Anonymous Bandits for Multi-User Systems

Hossein Esfandiari
Google Research
esfandiari@google.com

Vahab Mirrokni
Google Research
mirrokni@google.com

Jon Schneider
Google Research
jschnei@google.com

Abstract

In this work, we present and study a new framework for online learning in systems with multiple users that provide user anonymity. Specifically, we extend the notion of bandits to obey the standard k-anonymity constraint by requiring each observation to be an aggregation of rewards for at least k users. This provides a simple yet effective framework where one can learn a clustering of users in an online fashion without observing any user's individual decision. We initiate the study of *anonymous bandits* and provide the first sublinear regret algorithms and lower bounds for this setting.

1 Introduction

In many modern systems, the system learns the behavior of the users by adaptively interacting with the users and adjusting to their responses. For example, in online advertisement, a website may show a certain type of ads to a user, observe whether the user clicks on the ads or not, and based on this feedback adjust the type of ads it serves the user. As part of this learning process, the system faces a dilemma between "exploring" new types of ads, which may or may not seem interesting to the user, and "exploiting" its knowledge of what types of ads historically seemed interesting to the user. This concept is very well studied in the context of online learning.

In principle, users benefit from a system that automatically adapts to their preferences. However, users may naturally worry about a system that observes all of their actions, and worry that the system may use this personal information against them or mistakenly reveal it to hackers or untrustworthy third parties. There therefore arises an additional dilemma between providing a better personalized user experience and acquiring the users' trust.

We study the problem of online learning in multi-user systems under a version of anonymity inspired by k-*anonymity* [Sweeney, 2002]. In its most general form, k-anonymity is a property of anonymous data guaranteeing that for every data point, there exist at least $k-1$ other indistinguishable data points in the dataset. Although there exist more recent notions of privacy and anonymity with stronger guarantees (e.g. various forms of differential privacy), k-anonymity is a simple and practical notion of anonymity that remains commonly employed in practice and enforced in various legal settings [Goldsteen et al., 2021, Saf, November 2019, Slijepčević et al., 2021].

To explain our notion of anonymity, consider the online advertisement application mentioned earlier. In online advertisement, when a user visits a website the website selects a type of ad and shows that user an ad of that type. The user may then choose to click on that ad or not, and a reward is paid out based on whether the user clicks on the ad or not (this reward may represent either the utility of the user or the revenue of the online advertisement system). Note that the ad here is chosen by the website, and the fact that the website assigns a user a specific type of ad is not something we intend to hide from the website. However, the decision to click on the ad is made by the user. We intend to protect these individual decisions, while allowing the website to learn what general types of ads each user is interested in. In particular, we enforce a form of group level C-anonymity on these decisions, by forcing the system to group users into groups of at least C users and to treat each group equally by

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

assigning all users in the same group the same ad and only observing the total aggregate reward (e.g. the total number of clicks) of these users.

More formally, we study this version of anonymity in a simple model of online learning based on the popular multi-armed bandit setting. In the classic (stochastic) multi-armed bandits problem, there is a learner with some number K of potential actions ("arms"), where each arm is associated with an unknown distribution of rewards. In each round (for T rounds) the learner selects an arm and collects a reward drawn from a distribution corresponding to that arm. The goal of the learner is to maximize their total expected reward (or equivalently, minimize some notion of regret).

In our multi-user model, a centralized learner assigns N users to K arms, and each rewards for each user/arm pair are drawn from some fixed, unknown distribution. Each round the learner proposes an assignment of users to arms, upon which each user receives a reward from the appropriate user/arm distribution. However, the learner is only allowed to record feedback about these rewards (and use this feedback for learning) if they perform this assignment in a manner compatible with C-*anonymity*. This entails partitioning the users into groups of size at least C, assigning all users in each group to the same arm, and only observing this group's aggregate rewards for this arm. For example, if $C = 3$ in one round we may combine users 1, 2 and 4 into a group and assign them all to arm 3; we would then observe as feedback the aggregate reward $r_{13} + r_{23} + r_{43}$, where r_{ij} represents the reward that user i experienced from arm j this round. The goal of the learner is to maximize the total reward by efficiently learning the optimal action for each user, while at the same time preserving the anonymity of individual rewards users experienced in specific rounds. See Section 2 for a more detailed formalization of the model.

1.1 Our results

In this paper we provide low-regret algorithms for the anonymous bandit setting described above. We present two algorithms which operate in different regimes (based on how users cluster into their favorite arms):

- If for each arm j there are at least $U \geq C+1$ users for which arm j is that user's optimal arm, then Algorithm 1 incurs regret at most $\tilde{O}(NC\sqrt{\alpha KT})$, where $\alpha = \max(1, \lceil K(C+1)/U \rceil)$.
- If there is no such guarantee (but $N > C$), then Algorithm 2 incurs regret at most $\tilde{O}(C^{2/3}K^{1/3}T^{2/3})$.

We additionally prove the following corresponding lower bounds:

- We show (Theorem 3) that the regret bound of Algorithm 1 is tight for any algorithm to within a factor of $K\sqrt{C}$ (and to within a factor of $\sqrt{C}$ for $U \geq K(C+1)$).
- We show (Theorem 4) that the dependence on $T^{2/3}$ in Algorithm 2 is necessary in the absence of a lower bound on U.

The main technical contribution of this work is the development/analysis of Algorithm 1. The core idea behind Algorithm 1 is to use recent algorithms developed for the problem of *batched bandits* (where instead of T rounds of adaptivity, users are only allowed $B \ll T$ rounds of adaptivity) to reduce this learning problem to a problem in combinatorial optimization related to decomposing a bipartite weighted graph into a collection of degree-constrained bipartite graphs. We then use techniques from combinatorial optimization and convex geometry to come up with efficient approximation algorithms for this combinatorial problem.

1.2 Related work

The bandits problem has been studied for almost a century [Thompson, 1933], and it has been extensively studied in the standard single user version [Audibert et al., 2009a,b, Audibert and Bubeck, 2010, Auer et al., 2002, Auer and Ortner, 2010, Bubeck et al., 2013, Garivier and Cappé, 2011, Lai and Robbins, 1985]. There exists recent work on bandits problem for systems with multiple users [Bande and Veeravalli, 2019a,b, Bande et al., 2021, Vial et al., 2021, Buccapatnam et al., 2015, Chakraborty et al., 2017, Kolla et al., 2018, Landgren et al., 2016, Sankararaman et al., 2019]. These papers study this problem from game theoretic and optimization perspectives (e.g. studying coordination /

competition between users) and do not consider anonymity. To the best of our knowledge this paper is the first attempt to formalize and study multi-armed bandits with multiple users from an anonymity perspective.

Learning how to assign many users to a (relatively) small set of arms can also be thought of as a clustering problem. Clustering of users in multi-user multi-arm bandits has been previously studied. Maillard and Mannor [2014] first studied sequentially clustering users, which was later followed up by other researchers [Nguyen and Lauw, 2014, Gentile et al., 2017, Korda et al., 2016]. Although these works, similar to us, attempt to cluster the users, they are allowed to observe each individual's reward and optimize based on that, which contradicts our anonymity requirement.

One technically related line of work that we heavily rely on is recent work on batched bandits. In the problem of batched bandits, the learner is prevented from iteratively and adaptively making decisions each round; instead the learning algorithm runs in a small number of "batches", and in each batch the learner chooses a set of arms to pull, and observes the outcome at the end of the batch. Batched multi-armed bandits were initially studied by Perchet et al. [2016] for the particular case of two arms. Later Gao et al. [2019] studied the problem for multiple arms. Esfandiari et al. [2019] improved the result of Gao et al. and extended it to linear bandits and adversarial multi-armed bandits. Later this problem was studied for batched Thompson sampling [Kalkanli and Ozgur, 2021, Karbasi et al., 2021], Gaussian process bandit optimization [Li and Scarlett, 2021] and contextual bandits [Zhang et al., 2021, Zanette et al., 2021].

In another related line of work, there have been several successful attempts to apply different notions of privacy such as differential privacy to multi-armed bandit settings [Tossou and Dimitrakakis, 2016, Shariff and Sheffet, 2018, Dubey and Pentland, 2020, Basu et al., 2019]. While these papers provide very promising guarantees of privacy measures, they focus on single-user settings. In this work we take advantage of the fact that there are several similar users in the system, and use this to provide guarantees of anonymity. Anonymity and privacy go hand in hand, and in a practical scenario, both lines of works can be combined to provide a higher level of privacy.

Finally, our setting has some similarities to the settings of stochastic linear bandits [Dani et al., 2008, Rusmevichientong and Tsitsiklis, 2010, Abbasi-Yadkori et al., 2011] and stochastic combinatorial bandits [Chen et al., 2013, Kveton et al., 2015]. For example, the superficially similar problem of assigning N users to K arms each round so that each arm has at least C users assigned to it (and where you get to observe the total reward per round) can be solved directly by algorithms for these frameworks. However, although such assignments are C-anonymous, there are important subtleties that prevent us from directly applying these techniques in our model. First of all, we do not actually constrain the assignment of users to arms – rather, our notion of anonymity constrains what feedback we can obtain from such an assignment (e.g., it is completely fine for us to assign zero users to an arm, whereas in the above model no actions are possible when $N < CK$). Secondly, we obtain more nuanced feedback than is assumed in these frameworks (specifically, we get to learn the reward of each group of $\geq C$ users, instead of just the total aggregate reward). Nonetheless it is an interesting open question if any of these techniques can be applied to improve our existing regret bounds (perhaps some form of linear bandits over the anonymity polytopes defined in Section 3.4.2).

2 Model and preliminaries

Notation. We write $[N]$ as shorthand for the set $\{1, 2, \ldots, N\}$. We write $\tilde{O}(\cdot)$ to suppress any poly-logarithmic factors (in N, K, C, or T) that arise. We say a random variable X is σ^2-subgaussian if the mean-normalized variable $Y = X - \mathbb{E}[x]$ satisfies $\mathbb{E}[\exp(sY)] \leq \exp(\sigma^2 s^2/2)$ for all $s \in \mathbb{R}$.

Proofs of most theorems have been postponed to Appendix C of the Supplemental Material in interest of brevity.

2.1 Anonymous bandits

In the problem of *anonymous bandits*, there are N users. Each round (for T rounds), our algorithm must assign each user to one of K arms (multiple users can be assigned to the same arm). If user i plays arm j, they receive a reward drawn independently from a 1-subgaussian distribution $\mathcal{D}_{i,j}$ with (unknown) mean $\mu_{i,j} \in [0, 1]$. We would like to minimize their overall expected regret of our algorithm. That is, if user i receives reward $r_{i,t}$ in round t, we would like to minimize

$$\mathbf{Reg} = T \sum_{i=1}^{N} \max_j \mu_{i,j} - \mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{N} r_{i,t}\right].$$

Thus far, this simply describes n independent parallel instances of the classic multi-armed bandit problem. We depart from this by imposing an *anonymity constraint* on how the learner is allowed to observe the users' feedback. This constraint is parameterized by a positive integer C (the minimum group size). In a given round, the learner may partition a subset of the users into groups of size at least C, under the constraint that the users within a single group must all be playing the same arm during this round. For each group G, the learner then receives as feedback the total reward $\sum_{i \in G} r_{i,t}$ received by users in this group. Note that not all users must belong to a group (the learner simply receives no feedback on such users), and the partition into groups is allowed to change from round to round.

Without any constraint on the problem instance, it may be impossible to achieve sublinear regret (see Section 3.6). We therefore additionally impose the following *user-cluster assumption* on the users: each arm j is the optimal arm for at least U users. Such an assumption generally holds in practice (e.g., in the regime where there are many users but only a few classes of arms). This also prevents situations where, e.g., only a single user likes a given arm but it is hard to learn this without allocating at least C users to this arm and sustaining significant regret. Typically we will take $U > C$; when $U \leq C$ the asymptotic regret bounds for our algorithms may be worse (see Section 3.5).

2.2 Batched stochastic bandits

Our main tool will be algorithms for *batched stochastic bandits*, as described in Gao et al. [2019]. For our purposes, a batched bandit algorithm is an algorithm for the classical multi-armed bandit problem that proceeds in B stages ("batches") where the bth stage has a predefined length of D_b rounds (with $\sum_b D_b = T$). At the beginning of each stage b, the algorithm outputs a non-empty subset of arms A_b (representing the set of arms the algorithm believes might still be optimal). At the end of each stage, the algorithm expects at least $D_b/|A_b|$ independent instances of feedback from arm j for each $j \in S$; upon receiving such feedback, the algorithm outputs the subset of arms A_{b+1} to explore in the next batch.

In Gao et al. [2019], the authors design a batched bandit algorithm they call BaSE (batched successive-elimination policy); for completeness, we reproduce a description of their algorithm in Appendix A. When $B = \log\log T$, their algorithm incurs a worst-case expected regret of at most $\tilde{O}(\sqrt{KT})$. In our analysis, we will need the following slightly stronger bound on the behavior of BaSE:

Lemma 1. *Set $B = \log\log T$. Let $\mu^* = \max_{j \in [K]} \mu_j$, and for each $j \in [K]$ let $\Delta_j = \mu^* - \mu_j$. Then for each $1 \leq b \leq B$, we have that:*

$$D_b \cdot \mathbb{E}\left[\max_{j \in A_b} \Delta_j\right] = \tilde{O}(\sqrt{KT}).$$

In other words, Lemma 1 bounds the expected regret in each batch, even under the assumption that we receive the reward of the worst active arm each round (even if we ask for feedback on a different arm).

It will also be essential in the analysis that follows that the total number of rounds D_b in the bth batch only depends on b and is independent of the feedback received thus far (in the language of Gao et al. [2019], the grid used by the batched bandit algorithm is *static*). This fact will let us run several instances of this batched bandit algorithm in parallel, and guarantees that batches for different instances will always have the same size.

3 Anonymous Bandits

3.1 Feedback-eliciting sub-algorithm

Our algorithms for anonymous bandits will depend crucially on the following sub-algorithm, which allows us to take a matching from users to arms and recover (in $O(C)$ rounds) an unbiased estimate

of each user's reward (as long as that user is matched to a popular enough arm). More formally, let $\pi : [N] \to [K]$ be a matching from users to arms. We will show how to (in $2C + 2$ rounds) recover an unbiased estimate of $\mu_{i,\pi(i)}$ for each i such that $|\pi^{-1}(\pi(i))| \geq C + 1$ (i.e. i is matched to an arm that at least C other users are matched to).

The main idea behind this sub-algorithm is simple; for each user, we will get a sample of the total reward of a group containing the user, and a sample of the total reward of the same group but minus this user. The difference between these two samples is an unbiased estimate of the user's reward. More concretely, we follow these steps:

1. Each round (for $2C + 2$ rounds) the learner will assign user i to arm $\pi(i)$. However, the partition of users into groups will change over the course of these $2C + 2$ rounds.
2. For each arm j such that $|\pi^{-1}(j)| \geq C + 1$, partition the users in $\pi^{-1}(j)$ into groups of size at least $C + 1$ and of size at most $2C + 1$. Let $G_1, G_2, \ldots, G_S$ be the set of groups formed in this way (over all arms j). In each group, order the users arbitrarily.
3. In the first round, the learner reports the partition into groups $\{G_1, G_2, \ldots, G_S\}$. For each group G_s, let $r_{s,0}$ be the total aggregate reward for group G_s this round.
4. In the kth of the next $2C + 1$ rounds, the learner reports the partition into groups $\{G_{1,k}, G_{2,k}, \ldots, G_{S,k}\}$, where $G_{s,k}$ is formed from G_s by removing the kth element (if $k > |G_s|$, then we set $G_{s,k} = G_s$). Let $r_{s,k}$ be the total aggregate reward from $G_{s,k}$ reported this round.
5. If user i is the kth user in G_s, we return the estimate $\hat{\mu}_{i,\pi(i)} = r_{s,0} - r_{s,k}$ of the average reward for user i and arm $\pi(i)$.

Lemma 2. *In the above procedure,* $\mathbb{E}[\hat{\mu}_{i,\pi(i)}] = \mu_{i,\pi(i)}$ *and* $\hat{\mu}_{i,\pi(i)}$ *is an* $O(C)$*-subgaussian random variable.*

3.2 Anonymous decompositions of bipartite graphs

The second ingredient we will need in our algorithm is the notion of an anonymous decomposition of a weighted bipartite graph. Intuitively, by running our batched stochastic bandits algorithm, at the beginning of each batch we will obtain a demand vector for each user (representing the number of times that user would like feedback on each of the m arms). Based on this, we want to generate a collection of assignments (from users to arms) which guarantee that we obtain (while maintaining our anonymity guarantees) the requested amount of information for each user/arm pair.

Formally, we represent a weighted bipartite graph as a matrix of nm non-negative entries $w_{i,j}$ (representing the number of instances of feedback user i desires from arm j). We will assume that for each i, $\sum_j w_{i,j} > 0$ (each user is interested in at least one arm). A *C-anonymous decomposition* of this graph is a collection of R assignments $M_1, M_2, \ldots, M_R$ from users to arms (i.e., functions from $[N]$ to $[K]$) that satisfies the following properties:

1. A user is never assigned to an arm for which they have zero demand. That is, if $M_r(i) = j$, then $w_{i,j} \neq 0$.
2. If $M_r(i) = j$, and $|M_r^{-1}(j)| \geq C + 1$, we say that matching M_r is *informative* for the user/arm pair (i, j). (Note that this is exactly the condition required for the feedback-eliciting sub-algorithm to output the unbiased estimate $\hat{\mu_{i,j}}$ when run on assignment M_r.) For each user/arm pair (i, j), there must be at least $w_{i,j}$ informative assignments.

The weighted bipartite graphs that concern us come from the parallel output of N batched bandit algorithms (described in Section 2.2) and have additional structure. These graphs can be described by a positive total demand D and a non-empty demand set $A_i \subseteq [K]$ for each user i (describing the arms of interest to user i). If $j \in A_i$, then $w_{i,j} = D/|A_i|$; otherwise, $w_{i,j} = 0$. To distinguish graphs with the above structure from generic weighted bipartite graphs, we call such graphs *batched graphs*.

Moreover, for each user i, let $j^*(i)$ be the optimal arm for user i (i.e., $j^*(i) = \arg\max_i \mu_i$). In the algorithm we describe in the next section, with high probability, $j^*(i)$ will always belong to A_i. Moreover, a user-cluster assumption of U implies that, for any arm j, $|j^{*-1}(j)| \geq U$. This means that in batched graphs that arise in our algorithm, there will exist an assignment where each

user i is assigned to an arm in A_i, and each arm j has at least U assigned users. We therefore call graphs that satisfy this additional assumption *U-batched graphs*. Note that this assumption also allows us to lower bound the degrees of arms in this bipartite graph. Specifically, for each arm j, let $B_j = \{i \in [N] \mid j \in A_i\}$. Then a user-cluster assumption of U directly implies that $|B_j| \geq U$.

In general, our goal is to minimize the number of assignments R required in such a decomposition (since each assignment corresponds to some number of rounds required). We call an algorithm that takes in a U-batched graph and outputs a C-anonymous decomposition of that graph an *anonymous decomposition algorithm*, and say that it has approximation ratio $\alpha(C, U) \geq 1$ if it generates an assignment with at most $\alpha(C, U) \cdot D + O(NK)$ total assignments. This additive $O(NK)$ is necessary for technical reasons, but in our algorithm, D will always be much larger than K (we will have $D \geq \sqrt{T}$), so this can be thought of as an additive $o(D)$ term.

Later, in Section 3.4, we explicitly describe several anonymous decomposition algorithms and their approximation guarantees. In the interest of presenting the algorithm, we will assume for now we have access to a generic anonymous decomposition algorithm Decompose with approximation ratio $\alpha(C, U)$.

3.3 An algorithm for anonymous bandits

We are now ready to present our algorithm for anonymous bandits. The main idea behind this algorithm (detailed in Algorithm 1) is as follows. Each of the N users will run their own independent instance of BaSE with $B = \log \log T$ synchronized batches. During batch b, BaSE requires each user i to get a total of D_b instances of feedback on a set $A_{i,b}$ of arms which are alive for them. These sets $A_{i,b}$ (with high probability) define a U-batched graph, so we can use an anonymous decomposition algorithm to construct a C-anonymous decomposition of this graph into at most $\alpha(C, U) D_b$ assignments. We then run the feedback-eliciting sub-algorithm on each assignment, getting one unbiased estimate of the reward for each user/arm pair for which the assignment is informative.

The guarantees of the C-anonymous decomposition mean that this process gives us D_b total pieces of feedback for each user, evenly split amongst the arms in $A_{i,b}$. We can therefore pass this feedback along to BaSE, which will eliminate some arms and return the set of alive arms for user i in the next batch.

Algorithm 1: Low-regret algorithm for anonymous bandits.

Input: Anonymity parameter C, a lower bound on the user-cluster assumption U, time-horizon T, number of users N, and number of arms K. We additionally assume we have access to an anonymous decomposition algorithm Decompose with approximation factor $\alpha = \alpha(C, U)$.

For each user $i \in [N]$, initialize an instance of BaSE with a time horizon of $T' = \frac{T}{\alpha(2C+2)}$ and $B = \log \log T$ batches. Let $D_1, D_2, \ldots, D_b$ be the corresponding batch sizes (with $\sum_b D_b = T'$). Let $A_{i,b}$ be the set of active arms for user i during batch b.

for $b \leftarrow 1$ **to** B **do**

- If D_b and $(A_{1,b}, A_{2,b}, \ldots, A_{n,b})$ do not define a U-batched graph, abort the algorithm (this means we have eliminated the optimal arm for a user, which can only happen with negligible probability).
- Run Decompose on D_b and $(A_{1,b}, A_{2,b}, \ldots, A_{n,b})$ to get a C-anonymous decomposition into at most $R_b = \alpha D_b$ assignments. Let M_r be the rth such assignment.
- **for** $r \leftarrow 1$ **to** R_b **do**
  - Over $(2C + 2)$ rounds, run the feedback-soliciting sub-algorithm (Section 3.1) on M_r to get estimates $\hat{\mu}_{i,M_r(i)}$ for all users i for which M_r is informative.
- **end**
- For each user i, we are guaranteed to receive (by the guarantees of Decompose) at least $\frac{D_b}{|A_{i,b}|}$ independent samples of feedback $\hat{\mu}_{i,j}$ for each arm $j \in A_{i,b}$. Pass these samples to user i's instance of BaSE and receive $A_{i,b+1}$ in response (unless this is the last batch).

end

Theorem 1. *Algorithm 1 incurs an expected regret of at most $\tilde{O}(NC\sqrt{\alpha KT})$ for the anonymous bandits problem.*

One quick note on computational complexity: note that we only run Decompose once every *batch*; in particular, at most $\log \log T$ times. This allows us to efficiently implement Algorithm 1 even for complex choices of Decompose that may require solving several linear programs.

3.4 Algorithms for constructing anonymous decompositions

3.4.1 A greedy method

We begin with perhaps the simplest method for constructing an anonymous decomposition, which achieves an approximation ratio $\alpha(C, U) = K$ as long as $U \geq C + 1$. To do this, for each arm $j \in [K]$, consider the assignment M_j where all users i with $j \in A_i$ (i.e., users with any interest in arm j) are matched to arm j, and other users are arbitrarily assigned to arms in their active arm set. Our final decomposition contains D copies of M_j for each $j \in [K]$ (for a total of KD assignments).

Note that since $U \geq C + 1$, there will be at least $C + 1$ users matched to arm j in M_j, and therefore M_j will be informative for all users i with $j \in A_i$. Since we repeat each assignment D times, we will have at least D informative assignments for every valid user/arm pair, and therefore this is a valid C-anonymous decomposition for the original batched graph.

Substituting this guarantee into Theorem 1 gives us an anonymous bandit algorithm with expected regret $\tilde{O}(NCK\sqrt{T})$.

3.4.2 The anonymity polytope

As U grows larger than C, it is possible to attain even better approximation guarantees. In this section we will give an anonymous decomposition algorithm that applies techniques from combinatorial optimization to attain the following guarantees:

- If $U \geq K(C + 1)$, then $\alpha(C, U) = 1$.
- If $(C + 1) \leq U \leq K(C + 1)$, then $\alpha(C, U) = \left\lceil \frac{K(C+1)}{U} \right\rceil$.

To gain some intuition for how this is possible, assume $U = K(C + 1)$, and consider the randomized algorithm which matches each user i to a random arm in A_i each turn. In expectation, after D rounds of this, user i will be matched to each arm $j \in A_i$ exactly $D/|A_i|$ times (as user i desires). Moreover, since $U \geq K(C + 1)$, each arm j has at least $K(C + 1)$ candidate users that can match to it. Each of these users matches to arm j with probability at least $1/K$, so in expectation at least $C + 1$ users match to arm j, and therefore the feedback from arm j is informative "in expectation".

The catch with this method is that it is possible (and even reasonably likely) for fewer than $C+1$ users to match to a given arm j, and in this case we receive no feedback for user i. While it is possible to adapt this method to work with high probability, this requires additional logarithmic factors in either α or the user-cluster bound U, and even then has some probability of failure. Instead, we present a deterministic algorithm which can exactly achieve the guarantees above by geometrically "rounding" the above randomized matching into a small weighted collection of deterministic matchings.

We define the *C-anonymity polytope* $\mathcal{P}_C \subseteq \mathbb{R}^{N \times K}$ to be the convex hull of all binary vectors $v \in \{0, 1\}^{N \times K}$ that satisfy the following conditions:

- For each i, $\sum_j v_{ij} \in \{0, 1\}$.
- For each j, either $\sum_i v_{ij} \geq C + 1$ or $\sum_i v_{ij} = 0$.

We can interpret each such vertex v as a single assignment in a C-anonymous decomposition, where $v_{ij} = 1$ iff we get feedback on the user/arm pair (i, j) (so we must match user i to j, and at least $C + 1$ users must be matched to arm j).

Now, for a fixed U-batched graph G, let $w \in [0, 1]^{N \times K}$ be the weights of this G normalized by the demand D: so $w_{ij} = 1/|A_i|$ if $j \in A_i$, and $w_{ij} = 0$ otherwise. It turns out that we can reduce (via Caratheodory's theorem) the problem of finding a C-anonymous decomposition of G into finding the maximal β for which $\beta w \in \mathcal{P}_C$.

Lemma 3. *If for some $\beta > 0$, $\beta w \in \mathcal{P}_C$, then there exists a C-anonymous decomposition of G into at most $\frac{1}{\beta}D + NK + 1$ assignments. Similarly, if there exists a C-anonymous decomposition of G into αD assignments, then $\frac{1}{\alpha}w \in \mathcal{P}_C$.*

In a sense, Lemma 3 provides an "optimal" algorithm for the problem of finding C-anonymous decompositions. There are two issues with using this algorithm in practice. The first – an interesting open question – is that we do not understand the approximation guarantees of this decomposition algorithm (although they are guaranteed to be at least as good as every algorithm we present here).

Open Problem 1. *For a U-batched graph G, let $\beta(G)$ be the maximum value of β such that $\beta w \in \mathcal{P}_C$ (where w is the weight vector associated with C). What is $\max \beta(G)$ over all U-batched graphs? Is it $\Omega(1)$?*

The second is that, computationally, it is not clear if there is an efficient way to check whether a point belongs to $\mathcal{P}_C$ (let alone write it as a convex combination of the vertices of $\mathcal{P}_C$). We will now decompose $\mathcal{P}_C$ into the convex hull of a collection of more tractable polytopes; while it will still be hard to e.g. check membership in $\mathcal{P}_C$, this will help us efficiently compute decompositions that provide the guarantees at the beginning of this section.

For a subset $S \subseteq [K]$ of arms, let $\mathcal{P}_C(S) \subseteq \mathbb{R}^{N\times K}$ be the convex hull of the binary vectors that satisfy the following conditions:

- For each i, $\sum_j v_{ij} \in \{0, 1\}$.
- If $j \in S$, then $\sum_i v_{ij} \geq C + 1$.
- If $j \notin S$, then $\sum_i v_{ij} = 0$.

By construction, each vertex of $\mathcal{P}_C$ appears as a vertex of some $\mathcal{P}_C(S)$ and each vertex of $\mathcal{P}_C(S)$ belongs to $\mathcal{P}_C$, so $\mathcal{P}_C = \mathrm{conv}(\{\mathcal{P}_C(S) \mid S \subseteq [K]\})$. We now claim that we can write each polytope $\mathcal{P}_C(S)$ as the intersection of a small number of halfspaces (and therefore check membership efficiently). In particular, we claim that $x \in \mathbb{R}^{N\times K}$ belongs to $\mathcal{P}_C(S)$ iff it satisfies the following linear constraints:

$$\begin{aligned}
0 \leq\ & x_{ij} \leq 1, && \forall\, i \in [N], j \in [K] \\
& \sum_{j\in[K]} x_{ij} \leq 1, && \forall\, i \in [N] \\
& \sum_{i\in[N]} x_{ij} \geq C + 1, && \forall\, j \in S \\
& \sum_{i\in[N]} x_{ij} = 0, && \forall\, j \in [K] \setminus S.
\end{aligned} \tag{1}$$

Lemma 4. *A point $x \in \mathbb{R}^{N\times K}$ belongs to $\mathcal{P}_C(S)$ iff it satisfies the constraints in* (1).

We can now prove the guarantees at the beginning of the section. We start with the case where $U \geq K(C + 1)$. Here we show (via similar logic to the initial randomized argument) that $w \in \mathcal{P}_C([K])$.

Lemma 5. *If $U \geq K(C + 1)$, then $w \in \mathcal{P}_C([K])$.*

Applying Caratheodory's theorem, we immediately obtain a C-anonymous decomposition from Lemma 5

Corollary 1. *If $U \geq K(C+1)$, there exists a C-anonymous decomposition into at most $D+KC+1$ assignments. Moreover, it is possible to find this decomposition efficiently.*

When $C + 1 \leq U \leq K(C + 1)$, we first arbitrarily partition our arms into $\alpha = \lceil K(C + 1)/U\rceil$ blocks $S_1, S_2, \ldots, S_\alpha$ of at most $U/(C + 1)$ vertices each. We then show how to write w as a linear combination of α points, one in each of the polytopes $\mathcal{P}_C(S_a)$.

Lemma 6. *There exist points $w^{(a)} \in \mathcal{P}_C(S_a)$ for $1 \leq a \leq \alpha$ such that $w \leq \sum_{i=1}^{\alpha} w^{(a)}$.*

Likewise, we can again apply Caratheodory's theorem to obtain a C-anonymous decomposition from Lemma 6.

Corollary 2. *If $(C+1) \leq U \leq K(C+1)$, there exists a C-anonymous decomposition into at most $\alpha D + NK + 1$ assignments, where $\alpha = \lceil K(C+1)/U \rceil$. Moreover, it is possible to find this decomposition efficiently.*

3.5 Settings without user-clustering

The previous algorithms we have presented for anonymous bandits rely heavily on the existence of a user-cluster assumption U. In this section we present an algorithm (Algorithm 2) for anonymous bandits which works in the absence of any user-cluster assumption as long as $N > C$. This comes at the cost of a slightly higher regret bound which scales as $T^{2/3}$ – as we shall see in Section 3.6, this dependence is in some sense necessary.

Algorithm 2: Explore-then-commit algorithm for anonymous bandits without a user-clustering assumption.

Input: Anonymity parameter C, time-horizon T, number of users N, and number of arms K.
Set $T_{exp} = 10C^{2/3}K^{1/3}T^{2/3}(\log NKT)^{1/3}$.
For each user $i \in [N]$ and arm $j \in [K]$ initialize two variables $n_{ij} = 0$ and $\sigma_{ij} = 0$.
for $r \leftarrow 1$ **to** $\lceil T_{exp}/(2C+2) \rceil$ **do**
  Divide $[N]$ arbitrarily into S groups $G_1, \ldots, G_S$ of size $C+1$ (adding all remaining users to the last group G_S).
  For each group G_s, set $j_s = (r \bmod K)$.
  Run the feedback-eliciting sub-algorithm on the assignment π induced by the groups G_s (where if $i \in G_s$ then $\pi(i) = j_s$).
  for $i \leftarrow 1$ **to** N **do**
    $n_{i,\pi(i)} \leftarrow n_{i,\pi(i)} + 1$
    $\sigma_{i,\pi(i)} \leftarrow \sigma_{i,\pi(i)} + \hat{\mu}_{i,\pi(i)}$. Here $\hat{\mu}_{i,\pi(i)}$ is the unbiased estimate for $\mu_{i,\pi(i)}$ produced by the feedback-eliciting sub-algorithm.
  end
end
for *remaining rounds* t **do**
  Match user i to $\arg\max_j \sigma_{ij}/n_{ij}$.
end

Algorithm 2 follows the standard pattern of Explore-Then-Commit algorithms (see e.g. Chapter 6 of Lattimore and Szepesvári [2020]). For approximately $O(T^{2/3})$ rounds, we run the feedback-eliciting sub-algorithm on random assignments from users to arms (albeit ones which are chosen to guarantee each arm with any users matched to it has at least $C+1$ users matched to it), getting unbiased estimates of the means $\mu_{i,j}$. For the remaining arms, we match each user to their historically best-performing arm.

Theorem 2. *Algorithm 2 incurs an expected regret of at most $\tilde{O}(NC^{2/3}K^{1/3}T^{2/3})$ for the anonymous bandits problem.*

3.6 Lower bounds

We finally turn our attention to lower bounds. In all of our lower bounds, we exhibit a family of hard distributions over anonymous bandits problem instances, where any algorithm facing a problem instance randomly sampled from this distribution incurs at least the regret lower bound in question.

We begin by showing an $\Omega(N\sqrt{CKT})$ lower bound that holds even in the presence of a user-cluster assumption (and in fact, even when $U \geq K(C+1)$). Since Algorithm 1 incurs at most $\tilde{O}(NC\sqrt{KT})$ regret for $U \geq K(C+1)$, this shows the regret bound of our algorithm is tight in this regime up to a factor of $\sqrt{C}$ (and additional polylogarithmic factors).

Theorem 3. *Every learning algorithm for the anonymous bandits problem must incur expected regret at least $\Omega(N\sqrt{CKT})$, even when restricted to instances satisfying $U \geq K(C+1)$.*

We now shift our attention to settings where there is no guaranteed user-cluster assumption. We first show that the $O(T^{2/3})$ dependency of the algorithm in Section 3.5 is necessary.

Theorem 4. *There exists a family of problem instances satisfying $N = C + 1$ where any learning algorithm must incur regret at least $\Omega(T^{2/3})$.*

Finally, we show the assumption in Section 3.5 that $N > C$ is in fact necessary; if $N = C$ then it is possible that no algorithm obtains sublinear regret.

Theorem 5. *There exists a family of problem instances satisfying $N = C$ where any learning algorithm must incur regret at least $\Omega(T)$.*

4 Simulations

Finally, we perform simulations of our anonymous bandits algorithms – the explore-then-commit algorithm (Algorithm 2) and several variants of Algorithm 1 with different decomposition algorithms – on synthetic data. We observe that both the randomized decomposition and LP decomposition based variants of Algorithm 1 significantly outperform the explore-then-commit algorithm and the greedy decomposition variant, as predicted by our theoretical bounds. We discuss these in more detail in Section B of the Supplemental Material.

References

Safari privacy overview. Apple, November 2019. `https://www.apple.com/safari/docs/Safari_White_Paper_Nov_2019.pdf`.

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

J.-Y. Audibert and S. Bubeck. Regret bounds and minimax policies under partial monitoring. *The Journal of Machine Learning Research*, 11:2785–2836, 2010.

J.-Y. Audibert, S. Bubeck, et al. Minimax policies for adversarial and stochastic bandits. In *COLT*, volume 7, pages 1–122, 2009a.

J.-Y. Audibert, R. Munos, and C. Szepesvári. Exploration–exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410(19):1876–1902, 2009b.

P. Auer and R. Ortner. Ucb revisited: Improved regret bounds for the stochastic multi-armed bandit problem. *Periodica Mathematica Hungarica*, 61(1-2):55–65, 2010.

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

M. Bande and V. V. Veeravalli. Adversarial multi-user bandits for uncoordinated spectrum access. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4514–4518. IEEE, 2019a.

M. Bande and V. V. Veeravalli. Multi-user multi-armed bandits for uncoordinated spectrum access. In *2019 International Conference on Computing, Networking and Communications (ICNC)*, pages 653–657. IEEE, 2019b.

M. Bande, A. Magesh, and V. V. Veeravalli. Dynamic spectrum access using stochastic multi-user bandits. *IEEE Wireless Communications Letters*, 10(5):953–956, 2021.

D. Basu, C. Dimitrakakis, and A. Tossou. Differential privacy for multi-armed bandits: What is it and what is its cost? *arXiv preprint arXiv:1905.12298*, 2019.

S. Bubeck, V. Perchet, and P. Rigollet. Bounded regret in stochastic multi-armed bandits. In *Conference on Learning Theory*, pages 122–134. PMLR, 2013.

S. Buccapatnam, J. Tan, and L. Zhang. Information sharing in distributed stochastic bandits. In *2015 IEEE Conference on Computer Communications (INFOCOM)*, pages 2605–2613. IEEE, 2015.

M. Chakraborty, K. Y. P. Chua, S. Das, and B. Juba. Coordinated versus decentralized exploration in multi-agent multi-armed bandits. In *IJCAI*, pages 164–170, 2017.

W. Chen, Y. Wang, and Y. Yuan. Combinatorial multi-armed bandit: General framework and applications. In *International conference on machine learning*, pages 151–159. PMLR, 2013.

V. Dani, T. P. Hayes, and S. M. Kakade. Stochastic linear optimization under bandit feedback. 2008.

A. Dubey and A. Pentland. Differentially-private federated linear bandits. *arXiv preprint arXiv:2010.11425*, 2020.

H. Esfandiari, A. Karbasi, A. Mehrabian, and V. Mirrokni. Regret bounds for batched bandits. *arXiv preprint arXiv:1910.04959*, 2019.

Z. Gao, Y. Han, Z. Ren, and Z. Zhou. Batched multi-armed bandits problem. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 503–513, 2019.

A. Garivier and O. Cappé. The kl-ucb algorithm for bounded stochastic bandits and beyond. In *Proceedings of the 24th annual conference on learning theory*, pages 359–376. JMLR Workshop and Conference Proceedings, 2011.

C. Gentile, S. Li, P. Kar, A. Karatzoglou, G. Zappella, and E. Etrue. On context-dependent clustering of bandits. In *International Conference on Machine Learning*, pages 1253–1262. PMLR, 2017.

A. Goldsteen, G. Ezov, R. Shmelkin, M. Moffie, and A. Farkash. Data minimization for gdpr compliance in machine learning models. *AI and Ethics*, pages 1–15, 2021.

R. Hoeksma, B. Manthey, and M. Uetz. Efficient implementation of carathéodory's theorem for the single machine scheduling polytope. *Discrete applied mathematics*, 215:136–145, 2016.

C. Kalkanli and A. Ozgur. Batched thompson sampling. *Advances in Neural Information Processing Systems*, 34, 2021.

A. Karbasi, V. Mirrokni, and M. Shadravan. Parallelizing thompson sampling. *arXiv preprint arXiv:2106.01420*, 2021.

R. K. Kolla, K. Jagannathan, and A. Gopalan. Collaborative learning of stochastic bandits over a social network. *IEEE/ACM Transactions on Networking*, 26(4):1782–1795, 2018.

N. Korda, B. Szorenyi, and S. Li. Distributed clustering of linear bandits in peer to peer networks. In *International conference on machine learning*, pages 1301–1309. PMLR, 2016.

B. Kveton, Z. Wen, A. Ashkan, and C. Szepesvari. Tight regret bounds for stochastic combinatorial semi-bandits. In *Artificial Intelligence and Statistics*, pages 535–543. PMLR, 2015.

T. L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

P. Landgren, V. Srivastava, and N. E. 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.

T. Lattimore and C. Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Z. Li and J. Scarlett. Gaussian process bandit optimization with few batches. *arXiv preprint arXiv:2110.07788*, 2021.

O.-A. Maillard and S. Mannor. Latent bandits. In *International Conference on Machine Learning*, pages 136–144. PMLR, 2014.

T. T. Nguyen and H. W. Lauw. Dynamic clustering of contextual multi-armed bandits. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 1959–1962, 2014.

V. Perchet, P. Rigollet, S. Chassang, and E. Snowberg. Batched bandit problems. *The Annals of Statistics*, 44(2):660–681, 2016.

P. Rusmevichientong and J. N. Tsitsiklis. Linearly parameterized bandits. *Mathematics of Operations Research*, 35(2):395–411, 2010.

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.

R. Shariff and O. Sheffet. Differentially private contextual linear bandits. *arXiv preprint arXiv:1810.00068*, 2018.

D. Slijepčević, M. Henzl, L. D. Klausner, T. Dam, P. Kieseberg, and M. Zeppelzauer. k-anonymity in practice: How generalisation and suppression affect machine learning classifiers. *Computers & Security*, 111:102488, 2021. ISSN 0167-4048. doi: https://doi.org/10.1016/j.cose.2021.102488. URL `https://www.sciencedirect.com/science/article/pii/S0167404821003126`.

L. Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570, 2002.

W. R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.

A. C. Tossou and C. Dimitrakakis. Algorithms for differentially private multi-armed bandits. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

D. Vial, S. Shakkottai, and R. Srikant. Robust multi-agent multi-armed bandits. In *Proceedings of the Twenty-second International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing*, pages 161–170, 2021.

A. Zanette, K. Dong, J. Lee, and E. Brunskill. Design of experiments for stochastic contextual linear bandits. *Advances in Neural Information Processing Systems*, 34, 2021.

Z. Zhang, X. Ji, and Y. Zhou. Almost optimal batch-regret tradeoff for batch linear contextual bandits. *arXiv preprint arXiv:2110.08057*, 2021.

Checklist

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] This is primarily a theory paper, and we do not foresee any negative societal impacts of this work.
   (d) Have you read the ethics review guidelines and ensured that your 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)? [Yes] Code for the simulations is included in the Supplemental Material
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (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)? [No] The simulations performed in this work were not computationally intensive, so this is not of particular interest (combined, they took about an hour on a laptop).
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]

A Batched Successive-Elimination Policy

For completeness, we include a description of the BaSE algorithm from Gao et al. [2019] below. The algorithm itself is quite simple: it just eliminates all arms that are suboptimal with high probability after each batch.

Algorithm 3: Batched Successive Elimination (BaSE) from Gao et al. [2019]

Input: Number K of arms, time horizon T, number of batches B, grid $t_0 = 0 < t_1, \ldots, t_B = T$, parameter γ.
Set $A = [K]$ (this will be the set of alive arms).
For each $j \in [K]$, initialize $n_j = 0$ and $\sigma_j = 0$.
for $b \leftarrow 1$ **to** $B - 1$ **do**
  for $t \leftarrow t_{b-1} + 1$ **to** t_b **do**
    Let j be the $(t \bmod |A|)$th element of A.
    Play arm j and receive reward x_t.
    $n_j \leftarrow n_j + 1$. $\sigma_j \leftarrow \sigma_j + x_t$.
  end
  Let $j^* = \arg\max_{j \in A} \sigma_j / n_j$.
  Let $\tau = \max_{j \in A} n_j$ (by construction, all n_j for $j \in A$ should be approximately equal).
  for $j \in A$ **do**
    if $\sigma_{j^*}/n_{j^*} - \sigma_j/n_j \geq \sqrt{\gamma/\tau}$ **then**
      Remove j from A.
    end
  end
end
for $t \leftarrow t_{b-1} + 1$ **to** t_b **do**
  Pull arm $j^* = \arg\max_{j \in A} \sigma_j / n_j$.
end

The authors show that when $B = \log\log T$, $\gamma = O(\log(KT))$, and $t_b = O(T^{1-1/2^b})$ the above algorithm incurs at most $\tilde{O}(\sqrt{KT})$ regret. In our applications, we will set $\gamma = O(\log(NKT))$; this guarantees that the probability the best arm is ever eliminated from A is at most $1/\mathrm{poly}(N, K, T)$ (see Lemma 1 of Gao et al. [2019]), and therefore with high probability Algorithm 1 will never have to abort.

B Simulations

In this appendix we empirically evaluate the learning algorithms we introduce on this paper on several synthetic problem instances. We implement the following algorithms:

1. The Explore-Then-Commit algorithm (Algorithm 2) of Section 3.5.
2. Algorithm 1 with *greedy decomposition*, as defined in Section 3.4.1. Recall that this decomposition algorithm simply assigns each of the users to the same arm, cycling through the arms until all the demand is met.
3. Algorithm 1 with *random decomposition*, as defined in the beginning of Section 3.4.2. This decomposition algorithm assigns each user to a random one of their active arms, and generates such assignments until all demand is met.
4. Algorithm 1 with *LP-based decomposition*, as defined in Corollary 1 of Section 3.4.2. This decomposition algorithm writes the demand vector as a convex combination of the vertices of the polytope $\mathcal{P}_C([K])$ (each of which corresponds to a specific assignment).
5. A non-anonymous UCB algorithm, where each user independently runs UCB over the K arms.

We evaluate these algorithms on two classes of instances, both with $N = 50$ users, $K = 5$ arms, anonymity parameter $C = 4$, and $T = 10^5$ rounds. In the first class of instances (Figure 1), the

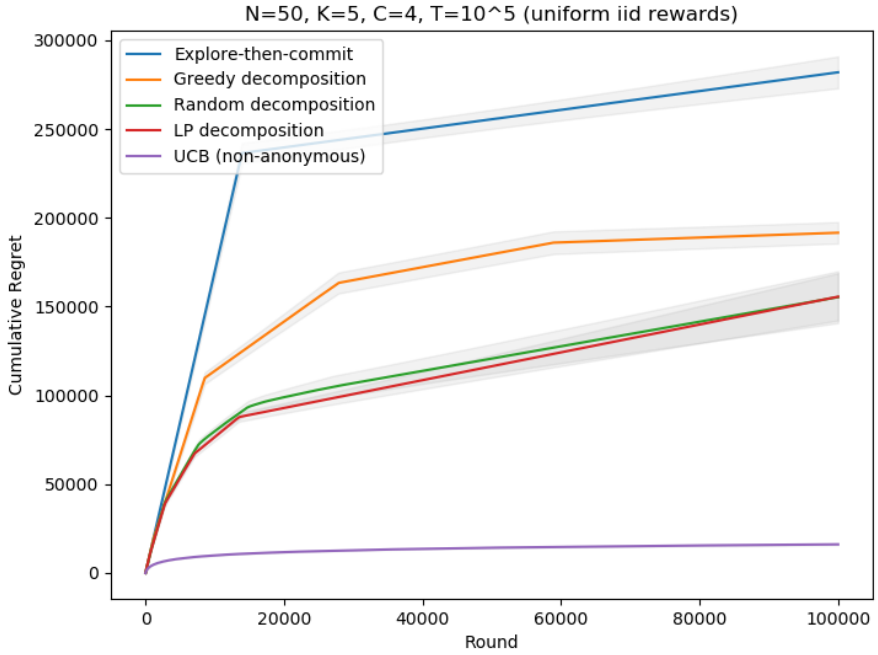

Figure 1: Cumulative regrets over time of five different learning algorithms on instances with uniform iid rewards (i.e., each user/arm distribution is a Bernoulli distribution with parameter independently drawn from $U([0,1])$). Grey regions represent 95% confidence intervals.

rewards for user i and arm j are drawn from a Bernoulli μ_{ij} distribution, where μ_{ij} is sampled from the uniform distribution $U([0,1])$ independently from all other means. The second class of instances (Figure 2) is generated by the following linear model: each user i is assigned a random unit norm 10-dimensional vector v_i and each action j is assigned a random unit norm 10-dimensional vector w_j. The rewards for user i and arm j are drawn from a Bernoulli μ_{ij} distribution, where $\mu_{ij} = 0.5(\theta_{ij} + 1)$, and where θ_{ij} is the cosine similarity between v_i and w_j.

With these parameters, both classes of instances almost always satisfy the user-cluster assumption for $U \geq C + 1$, so the preconditions to apply Algorithm 1 (and the various decomposition algorithms) almost always hold. Nonetheless, we implement Algorithm 1 semi-robustly: e.g., instead of aborting when we don't see a U-batched graph, we continue assigning users to random active arms. Similarly, we implement all algorithms so that they are agnostic to U (essentially, we set $U = C + 1$ in Algorithm 1 and compute an appropriate lower-bound on the approximation factor α). We optimize various hyperparameters of our algorithm (e.g. the learning rate of BaSE) by validation on an independent set of learning instances. See the attached code for additional details.

Figure 1 shows the average cumulative rewards (and confidence intervals) of twenty independent runs of these algorithms on the class of instances with uniformly generated rewards. Unsurprisingly, the one algorithm which violates anonymity (parallel UCB) does much better than all the C-anonymous algorithms, obtaining a total regret about 10 times smaller than the next better (this is consistent with our regret bounds, which indicate that at best our anonymous algorithms incur at least $\Omega(C)$ more regret than non-anonymous algorithms). The ordering of the various anonymous algorithms is also as expected. Explore-then-commit (which works without any guarantee on U and only uses one "batch") performs significantly worse than the variations of Algorithm 1 based off batched bandit algorithms. Within the variations of Algorithm 1, the naive greedy decomposition performs worst. The variations with random decomposition and LP decomposition perform similarly; this is not too surprising given that the LP decomposition is in some sense a derandomization of the random decomposition.

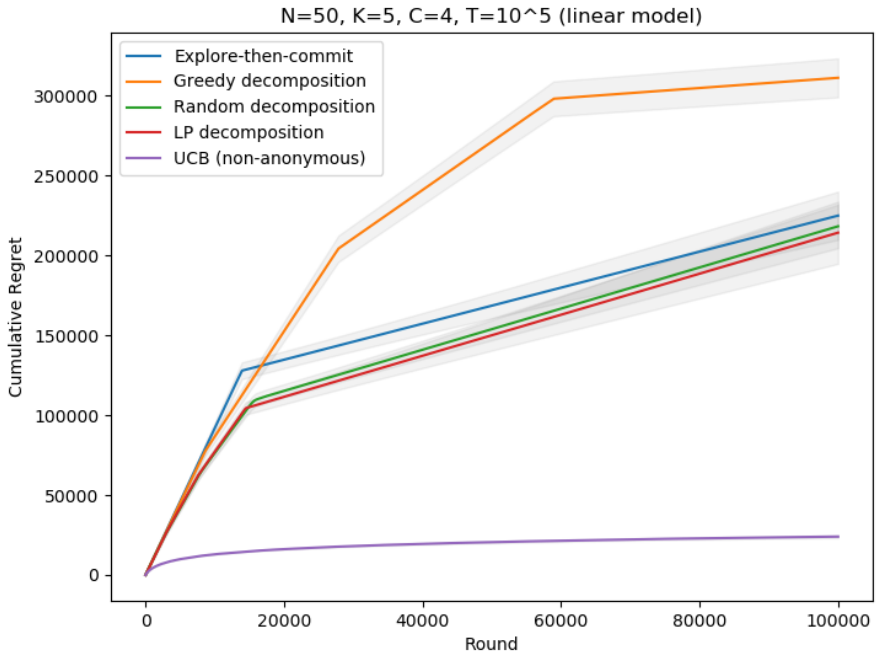

Figure 2: Same as Figure 1, but where reward distributions arise from a linear model instead of being generated uniformly (see text).

Figure 2 shows the average cumulative rewards (and confidence intervals) of twenty independent runs of these algorithms on instances generated by the linear model defined above. In most aspects, it is very similar to Figure 1; one major qualitative difference, however, is that greedy decomposition appears to perform much worse on this class of instances (whereas greedy decomposition outperformed explore-then-commit in Figure 1, it does markedly worse than explore-then-commit here).

C Omitted proofs

C.1 Proof of Lemma 1

Proof of Lemma 1. First note that with a simple Hoeffding bound for an i and a b we have

$$\begin{aligned}
&Pr\Big(|\bar{Y}^i(t_b) - \mu_i| \leq \frac{1}{2}\sqrt{\frac{\gamma \log(TK)}{\tau_b}}\Big) \\
&< 2\exp\Big(-2\tau_b \frac{1}{4}\frac{\gamma \log(TK)}{\tau_b}\Big) \\
&= 2\exp\Big(-0.5\gamma \log(TK)\Big) \\
&\leq \frac{1}{(TK)^2}. \qquad \text{for large enough constant } \gamma \text{ and } K \geq 2
\end{aligned}$$

By union bound this holds for all i and b with probability at least $1 - \frac{1}{TK}$. In the rest, w.l.g. we assume this holds, since $D_b \cdot \mathbb{E}\left[\max_{j \in A_b} \Delta_j\right] \times \frac{1}{TK} \leq \frac{D_b}{TK} \leq 1$. This means that for any j that

survives round $b-1$, (i.e. $\forall j \in A_b$) we have

$$\begin{aligned}\Delta_j = \mu^* - \mu_j &\leq 2\sqrt{\frac{\gamma \log(TK)}{\tau_{b-1}}} \\ &\leq 2\sqrt{\frac{\gamma \log(TK)}{\frac{u_{b-1}}{K}}} \\ &= 2\sqrt{\frac{\gamma K \log(TK)}{a^{2-2^{2-b}}}}\end{aligned}$$

Therefore we have

$$\begin{aligned}D_b \cdot \mathbb{E}\left[\max_{j \in A_b} \Delta_j\right] &\leq D_b \times 2\sqrt{\frac{\gamma K \log(TK)}{a^{2-2^{2-b}}}} \\ &\leq u_b \times 2\sqrt{\frac{\gamma K \log(TK)}{a^{2-2^{2-b}}}} \\ &= a^{2-2^{1-b}} \times 2\sqrt{\frac{\gamma K \log(TK)}{a^{2-2^{2-b}}}} \\ &= \frac{a^{2-2^{1-b}}}{a^{1-2^{1-b}}} \times 2\sqrt{\gamma K \log(TK)} \\ &= 2a\sqrt{\gamma K \log(TK)} \\ &= \Theta\Big(T^{\frac{1}{2-2^{1-B}}}\sqrt{\gamma K \log(TK)}\Big) \\ &= \Theta\big(\sqrt{\gamma TK \log(TK)}\big).\end{aligned}$$

□

C.2 Proof of Lemma 2

Proof of Lemma 2. Assume (as above) that user i is the kth user in G_s. Let $G'_s = G_{s,k} = G_s \setminus \{i\}$. For each user $i' \in G_s$, let $X_{i'}$ denote the r.v. representing the reward user i' contributed in round 0 of the above procedure, and for each user $i' \in G'_s$, let $Y_{i'}$ denote the r.v. representing the reward user j contributed in round k of the above procedure. Note that (by the definition of our setting), the value $X_{i'}$ and $Y_{i'}$ are independent r.v.s with $\mathbb{E}[X_{i'}] = \mathbb{E}[Y_{i'}] = \mu_{i',\pi(i')}$. In addition, $|X_{i'}|, |Y_{i'}| \in [0,1]$, so $\mathrm{Var}(X_{i'}), \mathrm{Var}(Y_{i'}) \leq \frac{1}{4}$.

Then, since

$$\hat{\mu}_{i,\pi(i)} = r_{s,0} - r_{s,k} = X_i + \sum_{i' \in G'} (X_{i'} - Y_{i'}),$$

it follows that

$$\mathbb{E}[\hat{\mu}_{i,\pi(i)}] = \mathbb{E}[X_i] + \sum_{i' \in G'_s} (\mathbb{E}[X_{i'}] - \mathbb{E}[Y_{i'}]) = \mathbb{E}[X_i] = \mu_{i,\pi(i)}.$$

Furthermore, note that each of the r.v.s $X_{i'}$ and $Y_{i'}$ are 1-subgaussian. Since $\hat{\mu}_{i,\pi(i)}$ is the sum of at most $4C+1$ independent 1-subgaussian variables, $\hat{\mu}_{i,\pi(i)}$ itself is $(4C+1)$-sub-Gaussian (and hence $O(C)$-subgaussian).

□

C.3 Proof of Theorem 1

Proof of Theorem 1. Fix a user i. We will show that the expected regret incurred by user i is at most $\tilde{O}(C\sqrt{\alpha KT})$, thus implying the theorem statement. As mentioned in Appendix A, the guarantees

of BaSE imply that with high probability this algorithm will never abort, so from now on we will condition on the algorithm not aborting.

Consider the expected regret incurred by user i during batch b of Algorithm 1. Since i is only ever assigned to arms in $A_{i,b}$ during this batch, and since user i gets matched a total of $\alpha(2C+2)D_b$ times during this batch, this expected regret is at most

$$\alpha(2C+2)D_b \cdot \mathbb{E}\left[\max_{j \in A_{i,b}} \Delta_{i,j}\right],$$

where $\Delta_{i,j} = (\max_{j'} \mu_{i,j'}) - \mu_j$. On the other hand, by Lemma 1 (and using the fact that the feedback we provide to BaSE is $O(C)$-subgaussian by Lemma 2), this is at most

$$\alpha(2C+2) \cdot \tilde{O}\left(\sqrt{C} \cdot \sqrt{KT'}\right) = \tilde{O}(C\sqrt{\alpha KT}),$$

as desired. □

C.4 Proof of Lemma 3

Proof of Lemma 3. By Caratheodory's theorem, if $\beta w \in \mathcal{P}_C$, we can write βw as the convex combination of at most $NK+1$ vertices $v^{(1)}, v^{(2)}, \dots v^{(NK+1)} \in \mathcal{P}_C$. Let us write

$$\beta w = \sum_{\ell=1}^{NK+1} \lambda_\ell v^{(\ell)} \tag{2}$$

where $\sum_\ell \lambda_\ell = 1$. In particular, we have that:

$$Dw = \sum_{\ell=1}^{NK+1} \frac{D\lambda_\ell}{\beta} v^{(\ell)}. \tag{3}$$

Consider the decomposition which contains $\lceil D\lambda_\ell/\beta \rceil$ instances of the assignment $v^{(\ell)}$ for each $1 \le \ell \le NK+1$ (here, "the assignment $v^{(\ell)}$" refers to any assignment which sends i to j if $v_{ij}^{(\ell)} = 1$). By (3), this assignment is guaranteed to contain at least $Dw_{ij} = D/|A_i|$ assignments which are informative for the pair (i,j), so this is a valid C-anonymous decomposition. Moreover, the number of assignments in this decomposition is at most

$$\sum_{\ell=1}^{NK+1} \left\lceil \frac{D\lambda_\ell}{\beta} \right\rceil \le (NK+1) + \sum_{\ell=1}^{NK+1} \frac{D\lambda_\ell}{\beta} = \frac{1}{\beta} D + NK + 1.$$

Likewise, if there are $R = \alpha D$ assignments $M_1, M_2, \dots M_R$ which form a C-anonymous decomposition of G, let $v^{(1)}, v^{(2)}, \dots, v^{(R)}$ be the vertices of $\mathcal{P}_C$ corresponding to these assignments. Then we are guaranteed that

$$Dw \le v^{(1)} + v^{(2)} + \cdots + v^{(R)},$$

so in particular we must have

$$\frac{1}{\alpha} w = \lambda \cdot \frac{v^{(1)} + v^{(2)} + \cdots + v^{(R)}}{R} + (1-\lambda) \cdot 0$$

for some $\lambda \in [0,1]$. Note that the RHS is a convex combination of vertices of $\mathcal{P}_C$ (in particular, $0 \in \mathcal{P}_C$), so $\frac{1}{\alpha} w \in \mathcal{P}_C$. □

C.5 Proof of Lemma 4

Proof of Lemma 4. Note that the vertices of $\mathcal{P}_C(S)$ satisfy all the constraints in (1), and every integral point satisfying (1) satisfies the conditions to be a vertex of $\mathcal{P}_C(S)$. It therefore suffices to show that the polytope defined by (1) is integral.

But this immediately follows, since the matrix of constraints in (1) are equivalent to the totally unimodular matrix defining the bipartite matching polytope between $[N]$ and $[K]$. □

C.6 Proof of Lemma 5

Proof of Lemma 5. We show w satisfies the constraints of (1) for $S = [K]$. The only nontrivial constraint is showing that

$$\sum_{i \in [N]} w_{ij} \geq C + 1, \ \forall j \in [K].$$

Note that by the definition of the user-cluster assumption, there are at least U choices of i for which $w_{ij} > 0$. By the definition of w_{ij}, if $w_{ij} > 0$, $W_{ij} \geq 1/K$. Since $U \geq K(C+1)$, the above inequality immediately follows. □

C.7 Proof of Theorem 2

Proof of Theorem 2. Note that, by construction, in the first T_{exp} rounds of this algorithm, we obtain $\approx T_{exp}/(K(2C+2))$ independent unbiased estimators $\hat{\mu_{i,j}}$ from the feedback-eliciting sub-algorithm for each user/arm pair (i, j). That is, for each (i, j) we are guaranteed that each $n_{ij} \geq T_{exp}/(K(2C+2))$ and that σ_{ij} is the sum of n_{ij} independent copies of an unbiased estimator for μ_{ij} with C-subgaussian noise.

It follows that $\sigma_{ij}/n_{ij} = m_{ij}$ is an unbiased estimator for μ_{ij} with C/n_{ij}-subgaussian noise. Let $\max_{ij} C/n_{ij} = \lambda$; note that $\lambda \leq KC(2C+2)/T_{exp} = \tilde{O}(K^{2/3}C^{4/3}T^{-2/3})$.

Since $m_{ij} - \mu_{ij}$ is λ-subgaussian, it follows from Hoeffding's inequality that

$$\Pr\left[|m_{ij} - \mu_{ij}| \geq \varepsilon\right] \leq 2\exp\left(-\frac{\varepsilon^2}{2\lambda}\right).$$

We will set $\varepsilon = \sqrt{4\lambda \log(NKT)}$; it then follows that

$$\Pr\left[|m_{ij} - \mu_{ij}| \geq \varepsilon\right] \leq \frac{2}{(NKT)^2}.$$

Union-bounding over all NK pairs $(i, j) \in [N] \times [K]$, with high probability (at least $1 - 1/T^2$) all estimates m_{ij} are within ε of μ_{ij}. Fix $i \in [N]$. It then follows that if we let $\hat{j} = \arg\max_j \sigma_{ij}/n_{ij}$, that $\mu_{i\hat{j}} \geq \max_j \mu_{ij} - 2\varepsilon$. In particular, for each remaining round after round T_{exp}, user i incurs at most ε regret.

The total expected regret from rounds up to T_{exp} is at most NT_{exp}. The total expected regret from rounds after T_{exp} is at most $NT\varepsilon$. It follows that the overall expected regret is at most $N(T_{exp} + T\varepsilon) = \tilde{O}(NC^{2/3}K^{1/3}T^{2/3})$ as desired. □

C.8 Proof of Corollary 1

Proof of Corollary 1. Follows from Lemmas 3 and 5. The decomposition of w into vertices of $\mathcal{P}_c([K])$ can be found efficiently from formulation (1) of $\mathcal{P}_c([K])$ as the intersection of a small number of half-spaces and following the algorithmic proof of Caratheodory's theorem (see e.g. Hoeksma et al. [2016] for details). □

C.9 Proof of Lemma 6

Proof of Lemma 6. Fix a, and set $w_{ij}^{(a)} = 1/|A_i \cap S_a|$ if $j \in A_i \cap S_a$, and $w_{ij}^{(a)} = 0$ otherwise. We first claim $w^{(a)} \in \mathcal{P}_c(S_a)$. As in Lemma 5, the only nontrivial condition to check is whether

$$\sum_{i \in [N]} w_{ij}^{(a)} \geq C + 1$$

holds for all $j \in S_a$. As before, for each $j \in S_a$, there must be at least U users i for which $j \in A_i$, and thus at least U users i where $w_{ij}^{(a)} > 0$. But now, if $w_{ij}^{(a)} > 0$, then $w_{ij}^{(a)} \geq 1/(U/(C+1)) = (C+1)/U$, and the above inequality follows.

To see that $w \leq \sum_{i=1}^{\alpha} w^{(a)}$, note that if $j \in A_i \cap S_a$, then $w_{ij} = 1/|A_i| \leq 1/|A_i \cap S_a| = w_{ij}^{(a)}$. □

C.10 Proof of Corollary 2

Proof of Corollary 2. By Lemma 6, we can partition K into α sets S_a and write

$$\frac{1}{\alpha} w = \lambda \cdot \left(\frac{1}{\alpha} \sum_{a=1}^{\alpha} w^{(a)} \right) + (1 - \lambda) \cdot 0$$

where $\lambda \in [0, 1]$ and each $w^{(a)} \in \mathcal{P}_C(S_a)$. This proves that $\frac{1}{\alpha} w \in \mathcal{P}_C$, and thus such a decomposition exists by Lemma 3. Moreover, we can efficiently find such a decomposition by decomposing each of the terms $w^{(a)}$ into a convex combination of 0 (which lies in $\mathcal{P}_C(S)$ for all S) and at most $N|S_a|$ other vertices of $\mathcal{P}_C(S_a)$ (since $\mathcal{P}_C(S_a)$ is only $N|S_a|$-dimensional). □

C.11 Proof of Theorem 3

Proof of Theorem 3. Consider the following variant of the anonymous bandits problem. As in the anonymous bandits problem, an instance of this problem is specified by a number of users N, a number of arms K, an anonymity parameter C, a time horizon T, and for every pair of user i and arm j a 1-subgaussian reward distribution $\mathcal{D}_{i,j}$ with mean $\mu_{i,j} \in [0, 1]$. However, unlike the anonymous bandits problem which has a centralized learner, in this variant each user is an independent learner – moreover, we prohibit users from communicating with one another or observing the feedback of other users (in this problem, users' actiuns will not interact). On the other hand, we will tell each user i all distributions $\mathcal{D}_{i',j}$ belonging to users $i' \neq i$ (so they only need to learn their own reward distributions).

During each round t each user i will choose both a target arm $j_{i,t}$ (as in anonymous bandits), *and also* a subset $F_{i,t} \subseteq [N] \setminus \{i\}$ of at least $C - 1$ other users. User i (indirectly) obtains reward $r_{i,j_{i,t}}$, but only observes as feedback the sum

$$r_{i,\pi_t(i)} + \sum_{i' \in F_{i,t}} r_{i',j_{i,t}},$$

where each $r_{i,j}$ is independently drawn from $\mathcal{D}_{i,j}$. The goal is to maximize the total reward among all users, and regret is defined analogously to how it is in the anonymous bandits problem.

We first claim the above problem is strictly easier than the anonymous bandits problem, in the sense that any algorithm for the anonymous bandits problem that achieves expected regret R on a specific instance can be converted to an algorithm for the above problem that achieves expected regret R on the analogous instance. To see this, fix an algorithm $\mathcal{A}$ for the anonymous bandits problem and a user i in our variant of the problem. Note that user i can accurately simulate $\mathcal{A}$ with their knowledge of other distributions $\mathcal{D}_{i',j}$ and the feedback provided in the variant: in particular, if $\mathcal{A}$ plays assignment π_t in round t, user i should set $j_t = \pi_t(i)$ and set $F_{i,t} = \pi_t^{-1}(i)$ (if $|\pi_t^{-1}(i)| < C$, $\mathcal{A}$ got no information on i in round t and the user can set $F_{i,t}$ arbitrarily). User i then receives as feedback the aggregate reward from their group in the current execution of $\mathcal{A}$, and can simulate

aggregate rewards from other groups by sampling from $\mathcal{D}_{i',j'}$. By doing this, user i receives the same expected reward in this problem as user i would in the analogous anonymous bandits problem.

We now show the above problem is hard. Consider the following distribution over instances of the above problem. Fix K and C, and choose an $N \gg K^2(C+1)$ and $T \gg \max(K, C, N)$. For each $i \in [N]$, choose an arm $j^*(i)$ uniformly at random from $[K]$ (this will be user i's unique favorite arm; since $N \gg K \cdot K(C+1)$, with high probability this assignment will satisfy the user-cluster assumption for $U = K(C+1)$). For each pair $i \in [N]$ and $j \in [K]$, let $\mathcal{D}_{i,j} = \mathcal{N}(0,1)$ if $j \neq j^*(i)$, and let $\mathcal{D}_{i,j} = \mathcal{N}\left(\sqrt{C/T}, 1\right)$ if $j = j^*(i)$.

Consider the problem faced by user i when faced with this distribution. If in round t user i selects arm j, then regardless of their choice of set $F_{i,t}$, they will observe a random variable drawn from $\mathcal{N}(\mu_{i,j} + M, C)$, where $M = \sum_{i' \in F_{i,t}} \mu_{i',j}$ is an offset term known to user i. After subtracting out M, this means that if user i selects arm j, they get an independent random variable from $\mathcal{N}(\mu_{i,j}, C)$. Since exactly one of the $\mu_{i,j}$ (as j ranges over $[K]$) equals $\sqrt{C/T}$ and all other $\mu_{i,j} = 0$, this is exactly the hard distribution for classic C-Gaussian bandits. It is known (see Chapter 15 of Lattimore and Szepesvári [2020]) that any algorithm must incur at least $\Omega(\sqrt{CKT})$ regret on this distribution of problem instances. It therefore follows that each user i incurs at least $\Omega(\sqrt{CKT})$, and therefore overall we incur at least $\Omega(N\sqrt{CKT})$ regret over all N users. □

C.12 Proof of Theorem 4

Proof of Theorem 4. For a fixed value of T, we will randomize uniformly between the following two instances. In both instances we will set $N = K = 3$, $C = 2$, and all reward distributions $\mathcal{D}_{i,j}$ will be Bernoulli distributions (defined by their mean $\mu_{i,j}$). Let $\varepsilon = T^{-1/3}$. In the first instance we set

$$\begin{bmatrix} \mu_{11} & \mu_{12} & \mu_{13} \\ \mu_{21} & \mu_{22} & \mu_{23} \\ \mu_{31} & \mu_{32} & \mu_{33} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{2} - \varepsilon & \frac{1}{2} + \varepsilon \\ 0 & \frac{1}{2} + \varepsilon & \frac{1}{2} - \varepsilon \end{bmatrix},$$

and in the second instance we set

$$\begin{bmatrix} \mu_{11} & \mu_{12} & \mu_{13} \\ \mu_{21} & \mu_{22} & \mu_{23} \\ \mu_{31} & \mu_{32} & \mu_{33} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{2} + \varepsilon & \frac{1}{2} - \varepsilon \\ 0 & \frac{1}{2} - \varepsilon & \frac{1}{2} + \varepsilon \end{bmatrix}.$$

Intuitively, user 1 likes only arm 1 (and arm 1 is only liked by user 1). User 2 either slightly prefers arm 2 to arm 3 or arm 3 to arm 2, and user 3 has the opposite preferences of user 2. We will let the random variable X denote which instance we are in, with $X = 1$ denoting the first instance and $X = -1$ denoting the second.

Consider the set of actions that reveal information about the value of X. Since $C = 2$, we must allocate at least 2 people to an arm. We know all the rewards (deterministically) for arm 1, so we must allocate this group of people to either arm 2 or 3. Finally, if we allocate both user 2 and user 3 to arm 2 or 3, we learn nothing about the instance we belong to (since they have symmetric actions). Therefore the only way to gain any information about X is to either allocate users $\{1, 2\}$ or $\{1, 3\}$ to one of arms 2 or 3.

Any of these allocations gives us equivalent information about X; in one of the two instances, we will receive a random variable $1 + \mathrm{Bern}(1/2 + \varepsilon)$, and in the other instance we will receive a random variable $1 + \mathrm{Bern}(1/2 - \varepsilon)$. In addition, in any of these allocations, we incur at least a net regret of 1 from querying this allocation (since we assign user 1 to an arm with reward 0). We call any assignment involving one of these allocations an "information-revealing assignment".

Assume we have an algorithm which sustains expected regret at most R. Since each information-revealing assignment incurs regret at least 1, by Markov's inequality the probability the algorithm performs more than $2R$ information-revealing assignments is at most $1/2$. Now, assume that at round t, the algorithm has performed only r information-revealing assignments. By the discussion above, this means that the only information the algorithm has regarding X is r samples from $\mathrm{Bern}(1/2 + X\varepsilon)$.

In general, the best statistical distinguisher between $\mathrm{Bern}(1/2+\varepsilon)$ and $\mathrm{Bern}(1/2-\varepsilon)$ requires at least C/ε^2 samples (for some constant $C > 0$) to distinguish these two distributions with probability at least $3/4$. Therefore, if $r \leq C/\varepsilon^2$, our learning algorithm cannot distinguish between $X = 0$ and $X = 1$ with probability greater than $3/4$, and is therefore guaranteed to incur at least $\Omega(\varepsilon)$ regret (by e.g. allocating user 2 to the wrong arm).

Now, if $2R < C/\varepsilon^2$, then in every round we have performed fewer than C/ε^2 information-revealing assignments, so we incur at least $\Omega(\varepsilon T) = \Omega(T^{2/3})$ regret. On the other hand, if $2R \geq C/\varepsilon^2$, then $R \geq C/2\varepsilon^2 = \Omega(T^{2/3})$, so in this case the algorithm also incurs at least $\Omega(T^{2/3})$ regret. □

C.13 Proof of Theorem 5

Proof of Theorem 5. For a fixed value of T, we will randomize uniformly between the following two instances. In both instances $N = K = C = 2$, and in both instances all reward distributions are completely deterministic. In the first instance we set $(\mu_{11}, \mu_{12}, \mu_{21}, \mu_{22}) = (1, 0, 0, 1)$ (user 1 likes arm 1 and user 2 likes arm 2); in the second instance we set $(\mu_{11}, \mu_{12}, \mu_{21}, \mu_{22}) = (0, 1, 1, 0)$ (user 1 likes arm 2 and user 2 likes arm 1).

Since $C = 2$, the only actions we can take which result in feedback are assigning both users to the same arm, but in both problem instances this results in an aggregate reward of 1 (and hence provides no information as to which instance we have chosen). By the choice of rewards, if an assignment receives reward x in the first instance, it receives reward $2 - x$ in the second instance – since it is impossible to learn any information about the choice of instance, this means any algorithm receives expected reward 1 per round. On the other hand, the optimal algorithm for each instance receives an expected reward of 2 per round. This implies the expected regret is at least T, as desired. □