MULTI-AGENT LINEAR CONTEXTUAL BANDITS WITH BOUNDED $O(1)$ REGRET

Anonymous authors
Paper under double-blind review

ABSTRACT

Asymptotically unbounded regret of order $O(\sqrt{T})$ has been proved to be the lowest possible regret order that can be achieved in typical linear contextual bandit settings. Here we present a linear contextual bandit setting with repetitive arrivals of a set of agents where bounded, i.e., $O(1)$, expected regret can be achieved for each agent. We provide a novel Counterfactual UCB (CFUCB) policy where agents benefit from the experiences of other agents. It is shown that sharing of information is a Subgame Perfect Nash Equilibrium for the agents with respect to the order of the regret, which results in each agent realizing bounded regret. Personalized recommender systems and adaptive experimentation are two important applications.

1 INTRODUCTION

In personalized recommender platforms, users (hereafter called agents) repeatedly explore the available choices (hereafter called arms). The platform then obtains increasing information with time about the rewards each agent gets from its choices. Moreover, the platform learns from the experiences of not just one agent but from the experiences of *all agents*. Could it put that totality of information to good use in helping each agent's exploration? The standard linear contextual bandit model does not feature repeated arrivals of the same agents and has an unbounded regret of $\Theta(\sqrt{T})$ (Auer (2002); Chu et al. (2011); Abbasi-Yadkori et al. (2011); Li et al. (2017)). We propose a model of *recurring* linear contextual bandits in which the expected regret is bounded, i.e., $O(1)$, under the condition that the number of agents is relatively large compared to the number of arms.

In the recurring linear contextual bandit framework each agent/each arm in a set A has its own fixed feature vector where the arm feature vectors are unknown, and the mean reward of agent j playing arm m is the inner product of their two feature vectors. Adaptive medical prescriptions is a good example of a linear contextual bandit problem where the number of agents (patients) is much larger than the number of arms (possible treatments). Each patient (agent) is represented by his/her medical record that is embedded into a known feature vector. There is a set of drugs (arms), with the individual effect of each drug unknown. As in other contextual linear bandit problems, the physician learns about other arms beyond the one played played, due to the linear dependences among their feature vectors. We show that after some time agents become fully relieved from further exploration, leading to bounded regret for all agents.

Another issue is that while a mediator (e.g., a physician, YouTube) may coordinate agents' exploration (e.g., by recommendations), agents are usually selfish and strategic, so they might not follow the suggested policy or might not report properly (e.g., refusing to provide the physician their previous medical history, or turn on the privacy mode while using YouTube). We show that all agents truthfully reporting their private information initially, conforming to the policy provided, and truthfully reporting their subsequent arm-pulling experiences to the mediator, constitutes a Subgame Perfect Nash Equilibrium (SPNE) for the agents with asymptotically indifferent preference. This means that the first-best outcome of bounded regret is achieved under this equilibrium.

The rest of this paper is organized as follows. In Section 2, we propose the recurring linear contextual bandit problem and explore the condition under which bounded regret is shown to be achieved. Section 3 then explores how our problem relates to other existing problems. The main algorithm is proposed in Section 4 and analyzed in Section 5. We then investigate the proposed algorithm's

robustness with respect to incentive constraints in Section 6 and noise in Section 7. Finally, we empirically demonstrate the bounded regret result using a simulation experiment in Section 8.

2 THE PROBLEM SETTING

Agent arrivals Let A denote the set of agents. Each agent repeatedly arrives according to a renewal process where the inter-arrival times are independent and identically distributed (i.i.d.). Two cases are considered: When all inter-arrival times are (i) i.i.d. subgaussian with a density on the real line, or (ii) i.i.d. exponential. Denote the time of the n-th 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 arrivals by $N^{(j)}(t)$. That is, $\{S_n^{(j)} \leq t\} = \{N^{(j)}(t) \geq n\}$.

Feature Vectors of Agents and Arms Denote the set arms by M. Each of the agents and each of the arms is associated with a feature vector of dimension d. Denote the feature vector of agent j by $\alpha^{(j)}$, and the feature vector of arm m by β_m. 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). Under the cooperative setting, we assume, as of now, that each agent's feature vector is common knowledge. Later we will show that sharing this constitutes a SPNE with respect to the order of the regret for each agent.

Rewards and objective An agent gets to pull an arm every time that it arrives. Therefore $N^{(j)}(t)$ is also the total number of pulls over all arms by agent j until time t. We further denote by $N_m^{(j)}(t)$ the number of agent j's pulls of arm m until time t. Agent j receives a random reward with mean $\mu_m^{(j)} := \alpha^{(j)}\beta_m$ when it pulls arm m. The kth reward of agent j from arm m 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)). Under the cooperative setup assumed for now, and later shown to be a SPNE, the reward observation immediately becomes common knowledge of all agents.

For each agent $j \in A$, define $m_j^* \in M$ as an arbitrarily chosen arm that satisfies $\mu_{m_j^*}^{(j)} \geq \mu_m^{(j)} \ \ \forall m \in M$. We define $\Delta_m^{(j)} := \mu_{m_j^*}^{(j)} - \mu_m^{(j)}$ and $A_n := \{j \in A : m_j^* = n\}$. Note that $\{A_n\}_{n \in M}$ partitions A. Denote the arm pulled by agent j at its n-th arrival by $m_j(n)$. Then the finite time regret of agent j until time T is $Regret^{(j)}(T) := \sum_{n=1}^{N^{(j)}(T)} \Delta_{m_j(n)}^{(j)} = \sum_{n=1}^{N^{(j)}(T)} (\mu_{m_j^*}^{(j)} - \mu_{m_j(n)}^{(j)})$.

Agent set size constraint 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:

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

How large should the number of agents be in order to make this condition as probable as desired? Theorem 1 answers this question which is a previously unaddressed version of the Double Dixie cup problem Newman (1960):

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 parameter η is a parameter to be tuned. 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.

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.

3 COMPARISON TO RELATED WORKS ON LINEAR CONTEXTUAL BANDITS AND MULTI-AGENT EXPLORATION

First we point out some differences with usual formulations and the motivation. In typical linear contextual bandits Auer (2002); Chu et al. (2011); Abbasi-Yadkori et al. (2011), at each time-step, an arrival of agent is modeled as a *context* newly born, which actually contains one feature vector for each arm for that agent. In our setup, there is a finite set of agents A, and an agent $j \in A$ has a known feature vector $\alpha^{(j)}$. Typically, there is no concept of agents' repeated arrivals or record of past plays. In our setup, each agent in A repeatedly arrives, and the platform can thus keep track of an individual agent's past record, as many real-world platforms do, e.g., YouTube. In typical linear contextual bandits, there is only a *single* vector θ that is unknown. In our setup, there is a whole set of vectors, one for each arm, that is unknown. For drug prescription, for example, there is a separate unknown feature vector for every drug.

The closest to our problem's formulation are cooperative linear bandits and multi-task linear bandits Soare & Pineau (2018) and Moradipari et al. (2022), where the agent side vectors are unknown, and the arm side vectors are known. In our formulation, it is the opposite. In Soare & Pineau (2018), only an empirical analysis is provided. In Moradipari et al. (2022), a special case where the agents share the effect of the arm pulled at each time-step is analyzed to provide a regret of $\mathcal{O}\left(\sqrt{T/N}(\log T)^2\right)$. In our paper, we suggest a linear contextual bandit problem and achieve $O(1)$ regret.

The decentralized version of the proposed algorithm (see Appendix B) that conveys the concept of each agent uploading local information to a mediator that coordinates information is also related to Federated Learning Huang et al. (2021). A notable result proved in Huang et al. (2021) is that a tight minimax regret performance of $O(\sqrt{T \log T})$ is achieved in spite of keeping information private. Our paper is motivated by a different consideration: How can agents obtain a *bounded regret of* $O(1)$? We show that this is achieved by the fact that sharing information constitutes a Subgame Perfect Nash Equilibrium (SPNE). This is an equilibrium where at every stage of the game no agent can strictly benefit by lying or not conforming Fudenberg & Levine (1983); Selten (1965). We establish that everyone reporting their private context to the mediator at time 0, following the CFUCB policy and reporting the rewards truthfully afterward does constitute a Subgame Perfect Nash equilibrium (SPNE) since $O(1)$ regret cannot be improved. There are some previous works on incentive constraints in coordinating exploration: Shi et al. (2021) shows that identification of a social planner's best arm does not require extra payment when you must incentivize individually rational agents (Jackson (2014)) instead of forcing them. In comparison, our paper shows that the context information enables an individual agent's best arm to be chosen after a finite time. Immorlica et al. (2019) considers incentive-compatible coordination of agents' exploration with a setting opposite to ours: the context is private, but the mean reward associated with each arm is known.

4 THE COUNTERFACTUAL UCB ALGORITHM

We now introduce the Counterfactual-UCB (CFUCB) algorithm 1 that achieves bounded regret.

The high-level idea of this algorithm is as follows. In a typical UCB-based algorithm (e.g., Auer (2002) for the multi-armed bandit problem, an agent forms a confidence interval based solely on its own experience, which we call the *self-experienced confidence interval*. In our problem, an agent can also construct a confidence interval for each arm based on the experience of *other* agents in addition to itself. We call this the *counterfactual confidence interval*.

The main idea that enables the bounded expected regret result is somewhat related to the idea of a technique called imputation, which has recently been popularized in the causal inference community (Abadie & Imbens (2011); Stuart (2010); Athey & Imbens (2015); Cunningham (2021); Ye et al. (2020)): Suppose that arm m is not the optimal arm for agent j, but it is optimal for a set of other agents, say A_m. If the total number of agents is large enough, we can express the feature vector of agent j by a linear combination of feature vectors of agents in A_m. Then this linear combination relationship can be used to simulate a counterfactual estimate of agent j's experience on arm m using the experiences of A_m on arm m. If we can further impute the uncertainty of that estimate, the agent j may be fully exempt from the burden of further exploring arm m after a finite time.

Self-experienced Confidence interval. 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. Then the width $w_m^{(j)}(t)$ of the lone wolf confidence interval is chosen as $w_m^{(j)}(t) := \sqrt{\frac{\log N^{(j)}(t)}{N_m^{(j)}(t)}}$. Defining $\overline{X}_m^{(j)}(t) + w_m^{(j)}(t)$ as $ucb_m^{(j)}(t)$, and $\overline{X}_m^{(j)}(t) - w_m^{(j)}(t)$ as $lcb_m^{(j)}(t)$, the *lone wolf confidence interval* is $CI_m^{(j)}(t) := \Big(lcb_m^{(j)}(t), ucb_m^{(j)}(t)\Big)$.

Counterfactual Confidence interval. Define $A_m(d,t) := \{j \in A : |\{i \in A : N_m^{(i)}(t) > N_m^{(j)}(t)\}| < d\}$. This set includes the 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 $|A_m(d{+}1,t)| \geq d{+}1$. Now arbitrarily choose a d-size subset $E_m^{(j)}(t)$ of $A_m(d+1,t) \setminus j$. Since the feature vectors of the d-size subset of A are linearly independent, $\alpha^{(j)} = \sum_{i \in E_m^{(j)}(t)} a_i^{(j)} \alpha^{(i)}$ for some coefficients $\{a_i^{(j)}\}$, and consequently $\mu_m^{(j)} = \sum_{i \in E_m^{(j)}(t)} a_i^{(j)} \mu_m^{(i)}$. 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 reward of agent j for arm m*. The width $\widehat{w}_m^{(j)}(t)$ of the corresponding counterfactual confidence interval is chosen as $\widehat{w}_m^{(j)}(t) := \sqrt{\frac{\log(N^{(j)}(t)/d)}{N_m^{(\min)}(d,t,j)/c_{m,t}^2}}$, where $c_{m,t} := \sum_{i \in E_m^{(j)}(t)} |a_i^{(j)}|$, and $N_m^{(\min)}(d,t,j) := \min_{i \in E_m^{(j)}(t)} N_m^{(i)}(t)$. If we define $\widehat{X}_m^{(j)}(t) + \widehat{w}_m^{(j)}(t)$ as $\widehat{ucb}_m^{(j)}(t)$, and $\widehat{X}_m^{(j)}(t) - \widehat{w}_m^{(j)}(t)$ as $\widehat{lcb}_m^{(j)}(t)$, then the *counterfactual confidence interval* is defined as $\widehat{CI}_n^{(i)}(t) := (\widehat{lcb}_m^{(j)}(t), \widehat{ucb}_m^{(j)}(t))$.

The corresponding *Self-experienced upper confidence bound* and *counterfactual upper confidence bound* are

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

The Counterfactual UCB (CFUCB) algorithm We introduce a notion of *epochs*. Define $S := \cup_{i \in A} \{S_n^{(i)}\}_{n \in \mathbb{N}}$, the set of all arrival times of all agents. The elements of S can be ordered as 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, denote by s_k the time of the kth *arrival epoch*, or simply the kth *epoch*. 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 the identity of the agent that arrives at each epoch. Ties between agents arriving at the same time can be be broken arbitrarily, while the probability of simultaneous arrivals at subsequent times is zero due to the existence of a density for inter-arrival times. Given $\{a_k\}_{k \in \mathbb{N}}$, denote the index of the arm pulled by agent a_k at epoch k by m_k, and the corresponding accrued reward 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 at epoch k). Recall that $X_m^{(j)}(n)$ denotes the n-th reward of agent j from arm m. $N_m^{(j)}(t)$ denotes agent j's number of pulls of arm m until time t.

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 $ucb_m^{(a_k)}(s_k)$ (Self-experienced upper confidence bound) according to the Eq (3)
5 Compute $\widehat{ucb}_m^{(a_k)}(s_k)$ (counterfactual upper confidence bound) according to the Eq (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**

Algorithm 1 describes the pseudocode of the CFUCB Algorithm. The only difference between CFUCB and UCB is that arm j at time t chooses the arm with largest $\widetilde{ucb}_m^{(j)}(t)$, not $ucb_m^{(j)}(t)$.

5 THE ANALYSIS OF CFUCB

We first start by describing how the confidence intervals are chosen. 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. The proofs are deferred to Appendix A.

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

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

Now we are ready to derive the condition for the agent j to pull a non-optimal arm m in Lemma 4. Lemma 4 is the key result in that it provides the intuition about why bounded regret is achieved.

Note that as a consequence of Lemmas 2 and 3, at time t, for every arm n and every agent i, the Self-experienced confidence interval $CI_n^{(i)}(t)$ and the counterfactual confidence interval $\widehat{CI}_n^{(i)}(t)$ 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* $\mu_n^{(i)}$ *for all* $i \in A$ *and* $n \in M$*, 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*

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

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.

Lemma 4 is based on the following Lemma 5.

Lemma 5. *Under the same conditions as in Lemma 4, agent* j *pulls arm* m *only if* $\min\left(2\sqrt{\frac{\log N^{(j)}(t)}{N_m^{(j)}(t)}}, 2\sqrt{\frac{\log(N^{(j)}(t)/d)}{N_m^{(\min)}(d,t,j)/c_{m,t}^2}}\right) \geq \Delta_m^{(j)}$*. That is, both* $N_m^{(j)}(t) \leq \frac{4 \log N^{(j)}(t)}{{\Delta_m^{(j)}}^2}$ *and* $N_m^{(\min)}(d,t,j) \leq \frac{4c_{m,t}^2 \log(N^{(j)}(t)/d)}{{\Delta_m^{(j)}}^2}$ *must hold for agent* j *to pull arm* m*.*

Proof of Lemma 5. Denote the optimal arm for agent j as arm m_j^*. According to Algorithm 1, $\{$Agent j pulls arm $m\} \subseteq \{\widetilde{\text{ucb}}_m^{(j)}(t) \geq \widetilde{\text{ucb}}_{m_j^*}^{(i)}(t)\}$. Note that $\widetilde{\text{lcb}}_m^{(j)}(t) \leq \mu_m^{(j)} \leq \widetilde{\text{ucb}}_m^{(j)}(t)$ and $\widetilde{\text{lcb}}_{m_j^*}^{(j)}(t) \leq \mu_{m_j^*}^{(j)} \leq \widetilde{\text{ucb}}_{m_j^*}^{(j)}(t)$ holds according to the assumptions of Lemma 4. Therefore, under the assumption of Lemma 4, $\{$Agent j pulls arm $m\} \subseteq \{\widetilde{\text{lcb}}_m^{(j)}(t) \leq \mu_m^{(j)}, \mu_m^{(j)} \leq \mu_{m_j^*}^{(j)}, \mu_{m_j^*}^{(j)} \leq \text{ucb}_{m_j^*}^{(j)}(t), \widetilde{\text{ucb}}_{m_j^*}^{(j)}(t) \leq \widetilde{\text{ucb}}_m^{(j)}(t)\} = \{\widetilde{\text{lcb}}_m^{(j)}(t) \leq \mu_m^{(j)} \leq \mu_{m_j^*}^{(j)} \leq \widetilde{\text{ucb}}_m^{(j)}(t)\} = \{\mu_m^{(j)}, \mu_{m_j^*}^{(j)} \in CI_m^{(j)}(t) \cap \widehat{CI}_m^{(j)}(t)\}$. Note that $\{\mu_m^{(j)}, \mu_{m_j^*}^{(j)} \in CI_m^{(j)}(t) \cap \widehat{CI}_m^{(j)}(t)\} \subseteq \{\min(2w_m^{(j)}(t), 2\widehat{w}_m^{(j)}(t)) \geq \Delta_m^{(j)}\}$. Therefore, under the assumptions of Lemma 4, agent j pulls arm m only if $\min(2w_m^{(j)}(t), 2\widehat{w}_m^{(j)}(t)) \geq \Delta_m^{(j)}$ holds. Combining this with Lemma 2 and 3 yields the result. □

As can be seen in the proof of Lemma 5, by using Algorithm 1 it is assured that the arm m is pulled by agent j only if both $\mu_m^{(j)}$ and $\mu_{m_j^*}^{(j)}$ are included in the intersection of Self-experienced confidence

interval $CI_m^{(j)}(t)$ and the counterfactual confidence interval $\widehat{CI}_m^{(j)}(t)$. If any of them shrinks and cannot include both $\mu_m^{(j)}$ and $\mu_{m_j^*}^{(j)}$ anymore, agent j won't pull the arm m anymore.

Proof of Lemma 4. Fix agent j and arm m. Note that for any arm $i \in A$, $N_m^{(i)}(t) = N^{(i)}(t) - \sum_{n \in M \setminus m} N_n^{(i)}(t)$. Let t^n be the last time prior to t at which a non-optimal arm n is played by agent i. Then $N_n^{(i)}(t) = N_n^{(i)}(t^n) \leq \frac{4 \log N^{(i)}(t^n)}{\Delta_n^{(i)2}} \leq \frac{4 \log N^{(i)}(t)}{\Delta_n^{(i)2}}$ holds by Lemma 5. Therefore, for agent $i \in A_m$, for arm m, $N_m^{(i)}(t) \geq N^{(i)}(t) - (\sum_{n \neq m} \frac{4}{\Delta_n^{(i)2}}) \log N^{(i)}(t)$. By the assumption (1), $|A_m| \geq d+1$, and $N_m^{\min}(d,t,j) \geq N_m^{(i)}(t)$ for some $i \in A_m$. Therefore, $N_m^{\min}(d,t,j) \geq N_m^{(i)}(t) \geq N^{(i)}(t) - (\sum_{n \neq m} \frac{4}{\Delta_n^{(i)2}}) \log N^{(i)}(t)$ for some $i \in A_m$. That is, $N_m^{\min}(d,t,j) \geq \min_{i \in A_m} \{N^{(i)}(t) - (\sum_{n \neq m} \frac{4}{\Delta_n^{(i)2}}) \log N^{(i)}(t)\}$. Substituting this into $N_m^{(\min)}(d,t,j) \leq \frac{4c_{m,t}^2 \log(N^{(j)}(t)/d)}{\Delta_m^{(j)2}}$ from Lemma 5, it can be seen that arm m is pulled by agent j only when $\min_{i \in A_m} \{N^{(i)}(t) - (\sum_{n \neq m} \frac{4}{\Delta_n^{(i)2}}) \log N^{(i)}(t)\} \leq \frac{4c_{m,t}^2 \log(N^{(j)}(t)/d)}{\Delta_m^{(j)2}}$. □

Lemma 6 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 of the following Lemma 6 is deferred to Appendix A.

Lemma 6. *Denote the event $\{$Agent j arrives at time t and pulls a non-optimal arm $m\}$ by $G_m^{(j)}(t)$, and the event $\{\mu_n^{(i)} \in CI_n^{(i)}(t) \cap \widehat{CI}_n^{(i)}(t)\ \forall i \in A_m, n \in M\}$ as $V(t)$. Suppose that there is a function $g_m^{(j)}(t)$ such that $P(G_m^{(j)}(t)|V(t)) \leq g_m^{(j)}(t)$. 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 $F_n^{(j)}(t) := P(S_n^{(j)} \leq t)$.*

Showing $\sum_{n=1}^{\infty} \int_0^{+\infty} g_m^{(j)}(t) dF_n^{(j)}(t) < \infty$ will yield the result on bounded expected regret. The strategy of the proof is to show that $\int_0^{+\infty} g_m^{(j)}(t) dF_n^{(j)}(t) = O(\frac{1}{n^2})$ holds for the two arrival process models: 1) agents arrive according to sub-Gaussian inter-arrival times (Section 5.1) and 2) agents arrive according to exponential inter-arrival times (Section 5.2). Before discussing how (4) of Lemma 4 can be used, in Lemma 7 we make an observation on the functional form of (4).

Lemma 7. *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 of the Lambert W-function (Corless et al. (1996)).*

Proof of Lemma 7. For $A, B, C > 0$, $\frac{A}{C} y - \frac{B}{C} \ln y < \ln(\frac{x}{d}) \iff y^{-\frac{B}{C}} e^{\frac{A}{C} y} < (\frac{x}{d}) \iff y e^{-\frac{A}{B} y} > (\frac{x}{d})^{-\frac{C}{B}} \iff -\frac{A}{B} y e^{-\frac{A}{B} y} < -\frac{A}{B} (\frac{x}{d})^{-\frac{C}{B}} \iff -\frac{B}{A} \mathcal{W}_0 \left(-\frac{A}{B} (\frac{x}{d})^{-\frac{C}{B}} \right) < y < -\frac{B}{A} \mathcal{W}_{-1} \left(-\frac{A}{B} (\frac{x}{d})^{-\frac{C}{B}} \right)$ 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 the present case, $y = N^{(i)}(t)$, $x = N^{(j)}(t)$, $A = 1$, $B = \sum_{n \neq m} \frac{4}{\Delta_n^{(i)2}}$ and $C = \frac{4c_{m,t}^2}{\Delta_m^{(j)2}}$. 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 the 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.

5.1 Bounded expected regret result for the agents with sub-Gaussian inter-arrival times

Lemma 8. *Suppose that each agent $i \in A$ arrives independently with i.i.d. 1-subgaussian inter-arrival times with mean θ_i, plays according to CFUCB. Then $P(G_m^{(j)}(t)|V(t)) \leq g_m^{(j)}(t)$ holds,*

where $g_m^{(j)}(t) = |A|(\exp(-2\frac{(t-q_{ij}(\lceil \frac{t}{\theta^j-\epsilon^j} \rceil)\theta_{\max})^2}{q_{ij}(\lceil \frac{t}{\theta^j-\epsilon^j} \rceil)}) + \exp(-2\frac{{\epsilon^j}^2}{\theta^j-\epsilon^j}t))$*, with* $\theta_{\max} =: \max_{i\in A}\theta_i$ *and* ϵ^j *is a parameter to be tuned later.*

The proof of Lemma 8 is a bit technical is deferred to Appendix A.

Theorem 9. *Suppose that each agent* $i \in A$ *arrives independently with i.i.d. 1-subgaussian inter-arrival times with mean* θ_i*. Then with* $g_m^{(j)}$ *defined as in Lemma 8,* $E[Regret(T)] \leq \sum_{m\in M}\Delta_m\left(\frac{\pi^2|A||M|}{6} + \sum_{n=1}^{\infty}\int_0^{+\infty} g_m^{(j)}(t)dF_n^{(j)}(t)\right) < \infty$ *for all* T *under CFUCB.*

Proof. From $\int_0^{+\infty} g_m^{(j)}(t)dF_n^{(j)}(t) = 2|A|(2\exp(-2n\epsilon^2) + \exp(-2\frac{(n(\theta^j-\epsilon))-q_{ij}(n)\theta_{\min})^2}{q_{ij}(n)})) = O(\frac{1}{n^2})$ where $\theta_{\min} = \min_{i\in A}\theta_i$, the result follows. See Appendix A for the details. □

5.2 Bounded expected regret result for the agents with exponential inter-arrival times

Lemma 10. *Suppose that each agent* i *of* A *arrives independently with* $i.i.d.$ *exponentially distributed inter-arrival times with Mean* $\frac{1}{\lambda_i}$*. Every time an agent arrives, it plays according to CFUCB. Then* $P(G_m^{(j)}(t)|V(t)) \leq g_m^{(j)}(t)$ *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.*

The proof of Lemma 10 is a bit technical and so we defer it to Appendix A.

Theorem 11. *Suppose that each agent* $i \in A$ *arrives independently with* $i.i.d.$ *exponentially distributed inter-arrival times with* λ_i*, and employs the CFUCB Policy. Then with* $g_m^{(j)}$ *defined as in Lemma 10,* $E[Regret(T)] \leq \sum_{m\in M}\Delta_m\left(\frac{\pi^2|A||M|}{6} + \sum_{n=1}^{\infty}\int_0^{+\infty} g_m^{(j)}(t)dF_n^{(j)}(t)\right) < \infty$ *for all* T*.*

Proof. The result follows from $\int_0^{+\infty} g_m^{(j)}(t)dF_n^{(j)}(t) \leq 3|A|\exp(-\frac{\epsilon_j^2}{2\lambda_j}\frac{n-1}{\lambda_j+\epsilon_j}) + |A|\exp(-\frac{(\lambda_{\min}\frac{n-1}{\lambda_j+\epsilon_j}-q_{ij}(n-1))^2}{2\lambda_{\min}\frac{n-1}{\lambda_j+\epsilon_j}}) = O(\frac{1}{n^2})$ where $\lambda_{\min} = \min_{i\in A}\lambda_i$. See the Appendix A for the details. □

6 Non-cooperative agents and Truthtelling

Now we turn to the non-cooperative case. Users of most real-world platforms are generally selfish and not necessarily cooperative. So far in the cooperative setup, there were the following two implicit truthfulness assumptions on each agent's behavior which will not be true in the non-cooperative case:

T1. Each agent truthfully shares its feature vector (to the mediator) at the very beginning.

T2. Each agent follows the CFUCB policy and truthfully shares every arm-pulling result (to the mediator) as it happens.

The idea of the decentralized CFUCB algorithm (see Appendix B for details) is as follows. When agent j arrives, the mediator assumes that all the agents have been conforming to $T1$ and $T2$ (e.g., all patients disclosing their previous medical records to the physician and reporting their progress accurately afterwards). The mediator computes $\{\widehat{ucb}_m^{(j)}(t)\}_{m\in M}$ (agent j's counterfactual UCBs for all arms $m \in M$) and lets agent j know. (Intuitively, this is a form of negative recommendation that reduces confidence bound). As we saw from equation (3) of Algorithm 1, agent j can compute $\widetilde{ucb}_m^{(j)}(t) = min(ucb_m^{(j)}(t), \widehat{ucb}_m^{(j)}(t))$ for all arms $m \in M$. (Intuitively, a patient updates her expectation of a relatively self-unexplored drug's side effects due to the physician's recommendation).

The question we address is the following: Fix an agent i, and suppose that all other agents in $A \setminus i$ don't violate the truthfulness assumptions. Would there be any incentive for the agent i to choose a

behavior that violates $T1$ and $T2$ at any time? This,is a dynamic game, and the question relates to whether truthtelling constitutes a Subgame Perfect Nash Equilibrium Fudenberg & Tirole (1991).

The answer, in plain English, is as follows. Suppose that an agent i only cares about the asymptotic order of the regret. That is, the agent i is indifferent between an $f(T)$ regret and an $g(T)$ regret if $f(T) = \Theta(g(T))$. Then we say that the agent i has an *asymptotically indifferent preference,* (defined formally in Appendix B). If all the agents of A have asymptotically indifferent preferences, it is trivial that no agent can strictly improve herself by violating $T1$ and $T2$ since she already has $O(1)$ regret. Hence one has the following result.

Theorem 12. *If all agents have asymptotically indifferent preferences, then the strategy where every agent conforms to T1 and T2 is a Subgame Perfect Nash Equilibrium.*

The formal formulation of this game and result are provided in the Appendix B.

7 Robustness to noise

The results in Section 4 implicitly assume that all measurements and reportings of rewards are perfectly accurate. That is, there is no measurement/communication noise. The following Theorem 13 shows that the algorithm 4 is robust to sub-Gaussian noise, and still achieves bounded regret.

Theorem 13. *Let $X'^{(j)}_m(k) = X^{(j)}_m(k) + e^{(j)}_m(k)$ be the noisy observation of $X^{(j)}_m(k)$. If the noise $e^{(j)}_m(k)$ is i.i.d. and follows a sub-Gaussian distribution, then each agent still has only bounded regret under the same conditions as Theorem 9.*

The proof is deferred to Appendix A.

8 Simulation experiments

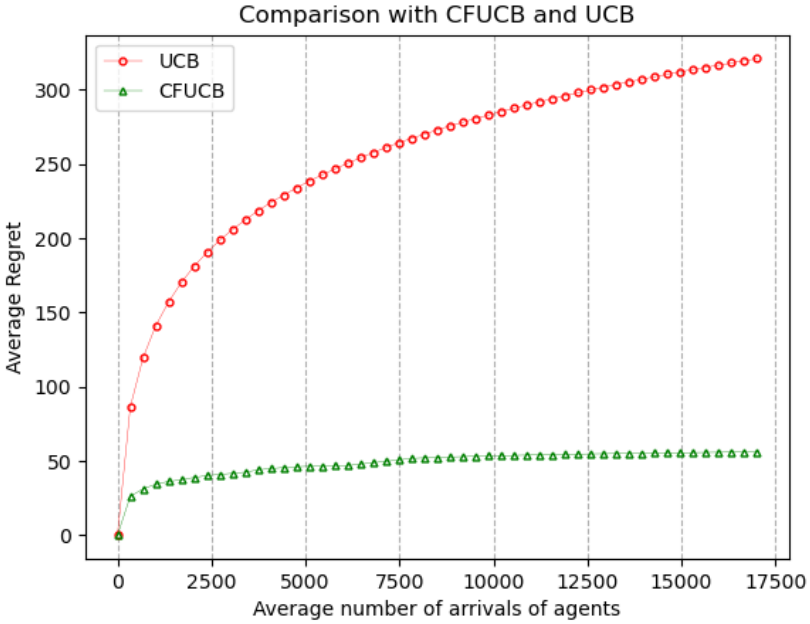

Figure 1: The regret of the CFUCB algorithm compared to that of the UCB algorithm, for the problem with 200 agents and 20 arms of feature vector dimension 5.

We conduct a simulation experiment to empirically demonstrate that the CFUCB algorithm indeed achieves $O(1)$ expected regret for the linear contextual bandit problem introduced in Section 2.

In this experiment, there are 200 agents repeatedly arriving to explore 20 arms. Each agent independently arrives according to its own renewal process with positively truncated $i.i.d.$ Normally distributed inter-arrival times. Both agent and arm feature vectors are randomly and uniformly generated as vectors on the surface of the 0-centered unit sphere in $\mathbb{R}^5$ (also known as unit 4-sphere). The inner product of the agent's pulled arm's feature vectors, plus noise that is i.i.d $N(0, 0.1)$, is the reward resulting from an arm pull. As a baseline for comparing the CFUCB algorithm's performance, we consider the same system, with the same arrival sequences, but with the agents following the vanilla UCB algorithm Auer (2002). As can be seen in Figure 1, the regret graph of CFUCB levels off, indicating that the regret does not increase further after a finite number of arrivals, showing that a regret of $O(1)$ is indeed achieved. In contrast, the average regret of the UCB algorithm is $O(\log T)$. Figure 1 averages the result of ten experiments in which arrivals and feature vectors are newly generated each time. For the codes, refer to Supplementary materials or Appendix C.

9 Concluding remarks

In many applications, multiple agents are simultaneously exploring choices. This paper proposes a new contextual bandit framework for which a policy that we call Counterfactual-UCB (CFUCB) guarantees that the expected regret of the totality of all agents is $O(1)$, i.e., it is bounded. The key idea enabling this result is to take advantage of the exploitation results of other agents to give every agent relief from its own exploration requirements on its bad arms.

References

Alberto Abadie and Guido W Imbens. Bias-corrected matching estimators for average treatment effects. Journal of Business & Economic Statistics, 29(1):1–11, 2011.

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

Susan Athey and Guido W Imbens. Machine learning methods for estimating heterogeneous causal effects. stat, 1050(5):1–26, 2015.

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

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, pp. 208–214. JMLR Workshop and Conference Proceedings, 2011.

Robert M Corless, Gaston H Gonnet, David EG Hare, David J Jeffrey, and Donald E Knuth. On the lambertw function. Advances in Computational mathematics, 5(1):329–359, 1996.

Scott Cunningham. Causal inference. In Causal Inference. Yale University Press, 2021.

Drew Fudenberg and David Levine. Subgame-perfect equilibria of finite-and infinite-horizon games. Journal of Economic Theory, 31(2):251–268, 1983.

Drew Fudenberg and Jean Tirole. Game theory. MIT press, 1991.

Ruiquan Huang, Weiqiang Wu, Jing Yang, and Cong Shen. Federated linear contextual bandits. Advances in Neural Information Processing Systems, 34:27057–27068, 2021.

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

Matthew O Jackson. Mechanism theory, page 2. DOI: 10.2139/ssrn.2542983, 2014. Lecture note.

Lihong Li, Yu Lu, and Dengyong Zhou. Provably optimal algorithms for generalized linear contextual bandits. In International Conference on Machine Learning, pp. 2071–2080. PMLR, 2017.

Ahmadreza Moradipari, Mohammad Ghavamzadeh, and Mahnoosh Alizadeh. Collaborative multi-agent stochastic linear bandits. arXiv preprint arXiv:2205.06331, 2022.

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.

Reinhard Selten. Spieltheoretische behandlung eines oligopolmodells mit nachfrageträgheit: Teil i: Bestimmung des dynamischen preisgleichgewichts. Zeitschrift für die gesamte Staatswissenschaft/Journal of Institutional and Theoretical Economics, (H. 2):301–324, 1965.

Chengshuai Shi, Haifeng Xu, Wei Xiong, and Cong Shen. (almost) free incentivized exploration from decentralized learning agents. Advances in Neural Information Processing Systems, 34:560–571, 2021.

Marta Soare and Joelle Pineau. Multi-task linear bandits. 2018. http://www.stat.yale.edu/ pollard/Courses/600.spring2017/Handouts/Basic.pdf.

Elizabeth A Stuart. Matching methods for causal inference: A review and a look forward. Statistical science: a review journal of the Institute of Mathematical Statistics, 25(1):1, 2010.

Li Ye, Yishi Lin, Hong Xie, and John Lui. Combining offline causal inference and online bandit learning for data driven decision. arXiv preprint arXiv:2001.05699, 2020.