j)}(S_n^{(j)})^c | S_n^{(j)})] + \sum_{n=1}^{\infty} E[E[1_{G_m^{(j)}(S_n^{(j)})} 1_{S_n^{(j)}\le T} | V_m^{(j)}(S_n^{(j)}), S_n^{(j)}]]\right)$
$\le \sum_{m\in M\setminus m_j^*} \Delta_m \left(\frac{\pi^2}{6} + \sum_{n=1}^{\infty} E[E[1_{G_m^{(j)}(S_n^{(j)})} | V_m^{(j)}(S_n^{(j)}), S_n^{(j)}]\right)$
$= \sum_{m\in M\setminus m_j^*} \Delta_m \left(\frac{\pi^2}{6} + \sum_{n=1}^{\infty} E[P(G_m^{(j)}(S_n^{(j)}) | V_m^{(j)}(S_n^{(j)}), S_n^{(j)}]\right)$
$\le \sum_{m\in M\setminus m_j^*} \Delta_m \left(\frac{\pi^2}{6} + \sum_{n=1}^{\infty} \int_0^{+\infty} g_m^{(j)}(t) dF_n^{(j)}(t)\right)$. ■

Lemma 9. *For $A, B, C > 0$, $Ay - B\ln y < C\ln(\frac{x}{d})$ is satisfied only if $y < -\frac{B}{A}\mathcal{W}_{-1}\left(-\frac{A}{B}(\frac{x}{d})^{-\frac{C}{B}}\right)$ where $\mathcal{W}_{-1}$ denotes the lower branch Lambert W-function.*

Proof [Proof of Lemma 9] For $A, B, C > 0$, $\frac{A}{C}y - \frac{B}{C}\ln y < \ln(\frac{x}{d}) \iff y^{-\frac{B}{C}} e^{\frac{A}{C}y} < (\frac{x}{d}) \iff ye^{-\frac{A}{B}y} > (\frac{x}{d})^{-\frac{C}{B}} \iff -\frac{A}{B}ye^{-\frac{A}{B}y} < -\frac{A}{B}(\frac{x}{d})^{-\frac{C}{B}} \iff -\frac{B}{A}\mathcal{W}_0\left(-\frac{A}{B}(\frac{x}{d})^{-\frac{C}{B}}\right) < y < -\frac{B}{A}\mathcal{W}_{-1}\left(-\frac{A}{B}(\frac{x}{d})^{-\frac{C}{B}}\right)$ where $\mathcal{W}_0$ denotes the principal branch of the Lambert W-function. Therefore, $Ay - B\ln y < C\ln(\frac{x}{d})$ holds only if $y < -\frac{B}{A}\mathcal{W}_{-1}\left(-\frac{A}{B}(\frac{x}{d})^{-\frac{C}{B}}\right)$. ■

In our case, for Lemma 9, $y = N^{(i)}(t)$, $x = N^{(j)}(t)$, $A = 1$, $B = \sum_{n\neq m} \frac{4}{{\Delta_n^{(i)}}^2}$ and $C = \frac{4c_{m,t}^2}{{\Delta_m^{(j)}}^2}$. We define q_{ij} as $q_{ij}(x) = -\frac{B}{A}\mathcal{W}_{-1}\left(-\frac{A}{B}(\frac{x}{d})^{-\frac{C}{B}}\right)$ where we use above parameter values. One can easily check that $\frac{B}{A}\mathcal{W}_{-1}\left(-\frac{A}{B}x^{-\frac{C}{B}}\right)$ is a function growing faster than $\log x$ and slower than x.

Proof [Proof of Theorem 6.] Before proving Theorem 6, we show Lemma 10 first.

Lemma 10. *Suppose that each agent of $i \in A$ arrives independently with i.i.d. 1-subgaussian inter-arrival times with mean θ_i. Every time an agent arrives, it plays according to CFUCB. Then $P(G_m^{(j)}(t)|V_m^{(j)}(t)) \leq g_m^{(j)}(t)$ holds, where $g_m^{(j)}(t) = |A|(\exp(-2\frac{(t-q_{ij}(\lceil \frac{t}{\theta^j-\epsilon^j}\rceil)\theta_{\min})^2}{q_{ij}(\lceil \frac{t}{\theta^j-\epsilon^j}\rceil)}) + \exp(-2\frac{{\epsilon^j}^2}{\theta^j-\epsilon^j}t))$, with $\theta_{\min} =: \min_{i\in A}\theta_i$ and ϵ^j is a parameter to be tuned later.*

Proof [Proof of Lemma 10.]

$$
\begin{aligned}
&P(G_m^{(j)}(t)|V_m^{(j)}(t)) \\
&\quad = P(\{\text{Agent } j \text{ pulls arm } m \text{ when it arrives at time } t\}|V_m^{(j)}(t)) \\
&\quad \overset{(b)}{\leq} P(\min_{i\in A_m}\{N^{(i)}(t) - (\sum_{n\neq m}\frac{4}{{\Delta_n^{(i)}}^2})\log N^{(i)}(t)\} < \frac{4c_{m,t}^2\log(N^{(j)}(t)/d)}{{\Delta_m^{(j)}}^2}) \qquad (9) \\
&\quad \leq \sum_{i\in A_m} P(N^{(i)}(t) - (\sum_{n\neq m}\frac{4}{{\Delta_n^{(i)}}^2})\log N^{(i)}(t) < \frac{4c_{m,t}^2\log(N^{(j)}(t)/d)}{{\Delta_m^{(j)}}^2}) \\
&\quad \leq \sum_{i\in A_m} P(N^{(i)}(t) < q_{ij}(N^{(j)}(t))) \text{ (because of Lemma 9)} \qquad (10) \\
&\quad = \sum_{i\in A_m}\int P(N^{(i)}(t) < q_{ij}(n))\, dF_{N^{(j)}(t)}(n) \\
&\quad \leq \sum_{i\in A}\int P(N^{(i)}(t) < q_{ij}(n))\, dF_{N^{(j)}(t)}(n) \\
&\quad = \sum_{i\in A}\int P(S^{(i)}_{\lceil q_{ij}(n)\rceil} > t)\, dF_{N^{(j)}(t)}(n) \\
&\quad = \sum_{i\in A}(\int_{[0,\frac{t}{\theta^j-\epsilon^j}]} P(S^{(i)}_{\lceil q_{ij}(n)\rceil} > t)\, dF_{N^{(j)}(t)}(n) \\
&\qquad\qquad + \int_{(\frac{t}{\theta^j-\epsilon^j},\infty)} P(S^{(i)}_{\lceil q_{ij}(n)\rceil} > t)\, dF_{N^{(j)}(t)}(n)) \\
&\quad \overset{(c)}{\leq} \sum_{i\in A}\left(P(S^{(i)}_{\lceil q_{ij}(\frac{t}{\theta^j-\epsilon^j})\rceil} > t)\times 1 + 1\times\exp(-2\frac{{\epsilon^j}^2}{\theta^j-\epsilon^j}t)\right) \qquad (11) \\
&\quad = \sum_{i\in A}\left(\exp(-2\frac{(t-\lceil q_{ij}(\frac{t}{\theta^j-\epsilon^j})\rceil\theta_i)^2}{\lceil q_{ij}(\frac{t}{\theta^j-\epsilon^j})\rceil}) + \exp(-2\frac{{\epsilon^j}^2}{\theta^j-\epsilon^j}t)\right) \\
&\quad \overset{(d)}{\leq} \exp(-2\frac{(t-\lceil q_{ij}(\frac{t}{\theta^j-\epsilon^j})\rceil\theta_i)^2}{\lceil q_{ij}(\frac{t}{\theta^j-\epsilon^j})\rceil}) \qquad (12) \\
&\quad \leq |A|\left(\exp(-2\frac{(t-\lceil q_{ij}(\frac{t}{\theta^j-\epsilon^j})\rceil\theta_{\max})^2}{\lceil q_{ij}(\frac{t}{\theta^j-\epsilon^j})\rceil}) + \exp(-2\frac{{\epsilon^j}^2}{\theta^j-\epsilon^j}t)\right). \qquad (13)
\end{aligned}
$$

Above, $\theta_{\max} := \max_{i\in A}\theta_i$ and

- The inequality (b) of (9) follows from Lemma 4.

- The inequality (c) of (11) holds because we apply left tail Hoeffding inequality, i.e.,

$$P\left(S_n^{(j)} \leq n(\theta^j - \epsilon^j)\right) = P\left(N^{(j)}(n(\theta^j - \epsilon^j)) \geq n\right) \leq e^{-2n\epsilon^{j2}}$$
$$\Leftrightarrow P\left(N^{(j)}(t) \geq \frac{t}{\theta^j - \epsilon^j}\right) \leq e^{-2\frac{\epsilon^{j2}}{\theta^j - \epsilon^j}t}$$

  and $\lceil q_{ij}(n) \rceil$ is a increasing function of n.

- The inequality (d) of (12) holds by applying an another version of right tail Hoeffding inequality $P\{S_n \geq n\theta + a\} \leq e^{-2a^2/n}$ for

$$P(S^{(i)}_{\lceil q_{ij}(\frac{t}{\theta^j - \epsilon^j})\rceil} > t) = P(S^{(i)}_{\lceil q_{ij}(\frac{t}{\theta^j - \epsilon^j})\rceil} - \lceil q_{ij}(\frac{t}{\theta^j - \epsilon^j})\rceil\theta_i > t - \lceil q_{ij}(\frac{t}{\theta^j - \epsilon^j})\rceil\theta_i)).$$

■

Now we can prove **Theorem 6**, starting as follows:

$$\begin{aligned}
\int g_m^{(j)} F_{S_n^{(j)}} &= \int_{[0, n(\theta^j - \epsilon))} g_m^{(j)} F_{S_n^{(j)}} + \int_{[n(\theta^j - \epsilon), \infty)} g_m^{(j)} F_{S_n^{(j)}} \\
&\leq g_m^{(j)}(0^+) \times e^{-2n\epsilon^2} + g_m^{(j)}(n(\theta^j - \epsilon)) \times 1 \qquad (14) \\
&= 2|A| e^{-2n\epsilon^2} + g_m^{(j)}(n(\theta^j - \epsilon)) \\
&= 2|A|(\exp(-2n\epsilon^2) + \exp(-2\frac{(n(\theta^j - \epsilon) - \lceil q_{ij}(\frac{n(\theta^j - \epsilon)}{\theta^j - \epsilon^j})\rceil\theta_{\max})^2}{\lceil q_{ij}(\frac{n(\theta^j - \epsilon)}{\theta^j - \epsilon^j})\rceil}) \\
&\qquad + \exp(-2\frac{\epsilon^{j2}}{\theta^j - \epsilon^j} n(\theta^j - \epsilon))) \\
&= 2|A| \left(2\exp(-2n\epsilon^2) + \exp(-2\frac{(n(\theta^j - \epsilon) - \lceil q_{ij}(n)\rceil\theta_{\max})^2}{\lceil q_{ij}(n)\rceil})\right) \\
&\qquad \text{(for simplicity, we fix } \epsilon^j = \epsilon) \\
&= O(\frac{1}{n^2}) \qquad (15)
\end{aligned}$$

Above, (14) holds because $P(S_n^{(j)} \leq n(\theta^j - \epsilon)) \leq e^{-2n\epsilon^2}$ and that $g_m^{(j)}$ is decreasing function), (15) holds because

$$\begin{aligned}
&\left(n(\theta^j - \epsilon)) - \lceil q_{ij}(n)\rceil\theta_{\max}\right)^2 \geq \lceil q_{ij}(n)\rceil^2 \text{ for all } n \geq N \text{ for some } N \qquad (16) \\
&\quad (\Rightarrow)(n(\theta^j - \epsilon)) - \lceil q_{ij}(n)\rceil\theta_{\max})^2 \geq \log(n)\lceil q_{ij}(n)\rceil \text{ for all } n \geq N \qquad (17) \\
&\quad (\Rightarrow) \exp(-2\frac{(n(\theta^j - \epsilon) - \lceil q_{ij}(n)\rceil\theta_{\max})^2}{\lceil q_{ij}(n)\rceil}) = O(\frac{1}{n^2})
\end{aligned}$$

where (16) follows from $\lceil q_{ij}(n)\rceil = o(n))$ and (17) follows from $\log(n) = o(q_{ij}(n))$

■

Proof [Proof of Theorem 7.] Before proving Theorem 7, we prove the Lemma 11.

Lemma 11. *Suppose that each agent i of A arrive independently with iid exponentially distributed inter-arrival times with λ_i. Every time an agent arrives, we play CFUCB. Then $P(G_m^{(j)}(t)|V_m^{(j)}(t)) \leq g_m^{(j)}(t)$ holds, where $g_m^{(j)}(t) = |A|(\exp(-\frac{(\lambda_{\min}t - q_{ij}((\lambda_j+\epsilon_j)t))^2}{2\lambda_{\min}t}) + \exp(-\frac{{\epsilon_j}^2}{2\lambda_j}t))$ and $\lambda_{\min} = \min_{i\in A}\lambda_i$ and ϵ_j is a parameter to be tuned later.*

Proof [Proof of Lemma 11.] Again, as in Lemma 10,

$$
\begin{aligned}
&P(G_m^{(j)}(t)|V_m^{(j)}(t)) \\
&\quad = P(\{ \text{ Agent } j \text{ pulls arm } m \text{ when it arrives at time } t\}|V_m^{(j)}(t)) \\
&\quad \leq P(\min_{i\in A_m}\{N^{(i)}(t) - (\sum_{n\neq m}\frac{4}{{\Delta_n^{(i)}}^2})\log N^{(i)}(t)\} < \frac{4c_{m,t}^2\log(N^{(j)}(t)/d)}{{\Delta_m^{(j)}}^2}) \quad (18) \\
&\quad \leq \sum_{i\in A_m} P(N^{(i)}(t) - (\sum_{n\neq m}\frac{4}{{\Delta_n^{(i)}}^2})\log N^{(i)}(t) < \frac{4c_{m,t}^2\log(N^{(j)}(t)/d)}{{\Delta_m^{(j)}}^2}) \\
&\quad = \sum_{i\in A_m} P(N^{(i)}(t) < q_{ij}(N^{(j)}(t)))(because of Lemma 9) \quad (19) \\
&\quad = \sum_{i\in A_m}\int P(N^{(i)}(t) < q_{ij}(n))\, dF_{N^{(j)}(t)}(n) \\
&\quad \leq \sum_{i\in A}\int P(N^{(i)}(t) < q_{ij}(n))\, dF_{N^{(j)}(t)}(n) \\
&\quad = \sum_{i\in A}(\int_{[0,(\lambda_j+\epsilon_j)t]} P(N^{(i)}(t) < q_{ij}(n))\, dF_{N^{(j)}(t)}(n) \\
&\qquad\qquad + \int_{((\lambda_j+\epsilon_j)t,\infty)} P(N^{(i)}(t) < q_{ij}(n))\, dF_{N^{(j)}(t)}(n)) \\
&\quad \leq \sum_{i\in A}\left(P(N^{(i)}(t) < q_{ij}((\lambda_j+\epsilon_j)t)) \times 1 + 1 \times e^{-\frac{{\epsilon^j}^2}{2\lambda_j}t}\right) \quad (20) \\
&\quad \leq \sum_{i\in A}\left(\exp(-\frac{(\lambda_i t - q_{ij}((\lambda_j+\epsilon_j)t))^2}{2\lambda_i t}) + \exp(-\frac{{\epsilon^j}^2}{2\lambda_j}t)\right) \quad (21) \\
&\quad \leq |A|\left(\exp(-\frac{(\lambda_{\min} t - q_{ij}((\lambda_j+\epsilon_j)t))^2}{2\lambda_{\min} t}) + \exp(-\frac{{\epsilon^j}^2}{2\lambda_j}t)\right) \quad (22)
\end{aligned}
$$

Above, $\lambda_{\min} = \min_{i\in A}\lambda_i$ and

- (18) follows from Lemma 4.
- (20) holds because $P\left(N^{(j)}(t) \geq (\lambda_j+\epsilon_j)t\right) \leq e^{-\frac{{\epsilon^j}^2}{2\lambda_j}t}$ from the Poisson concentration right tail bound $P(X \geq \lambda + x) \leq e^{-\frac{x^2}{2\lambda}}$ Pollard (2015) and $q_{ij}(n)$ being an increasing function of n

- (21) holds because $P(X \leq \lambda - x) \leq e^{-\frac{x^2}{2\lambda}}$ from the Poisson concentration left tail bound Pollard (2015)

■

Now we can show **Thoerem 7**, starting as follows:

$$
\begin{aligned}
&\int g_m^{(j)} F_{S_n^{(j)}} \\
&= \int_{[0, \frac{n-1}{\lambda_j+\epsilon_j})} g_m^{(j)} F_{S_n^{(j)}} + \int_{[\frac{n-1}{\lambda_j+\epsilon_j}, \infty)} g_m^{(j)} F_{S_n^{(j)}} \\
&\leq g_m^{(j)}(0^+) \times \exp(-\frac{\epsilon_j^2}{2\lambda_j} \frac{n-1}{\lambda_j+\epsilon_j}) + g_m^{(j)}(\frac{n-1}{\lambda_j+\epsilon_j}) \times 1 \qquad (23) \\
&= 2|A| \exp(-\frac{\epsilon_j^2}{2\lambda_j} \frac{n-1}{\lambda_j+\epsilon_j}) + g_m^{(j)}(\frac{n-1}{\lambda_j+\epsilon_j}) \\
&= 3|A| \exp(-\frac{\epsilon_j^2}{2\lambda_j} \frac{n-1}{\lambda_j+\epsilon_j}) + |A| \exp(-\frac{(\lambda_{\min} \frac{n-1}{\lambda_j+\epsilon_j} - q_{ij}(n-1))^2}{2\lambda_{\min} \frac{n-1}{\lambda_j+\epsilon_j}}) \\
&= O(\frac{1}{n^2}) \qquad (24)
\end{aligned}
$$

Above, $\lambda_{\min} = \min_{i \in A} \lambda_i$ and

- (23) holds because $P(S_n^{(j)} < \frac{n-1}{\lambda_j+\epsilon_j})$
  $\leq P(S^{(j)}_{\lceil(\lambda_j+\epsilon_j)T\rceil} \leq T) = P(N^{(j)}(T) \geq \lceil(\lambda_j+\epsilon_j)T\rceil) = P(N^{(j)}(T) \geq (\lambda_j+\epsilon_j)T)$
  $\leq \exp\left(-\frac{{\epsilon_j}^2}{2\lambda_j}T\right) \leq \exp\left(-\frac{\epsilon_j^2}{2\lambda_j}\frac{n-1}{\lambda_j+\epsilon_j}\right)$ and $g_m^{(j)}(t)$ is a decreasing function of t.

- (24) holds because of the following:
  We want to find N such that $\frac{(\lambda_{\min}\frac{n-1}{\lambda_j+\epsilon_j} - q_{ij}(n-1))^2}{2\lambda_{\min}\frac{n-1}{\lambda_j+\epsilon_j}} \geq 2\ln(n)$ for all $n \geq N$, i.e., $(\frac{\lambda_{\min}}{\lambda_j+\epsilon_j}(n-1) - q_{ij}(n-1))^2 \geq 4\frac{\lambda_{\min}}{\lambda_j+\epsilon_j}(n-1)\ln(n)$ for all $n \geq N$. We can show this by instead showing $\frac{\lambda_{\min}}{\lambda_j+\epsilon_j}((n-1) - q_{ij}(n-1))^2 \geq 4n\ln(n)$, or $\frac{((n-1)-q_{ij}(n-1))^2}{n\ln(n)} \geq 4\frac{\lambda_j+\epsilon_j}{\lambda_{\min}}$ for all $n \geq N$. Note that for some $\beta > 1$, $\beta(n-1) > n$ holds for all $n \geq N_1$ for some N_1. Therefore showing $\frac{(n-q_{ij}(n))^2}{\beta n \ln(\beta n)} \geq 4\frac{\lambda_j+\epsilon_j}{\lambda_{\min}}$ for all $n \geq N'$ is what we want. Now note that $\frac{d}{dn}\left(\frac{\left(n+\mathcal{W}_{-1}\left(-\frac{1}{n}\right)\right)^2}{\beta n \log(\beta n)}\right) =$
  $\left(\left(\mathcal{W}_{-1}\left(-\frac{1}{n}\right)+n\right)\left(\mathcal{W}_{-1}\left(-\frac{1}{n}\right)^2(-(\log(n\beta)+1)) + \mathcal{W}_{-1}\left(-\frac{1}{n}\right)((n-3)\log(n\beta) - n - 1) + n(\log(n\beta) - 1)\right)\right/ \left(n^2\beta\left(\mathcal{W}_{-1}\left(-\frac{1}{n}\right)+1\right)\log^2(n\beta)\right) > 0$ for $n > 5$. This means that $\frac{(n-q_{ij}(n))^2}{\beta n \ln(\beta n)}$ is monotone strictly increasing, and therefore there exists some N' such that $\frac{(n-q_{ij}(n))^2}{\beta n \ln(\beta n)} \geq 4\frac{\lambda_j+\epsilon_j}{\lambda_{\min}}$ for all $n \geq N'$. We achieve $\exp(-\frac{(\lambda_{\min}\frac{n-1}{\lambda_j+\epsilon_j} - q_{ij}(n-1))^2}{2\lambda_{\min}\frac{n-1}{\lambda_j+\epsilon_j}}) = O(\frac{1}{n^2})$.

■