A Linear Contextual Bandit Problem with a Bounded (O(1)) Regret Policy

Anonymous Author(s)
Affiliation
Address
email

Abstract

Asymptotically unbounded regret bounds, of the form $O(\sqrt{T})$ have been proven to be the lowest possible regret upper bounds that can be achieved in typical linear contextual bandit settings. In this paper, we suggest a recurring linear contextual bandit formulation where bounded, i.e., $O(1)$ expected regret can be achieved. It features repetitive arrivals of a large enough set of agents, and provides a novel Counterfactual UCB (CFUCB) policy where agents benefit from the experiences of other agents. It is shown that sharing of information between agents is a Subgame Perfect Nash Equilibrium in regret order, which results in each agent realizing bounded regret. Recommender systems and adaptive experimentation are two important applications.

1 Introduction

In many personalized recommender platforms, agents repeatedly interact with the platform. Through such repeated interactions, the platform obtains increasing information *with time* about each agent. 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 bandits model does not feature repeated arrivals of the same agents and has an asymptotically unbounded regret of $O(\sqrt{T})$ [4, 9, 2, 15]. We propose a model of *recurring* linear contextual bandits in which the regret is bounded, i.e., $O(1)$, under the condition that the number of agents is relatively large compared with the number of arms. Adaptive medical experimentation, is a good example of satisfying this condition, where the number of patients is much larger than the number of possible treatments.

Recurring linear contextual bandit framework proposed in this paper specifies the set of agents A and keeps track of each individual agent's repeated arrivals. Each agent/each arm has its own fixed feature vector where the arm feature vectors are unknown. Since everyone finds its optimal arm within a finite time, the change in the feature vectors within this short time may be negligible in practice. In contrast, typical contextual bandits address this issue of change in feature vectors over time not by tracking the individuals but instead by solving for the worst-case arrival of a sequence of features.

The main idea that enables the bounded expected regret result is somewhat related to the idea of a technique called matching, which has recently been popularized in the causal inference community [1, 27, 3, 10, 30]: Suppose that the arm m is not the optimal arm for the agent j, but it is optimal for a set of other agents, say A_m. If the total number of agents is 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. Then the agent j may be fully exempt from the exploration of arm m after a finite time, achieving bounded expected regret.

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

An intuitive medical experiment example of how bounded regret is achieved is as follows. Each patient (agent) is represented by his/her medical record, a known feature vector. There is a set of drugs (arms), with the individual side effect of each drug unknown. Note that it is reasonable to assume that the short term intake of drugs would not make the medical record obsolete. According to the feature vector, the doctor finds out that your reactions to the drugs can be expressed by 'You $= (-A) + (-B) + (-C)$' where A, B, and C are the ones for whom the drug m was highly reactive. Then the physician can avoid the risk of prescribing you drug m - because you are the sum of $-A, -B$ and $-C$ - without making you actually explore it. As A, B, C's frequency of taking drug m is high, the precision of prediction that you can go bad with drug m goes high fast. If you can find those kinds of people for all other drugs for you, you can be fully exempt from exploration at all.

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 previous medical history, turn on the privacy mode while using YouTube). In Section 6, we illustrate that everyone truthfully reporting their private information initially, conforming to the policy provided, and truthfully reporting the arm pulling experience to the mediator afterward, trivially 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 incentive constraints.

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 (iid). 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 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, $\{S_n^{(j)} \leq t\} = \{N^{(j)}(t) \geq n\}$.

An agent gets to pull an arm when it arrives. 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.

Rewards and objective Denote the set of agents' indices by A and the set of arms' indices as M. We assume that each of the agents and each of the arms are 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.

An agent receives a random reward when it pulls an arm. Denote the kth reward of agent j from arm m as $X_m^{(j)}(k)$. 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 [23]. We denote $E[X_{m,k}^{(j)}] = \alpha^{(j)}\beta_m$ by $\mu_m^{(j)}$. Under a cooperative setup, we assume that a reward observation immediately becomes a common knowledge.

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 $Regret^{(j)}(T) := \sum_{n=1}^{N^{(j)}(T)}(\mu_{m_j^*}^{(j)} - \mu_{m_j(n)}^{(j)})$. Under the cooperative setting (which will be lifted in Section 6), the common objective is to upper bound $E[Regret^{(j)}(T)]$ by a constant for all $j \in A$.

Agent set size constraint 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.

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:

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

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 [21]. 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 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.

3 Comparison to related works

We first show the commonalities and differences with the usual formulation of linear contextual bandits. Subsequently we provide an account of past work on multi-agent bandits.

3.1 Comparison to typical linear contextual bandit problems

1. In typical linear contextual bandits [4, 9, 2, 15], at each time-step, a new agent is born and arrives carrying known information called *context*, which actually contains one feature vector for each arm for that agent. In our setup, there is a set of agents A. An agent $i \in A$ has a known feature vector ϕ_i.

2. In typical contextual bandits, there is no concept of agents' repeated arrivals or record of plays and performances. In our setup, each agent in A repeatedly arrives. We can thus keep track of an individual agent's records, as in many real-world platforms (e.g., as a physician or YouTube does).

3. In typical linear contextual bandits, there is only a *single* vector θ that is unknown. In contrast, 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 feature vector for every drug).

4. In both formulations, the reward is the dot product of two feature vectors plus noise.

5. In typical linear contextual bandits, the best regret achievable is $O(\sqrt{T})$. In our setup, we show that agents can achieve bounded regret $O(1)$ by learning from the experiences of other agents.

6. This raises the question of why agents would share their experiences. We show that all agents gain when all share their experiences. Moreover, we show it is a Subgame Perfect Nash Equilibrium to share experiences. Thus all agents are implicitly motivated to share their experiences. (This is reminiscent of a happy family where all family members help each other, and everyone benefits).

3.2 Related works on linear contextual bandits and multi-agent exploration

The closest to our problem's formulation are cooperative linear bandits and multi-task linear bandits [26] and [20], where the agent side vectors are unknown, and the arm side vectors are known. In our formulation, it is the opposite. In [26], only an empirical analysis is provided. In [20], 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 contextual bandit problem and achieve $O(1)$ regret.

Coordination of multiple agents' exploration has been mainly studied with the focus of addressing decentralized communication. [24, 17, 29] studies the coordinating decentralized homogeneous

agents considering communication cost. [6] further considers how to achieve the same performance as a centralized one. As an extension of linear contextual bandit problems in this direction, [7] studies the coordination of multiple communicating linear contextual bandit problem instances sharing the same unknown vector θ and achieved $O(\sqrt{T}\log T)$ regret. For heterogeneous agents, constraints such as restriction in arm observation [19, 16] or matching to avoid collision [5, 18] were considered.

On the incentive constraint, in this paper, we illustrate that everyone reporting their private context to the mediator at time 0, following the CFUCB policy and reporting the rewards truthfully afterward trivially constitutes Subgame Perfect Nash equilibrium (SPNE) since $O(1)$ regret cannot be improved. There are related works on incentive constraints in coordinating exploration: [8] studies Bayesian perspectives of paying myopic agents with private context for the exploration to achieve $O\left(\log^3(T)\right)$ regret. [25] shows that identification of a social planner's best arm does not require extra payment when you must incentivize individually rational [14] agents 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. [13] 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.

The concept of each agent uploading local information to a mediator where the server coordinates information is also related to Federated Learning [12] and batch decision-making [28]. A notable result proved in [12] 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 can be done by sharing information. Moreover, we show that all agents behaving so is a *selfish equilibrium*, where *every* selfish agent's optimal reaction is to share information at every time-step. Moreover, it is a very strong equilibrium in that it is a selfish equilibrium at *every time* t, which is called a Subgame Perfect Nash equilibrium (SPNE). One common aspect though is that both the results in [12] and our paper are robust to some lying/mistakes in reporting; in our case, it is because the Confidence Interval changes only slightly by occasional lies/mistakes in reporting, and the gap between the non-optimal arm and optimal arm provides robustness.

4 Counterfactual UCB Algorithm (CFUCB)

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., [4] for the multi-armed bandit problem, an agent forms a confidence interval based on its own experience, which we call the *original 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*.

Original Confidence interval Recall that we defined $N_m^{(j)}(t)$ as the number of agent j's pulls of arm m until time t. We denote $\overline{X}_m^{(j)}(t) = \frac{\sum_{k=1}^{N_m^{(j)}(t)} X_{m,k}^{(j)}}{N_m^{(j)}(t)}$ as the empirical mean reward of agent j on arm m. Then the width of original confidence interval we denote by $w_m^{(j)}(t)$ is chosen as $w_m^{(j)}(t) = \sqrt{\frac{\log N^{(j)}(t)}{N_m^{(j)}(t)}}$. If we define $\overline{X}_m^{(j)}(t) + w_m^{(j)}(t)$ as $ucb_m^{(j)}(t)$, $\overline{X}_m^{(j)}(t) - w_m^{(j)}(t)$ as $lcb_m^{(j)}(t)$, then *original confidence interval* is defined as $CI_m^{(j)}(t) := \left(lcb_m^{(j)}(t), ucb_m^{(j)}(t)\right)$.

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 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 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 reward of agent j for arm m. Then the width of counterfactual confidence interval we denote $\widehat{w}_m^{(j)}(t)$ 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 $\widehat{ucb}_m^{(j)}(t)$, $\widehat{X}_m^{(j)}(t) - \widehat{w}_m^{(j)}(t)$ as $\widehat{lcb}_m^{(j)}(t)$, then *counterfactual confidence interval* is defined as $\widehat{CI}_n^{(i)}(t) := (\widehat{lcb}_m^{(j)}(t), \widehat{ucb}_m^{(j)}(t))$.

The corresponding *original 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 first introduce the concept of *epoch*s 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 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 $\{a_k\}_{k \in \mathbb{N}}$, 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 at epoch k). Recall that $X_m^{(j)}(n)$ denotes the nth 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 describes the pseudocode for 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 $ucb_m^{(j)}(t)$.

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 $ucb_m^{(a_k)}(s_k)$ (original 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**

5 The analysis of CFUCB

We first start by describing how confidence intervals were chosen. We follow the spirit of [4] - 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 ([4]). *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 of this paper in that it gives 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 original 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 we can see in the proof of Lemma 5, by using the Algorithm 1 we assure 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 original 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 an agent j and arm m. Now 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, we see 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 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 bring us the bounded expected regret result. Our strategy is to show that $\int_0^{+\infty} g_m^{(j)}(t)dF_n^{(j)}(t) = O(\frac{1}{n^2})$ holds for the two most representative settings: 1) agents arriving according to sub-Gaussian inter-arrival times (Section 5.1) and 2) agents arriving 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 Lambert W-function.*

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

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. 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 $g_m^{(j)}(t) = |A|(\exp(-2\frac{(t-q_{ij}(\lceil\frac{t}{\theta^j-\epsilon^j}\rceil)\theta_{\max})^2}{q_{ij}(\lceil\frac{t}{\theta^j-\epsilon^j}\rceil)}) + \exp(-2\frac{\epsilon^{j^2}}{\theta^j-\epsilon^j}t))$, 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 so we defer it 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 the 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 arrive independently with iid exponentially distributed inter-arrival times with λ_i. Every time an agent arrives, we play CFUCB. Then $P(G_m^{(j)}(t)|V(t)) \leq g_m^{(j)}(t)$ 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 so we defer it to Appendix A.

Theorem 11. *Suppose that each agent i of A arrive independently with iid exponentially distributed inter-arrival times with λ_i. 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 under CFUCB.*

Proof. 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 $\lambda_{\min} = \min_{i \in A} \lambda_i$, the result follows. 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 therefore not 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 anymore:

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 first assumes that all the agents have been conforming to $T1$ and $T2$ (A patient disclosing her previous medical records to the physician and reports her progresses precisely afterwards). Mediator computes $\{\widehat{ucb}_m^{(j)}(t)\}_{m \in M}$ (agent j's counterfactual UCBs for all arm $m \in M$) and let agent j know (Intuitively, this is a form of recommendation). As we saw from the 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 arm $m \in M$ (Intuitively, a patient updating her expectation of a drug's side effect due to physician's recommendation).

Fix an agent i, and suppose that all other agents of $A \setminus i$ don't violate such assumptions. Would there be an incentive for the agent i to choose a behavior that violates $T1$ and $T2$ at any time? This is related to the question of a Subgame Perfect Nash Equilibrium [11] in game theory.

The answer, in plain English, is as follows. Suppose that an agent i only cares about the asymptotics of the regret outcome. 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 an agent i has an *asymptotically indifferent preference.* (We precisely define this notion of preference in Appendix B). If all the agents of A have asymptotically indifferent preferences, it is trivial that no one can strictly improve herself by violating the $T1$ and $T2$ above since she already has $O(1)$ regret. Hence one has the following result.

Theorem 12. *The strategy where every agent conforms to T1 and T2 is a Subgame Perfect Nash Equilibrium when all agents have asymptotically indifferent preferences on their regrets.*

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 perfect. That is, there is no measurement/communication noise. However, Theorem 13 says that algorithm 4 is robust to sub-Gaussian noise, and still achieves bounded regret. The proof of Theorem 13 is deferred to Appendix A.

Theorem 13. *Let $X'^{(j)}_m(k) = X_m^{(j)}(k) + e_m^{(j)}(k)$ be the noisy observation of $X_m^{(j)}(k)$. If the noise $e_m^{(j)}(k)$ 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.*

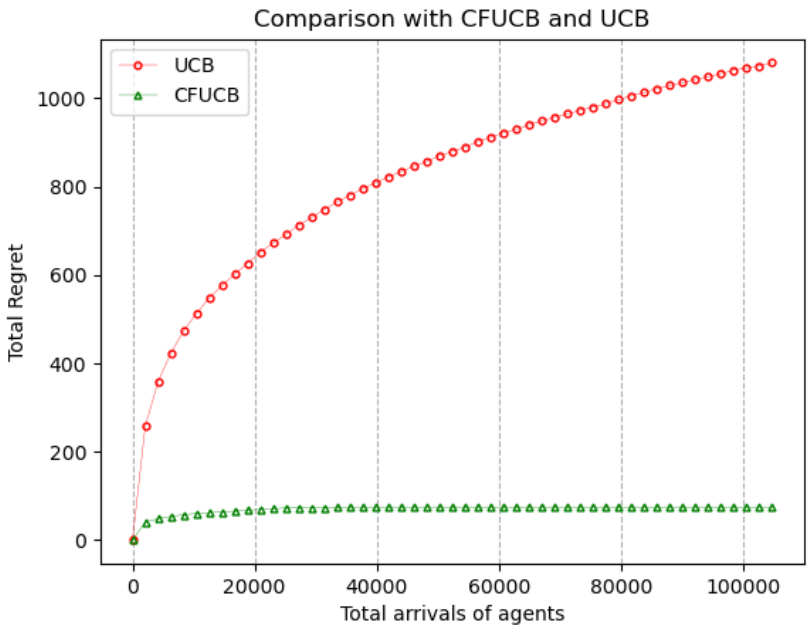

Figure 1: CFUCB algorithm's regret for the problem with 32 agents and 4 arms with dimension 2.

8 Simulation experiments

We conduct a simulation experiment to empirically demonstrate that, the CFUCB algorithm indeed achieves $O(1)$ expected regret bound for the recurring linear contextual formulation introduced in Section 2. We would want to provide a minimum baseline, of which the only purpose is to highlight CFUCB's $O(1)$ regret result. One such baseline is the case when agents had not helped each other; everyone arriving according to the same sequence as the CFUCB experiment but this time independently following the vanilla UCB algorithm [4]. In terms of the total system regret, $O(\log N)$ total regret is to be achieved in this case, where N is the total arrival number of all agents.

As can be seen in Figure 1, the regret graph of CFUCB levels off, indicating that the regret does not increase at all after a finite number of arrivals (In this particular experiment, no increase after 40000 total arrivals). This means that the regret of $O(1)$ is indeed achieved. Figure 1 averaged the result of ten experiments in which arrivals and feature vectors were newly generated each time. In this experiment, 32 agents independently and repeatedly arrived to pull four arms. Each agent arrived according to its own renewal process with $i.i.d.$ Normally distributed inter-arrival times. Both agent and arm feature vectors were randomly and uniformly generated as vectors on a circle of radius one in $\mathbb{R}^2$. The inner product of the arrived agent and pulled arm plus noise sampled from $N(0, 0.3)$ constituted each reward from arm pulling. For the codes, please refer to the Appendix C.

9 Concluding remarks

This paper suggests a new contextual bandit framework for which a policy that we call Counterfactual-UCB (CFUCB) guarantees that the expected regret 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 an exemption from its exploration requirements on bad arms.

References

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

[2] Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári. Improved algorithms for linear stochastic bandits. Advances in neural information processing systems, 24, 2011.

[3] S. Athey and G. W. Imbens. Machine learning methods for estimating heterogeneous causal effects. stat, 1050(5):1–26, 2015.

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

[5] I. Bistritz and A. Leshem. Game of thrones: Fully distributed learning for multiplayer bandits. Mathematics of Operations Research, 46(1):159–178, 2021.

[6] E. Boursier and V. Perchet. Sic-mmab: Synchronisation involves communication in multiplayer multi-armed bandits. Advances in Neural Information Processing Systems, 32, 2019.

[7] R. Chawla, A. Sankararaman, and S. Shakkottai. Multi-agent low-dimensional linear bandits. arXiv preprint arXiv:2007.01442, 2020.

[8] B. Chen, P. Frazier, and D. Kempe. Incentivizing exploration by heterogeneous users. In Conference On Learning Theory, pages 798–818. PMLR, 2018.

[9] W. Chu, L. Li, L. Reyzin, and R. 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.

[10] S. Cunningham. Causal inference. In Causal Inference. Yale University Press, 2021.

[11] D. Fudenberg and J. Tirole. Game theory. MIT press, 1991.

[12] R. Huang, W. Wu, J. Yang, and C. Shen. Federated linear contextual bandits. Advances in Neural Information Processing Systems, 34:27057–27068, 2021.

[13] N. Immorlica, J. Mao, A. Slivkins, and Z. S. Wu. Bayesian exploration with heterogeneous agents. In The World Wide Web Conference, pages 751–761, 2019.

[14] M. O. Jackson. Mechanism theory, page 2. Available at SSRN 2542983, 2014.

[15] L. Li, Y. Lu, and D. Zhou. Provably optimal algorithms for generalized linear contextual bandits. In International Conference on Machine Learning, pages 2071–2080. PMLR, 2017.

[16] A. Magesh and V. V. Veeravalli. Decentralized heterogeneous multi-player multi-armed bandits with non-zero rewards on collisions. IEEE Transactions on Information Theory, 68(4):2622–2634, 2021.

[17] D. Martínez-Rubio, V. Kanade, and P. Rebeschini. Decentralized cooperative stochastic bandits. Advances in Neural Information Processing Systems, 32, 2019.

[18] A. Mehrabian, E. Boursier, E. Kaufmann, and V. Perchet. A practical algorithm for multi-player bandits when arm means vary among players. In International Conference on Artificial Intelligence and Statistics, pages 1211–1221. PMLR, 2020.

[19] A. Mitra, H. Hassani, and G. Pappas. Exploiting heterogeneity in robust federated best-arm identification. arXiv preprint arXiv:2109.05700, 2021.

[20] A. Moradipari, M. Ghavamzadeh, and M. Alizadeh. Collaborative multi-agent stochastic linear bandits. arXiv preprint arXiv:2205.06331, 2022.

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

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

[23] O. Rivasplata. Subgaussian random variables: An expository note. Internet publication, PDF, 5, 2012.

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

[25] C. Shi, H. Xu, W. Xiong, and C. Shen. (almost) free incentivized exploration from decentralized learning agents. Advances in Neural Information Processing Systems, 34:560–571, 2021.

[26] M. Soare and J. Pineau. Multi-task linear bandits.

[27] E. 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.

[28] C.-H. Wang and G. Cheng. Online batch decision-making with high-dimensional covariates. In International Conference on Artificial Intelligence and Statistics, pages 3848–3857. PMLR, 2020.

[29] P.-A. Wang, A. Proutiere, K. Ariu, Y. Jedra, and A. Russo. Optimal algorithms for multiplayer multi-armed bandits. In International Conference on Artificial Intelligence and Statistics, pages 4120–4129. PMLR, 2020.

[30] L. Ye, Y. Lin, H. Xie, and J. Lui. Combining offline causal inference and online bandit learning for data driven decision. arXiv preprint arXiv:2001.05699, 2020.

Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [Yes] See Section **??**.
- Did you include the license to the code and datasets? [No] The code and the data are proprietary.
- Did you include the license to the code and datasets? [N/A]

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your [N/A] paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]