DIFFERENTIALLY PRIVATE ALGORITHMS FOR EFFICIENT ONLINE MATROID OPTIMIZATION

Kushagra Chandak
Department of Computing Science, University of Alberta
Edmonton, Canada
kchandak@ualberta.ca

Bingshan Hu
Department of Computing Science, University of Alberta
Alberta Machine Intelligence Institute (Amii)
Edmonton, Canada
bingsha1@ualberta.ca

Nidhi Hegde
Department of Computing Science, University of Alberta
Canada CIFAR AI Chair, Alberta Machine Intelligence Institute (Amii)
Edmonton, Canada
nidhi.hegde@ualberta.ca

ABSTRACT

A matroid bandit is the online version of combinatorial optimization on a matroid, in which the learner chooses K actions from a set of L actions that can form a matroid basis. Many real-world applications such as recommendation systems can be modeled as matroid bandits. In such learning problems, the revealed data may involve sensitive user information. Therefore, privacy considerations are crucial. We propose two simple and practical differentially private algorithms for matroid bandits built on the well-known Upper Confidence Bound algorithms and Thompson Sampling. The key idea behind our first algorithm, Differentially Private Upper Confidence Bound for Matroid Bandits (DPUCB-MAT), is to construct differentially private upper confidence bounds. The second algorithm, Differentially Private Thompson Sampling for Matroid Bandits (DPTS-MAT), is based on the idea of drawing random samples from differentially private posterior distributions. Both algorithms achieve $O\left(L\ln(n)/\Delta + LK\ln(n)\min\left\{K, \ln(n)\right\}/\varepsilon\right)$ regret bounds, where Δ denotes the mean reward gap and ε is the required privacy parameter. Our derived regret bounds rely on novel technical arguments that deeply explore the special structure of matroids. We show a novel way to construct ordered pairs between the played actions and the optimal actions, which contributes to decomposing a matroid bandit problem into K stochastic multi-armed bandit problems. Finally, we conduct experiments to demonstrate the empirical performance of our proposed learning algorithms on both a synthetic dataset and a real-world movie-recommendation dataset.

1 INTRODUCTION

We study the learning problem of stochastic matroid bandits first proposed by Kveton et al. (2014) with differential privacy. In a stochastic matroid bandit problem, we have a matroid $(E, \mathcal{I})$ and a stochastic environment $\nu(E)$, where E is a set of L *base arms* and $\mathcal{I}$ is a set of *independent sets*. Each base arm $e \in E$ is associated with a weight that is independently drawn from a fixed but unknown probability distribution P_e collected by $\nu(E) = (P_e : e \in E)$. As matroids generalize the notion of linear independence in vector spaces, we are interested in a *basis* of a matroid which is a maximal independent set $A \in \mathcal{I}$. All the bases of a matroid have the same size denoted by K. Each basis can be viewed as a *super arm* which consists of exactly K base arms. An important problem for matroid bandits is to learn the basis with the maximum total weight by interacting with environment $\nu(E)$ sequentially. In each round $t \in [n]$, the learner chooses a basis A^t and simultaneously, the environment generates a random weight vector $w_t := (w_t(e_1), w_t(e_2), \ldots, w_t(e_L))$ for all the base arms $e_i \in E$. At the end of round t, the learner observes each individual weight $w_t(e)$ for all $e \in A^t$ and obtains as *return* the sum of the weights associated with the played basis. The goal of the learner is to choose bases sequentially to maximize the total return over a finite number of n rounds.

The application of recommending a set of diverse movies to users can be framed as a matroid bandit learning problem (Kveton et al., 2014), where the learning algorithm recommends a set of carefully chosen movies to the users. In this example, each movie can be characterized by a feature vector denoting the genres of that movie. In each round, the set of recommended movies can be viewed as a super arm. Since we would like the recommended movies to

represent diverse genres, the matroid structure ensures that the feature vectors of the recommended movies are linearly independent, which further indicates that the recommended movies do not include movies with similar genres. At the end of the round, based on the feedback from the users, the learning algorithm adjusts its future recommendations.

Since the learner only observes the weights associated with the played super arm, the central challenge that the learner faces is the *exploitation-versus-exploration* dilemma. In each round, the learner needs to decide whether to choose a super arm with the highest empirical return based on the past information (exploitation) or choose an under-explored super arm to gain information about the unknown environment (exploration). Upper Confidence Bound (UCB)-based (Auer et al., 2002; Garivier & Cappé, 2011) and Thompson Sampling-based learning algorithms (Kaufmann et al., 2012; Agrawal & Goyal, 2017) are both successful in balancing exploitation-versus-exploration. In bandit problems, the UCB-based algorithms are motivated by the principle of *optimism in the face of uncertainty* and can be justified by well-known concentration inequalities. In the UCB-based algorithms, each arm maintains a UCB index, which is an upper bound of the constructed confidence interval. Then, the learning algorithm makes a decision based on UCB indices for all the arms. Different from the UCB-based algorithms that rely on the UCB indices to tackle the exploitation-versus-exploration dilemma, Thompson Sampling-based algorithms maintain a posterior distribution for each arm. For each arm, a random sample is drawn from the posterior distribution. Then, the learning algorithm makes a decision based on the random samples for all the arms. If the learning algorithm knows the mean rewards of all the arms, pulling the arm with the highest mean reward is always the best strategy to achieve the highest expected reward. The performance of bandit algorithms is commonly measured by *Regret*, which represents the expected cumulative performance gap between always choosing the arm with the highest expected reward (fixed but unknown) and the learning algorithm's actual pulled arms.

The example of recommending diverse movies highlights the necessity for preserving the privacy of users. The feedback information from users (e.g. movie ratings) also reveals their watch history or preferences towards the recommended movies. Being users, we may wish to keep this information private. Take the Netflix Prize dataset for example. As shown in Narayanan & Shmatikov (2008), just anonymizing the ratings is not enough for preserving privacy. In this paper, we focus on the learning problem of matroid bandits with differential privacy. Differential privacy is the most commonly-used notion of privacy for machine learning algorithms (Dwork et al., 2014). If a learning algorithm is implemented in a differentially private manner, the information associated with an individual has almost no impact on the output of the learning. In other words, private learning algorithms are not sensitive to information from a single individual. In this work, we focus on ε-differential privacy, where ε is the privacy parameter. The parameter ε can be viewed as the privacy budget that can be distributed among different components of the learning algorithm.

Now, we summarize the key contributions of this work.

1. In this work, we propose two *sample efficient* and *computationally fast* differentially private algorithms for matroid bandits. Our first algorithm, `DPUCB-MAT`, is built upon the well-established UCB1 policy of Auer et al. (2002). The regret bound of `DPUCB-MAT` is $O\left(L\ln(n)/\Delta + \min\left\{K, \log(n)\right\} LK\ln(n)/\varepsilon\right)$, where Δ is the mean reward gap and ε is the required privacy parameter. Our second algorithm, `DPTS-MAT`, is built on Thompson Sampling which has demonstrated competitive practical performances (Chapelle & Li, 2011). The regret bound of `DPTS-MAT` is $O\left(L\ln(n)/\Delta + \min\left\{K, \log(n)\right\} LK\ln(n)/\varepsilon\right) + O\left(K\ln(n)/\Delta\right)$.
2. Regarding the regret analysis, we propose a unified approach to decompose the regret of `DPUCB-MAT` and `DPTS-MAT`. Our novel regret decomposition relies on the introduction of a round-dependent permutation on the order of the base arms in the optimal super arm. The permutation contributes to decomposing the regret of matroid bandits into K different stochastic bandit problems.
3. We conduct experiments to evaluate the empirical performance of our proposed algorithms by using both synthetic and real-world movie-recommendation datasets. The experimental results demonstrate that our proposed differentially private algorithms are efficient.

2 Differentially Private Matroid Bandits

In this section, we present our learning problem formally. We first introduce stochastic matroid bandits. Then, we introduce the definition of differential privacy for matroid bandit learning algorithms.

2.1 Matroid Bandits

A stochastic matroid bandit problem can be specified by $((E, \mathcal{I}); \nu(E))$, where $(E, \mathcal{I})$ defines a matroid and $\nu(E)$ defines an environment that a learner interacts with. In a matroid bandit problem, E is a set of L items, also called the *set of base arms*, and $\mathcal{I}$ is a family of subsets of E, also called the family of *independent sets*, defined by the

following three properties. (1) $\varnothing \in \mathcal{I}$. (2) If $X \subset Y$ and $Y \in \mathcal{I}$, then $X \in \mathcal{I}$. (3) If $X, Y \in \mathcal{I}$ such that $|X| = |Y| + 1$, then there exists $e \in X \setminus Y$ such that $Y \cup \{e\} \in \mathcal{I}$. Each base arm $e \in E$ is associated with a weight that is independently drawn from a fixed but unknown probability distribution P_e with mean $\bar{w}(e)$ collected by $\nu(E) = (P_e : e \in E)$. As matroids generalize the notion of linear independence in vector spaces, we are interested in a *basis* of a matroid which is a maximal independent set $A \in \mathcal{I}$. Similar to vector spaces, all bases of a matroid have the same cardinality denoted by K. Each basis $A \subset E$ can be viewed as a *super arm* which consists of exactly K base arms. An important problem for matroid bandits is to learn the basis with the maximum total weight, i.e., learn the super arm $A^* = \arg\max_{A \in \mathcal{I}} \sum_{e \in A} \bar{w}(e)$. If we knew $\bar{w}(e)$ for all $e \in E$, the optimal basis A^* would be found greedily as shown in Algorithm 1.

In reality, we cannot assume we know $\bar{w}(e)$ in advance. Instead, the learner can interact with environment $\nu(E)$ sequentially to learn $\bar{w}(e)$ with the following learning protocol. In each round $t \in [n]$, the environment generates a random weight vector $w_t = (w_t(e_1), \ldots, w_t(e_L))$ with each entry $w_t(e_i)$ i.i.d over time according to P_{e_i}. Simultaneously, the learner selects a super arm $A^t = \{a_1^t, \ldots, a_K^t\}$, a basis of the matroid. At the end of round t, the learner observes the random weights $\{w_t(e) : e \in A^t\}$ and receives as return the sum of all the weights in A^t, i.e., the learner obtains a return $f(A^t, w_t) = \sum_{e \in A^t} w_t(e)$. The goal of the learner is to choose super arms sequentially over a finite of n rounds to maximize the expected cumulative return.

We use (pseudo)-regret R_n to measure the quality of the learner's strategy for deciding which super arms to choose, which is defined as

$$R_n = \sum_{t=1}^{n} \mathbb{E}\left[f(A^*, \bar{w}) - f(A^t, \bar{w})\right] , \tag{1}$$

where the expectation is over the randomness in the learner's action selection strategy.[1] Without loss of generality, we assume $\bar{w}(a_1^*) \geq \ldots \geq \bar{w}(a_k^*) \geq \ldots \geq \bar{w}(a_K^*)$ and let $A^* := \{a_1^*, \ldots, a_k^*, \ldots, a_K^*\}$ be a descending ordered set based on the means $\bar{w}(a_k^*)$. For the regret bounds presentation, we also need the following notation. Let $\bar{A}^* = E \setminus A^*$ denote the set of all sub-optimal base arms. Let $\Delta_{e,k} = \bar{w}(a_k^*) - \bar{w}(e)$ denote the mean reward gap between a base arm $e \in E$ and an optimal base arm $a_k^* \in A^*$. Without loss of generality, we also assume $\Delta_{e,k} > 0$ for all $e \in \bar{A}^*$ and $k \in [K]$. For a sub-optimal base arm $e \in \bar{A}^*$, let $\Delta_{e,\min} = \min_{k \in [K]} \Delta_{e,k}$ denote the minimum mean reward gap.

Algorithm 1 Greedy algorithm to find a maximum weight basis of a matroid (Kveton et al., 2014)

Input: Matroid $(E, \mathcal{I})$ and weights $\bar{w}(e)$ for all $e \in E$; **Output**: A^*

1: Sort all weights in a descending order such that $\bar{w}(e_1) \geq \ldots \geq \bar{w}(e_L)$ for all $e_i \in E$
2: Initialize $A^* = \varnothing$
3: **for** $i = 1, \ldots, K$ **do**
4: $\quad \sigma(A^*) = \{e : e \in E \setminus A^*, A^* \cup \{e\} \in \mathcal{I}\}$ ▷ The set of base arms that can be added to A^*
5: $\quad A^* \leftarrow A^* \cup \{e_i\}$, where $e_i = \arg\max_{e \in \sigma(A^*)} \bar{w}(e)$. ▷ Add the base arm with the highest weight in $\sigma(A^*)$ to A^*
6: **end for**

2.2 Differential Privacy

Differential privacy (DP) is the most commonly-used notion for preserving the privacy of individuals in machine learning algorithms. It is based on the idea that the change of a single user's data in the dataset cannot impact the output of the learning algorithm too much. The insensitivity to a single user's data guarantees that from the algorithm's output, an adversary cannot learn the existence or anything useful about that user. Since in online learning, reward vectors are fed into learning algorithms in a streaming fashion and each reward vector can encode private information associated with an individual, we focus on the framework of event-level privacy (Dwork et al., 2010) in this work.

Before presenting the definition of differential privacy, we define what the neighbouring datasets are in bandit problems. In a bandit learning problem, a dataset is a sequence of reward vectors. Let $\mathbf{W}_{1:t} = (w_s : s \in [t])$ be the original reward vectors fed into the learning algorithm up to round t. Two reward sequences $\mathbf{W}_{1:t}$ and $\mathbf{W}'_{1:t}$ are called neighboring if they differ in at most one round $r \leq t$. That is, $w_r \neq w'_r$ and $w_s = w'_s$ for all $s \in [t] \setminus \{r\}$. With the definition of neighbouring reward sequences in mind, we now define differential privacy for matroid bandit algorithms.

Definition 1. (Differential privacy in matroid bandits). A matroid bandit algorithm $\mathcal{B}$ is ε-DP if, for all rounds $t \in [n]$, for every pair of neighboring datasets $\mathbf{W}_{1:t}$ and $\mathbf{W}'_{1:t}$, and for any set $\mathcal{D}$ of decisions, we have

$$\mathbb{P}\left\{\mathcal{B}(\mathbf{W}_{1:t}) \in \mathcal{D}\right\} \leq e^{\varepsilon} \cdot \mathbb{P}\left\{\mathcal{B}(\mathbf{W}'_{1:t}) \in \mathcal{D}\right\} . \tag{2}$$

[1] Note that $\bar{w}$ is the mean reward vector accounting for all the base arms.

The privacy parameter ε can be thought of as a privacy budget. It can be allocated among different components within the algorithm. An important property of DP is that it is immune to post-processing, which states that the privacy guarantee cannot get worse if we release a function of the output instead of the output itself (Smith & Ullman, 2021).

Fact 1. (Post-processing, (Dwork et al., 2014)). Let $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$ be arbitrary sets and n be a positive integer. Let $\mathcal{B} : \mathcal{X}^n \to \mathcal{Y}$ and $\mathcal{B}' : \mathcal{Y} \to \mathcal{Z}$ be randomized algorithms. If $\mathcal{B}$ is ε-DP, then the composition $\mathcal{B}'(\mathcal{B}(\cdot))$ is also ε-DP.

3 Related Works

Kveton et al. (2014) initiated the study of learning the maximum weight basis for matroid bandits. They proposed a UCB-based algorithm, Optimistic Matroid Maximization (OMM), that achieves the optimal $O(L\ln(n)/\Delta)$ regret bound. The key idea in OMM is to construct upper confidence bounds based on UCB1 of Auer et al. (2002) for all base arms. As OMM is built upon UCB1, OMM cannot be asymptotically optimal. Later, Talebi & Proutiere (2016) proposed another UCB-based algorithm, Efficient Sampling for Matroids (KL-OSM), that achieves the asymptotically optimal $(1+\epsilon)L\ln(n)\Delta/d_{\text{KL}}(\mu_* - \Delta, \mu_*) + o(\ln(n))$ regret bound, where $\epsilon > 0$ can be any value, μ_* denotes the mean reward of an optimal base arm and $d_{\text{KL}}(a, b)$ denotes the KL-divergence between two Bernoulli distributions with parameters $a, b \in (0, 1)$. The key idea to achieve asymptotic optimality is to construct upper confidence bounds based on KL-UCB of Garivier & Cappé (2011). Other than the aforementioned UCB-based algorithms, Wang & Chen (2018) proposed Combinatorial Thompson Sampling (CTS), a Thompson Sampling-based algorithm with Beta priors, for combinatorial bandits. Combinatorial bandits generalize the setting of matroid bandits with the following key features. In combinatorial bandits, the set of super arms does not have any special structure that we can utilize. Also, the size of a super arm is at most K instead of exactly K. Since matroid bandits are special cases of combinatorial bandits, by a refined regret analysis, CTS achieves an $O(L\ln(n)/\Delta) + \tilde{O}\left(L/\Delta^4\right)$ regret for matroid bandits.

Mishra & Thakurta (2015) initiated the study of stochastic multi-armed bandits with differential privacy and proposed both the UCB-based and Thompson Sampling-based learning algorithms. Their proposed algorithms rely on the post-processing property of differential privacy (Fact 1). More specifically, they first guarantee that the internal learning algorithm computing the empirical means is ε-differentially private. Then, from post-processing, they immediately conclude that the proposed bandit algorithms are ε-DP. However, their proposed learning algorithms are very sub-optimal due to the usage of the Tree-based Mechanism (Chan et al., 2011; Dwork et al., 2010) to inject noise to preserve privacy. Later, Chen et al. (2020) introduced differential privacy in combinatorial bandits and proposed a UCB-based algorithm, Differentially Private Combinatorial UCB (CUCB-DP), for differentially private combinatorial bandits. CUCB-DP achieves an $O\left(LK\ln^2(n)/\Delta\right) + \tilde{O}\left(LK\ln^3(n)/\varepsilon\right)$ regret upper bound and an $\Omega\left(LK\ln(n)/\Delta + LK\ln(n)/\varepsilon\right)$ regret lower bound.[2]

Recently, Hu et al. (2021); Azize & Basu (2022) devised optimal UCB-based algorithms and Hu & Hegde (2022) devised an optimal Thompson Sampling-based algorithm with Beta priors for differentially private stochastic bandits. These algorithms all achieve the optimal $O\left(L\ln(n)/\Delta + L\ln(n)/\varepsilon\right)$ regret bound. The key ideas to achieve optimality are the usage of "laziness" and "forgetfulness" along with the Laplace Mechanism (Dwork et al., 2014) to inject noise to mask the true empirical means instead of using the Tree-based Mechanism. The idea of laziness is to update the differentially private empirical mean of an arm in a delayed manner. We only update the DP empirical mean of an arm when a certain number of observations are available from that arm. The idea of forgetfulness is to use fresh observations to update the DP empirical mean. Once observations have been used, we abandon them. Since the change of one reward vector only impacts the aggregated reward by one at most, from Laplace Mechanism, we can add a noise drawn from $\text{Lap}\left(1/\varepsilon\right)$ to the aggregated reward of each arm when updating the DP empirical mean.[3].

Our proposed UCB-based algorithm `DPUCB-MAT` (Algorithm 2) can be viewed as a differentially private version of OMM. When setting $\varepsilon \to \infty$, i.e., in the non-private matroid bandit setting, the regret bound of `DPUCB-MAT` (Theorem 2) recovers the regret bound of OMM. When setting a suitable privacy parameter ε, our regret bound removes an extra $\log(n)$ factor as compared to the regret bound shown in Chen et al. (2020). Although our proposed Thompson sampling-based algorithm `DPTS-MAT` (Algorithm 3) could be seen as a differentially private version of CTS for matroid bandits, we use different proof techniques. Section 4.3 presents more detail. Table 1 summarizes the regret bounds for matroid bandits in both non-private and private settings.

[2]The $\tilde{O}(\cdot)$ notation hides an extra $\log\log T$ factor.

[3]The probability density function of a Laplace distribution $\text{Lap}\left(b\right)$ centered at 0 with b as the scale parameter is $h_b(y) = (1/2b)e^{-|y|/b}$ with $b > 0$.

Table 1: Regret upper bounds for UCB and TS-based algorithms for matroid bandits.

Algorithms	Non-private results	Private results (ours)
OMM (UCB1-based)	$O\left(\frac{L\ln(n)}{\Delta}\right)$ (Kveton et al., 2014)	$O\left(\frac{L\ln(n)}{\Delta}+\frac{\min\{K,\log(n)\}LK\ln(n)}{\varepsilon}\right)$
CTS (TS-based)	$\tilde{O}\left(\frac{L\ln(n)}{\Delta}+\frac{L}{\Delta^4}\right)$ (Wang & Chen, 2018)	$O\left(\frac{L\ln(n)}{\Delta}+\frac{\min\{K,\log(n)\}LK\ln(n)}{\varepsilon}\right)$

4 Algorithms and Analysis

In this section, we present our proposed learning algorithms and their theoretical guarantees. First, we introduce some notation used by Algorithm 2 and Algorithm 3. We denote by $T_e(t-1)$ the "effective" number of observations used to compute the empirical mean of a base arm $e \in E$ by the end of round $t-1$ and denote by $\hat{w}_{e,T_e(t-1)}(t-1)$ the empirical mean of a base arm e among these $T_e(t-1)$ observations. To have a differentially private version of $\hat{w}_{e,T_e(t-1)}(t-1)$ with a small amount of noise, we still use the ideas of "laziness" and "forgetfulness" introduced in Hu et al. (2021); Azize & Basu (2022); Hu & Hegde (2022). As the total privacy budget is ε and at the end of each round t there are exactly K observations revealed, from the basic composition property (Dwork et al., 2014), we know that each base arm has an ε/K privacy budget. Let $\varepsilon_0 := \varepsilon/K$. Let $\tilde{w}_{e,T_e(t-1)}(t-1) = \hat{w}_{e,T_e(t-1)}(t-1) + \frac{\mathrm{Lap}(1/\varepsilon_0)}{T_e(t-1)}$ denote the DP mean estimate of a base arm e. For each base arm e, we store the reward observations in the array $\mathcal{T}_e$, and $\Sigma\mathcal{T}_e$ denotes the sum of all the observations in $\mathcal{T}_e$.

4.1 Differentially Private UCB for Matroid Bandits

The UCB-based differentially private algorithm for matroid bandits DPUCB-MAT is shown in Algorithm 2. The general idea is to construct *differentially private upper confidence bound* $U_t(e)$ (Line 5) for each base arm $e \in E$. We construct $U_t(e)$ as

$$U_t(e) := \tilde{w}_{e,T_e(t-1)}(t-1) + \sqrt{\frac{3\ln(Kt)}{T_e(t-1)}} + \frac{3\ln(Kt)}{\varepsilon_0 \cdot T_e(t-1)} \quad . \tag{3}$$

With all these private upper confidence bounds $U_t(e)$ in hand, DPUCB-MAT selects the best super arm A^t in a greedy way, i.e., invoking Algorithm 1 with all $U_t(e)$ as input and A^t as output. In other words, DPUCB-MAT plays $A^t = \arg\max_{A\in\mathcal{I}} \sum_{e\in A} U_t(e)$. Then, the rewards $w_t(e)$ for $e \in A^t$ are revealed. For each individual $w_t(e)$, we add it to the corresponding $\mathcal{T}_e$ (Line 9). If the number of observations of any base arm e hits 2^{s_e+1}, then, it is the right time to update the DP mean estimate $\tilde{w}_{e,T_e(t-1)}(t-1)$ using the 2^{s_e+1} observations stored in $\mathcal{T}_e$ (Line 12). Since now all the observations in $\mathcal{T}_e$ have been processed, we increment the counter s_e by 1 and reset $\mathcal{T}_e$ (Line 13).

Algorithm 2 DPUCB-MAT

1: **Input**: Matroid $(E, \mathcal{I})$ and privacy budget ε
2: Observe $w_0(e) \sim P_e$, set $T_e \leftarrow 1, s_e \leftarrow 0, \tilde{w}_{e,T_e} \leftarrow w_0(e) + \mathrm{Lap}\left(\frac{1}{\varepsilon_0}\right), \mathcal{T}_e \leftarrow \{\}$, where $\varepsilon_0 = \frac{\varepsilon}{K}$ ▷ Initialization
3: **for** $t = 1, 2, \ldots$ **do**
4: **for** $e \in E$ **do**
5: $U_t(e) := \tilde{w}_{e,T_e} + \sqrt{\frac{3\ln(Kt)}{T_e}} + \frac{3\ln(Kt)}{\varepsilon_0 \cdot T_e}$
6: **end for**
7: Invoke Algorithm 1 with $U_t(e)$ for all $e \in E$ as input and A^t as output
8: Play super arm A^t
9: Observe $w_t(e) \sim P_e$ and add $w_t(e)$ to $\mathcal{T}_e$, $\quad \forall e \in A^t$
10: Find $B^t = \left\{e \in A^t : |\mathcal{T}_e| = 2^{s_e+1}\right\}$ ▷ Find all the base arms with the number of observations hitting 2^{s_e+1}
11: **for** $e \in B^t$ **do**
12: $T_e \leftarrow 2^{s_e+1}, \tilde{w}_{e,T_e} = \frac{\Sigma\mathcal{T}_e + \mathrm{Lap}(1/\varepsilon_0)}{T_e}$ ▷ $\Sigma\mathcal{T}_e$ denotes the sum of all the observations in $\mathcal{T}_e$
13: $s_e \leftarrow s_e + 1, \mathcal{T}_e \leftarrow \{\}$ ▷ Doubling the "effective" number of observations and reset $\mathcal{T}_e$
14: **end for**
15: **end for**

We now present theoretical guarantees for Algorithm 2.

Theorem 1. Algorithm 2 is ε-differentially private.

Proof. Since the learner makes decisions based on the differentially private upper confidence bounds and constructing differentially private upper confidence bounds (Lines 4 to 6) can be thought of as post-processing of the DP mean estimates, it suffices to show that as long as the empirical mean computation (lines 9 to 14) is ε-DP, then the matroid bandit learning algorithm itself is ε-DP. Now, we consider two neighboring reward vector sequences $\mathbf{W}$ and $\mathbf{W}'$ which differ in a reward vector in some round s, i.e., $w_s \neq w'_s$. Since the learner only observes the rewards associated with the base arms in the played super arm, changing from w_s to w'_s may only affect the DP mean estimates for all the base arms in A^s. In other words, the DP mean estimates of at most K base arms may be impacted. Since the learning algorithm uses fresh observations to update the DP mean estimates and for each base arm $e \in A^s$ the aggregated reward $\Sigma\mathcal{T}_e$ changes by at most one when working over $\mathbf{W}$ and $\mathbf{W}'$, from Laplace Mechanism, we know that adding a noise random variable sampled from $\mathrm{Lap}(1/\varepsilon_0)$ is enough to make the function computing the mean estimate of an individual base arm ε_0-DP. Since at most K base arms may be impacted, using the basic composition property of DP, we know that the learning algorithm to compute the DP mean estimates for all the base arms is ε-DP. □

Theorem 2. The regret of Algorithm 2 is

$$R_n \leq \sum_{e \in \bar{A}^*: \Delta_{e,\min} > 0} O\left(\frac{\ln(Kn)}{\Delta_{e,\min}} + \frac{\min\{K, \log(Kn)\} \cdot \ln(Kn)}{\varepsilon / K}\right), \tag{4}$$

where $\Delta_{e,\min} = \min_{k \in [K]: \Delta_{e,k} > 0} \Delta_{e,k}$.

Discussion Algorithm 2 can be viewed as a differentially private version of OMM of Kveton et al. (2014). When setting $\varepsilon \to \infty$, the differentially private matroid bandit problem boils down to the non-private matroid bandits. In this special setting, the regret bound shown in Theorem 2 matches both the regret upper bound and regret lower bound presented in Kveton et al. (2014). With a suitable privacy parameter ε, our regret bound improves the state-of-the-art regret bound presented in Chen et al. (2020) by removing an extra $\log(n)$ factor. [4]

Algorithm 3 DPTS-MAT

1: **Input**: Matroid $M = (E, \mathcal{I})$ and privacy parameter ε
2: Observe $w_0(e) \sim P_e$, set $T_e \leftarrow 1, s_e \leftarrow 0, \tilde{w}_e \leftarrow w_0(e) + \mathrm{Lap}\left(\frac{1}{\varepsilon_0}\right), \mathcal{T}_e \leftarrow \{\}$, where $\varepsilon_0 = \frac{\varepsilon}{K}$ ▷ Initialization
3: **for** $t = 1, 2, \ldots$ **do**
4: **for** $e \in E$ **do**
5: Set $w'_{e,T_e}(t) := \tilde{w}_{e,T_e} + \frac{3\ln(Kt)}{\varepsilon_0 \cdot T_e}$ ▷ Boost the parameters of the posterior distributions
6: Sample $\theta_e(t) \sim \mathcal{N}\left(w'_{e,T_e}(t), \frac{1}{T_e}\right)$ ▷ Draw random samples from a Gaussian distribution
7: **end for**
8: Invoke Algorithm 1 with $\theta_e(t)$ for all $e \in E$ as input and A^t as output
9: Play super arm A^t
10: Observe $w_t(e) \sim P_e$ and add $w_t(e)$ to $\mathcal{T}_e$, $\forall e \in A^t$
11: Find $B^t = \left\{e \in A^t : |\mathcal{T}_e| = 2^{s_e+1}\right\}$ ▷ Find all the base arms with the number of observations hitting 2^{s_e+1}
12: **for** $e \in B^t$ **do**
13: $T_e \leftarrow 2^{s_e+1}$, $\tilde{w}_{e,T_e} = \frac{\Sigma\mathcal{T}_e + \mathrm{Lap}(1/\varepsilon_0)}{T_e}$ ▷ $\Sigma\mathcal{T}_e$ denotes the sum of all the observations in $\mathcal{T}_e$
14: $s_e \leftarrow s_e + 1, \mathcal{T}_e \leftarrow \{\}$ ▷ Doubling the "effective" number of observations and reset $\mathcal{T}_e$
15: **end for**
16: **end for**

4.2 Differentially Private Thompson Sampling for Matroid Bandits

Different from the UCB-based algorithms where the exploration-exploitation trade-off is done by constructing confidence intervals centered at the mean rewards, Thompson sampling-based algorithms maintain posterior distributions that model the mean rewards of each base arm. As our goal is to design learning algorithms that have good regret guarantees with a finite time horizon, we can use Gaussian priors.[5]

Our Thompson Sampling-based differentially private algorithm for matroid bandits DPTS-MAT is shown in Algorithm 3. The general idea is to boost the parameter of the posterior distribution from $\tilde{w}_{e,T_e(t-1)}(t-1)$ to $w'_{e,T_e(t-1)}(t)$, where

[4] For the discussion, we already translate their regret bound (Theorem 8) from combinatorial bandits to matroid bandits.

[5] As proved in Agrawal & Goyal (2017), Thompson Sampling with Beta priors can be asymptotically optimal while Thompson Sampling with Gaussian priors may not be asymptotically optimal.

$w'_{e,T_e(t-1)}(t) = \tilde{w}_{e,T_e(t-1)}(t-1) + 3\ln(Kt)/(\varepsilon_0 \cdot T_e(t-1))$ for each base arm $e \in E$. Then, DPTS-MAT draws a random sample $\theta_e(t) \sim \mathcal{N}\left(w'_{e,T_e(t-1)}(t), 1/T_e(t-1)\right)$ for all $e \in E$. With all these differentially private posterior samples $\theta_e(t)$ in hand, DPTS-MAT selects the best super arm A^t in a greedy way, i.e., invoking Algorithm 1 with all $\theta_e(t)$ as input and A^t as output. That is also to say, DPTS-MAT plays $A^t = \arg\max_{A \in \mathcal{I}} \sum_{e \in A} \theta_e(t)$. DPTS-MAT uses the same way as Algorithm 2 to process the revealed observations, i.e., only update the DP mean estimate of a base arm $e \in A^t$ if the number of observations in $\mathcal{T}_e$ hits 2^{s_e+1}.

We now present theoretical guarantees for Algorithm 3.

Theorem 3. Algorithm 3 is ε-differentially private.

Proof. The proof is similar to the DP proof for Algorithm 2. It suffices to show that as long as the algorithm to compute the empirical mean of each base arm is ε_0-DP, then the learning algorithm is ε-DP. Note that lines 4 to 9 can be thought of as post-processing. Since Algorithm 3 and Algorithm 2 use the same way to process the obtained observations, by using the same arguments as in the proof of Theorem 1, we can conclude the proof. □

Theorem 4. The regret of Algorithm 3 is

$$R_n \leq \sum_{e \in \bar{A}^*: \Delta_{e,\min} > 0} O\left(\frac{\ln(Kn)}{\Delta_{e,\min}} + \frac{\min\{K, \log(Kn)\} \cdot \ln(Kn)}{\varepsilon/K}\right) + \sum_{k \in [K]: \Delta_{\min,k} > 0} O\left(\frac{\ln(Kn)}{\Delta_{\min,k}}\right), \tag{5}$$

where $\Delta_{e,\min} = \min_{k \in [K]: \Delta_{e,k} > 0} \Delta_{e,k}$ and $\Delta_{\min,k} = \min_{e \in \bar{A}^*: \Delta_{e,k} > 0} \Delta_{e,k}$.

Discussion Different from Theorem 2 where the regret bound only has one term which is linear in the size of the sub-optimal base arms, there are two terms in Theorem 4. The first term is the same as the regret bound shown in Theorem 2 and it captures the regret for introducing differential privacy. This term characterizes the regret in all the rounds when the posterior distributions of the sub-optimal base arms are not concentrated. It is not surprising why we have the second non-private term which is linear in the size of the optimal base arm set. The second term upper bounds the regret among all the rounds when the posterior distributions of the optimal base arms are not concentrated. As will be shown in Section 4.3, the core of our regret decomposition is to decompose a matroid bandit problem into K stochastic bandit problems. For each individual bandit problem, by slightly modifying the regret analysis in Agrawal & Goyal (2017), when using Gaussian priors, an additive $O\left(\ln(n)/\Delta_{\min}\right)$ regret occurs among all the rounds when the Gaussian posterior distributions of the optimal arm are not concentrated. In contrast, when using Beta priors, the additive term can be $\tilde{O}\left(1/\Delta^4_{\min}\right)$, where $\tilde{O}$ hides problem-dependent constants. Since the regret for a matroid bandit is composed of K different stochastic bandit problems, we have the second term in Theorem 4.

4.3 Regret Decomposition

In this section, we present a unified approach to decompose the regret of Algorithm 2 and Algorithm 3. The core is to introduce a round-dependent permutation π_t over the ordered set A^*. Then, we explore the special structures that matroids have, which are all the bases of a matroid have the same size and matroids have certain properties described in Section 2.1. After introducing π_t, the regret for a matroid bandit problem can be decomposed to upper bound the total regret of K different stochastic bandit problems. The construction of π_t is inspired by Lemma 1 in Kveton et al. (2014). Recall $A^t = \{a_1^t, a_2^t, \ldots, a_K^t\}$ is a descending ordered set based on the differentially private upper confidence bounds in Algorithm 2 (or the differentially private posterior samples in Algorithm 3) and $A^* = \{a_1^*, a_2^*, \ldots, a_K^*\}$ is a descending ordered set based on the mean rewards. The purpose of introducing permutation $\pi_t : \{1, \ldots, K\} \to \{1, \ldots, K\}$ over the ordered set A^* is to construct K ordered pairs between A^t and A^* with certain properties. We construct π_t in a backward order as follows.

Fix $B_K = \left\{a_1^t, \ldots, a_{K-1}^t\right\}$. If $a_K^t \in A^*$, i.e., $a_K^t = a_i^*$ for some $i \in [K]$, we set $\pi_t(K) = i$, i.e., we pair a_K^t to itself. If $a_K^t \notin A^*$, due to the augmentation property of matroids (Property (3) in Section 2.1), we can set $\pi_t(K) = \min\{i : a_i^* \in A^* \setminus B_K, B_K \cup \{a_i^*\} \in \mathcal{I}\}$, i.e., we can pair a_K^t with the optimal base arm with the smallest index that can be added to B_K to form a matroid basis. Now, fix $B_{K-1} = \left\{a_1^t, \ldots, a_{K-2}^t, a_{\pi_t(K)}^*\right\}$. If $a_{K-1}^t = a_i^*$ for some $i \in [K]$, we set $\pi_t(K-1) = i$. If $a_{K-1}^t \notin A^*$, we can set $\pi_t(K-1) = \min\{i : a_i^* \in A^* \setminus B_{K-1}, B_{K-1} \cup \{a_i^*\} \in \mathcal{I}\}$. The same idea is applied to all the remaining base arms $a_{K-2}^t, \ldots, a_1^t$ in A^t.

After applying π_t, all the optimal base arms in A^* will be ordered as $A^*_{\pi_t} = \left\{a^*_{\pi_t(1)}, \ldots, a^*_{\pi_t(k)}, \ldots, a^*_{\pi_t(K)}\right\}$. Still, the order of all the base arms in A^t is $\{a_1^t, \ldots, a_k^t, \ldots, a_K^t\}$. Now, we can construct the following ordered pairs

$\left(a_1^t, a^*_{\pi_t(1)}\right), \ldots, \left(a_k^t, a^*_{\pi_t(k)}\right), \ldots, \left(a_K^t, a^*_{\pi_t(K)}\right)$. It is not hard to verify that each $\left(a_k^t, a^*_{\pi_t(k)}\right)$ has the following two properties. First, if $a_k^t \in A^*$, its pair is itself, i.e., $\Delta_{a_k^t, \pi_t(k)} = 0$ (recall $\Delta_{e,k} = \bar{w}(a_k^*) - \bar{w}(e)$ for any two base arms a_k^* and e). If $a_k^t \notin A^*$, the differentially private upper confidence bound in Algorithm 2 (or the differentially private posterior sample in Algorithm 3) of a_k^t is no smaller than that of $a^*_{\pi_t(k)}$. Second, given $\{a_1^t, a_2^t, \ldots, a_{k-1}^t\}$, both a_k^t and its pair $a^*_{\pi_t(k)}$ can be added to the solution set without breaking the independence of the solution set. In other words, both $\{a_1^t, \ldots, a_{k-1}^t\} \cup \{a_k^t\} \in \mathcal{I}$ and $\{a_1^t, \ldots, a_{k-1}^t\} \cup \{a^*_{\pi_t(k)}\} \in \mathcal{I}$ hold.

To permute $A^*_{\pi_t}$ back to the original order $\{a_1^*, \ldots, a_K^*\}$, we define $\pi_t^{-1} : \{1, \ldots, K\} \to \{1, \ldots, K\}$, the inverse permutation of π_t. That is, for each $k \in [K]$, we have $\pi_t^{-1}(\pi_t(k)) = k$. We apply π_t^{-1} to permute the ordered sets $A^*_{\pi_t}$ and A^t separately. Applying π_t^{-1} over $A^*_{\pi_t}$ gives us the original order, which is $\{a_1^*, \ldots, a_K^*\}$. Applying π_t^{-1} over A^t gives $\left\{a^t_{\pi_t^{-1}(1)}, \ldots, a^t_{\pi_t^{-1}(k)}, \ldots, a^t_{\pi_t^{-1}(K)}\right\}$. Based on the new order, we construct the following ordered pairs

$$\left(a^t_{\pi_t^{-1}(1)}, a_1^*\right), \ldots, \left(a^t_{\pi_t^{-1}(k)}, a_k^*\right), \ldots, \left(a^t_{\pi_t^{-1}(K)}, a_K^*\right). \tag{6}$$

Since a single π_t^{-1} is used to permute both the ordered sets, π_t^{-1} can be viewed as a permutation that permutes $\left(a_1^t, a^*_{\pi_t(1)}\right), \ldots, \left(a_k^t, a^*_{\pi_t(k)}\right), \ldots, \left(a_K^t, a^*_{\pi_t(K)}\right)$ to $\left(a^t_{\pi_t^{-1}(1)}, a_1^*\right), \ldots, \left(a^t_{\pi_t^{-1}(k)}, a_k^*\right), \ldots, \left(a^t_{\pi_t^{-1}(K)}, a_K^*\right)$.

Now, we are ready to decompose the regret. Define $\{a \leftrightarrow a^*\}$ as an event that base arms a and a^* form a pair. We have

$$\begin{aligned}
R_n &= \sum_{t=1}^{n} \mathbb{E}\left[\sum_{e \in A^*} \bar{w}(e) - \sum_{e \in A^t} \bar{w}(e)\right] \\
&\overset{(a)}{=} \sum_{t=1}^{n} \mathbb{E}\left[\sum_{k=1}^{K} \left(\bar{w}(a_k^*) - \bar{w}\left(a^t_{\pi_t^{-1}(k)}\right)\right) \cdot \mathbb{1}\left\{a^t_{\pi_t^{-1}(k)} \leftrightarrow a_k^*\right\}\right] \\
&\leq \sum_{k=1}^{K} \sum_{t=1}^{n} \mathbb{E}\left[\Delta_{a^t_{\pi_t^{-1}(k)}, k} \cdot \mathbb{1}\left\{\Delta_{a^t_{\pi_t^{-1}(k)}, k} > 0\right\} \cdot \mathbb{1}\left\{a^t_{\pi_t^{-1}(k)} \leftrightarrow a_k^*\right\}\right] \\
&\leq \sum_{k=1}^{K} \sum_{e \in \bar{A}^*} \sum_{t=1}^{n} \mathbb{E}\left[\mathbb{1}\left\{a^t_{\pi_t^{-1}(k)} = e\right\} \cdot \mathbb{1}\left\{\Delta_{e,k} > 0\right\}\right] \cdot \mathbb{1}\left\{e \leftrightarrow a_k^*\right\} \cdot \Delta_{e,k} \\
&= \sum_{k=1}^{K} \underbrace{\sum_{e \in \bar{A}^* : \Delta_{e,k} > 0} \sum_{t=1}^{n} \mathbb{E}\left[\mathbb{1}\left\{a^t_{\pi_t^{-1}(k)} = e\right\}\right] \cdot \mathbb{1}\left\{e \leftrightarrow a_k^*\right\} \cdot \Delta_{e,k}}_{I_k} .
\end{aligned} \tag{7}$$

Equality (a) uses the ordered pairs shown in (6). Note that I_k can be viewed as the regret of a stochastic bandit problem with a_k^* as the optimal arm and $\left\{e \in \bar{A}^* : \Delta_{e,k} > 0\right\}$ as the set of sub-optimal arms. The regret bounds for both `DPUCB-MAT` (Theorem 2) and `DPTS-MAT` (Theorem 4) can be derived based on the regret decomposition shown in (7). We defer the details of the proof to the appendix.

5 Experiments

We perform experiments in two different settings. In Section 5.1, we evaluate our algorithms on a synthetic dataset and in Section 5.2, we use a real-world movie-rating dataset with the purpose of recommending diverse and popular movies to users in a differentially private manner. To measure the performance of our algorithms we use the *expected per-round return* as suggested in Kveton et al. (2014). The expected per-round return in round s is computed as $\frac{1}{s}\sum_{t=1}^{s}\sum_{e \in A^t} \bar{w}(e)$. For `DPUCB-MAT`, we compare its empirical performance with two baselines. The first baseline is the optimal return $f(A^*, \bar{w}) = \sum_{e \in A^*} \bar{w}(e)$ denoted as Optimal Policy in the plots. The second baseline is OMM in Kveton et al. (2014), which is the non-private UCB1-based algorithm for matroid bandits. Similarly, for `DPTS-MAT`, we compare its empirical performance with the optimal return and the non-private CTS in Wang & Chen (2018). To have a fair empirical performance comparison, we have already adapted CTS to the matroid bandit setting. We also show the performance of our algorithms with different values of $\varepsilon \in \left\{10^5, 2, 10^{-4}\right\}$.

Table 2: Syntetic dataset.

Base arm e	Mean reward $\bar{w}(e)$
(1, 0, 0)	0.80
(0, 1, 0)	0.75
(0, 0, 1)	0.60
(1, 0, 1)	0.20
(0, 1, 1)	0.30
(2, 0, 0)	0.40
(0, 0, 0)	0.70

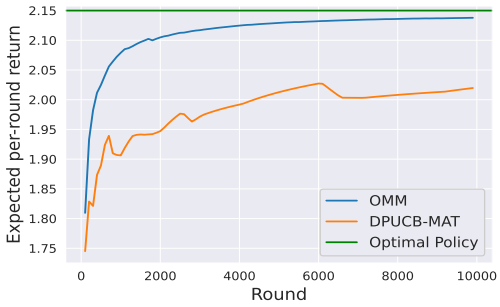
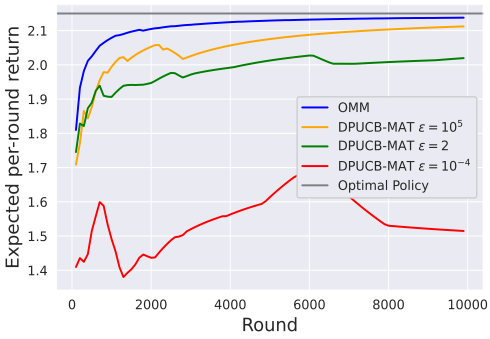
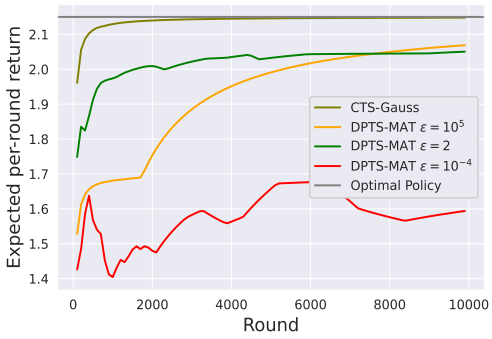


Figure 1: DPUCB−MAT on synthetic dataset.



(a) DPUCB−MAT



(b) DPTS−MAT

Figure 2: **Synthetic dataset**. Figure 2a shows the performance of DPUCB−MAT (Algorithm 2) for different values of ε on the synthetic dataset. We observe that the performance of DPUCB−MAT decreases as the value of ε decreases. We also observe that as ε increases, we can be in the non-private regime and the performance of DPUCB−MAT does not deteriorate by much. Figure 2b shows similar trends for DPTS−MAT (Algorithm 3).

5.1 Synthetic Dataset

In this section, we report the experimental results of our proposed algorithms on a set of 3-dimensional vectors taken from Neel & Neudauer (2009) with $\varepsilon = 2$. The base arm set is $E = \{e_1, \ldots, e_7\}$, where each base arm is a 3-dimensional vector in the Euclidean space. Table 2 shows the mean rewards of all the base arms. As matroid independence is defined by the linear independence of the vectors, we have $|A^*| = |A^t| = 3$ and $A^* = \{e_1, e_2, e_3\}$. Since the total privacy budget is ε, the privacy budget for each $e \in A^t$ is $\varepsilon_0 = \varepsilon/3 = 2/3$. The reward $w_t(e)$ for each $e \in E$ is generated from a Bernoulli distribution with mean $\bar{w}(e)$. The total number of rounds is $n = 10,000$. The experimental results for DPUCB−MAT, OMM, and Optimal Policy are presented in Figure 1. From the results, we can see that DPUCB−MAT and OMM have a similar growth rate in terms of the return and approach the optimal return. The results for DPTS−MAT, CTS, and Optimal Policy also show similar trends. We defer this set of plots to the appendix.

Figure 2 shows the expected per-round return for different values of the privacy parameter $\varepsilon \in \{10^5, 2, 10^{-4}\}$. In Figure 2a, we show the performance of DPUCB−MAT (Algorithm 2). We observe that when ε decreases, the performance of DPUCB−MAT deteriorates, and when ε increases, the performance of DPUCB−MAT becomes better. This is expected as a good differentially private learning algorithm should balance the privacy and regret guarantees. We also observe that when the privacy parameter ε is large, i.e., in the non-private regime, the performance of DPUCB−MAT approaches that of the non-private OMM. This is also expected as in the non-private regime, we do not pay any price for preserving privacy. Similar trends can also be seen for DPTS−MAT (Algorithm 3) in Figure 2b.

Table 3: Movies recommended by DPUCB−MAT that overlap with movies in A^* after 20k rounds.

Movie e	$\bar{w}(e)$	Genres
American Beauty	0.568	Comedy, Drama
Star Wars: Episode IV	0.496	Action, Adventure, Fantasy, Sci-Fi
Star Wars: Episode VI	0.478	Action, Adventure, Romance, Sci-Fi, War
Saving Private Ryan	0.440	Action, Drama, War
Men in Black	0.420	Action, Adventure, Comedy, Sci-Fi
L.A. Confidential	0.379	Crime, Film-Noir, Mystery, Thriller
Ghostbusters	0.361	Comedy, Horror
The Wizard of Oz	0.285	Animation, Children's, Comedy, Musical

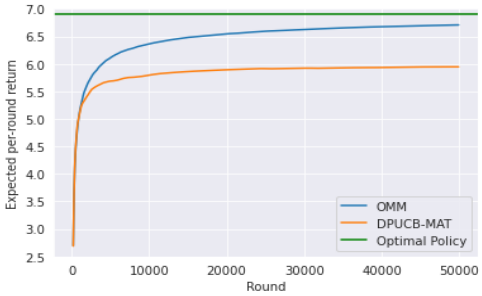


Figure 3: DPUCB−MAT on movie recommendation.

5.2 Movie Recommendation Dataset

In this experiment, we learn to recommend a set of *diverse* and *popular* movies from the MovieLens dataset (Harper & Konstan, 2015) with differential privacy. The experiment design is adopted from Kveton et al. (2014). Still, we report experimental results with the total privacy budget $\varepsilon = 2$. The total number of rounds is $n = 20,000$.

The entire dataset contains 1 million ratings from 6040 users. The total number of movies is 3883 from 18 different genres. To recommend popular movies, we select 100 movies that have received the most ratings. These 100 movies constitute the base arm set E of a matroid. To construct the independent sets $\mathcal{I}$ of the matroid, we select a one-hot feature vector u_e for each movie $e \in E$, which denotes the genres of that movie. If the feature vectors u_e for all $e \in A$ are linearly independent, all the movies collected in A form an independent set, which further indicates that these movies are *diverse*. The expected reward $\bar{w}(e)$ for each movie e is given by the total number of ratings for e divided by the total number of users in the dataset. The optimal solution A^* is computed greedily with respect to $\bar{w}$ using Algorithm 1. In each round, we recommend 17 movies, i.e., $|A^*| = |A^t| = 17$. The privacy budget for each $e \in A^t$ is $\varepsilon_0 = \varepsilon/17 = 2/17$. The randomness comes from the fact that in each round t, a random user is selected. For each movie $e \in A^t$, if e is rated by that selected random user, then the reward $w_t(e)$ is set to 1.

Figure 3 shows the results for DPUCB−MAT, OMM, and Optimal Policy. We observe that the expected per-round return of DPUCB−MAT is comparable to that of the non-private baseline (OMM) and approaches the optimal policy. Table 3 lists the overlapping movies learned by DPUCB−MAT and Optimal Policy at the end of learning. We can see that the movie genres of those movies appear to be diverse. Experimental results for DPTS−MAT are deferred to Appendix A.

6 Conclusions and Future Directions

In this work, we have shown that learning the maximum weight basis for matroid bandits can be done efficiently in a differentially private manner. We propose two simple differentially private algorithms DPUCB−MAT and DPTS−MAT and conduct experiments to evaluate their practical performance on both synthetic and real-world datasets. There are still some problems remaining for this work. So far, we only have an $\Omega\left(LK\ln(n)/\Delta + LK\ln(n)/\varepsilon\right)$ regret lower bound for differentially private combinatorial bandits (Chen et al., 2020). By modifying their proof for Theorem 9, we conjecture that an improved $\Omega\left(L\ln(n)/\Delta + LK\ln(n)/\varepsilon\right)$ regret lower bound for differentially private matroid bandits is achievable. However, our Theorem 2 is still an extra $\min\{K, \ln(n)\}$ factor far from this conjectured regret lower bound. We are not sure yet whether our derived regret bound is not tight or whether a better regret lower bound exists for differentially private matroid bandits.

Acknowledgements

This work was supported by the Alberta Machine Intelligence Institute (Amii).

References

Shipra Agrawal and Navin Goyal. Near-optimal regret bounds for Thompson Sampling. *Journal of the ACM (JACM)*, 64(5):1–24, 2017.

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

Achraf Azize and Debabrota Basu. When privacy meets partial information: A refined analysis of differentially private bandits. *arXiv preprint arXiv:2209.02570*, 2022.

T-H Hubert Chan, Elaine Shi, and Dawn Song. Private and continual release of statistics. *ACM Transactions on Information and System Security (TISSEC)*, 14(3):1–24, 2011.

Olivier Chapelle and Lihong Li. An empirical evaluation of thompson sampling. *Advances in neural information processing systems*, 24, 2011.

Xiaoyu Chen, Kai Zheng, Zixin Zhou, Yunchang Yang, Wei Chen, and Liwei Wang. (locally) differentially private combinatorial semi-bandits. In *International Conference on Machine Learning*, pp. 1757–1767. PMLR, 2020.

Cynthia Dwork, Moni Naor, Toniann Pitassi, and Guy N Rothblum. Differential privacy under continual observation. In *Proceedings of the forty-second ACM symposium on Theory of computing*, pp. 715–724, 2010.

Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

Aurélien Garivier and Olivier Cappé. The KL-UCB algorithm for bounded stochastic bandits and beyond. In *Proceedings of the 24th annual conference on learning theory*, pp. 359–376, 2011.

F. Maxwell Harper and Joseph A. Konstan. The movielens datasets: History and context, 2015.

Bingshan Hu and Nidhi Hegde. Near-optimal Thompson Sampling-based algorithms for differentially private stochastic bandits. In *Uncertainty in Artificial Intelligence*, pp. 844–852. PMLR, 2022.

Bingshan Hu, Zhiming Huang, and Nishant A Mehta. Optimal algorithms for private online learning in a stochastic environment. *arXiv preprint arXiv:2102.07929*, 2021.

Emilie Kaufmann, Nathaniel Korda, and Rémi Munos. Thompson Sampling: An asymptotically optimal finite-time analysis. In *International Conference on Algorithmic Learning Theory*, pp. 199–213. Springer, 2012.

Branislav Kveton, Zheng Wen, Azin Ashkan, Hoda Eydgahi, and Brian Eriksson. Matroid bandits: Fast combinatorial optimization with learning. *arXiv preprint arXiv:1403.5045*, 2014.

T.L Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985. ISSN 0196-8858. doi: https://doi.org/10.1016/0196-8858(85)90002-8. URL https://www.sciencedirect.com/science/article/pii/0196885885900028.

Nikita Mishra and Abhradeep Thakurta. (nearly) optimal differentially private stochastic multi-arm bandits. In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*, pp. 592–601, 2015.

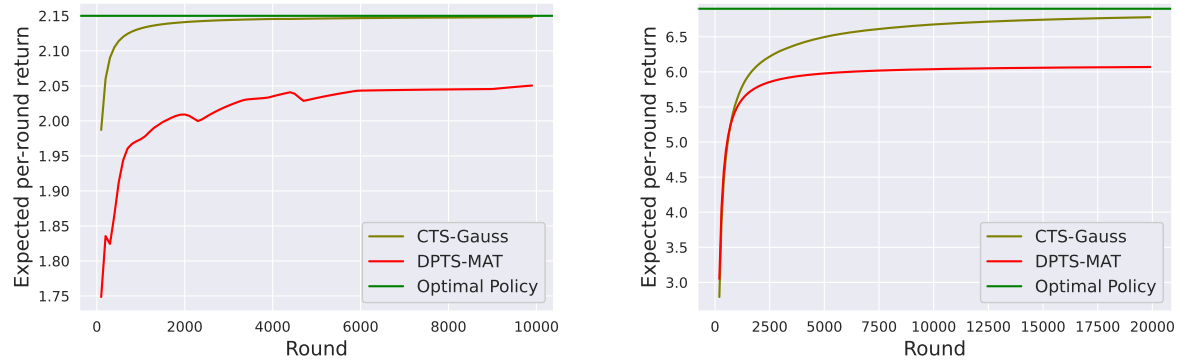
Arvind Narayanan and Vitaly Shmatikov. Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy (sp 2008)*, pp. 111–125. IEEE, 2008.

David L Neel and Nancy Ann Neudauer. Matroids you have known. *Mathematics magazine*, 82(1):26–41, 2009.

Adam Smith and Jonathan Ullman. Privacy in statistics and machine learning course. https://dpcourse.github.io/2021-spring/schedule.html, 2021.

Mohammad Sadegh Talebi and Alexandre Proutiere. An optimal algorithm for stochastic matroid bandit optimization. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*, pp. 548–556, 2016.

Siwei Wang and Wei Chen. Thompson Sampling for combinatorial semi-bandits. In *International Conference on Machine Learning*, pp. 5114–5122. PMLR, 2018.



(a) Synthetic dataset

(b) Movie-recommendation dataset

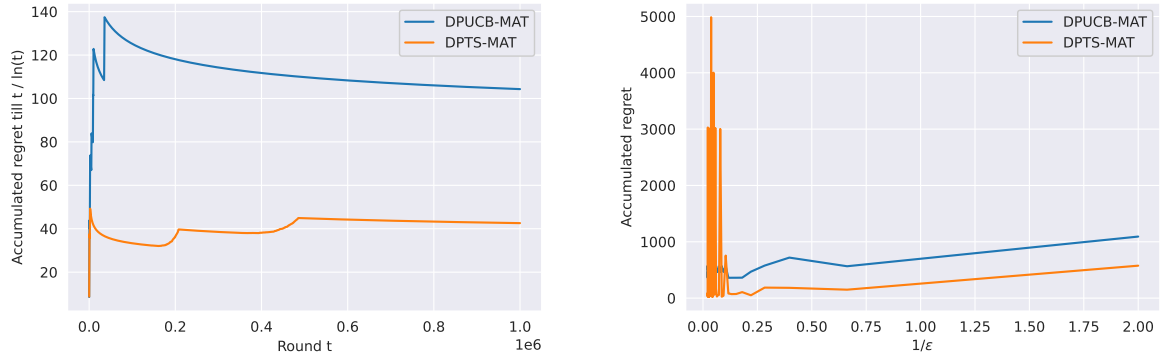
Figure 4: **Performance of `DPTS-MAT` on synthetic and movie-recommendation datasets.** Figure 4a shows the performances of `DPTS-MAT` (Algorithm 3), non-private Combinatorial Thompson Sampling (CTS) with Gaussian priors, and Optimal Policy on the synthetic dataset. Figure 4b shows the same algorithms on the movie-recommendation dataset. For both the plots, we again set $\varepsilon = 2$.

Appendices

A Additional Experiments

In this section, we provide more experimental results. Figure 4 shows the performance of `DPTS-MAT` (Algorithm 3) on both the synthetic datastet (Figure 4a) and the movie-recommendation dataset (Figure 4b) with $\varepsilon = 2$. We compare `DPTS-MAT` against the non-private Combinatorial Thompson Sampling (CTS) with Gaussian priors and the Optimal Policy. We observe that the expected per-round return of `DPTS-MAT` is comparable to that of the non-private baseline and approaches the optimal policy for both datasets.

Figure 5a shows the accumulated regret R_n till round n over $\ln(n)$ for $n = 1, 2, \ldots, 10^6$ rounds for both our private algorithms `DPUCB-MAT` and `DPTS-MAT` on the synthetic dataset. The quantity $\lim_{n\to\infty} R_n/\ln(n)$ characterizes the asymptotic rate of growth of the regret (Lai & Robbins, 1985). If a learning algorithm has a smaller rate, it suffers less regret. We observe that both `DPUCB-MAT` and `DPTS-MAT` converge but with different rates. In addition, we also observe that the Thompson Sampling-based algorithm, `DPTS-MAT`, empirically outperforms the UCB-based algorithm, `DPUCB-MAT`. Figure 5b studies the relationship between the accumulated regret R_n and $1/\varepsilon$ for `DPUCB-MAT` and `DPTS-MAT` on the synthetic dataset. The values of ε are 50 evenly spaced numbers between 0.5 and 50. We run our algorithms for $n = 10,000$ rounds for each value of ε and plot the accumulated regret. From the experimental results, we can see that for both the algorithms the regret trend is linear in $1/\varepsilon$ and the Thompson Sampling-based algorithm empirically outperforms the UCB-based one.



(a) Asymptotic regret rate.

(b) Regret versus $1/\varepsilon$.

Figure 5: **Regret of `DPUCB-MAT` and `DPTS-MAT` on the synthetic dataset**. Figure 5a shows the accumulated regret R_n till round n divided by $\ln(n)$ for $n = 1, 2, \ldots, 10^6$ rounds for both `DPUCB-MAT` and `DPTS-MAT`. The expression $\lim_{n\to\infty} R_n/\ln(n)$ characterizes the asymptotic growth rate of the regret. Figure 5b shows the accumulated regret versus $1/\varepsilon$ for both `DPUCB-MAT` and `DPTS-MAT`.

B Regret Analysis of Algorithm 2

Recall that the regret can be expressed as

$$R_n = \sum_{k=1}^{K} \sum_{e \in \bar{A}^* : \Delta_{e,k} > 0} \underbrace{\sum_{t=1}^{n} \mathbb{E}\left[\mathbb{1}\left\{a^t_{\pi_t^{-1}(k)} = e, e \leftrightarrow a^*_k\right\}\right] \cdot \Delta_{e,k}}_{I_{e,k}} \quad . \tag{8}$$

The indicator function $\mathbb{1}\left\{a^t_{\pi_t^{-1}(k)} = e, e \leftrightarrow a^*_k\right\}$ will be 1 when the learner selects a sub-optimal base arm e instead of the optimal base arm a^*_k. This further implies the differentially private upper confidence bound of e is no smaller than that of a^*_k in that round.

Proof. Proof of Theorem 2: We first show the regret is at most

$$\sum_{e \in \bar{A}^* : \Delta_{e,\min} > 0} O\left(\frac{\ln(Kn)}{\Delta_{e,\min}} + \frac{K\ln(Kn)}{\varepsilon/K}\right) \quad . \tag{9}$$

Then, we show the regret is also upper bounded by

$$\sum_{e \in \bar{A}^* : \Delta_{e,\min} > 0} O\left(\frac{\ln(Kn)}{\Delta_{e,\min}} + \frac{\ln^2(Kn)}{\varepsilon/K}\right) \quad . \tag{10}$$

Combining these two claims concludes the proof.

Recall $\varepsilon_0 = \varepsilon/K$ and let $l_{e,k} = O\left(\frac{\ln(Kn)}{\Delta_{e,k} \cdot \min\{\Delta_{e,k}, \varepsilon_0\}}\right)$, where the big-O notation only hides a universal constant.

We decompose $I_{e,k}$ in (8) as

$$I_{e,k} = \underbrace{\sum_{t=1}^{n} \mathbb{E}\left[\mathbb{1}\left\{a^t_{\pi_t^{-1}(k)} = e, e \leftrightarrow a^*_k, T_e(t-1) \le l_{e,k}\right\}\right] \Delta_{e,k}}_{\Gamma_{1,k,e}} + \underbrace{\sum_{t=1}^{n} \mathbb{E}\left[\mathbb{1}\left\{a^t_{\pi_t^{-1}(k)} = e, e \leftrightarrow a^*_k, T_e(t-1) > l_{e,k}\right\}\right] \Delta_{e,k}}_{\Gamma_{2,k,e}} . \tag{11}$$

From Hu et al. (2021); Azize & Basu (2022), we know that $\Gamma_{2,k,e} = O(1/K^2)$. Now, we use similar arguments as the one shown in Kveton et al. (2014) to complete the proof. As for a fixed $e \in \bar{A}^*$, we maintain a counter $T_e(t-1)$

during the learning. The counter counts the number of observations that are used to compute the differentially private empirical mean and the values of the counter double each time. For all the rounds when the values of the counter are in the range of $[0, l_{e,1}]$, the total regret among all these rounds is at most $\Delta_{e,1} \cdot O(l_{e,1})$. For all the rounds when the values of the counter are in the range of $[l_{e,1}+1, l_{e,2}]$, the total regret among all these rounds is upper bounded by $\Delta_{e,2} \cdot O\left(l_{e,2} - l_{e,1}\right) + \Gamma_{2,1,e} \leq \Delta_{e,2} \cdot O\left(l_{e,2} - l_{e,1}\right) + O(1/K)$. Finally, for all the rounds when the values of the counter are in the range of $[l_{e,K-1}+1, l_{e,K}]$, the total regret among all these rounds is at most $\Delta_{e,K} \cdot O\left(l_{e,K} - l_{e,K-1}\right) + \sum_{k=1}^{K-1} \Gamma_{2,k,e} \leq \Delta_{e,K} \cdot O\left(l_{e,K} - l_{e,K-1}\right) + O(1/K)$. For all the rounds after the counter hits $l_{e,K}$, the total regret is at most $\sum_{k=1}^{K} \Gamma_{2,k,e} \leq O(1/K)$.

By using Lemma 1, we have

$$\Delta_{e,1} \cdot O(l_{e,1}) + \sum_{k=2}^{K} \Delta_{e,k} \cdot O\left(l_{e,k} - l_{e,k-1}\right) + \sum_{k=2}^{K+1} \sum_{q=1}^{k-1} \Gamma_{2,q,e} \leq O\left(\frac{\ln(Kn)}{\Delta_{e,\min}} + \frac{K \ln(Kn)}{\varepsilon_0}\right) \quad , \tag{12}$$

which yields the result in the first claim, i.e.,

$$\sum_{e \in \bar{A}^* : \Delta_{e,\min} > 0} O\left(\frac{\ln(Kn)}{\Delta_{e,\min}} + \frac{K \ln(Kn)}{\varepsilon/K}\right) \quad .$$

We can also use the well-known "doubling-trick" in the bandit literature to have an upper bound.

Let $r_{\max,e} := \min\left\{\log\left(\frac{1}{\Delta_{e,\min}}\right), \log(Kn)\right\}$.

For any $1 \leq r \leq r_{\max,e}$, let $\Phi_r := \left\{k \in [K] : \Delta_{e,k} \in \left[0.5^r, 0.5^{r-1}\right)\right\}$ and $l_{e,r} = O\left(\frac{\ln(Kn)}{0.5^r \cdot \min\{0.5^r, \varepsilon_0\}}\right)$, where the big-O notation only hides a universal constant.

Then, we take the following regret decomposition. We have

$$\begin{aligned}
R_n &= \sum_{k=1}^{K} \sum_{e \in \bar{A}^* : \Delta_{e,k} > 0} \sum_{t=1}^{n} \mathbb{E}\left[\mathbb{1}\left\{a^t_{\pi_t^{-1}(k)} = e, e \leftrightarrow a^*_k\right\}\right] \cdot \Delta_{e,k} \\
&\leq 1 + \sum_{e \in \bar{A}^*} \sum_{k \in [K] : \Delta_{e,k} \geq \frac{1}{Kn}} \sum_{t=1}^{n} \mathbb{E}\left[\mathbb{1}\left\{a^t_{\pi_t^{-1}(k)} = e, e \leftrightarrow a^*_k\right\}\right] \cdot \Delta_{e,k} \\
&\leq 1 + \sum_{e \in \bar{A}^*} \sum_{r=1}^{r_{\max,e}} \sum_{k \in \Phi_r} \sum_{t=1}^{n} \mathbb{E}\left[\mathbb{1}\left\{a^t_{\pi_t^{-1}(k)} = e, e \leftrightarrow a^*_k\right\}\right] \cdot \Delta_{e,k} \\
&\leq 1 + \sum_{e \in \bar{A}^*} \sum_{r=1}^{r_{\max,e}} \sum_{k \in \Phi_r} \sum_{t=1}^{n} \mathbb{E}\left[\mathbb{1}\left\{a^t_{\pi_t^{-1}(k)} = e, e \leftrightarrow a^*_k\right\}\right] \cdot 0.5^r \\
&\leq 1 + \sum_{e \in \bar{A}^*} \sum_{r=1}^{r_{\max,e}} \underbrace{\sum_{t=1}^{n} \mathbb{E}\left[\mathbb{1}\left\{\exists k \in \Phi_r : a^t_{\pi_t^{-1}(k)} = e, e \leftrightarrow a^*_k\right\}\right] \cdot 0.5^r}_{I_{e,r}} \quad .
\end{aligned} \tag{13}$$

Let $\left\{a^t_{\pi_t^{-1}(r)} = e, e \leftrightarrow r\right\}$ denote the event that $\left\{\exists k \in \Phi_r : a^t_{\pi_t^{-1}(k)} = e, e \leftrightarrow a^*_k\right\}$.

Now, we decompose $I_{e,r}$ as

$$\begin{aligned}
I_{e,r} = &\underbrace{\sum_{t=1}^{n} \mathbb{E}\left[\mathbb{1}\left\{a^t_{\pi_t^{-1}(r)} = e, e \leftrightarrow r, T_e(t-1) \leq l_{e,r}\right\}\right] \cdot 0.5^r}_{\Gamma_{1,r,e}} \\
&+ \underbrace{\sum_{t=1}^{n} \mathbb{E}\left[\mathbb{1}\left\{a^t_{\pi_t^{-1}(r)} = e, e \leftrightarrow r, T_e(t-1) > l_{e,r}\right\}\right] \cdot 0.5^r}_{\Gamma_{2,r,e}} \quad .
\end{aligned} \tag{14}$$

Similarly, we have $\Gamma_{2,r,e} \leq O(1/K^2)$ and the total regret is at most

$$\sum_{e\in\bar{A}^*}\left(0.5^0\cdot O\left(l_{e,1}\right)+\sum_{r=2}^{r_{\max,e}}0.5^{r-1}\cdot O\left(l_{e,r}-l_{e,r-1}\right)+O(1)\right)\leq\sum_{e\in\bar{A}^*}O\left(\frac{\ln(Kn)}{\Delta_{e,\min}}+\frac{\ln^2(Kn)}{\varepsilon_0}\right)\quad, \tag{15}$$

where the inequality uses Lemma 2. □

Lemma 1. Let $\Delta_1 \geq \ldots \geq \Delta_K$ be a sequence of reals in $(0, 1]$. For any $\varepsilon_0 > 0$, we have

$$\Delta_1\frac{1}{\Delta_1\cdot\min\left\{\Delta_1,\varepsilon_0\right\}}+\sum_{k=2}^{K}\Delta_k\left(\frac{1}{\Delta_k\cdot\min\left\{\Delta_k,\varepsilon_0\right\}}-\frac{1}{\Delta_{k-1}\cdot\min\left\{\Delta_{k-1},\varepsilon_0\right\}}\right)\leq\frac{2}{\Delta_K}+\frac{K}{\varepsilon_0}\,. \tag{16}$$

Lemma 2. For any $\varepsilon_0 > 0$, we have

$$0.5^0\cdot l_{e,1}+\sum_{r=2}^{r_{\max,e}}0.5^{r-1}\cdot\left(l_{e,r}-l_{e,r-1}\right)\leq O(\ln(Kn))\cdot\left(\frac{2}{\Delta_{e,\min}}+\frac{\log(Kn)}{\varepsilon_0}\right)\quad. \tag{17}$$

Lemma 3. (Kveton et al., 2014, Lemma 3). Let $\Delta_1 \geq \ldots \geq \Delta_K$ be a sequence of reals in $(0, 1]$. Then we have

$$\Delta_1\frac{1}{\Delta_1^2}+\sum_{k=2}^{K}\Delta_k\left(\frac{1}{\Delta_k^2}-\frac{1}{\Delta_{k-1}^2}\right)\leq\frac{2}{\Delta_K}\,. \tag{18}$$

Proof. Proof of Lemma 1:

Case 1: When $\varepsilon_0 \geq \Delta_1$, we have

$$\text{LHS of (16)}=\Delta_1\frac{1}{\Delta_1^2}+\sum_{k=2}^{K}\Delta_k\left(\frac{1}{\Delta_k^2}-\frac{1}{\Delta_{k-1}^2}\right)\leq\frac{2}{\Delta_K}\leq\frac{2}{\Delta_K}+\frac{K}{\varepsilon_0}\,, \tag{19}$$

where the first inequality uses Lemma 3.

Case 2: When $\Delta_K \geq \varepsilon_0$, we have

$$\text{LHS of (16)}=\frac{1}{\varepsilon_0}+\frac{1}{\varepsilon_0}\sum_{k=2}^{K}\left(1-\frac{\Delta_k}{\Delta_{k-1}}\right)\leq\frac{1}{\varepsilon_0}+\frac{1}{\varepsilon_0}\cdot(K-1)\leq\frac{K}{\varepsilon_0}+\frac{2}{\Delta_K}\,.$$

Case 3: When $\Delta_1 \geq \ldots \Delta_j \geq \varepsilon_0 \geq \Delta_{j+1} \geq \ldots \geq \Delta_K$ for some $j \in [K-1]$. We rewrite the LHS of (16) as

$$\begin{aligned}
\sum_{k=1}^{K-1}\frac{\Delta_k-\Delta_{k+1}}{\Delta_k\cdot\min\left\{\Delta_k,\varepsilon_0\right\}}+\frac{1}{\min\left\{\Delta_K,\varepsilon_0\right\}}&=\sum_{k=1}^{j}\frac{\Delta_k-\Delta_{k+1}}{\Delta_k\cdot\varepsilon_0}+\sum_{k=j+1}^{K-1}\frac{\Delta_k-\Delta_{k+1}}{\Delta_k^2}+\frac{1}{\Delta_K}\\
&\leq\sum_{k=1}^{j}\left[\frac{1}{\varepsilon_0}-\frac{\Delta_{k+1}}{\varepsilon_0\Delta_k}\right]+\sum_{k=j+1}^{K-1}\frac{\Delta_k-\Delta_{k+1}}{\Delta_k\cdot\Delta_{k+1}}+\frac{1}{\Delta_K}\\
&=\left[\frac{j}{\varepsilon_0}-\sum_{k=1}^{j}\frac{\Delta_{k+1}}{\varepsilon_0\Delta_k}\right]+\left[\frac{1}{\Delta_K}-\frac{1}{\Delta_{j+1}}\right]+\frac{1}{\Delta_K}\\
&\leq\frac{j}{\varepsilon_0}+\frac{2}{\Delta_K}\\
&\leq\frac{K}{\varepsilon_0}+\frac{2}{\Delta_K}\,,
\end{aligned} \tag{20}$$

which concludes the proof. □

Proof of Lemma 2. We have the LHS in (17) is

$$\begin{aligned} & 0.5^0 \cdot l_{e,1} + \sum_{r=2}^{r_{\max,e}} 0.5^{r-1} \cdot (l_{e,r} - l_{e,r-1}) \\ = \; & O(\ln(Kn)) \cdot \underbrace{\left(0.5^0 \frac{1}{0.5^0 \cdot \min\{0.5^0, \varepsilon_0\}} + \sum_{r=2}^{r_{\max,e}} 0.5^{r-1} \left(\frac{1}{0.5^{r-1} \cdot \min\{0.5^{r-1}, \varepsilon_0\}} - \frac{1}{0.5^r \cdot \min\{0.5^r, \varepsilon_0\}} \right) \right)}_{I} \\ \leq \; & O(\ln(Kn)) \cdot \left(\frac{2}{\Delta_{e,\min}} + \frac{\log(Kn)}{\varepsilon_0} \right) \quad , \end{aligned} \tag{21}$$

where the last step uses Lemma 1.

Note that $0.5^0 \geq 0.5^1 \geq \ldots \geq 0.5^{r_{\max,e}}$ and from $r_{\max,e} = \min\{\log(1/\Delta_{e,\min}), \log(Kn)\}$, we have $2^{r_{\max,e}} \leq \Delta_{e,\min}$ and $r_{\max,e} \leq \log(Kn)$. Then, we have $I \leq \frac{2}{0.5^{r_{\max,e}}} + \frac{r_{\max,e}}{\varepsilon_0}$, which concludes the proof. □

C REGRET ANALYSIS OF ALGORITHM 3

Still, recall the regret can be expressed as

$$R_n = \sum_{k=1}^{K} \sum_{e \in \bar{A}^* : \Delta_{e,k} > 0} \underbrace{\sum_{t=1}^{n} \mathbb{E}\left[\mathbb{1}\left\{a^t_{\pi_t^{-1}(k)} = e, e \leftrightarrow a^*_k\right\}\right] \cdot \Delta_{e,k}}_{I_{e,k}} \quad . \tag{22}$$

The indicator function $\mathbb{1}\left\{a^t_{\pi_t^{-1}(k)} = e, e \leftrightarrow a^*_k\right\}$ will be 1 when the learner selects a sub-optimal base arm e instead of the optimal base arm a^*_k. This implies that the posterior sample of e is no smaller than the posterior sample of a^*_k in that round. For a fixed k and $e \in \bar{A}^*$, we define event $\mathcal{E}^{\theta}_{e,k}(t) := \{\theta_e(t) \leq y_{e,k}\}$, where $y_{e,k} := \bar{w}(a^*_k) - \frac{1}{3}\Delta_{e,k}$, to decompose the regret. By introducing $\mathcal{E}^{\theta}_{e,k}(t)$, the regret can be decomposed as

$$\begin{aligned} R_n &= \sum_{k=1}^{K} \sum_{e \in \bar{A}^*} \sum_{t=1}^{n} \mathbb{E}\left[\mathbb{1}\left\{a^t_{\pi_t^{-1}(k)} = e, e \leftrightarrow a^*_k\right\}\right] \cdot \Delta_{e,k} \\ &= \underbrace{\sum_{k=1}^{K} \sum_{e \in \bar{A}^*} \underbrace{\sum_{t=1}^{n} \mathbb{E}\left[\mathbb{1}\left\{a^t_{\pi_t^{-1}(k)} = e, \overline{\mathcal{E}^{\theta}_{e,k}(t)}, e \leftrightarrow a^*_k\right\}\right] \cdot \Delta_{e,k}}_{V_{e,k}}}_{V} + \underbrace{\sum_{k=1}^{K} \sum_{e \in \bar{A}^*} \underbrace{\sum_{t=1}^{n} \mathbb{E}\left[\mathbb{1}\left\{a^t_{\pi_t^{-1}(k)} = e, \mathcal{E}^{\theta}_{e,k}(t), e \leftrightarrow a^*_k\right\}\right] \cdot \Delta_{e,k}}_{U_{e,k}}}_{U} \quad . \end{aligned} \tag{23}$$

Bounding $V_{e,k}$ and V: The analysis of this term is very similar to the regret analysis of Algorithm 2. We first define two high probability events. Let $C_e(t) := \left\{|\hat{w}_{e,T_e(t-1)}(t-1) - \bar{w}(e)| \leq \sqrt{\frac{3\ln(Kt)}{T_e(t-1)}}\right\}$ be the event that the mean reward of $e \in E$ is within the confidence interval in round t and $G_e(t) := \left\{|\tilde{w}_{e,T_e(t-1)}(t-1) - \hat{w}_{e,T_e(t-1)}(t-1)| \leq \frac{3\ln(Kt)}{\varepsilon_0 \cdot T_e(t-1)}\right\}$ be the event that the noise added is not too much in round t. Let $\overline{C_e(t)}$ and $\overline{G_e(t)}$ denote the complements of events $C_e(t)$ and $G_e(t)$, respectively.

Let $l_{e,k} := \frac{72\ln(nK)}{\min\left\{\Delta^2_{e,k}, \varepsilon_0 \cdot \Delta_{e,k}\right\}}$. Then, $V_{e,k}$ can be further decomposed as

$$\begin{aligned} V_{e,k} = &\underbrace{\sum_{t=1}^{n} \mathbb{E}\left[\mathbb{1}\left\{a^t_{\pi_t^{-1}(k)} = e, \overline{\mathcal{E}^{\theta}_{e,k}(t)}, T_e(t-1) \leq l_{e,k}, e \leftrightarrow a^*_k\right\}\right] \cdot \Delta_{e,k}}_{\Gamma_{1,k,e}} \\ &+ \underbrace{\sum_{t=1}^{n} \mathbb{E}\left[\mathbb{1}\left\{a^t_{\pi_t^{-1}(k)} = e, \overline{\mathcal{E}^{\theta}_{e,k}(t)}, T_e(t-1) > l_{e,k}, e \leftrightarrow a^*_k\right\}\right] \cdot \Delta_{e,k}}_{\Gamma_{2,k,e}} \quad . \end{aligned} \tag{24}$$

We further decompose $\Gamma_{2,k,e}$ as

$$\begin{aligned} \Gamma_{2,k,e} &= \sum_{t=1}^{n} \mathbb{E}\left[\mathbb{1}\left\{a^t_{\pi_t^{-1}(k)} = e, \overline{\mathcal{E}^{\theta}_{e,k}(t)}, T_e(t-1) > l_{e,k}, e \leftrightarrow a^*_k\right\}\right] \cdot \Delta_{e,k} \\ &\leq \underbrace{\sum_{t=1}^{n} \mathbb{E}\left[\mathbb{1}\left\{a^t_{\pi_t^{-1}(k)} = e, \overline{\mathcal{E}^{\theta}_{e,k}(t)}, C_e(t), G_e(t), T_e(t-1) > l_{e,k}, e \leftrightarrow a^*_k\right\}\right] \cdot \Delta_{e,k}}_{\gamma} \\ &+ \underbrace{\sum_{t=1}^{n} \mathbb{E}\left[\mathbb{1}\left\{\overline{C_e(t)}\right\}\right] + \sum_{t=1}^{n} \mathbb{E}\left[\mathbb{1}\left\{\overline{G_e(t)}\right\}\right]}_{\leq O\left(\frac{1}{K^2 n^2}\right), \text{ Lemma 4}} . \end{aligned} \tag{25}$$

Let $d_{e,k} := \lceil \log(l_{e,k}) \rceil$. Let τ_s denote the round by the end of which there are exactly 2^s fresh observations that will be used to compute the differentially private empirical mean of a sub-optimal base arm e. Now, we upper bound γ.

We have

$$
\begin{aligned}
\gamma &= \sum_{t=1}^{n} \mathbb{E}\left[\mathbb{1}\left\{a^t_{\pi_t^{-1}(k)} = e, \overline{\mathcal{E}^{\theta}_{e,k}(t)}, C_e(t), G_e(t), e \leftrightarrow a^*_k, T_e(t-1) > l_{e,k}\right\}\right] \cdot \Delta_{e,k} \\
&\leq \sum_{s=d_e}^{\log n} \mathbb{E}\left[\sum_{t=\tau_s+1}^{\tau_{s+1}} \mathbb{1}\left\{a^t_{\pi_t^{-1}(k)} = e, \overline{\mathcal{E}^{\theta}_{e,k}(t)}, C_e(t), G_e(t), e \leftrightarrow a^*_k, T_e(t-1) > l_{e,k}\right\}\right] \cdot \Delta_{e,k} \\
&\leq \sum_{s=d_e}^{\log n} \sum_{t=1}^{n} \mathbb{E}\left[\mathbb{1}\left\{a^t_{\pi_t^{-1}(k)} = e, \overline{\mathcal{E}^{\theta}_{e,k}(t)}, C_e(t), G_e(t), e \leftrightarrow a^*_k, T_e(t-1) = 2^s\right\}\right] \cdot \Delta_{e,k} \\
&= \sum_{s=d_e}^{\log n} \sum_{t=1}^{n} \mathbb{E}\left[\mathbb{E}\left[\mathbb{1}\left\{a^t_{\pi_t^{-1}(k)} = e, \overline{\mathcal{E}^{\theta}_{e,k}(t)}, C_e(t), G_e(t), e \leftrightarrow a^*_k, T_e(t-1) = 2^s\right\} \mid \mathcal{F}_{t-1}\right]\right] \cdot \Delta_{e,k} \\
&\leq \sum_{s=d_e}^{\log n} \sum_{t=1}^{n} \underbrace{\mathbb{E}\left[\mathbb{1}\left\{C_e(t), G_e(t), T_e(t-1) = 2^s\right\} \underbrace{\mathbb{E}\left[\mathbb{1}\left\{\overline{\mathcal{E}^{\theta}_{e,k}(t)}\right\} \,\middle|\, \mathcal{F}_{t-1}\right]}_{\lambda}\right]}_{\Lambda} \cdot \Delta_{e,k} \quad .
\end{aligned}
\tag{26}
$$

To upper bound Λ, we divide all the instantiations F_{t-1} of $\mathcal{F}_{t-1}$ into two groups based on whether $\mathbb{1}\left\{C_e(t), G_e(t), T_e(t) = 2^s\right\}$ is 1 or 0.

Case 1: For the instantiation F_{t-1} such that $\mathbb{1}\left\{C_e(t), G_e(t), T_e(t-1) = 2^s\right\} = 0$, we have $\Lambda = 0$.

Case 2: For the instantiation F_{t-1} such that $\mathbb{1}\left\{C_e(t), G_e(t), T_e(t-1) = 2^s\right\} = 1$, we construct an upper bound for λ. To do so, recall that $\theta_e(t) \sim \mathcal{N}\left(w'_{e,T_e(t-1)}(t), \frac{1}{T_e(t-1)}\right)$, where $w'_{e,T_e(t-1)}(t) = \tilde{w}_{e,T_e(t-1)}(t-1) + \frac{3\ln(Kt)}{\varepsilon_0 \cdot T_e(t-1)}$ and $\mathcal{N}(\mu, \sigma^2)$ is a normal distribution with mean μ and variance σ^2. We have

$$
\begin{aligned}
w'_{e,T_e(t-1)}(t) &= \tilde{w}_{e,T_e(t-1)}(t-1) + \frac{3\ln(Kt)}{\varepsilon_0 \cdot T_e(t-1)} \\
&\leq \hat{w}_{e,T_e(t-1)}(t-1) + \frac{6\ln(Kt)}{\varepsilon_0 \cdot T_e(t-1)} \qquad (\text{since } \mathbb{1}\left\{G_e(t-1)\right\} = 1) \\
&\leq \bar{w}(e) + \frac{6\ln(Kt)}{\varepsilon_0 \cdot T_e(t-1)} + \sqrt{\frac{3\ln(Kt)}{T_e(t-1)}} \qquad (\text{since } \mathbb{1}\left\{C_e(t-1)\right\} = 1) \\
&= \bar{w}(e) + \frac{6\ln(Kt)}{\varepsilon_0 \cdot 2^s} + \sqrt{\frac{3\ln(Kt)}{2^s}} \qquad (\text{since } \mathbb{1}\left\{T_e(t-1) = 2^s\right\} = 1) \\
&\leq \bar{w}(e) + \frac{\Delta_{e,k}}{12} + \frac{\Delta_{e,k}}{\sqrt{18}} \qquad \left(\text{for } s \geq \left\lceil \log \frac{72\ln(nK)}{\min\left\{\Delta^2_{e,k}, \varepsilon_0 \cdot \Delta_{e,k}\right\}} \right\rceil\right) \\
&\leq \bar{w}(e) + \frac{1}{3}\Delta_{e,k} \,.
\end{aligned}
$$

From first-order stochastic dominance, we know a $\mathcal{N}\left(w'_{e,T_e(t-1)}(t), \frac{1}{T_e(t-1)}\right)$ distributed random variable is stochastically dominated by a $\mathcal{N}\left(\bar{w}(e) + \frac{1}{3}\Delta_{e,k}, \frac{1}{T_e(t-1)}\right)$ distributed random variable. Next, we slightly abuse the notation and use $\mathbb{P}\left\{\mathcal{N}(\mu, \sigma^2) > y_{e,k}\right\}$ to denote the probability that a $\mathcal{N}(\mu, \sigma^2)$ distributed random variable is drawn greater than $y_{e,k}$. Now, we construct an upper bound for $\mathbb{E}\left[\mathbb{1}\left\{\overline{\mathcal{E}^{\theta}_{e,k}(t)}\right\} \,\middle|\, \mathcal{F}_{t-1} = F_{t-1}\right]$ by using the first-order stochastic

dominance between two Gaussian distributed random variables. We have

$$\begin{aligned}
\mathbb{E}\left[\mathbb{1}\left\{\overline{\mathcal{E}^{\theta}_{e,k}(t)}\right\}\Big|\mathcal{F}_{t-1}=F_{t-1}\right] &= \mathbb{P}\left\{\theta_e(t) > \bar{w}(e)+\frac{2}{3}\Delta_{e,k}\right\} \\
&= \mathbb{P}\left\{\mathcal{N}\left(w'_{e,T_e(t-1)}(t), \frac{1}{T_e(t-1)}\right) > \bar{w}(e)+\frac{2}{3}\Delta_{e,k}\right\} \\
&\leq \mathbb{P}\left\{\mathcal{N}\left(\bar{w}(e)+\frac{1}{3}\Delta_{e,k}, \frac{1}{T_e(t-1)}\right) > \bar{w}(e)+\frac{2}{3}\Delta_{e,k}\right\} \\
&= \mathbb{P}\left\{\mathcal{N}\left(\bar{w}(e), \frac{1}{T_e(t-1)}\right) > \bar{w}(e)+\frac{1}{3}\Delta_{e,k}\right\} \\
&\overset{(\clubsuit)}{\leq} \exp\left(-\frac{1}{18}\cdot\Delta^2_{e,k}\cdot\frac{72\ln(nK)}{\min\left\{\Delta^2_{e,k},\varepsilon_0\cdot\Delta_{e,k}\right\}}\right) \\
&\leq O\left(\frac{1}{(nK)^4}\right),
\end{aligned} \tag{27}$$

where ($\clubsuit$) uses the concentration bounds for a normally distributed random variable (Fact 5) and the fact that $T_e(t-1) = 2^s \geq \frac{72\ln(nK)}{\min\left\{\Delta^2_{e,k},\varepsilon_0\cdot\Delta_{e,k}\right\}}$.

Putting all the pieces together, we have $\Gamma_{2,k,e} \leq O\left(\frac{1}{K^2}\right)$. Now, we use similar arguments that have been used for the proofs of Theorem 2 to upper bound term V in (23). Recall that $T_e(t-1)$ is a counter that counts the number of fresh observations that has been used to compute the differentially private empirical mean of a sub-optimal base arm e. For all the rounds when the counter is in the range of $[0, l_{e,1}]$, the total regret among all these rounds is at most $\Delta_{e,1}\cdot O\left(l_{e,1}\right)$. For all the rounds when the values of the counter are in the range of $[l_{e,1}+1, l_{e,2}]$, the total regret among all these rounds is upper bounded by $\Delta_{e,2}\cdot O\left(l_{e,2}-l_{e,1}\right)+\Gamma_{2,1,e} \leq \Delta_{e,2}\cdot O\left(l_{e,2}-l_{e,1}\right)+O(1/K)$. Finally, for all the rounds when the values of the counter are in the range of $[l_{e,K-1}+1, l_{e,K}]$, the total regret among all these rounds is at most $\Delta_{e,K}\cdot O\left(l_{e,K}-l_{e,K-1}\right)+\sum_{k=1}^{K-1}\Gamma_{2,k,e} \leq \Delta_{e,K}\cdot O\left(l_{e,K}-l_{e,K-1}\right)+O(1/K)$. For all the rounds after the counter hits $l_{e,K}$, the total regret is at most $\sum_{k=1}^{K}\Gamma_{2,k,e} \leq O(1/K)$.

By using Lemma 1, we have

$$V \leq \sum_{e\in\bar{A}^*}\left(\Delta_{e,1}\cdot O\left(l_{e,1}\right)+\sum_{k=2}^{K}\Delta_{e,k}\cdot O\left(l_{e,k}-l_{e,k-1}\right)+\sum_{k=2}^{K+1}\sum_{q=1}^{k-1}\Gamma_{2,q,e}\right) \leq \sum_{e\in\bar{A}^*} O\left(\frac{\ln(Kn)}{\Delta_{e,\min}}+\frac{K\ln(Kn)}{\varepsilon_0}\right). \tag{28}$$

Similarly, by using Lemma 2, we have

$$V \leq \sum_{e\in\bar{A}^*} O(\ln(Kn))\cdot\left(\frac{2}{\Delta_{e,K}}+\frac{\log(Kn)}{\varepsilon_0}\right) = \sum_{e\in\bar{A}^*} O\left(\frac{\ln(Kn)}{\Delta_{e,\min}}+\frac{\ln^2(Kn)}{\varepsilon_0}\right). \tag{29}$$

Combining (28) and (29), we have

$$V \leq \sum_{e\in\bar{A}^*} O(\ln(Kn))\cdot\left(\frac{2}{\Delta_{e,K}}+\frac{\log(Kn)}{\varepsilon_0}\right) = \sum_{e\in\bar{A}^*} O\left(\frac{\ln(Kn)}{\Delta_{e,\min}}+\frac{\min\left\{K,\ln(Kn)\right\}\cdot\ln(Kn)}{\varepsilon_0}\right). \tag{30}$$

Bounding $U_{e,k}$ and U: Recall

$$U = \sum_{k=1}^{K}\sum_{e\in\bar{A}^*}\underbrace{\sum_{t=1}^{n}\mathbb{E}\left[\mathbb{1}\left\{a^t_{\pi_t^{-1}(k)}=e, \mathcal{E}^{\theta}_{e,k}(t), e\leftrightarrow a^*_k\right\}\right]\cdot\Delta_{e,k}}_{U_{e,k}}. \tag{31}$$

Let $l^*_{e,k} := \left\lceil \frac{288\ln\left(L^2\left(n+e^{32}\right)\right)}{\Delta^2_{e,k}} \right\rceil$ and $d^*_{e,k} = \log\left(l^*_{e,k}\right)$. We first decompose $U_{e,k}$ as

$$\begin{aligned}
U_{e,k} &= \sum_{t=1}^{n} \mathbb{E}\left[\mathbb{1}\left\{a^t_{\pi_t^{-1}(k)} = e, \mathcal{E}^\theta_{e,k}(t), e \leftrightarrow a^*_k\right\}\right] \cdot \Delta_{e,k} \\
&\leq \sum_{t=1}^{n} \mathbb{E}\left[\mathbb{1}\left\{a^t_{\pi_t^{-1}(k)} = e, \mathcal{E}^\theta_{e,k}(t), G_{a^*_k}(t), e \leftrightarrow a^*_k\right\}\right] + \underbrace{\sum_{t=1}^{n} \mathbb{E}\left[\mathbb{1}\left\{\overline{G_{a^*_k}(t)}\right\}\right]}_{O\left(\frac{1}{K^2}\right), \text{Lemma 4}} \\
&= \underbrace{\sum_{t=1}^{n} \mathbb{E}\left[\mathbb{1}\left\{a^t_{\pi_t^{-1}(k)} = e, \mathcal{E}^\theta_{e,k}(t), G_{a^*_k}(t), e \leftrightarrow a^*_k, T_{a^*_k}(t-1) \leq l^*_{e,k}\right\}\right]}_{\Gamma_{1,k,e}} \cdot \Delta_{e,k} \\
&\quad + \underbrace{\sum_{t=1}^{n} \mathbb{E}\left[\mathbb{1}\left\{a^t_{\pi_t^{-1}(k)} = e, \mathcal{E}^\theta_{e,k}(t), G_{a^*_k}(t), e \leftrightarrow a^*_k, T_{a^*_k}(t-1) > l^*_{e,k}\right\}\right] \cdot \Delta_{e,k} + O\left(\frac{1}{K^2}\right)}_{\Gamma_{2,k,e}} .
\end{aligned} \tag{32}$$

Let $Y^\theta_{e,k}(t) := \mathbb{P}\left\{\theta_{a^*_k}(t) > y_{e,k} \,\middle|\, \mathcal{F}_{t-1}\right\}$. Then, we have

$$\begin{aligned}
\Gamma_{1,k,e} &= \sum_{t=1}^{n} \mathbb{E}\left[\mathbb{1}\left\{a^t_{\pi_t^{-1}(k)} = e, \mathcal{E}^\theta_{e,k}(t), G_{a^*_k}(t), e \leftrightarrow a^*_k, T_{a^*_k}(t-1) \leq l^*_{e,k}\right\}\right] \\
&= \sum_{t=1}^{n} \mathbb{E}\left[\mathbb{E}\left[\mathbb{1}\left\{a^t_{\pi_t^{-1}(k)} = e, \mathcal{E}^\theta_{e,k}(t), G_{a^*_k}(t), e \leftrightarrow a^*_k, T_{a^*_k}(t-1) \leq l^*_{e,k}\right\} \,\middle|\, \mathcal{F}_{t-1}\right]\right] \\
&= \sum_{t=1}^{n} \mathbb{E}\left[\mathbb{E}\left[\mathbb{1}\left\{a^t_{\pi_t^{-1}(k)} = e, \mathcal{E}^\theta_{e,k}(t), G_{a^*_k}(t), e \leftrightarrow a^*_k\right\} \cdot \mathbf{1}\left\{T_{a^*_k}(t-1) \leq l^*_{e,k}\right\} \,\middle|\, \mathcal{F}_{t-1}\right]\right] \\
&= \sum_{t=1}^{n} \mathbb{E}\left[\mathbf{1}\left\{T_{a^*_k}(t-1) \leq l^*_{e,k}, e \leftrightarrow a^*_k\right\} \cdot \mathbb{P}\left\{a^t_{\pi_t^{-1}(k)} = e, \mathcal{E}^\theta_{e,k}(t), G_{a^*_k}(t) \,\middle|\, \mathcal{F}_{t-1}\right\}\right] \\
&\overset{(a)}{\leq} \sum_{t=1}^{n} \mathbb{E}\left[\mathbf{1}\left\{T_{a^*_k}(t-1) \leq l^*_{e,k}, e \leftrightarrow a^*_k\right\} \cdot \frac{1 - Y^\theta_{e,k}(t)}{Y^\theta_{e,k}(t)} \mathbb{P}\left\{a^t_{\pi_t^{-1}(k)} = a^*_k, \mathcal{E}^\theta_{e,k}(t), G_{a^*_k}(t) \,\middle|\, \mathcal{F}_{t-1}\right\}\right] \\
&\leq \sum_{t=1}^{n} \mathbb{E}\left[\underbrace{\mathbf{1}\left\{T_{a^*_k}(t-1) \leq l^*_{e,k}, e \leftrightarrow a^*_k\right\} \cdot \frac{1 - Y^\theta_{e,k}(t)}{Y^\theta_{e,k}(t)} \mathbb{1}\left\{a^t_{\pi_t^{-1}(k)} = a^*_k, G_{a^*_k}(t)\right\}}_{\eta}\right] ,
\end{aligned} \tag{33}$$

where inequality (a) uses Lemma 5.

We now reduce the proof to the non-private setting. First, we divide all the instantiations F_{t-1} of $\mathcal{F}_{t-1}$ into two groups depending on whether $\mathbb{1}\left\{G_{a^*_k}(t), T_{a^*_k}(t-1) \leq l^*_{e,k}\right\}$ is 1 or 0.

Case 1: For F_{t-1} such that $\mathbb{1}\left\{G_{a^*_k}(t), T_{a^*_k}(t-1) \leq l^*_{e,k}\right\} = 0$, we have $\eta = 0$.

Case 2: For F_{t-1} such that $\mathbb{1}\left\{G_{a^*_k}(t), T_{a^*_k}(t-1) \leq l^*_{e,k}\right\} = 1$, we have $w'_{a^*_k, T_{a^*_k}(t-1)}(t) = \tilde{w}_{a^*_k, T_{a^*_k}(t-1)}(t-1) + \frac{3\ln(Kt)}{\varepsilon_0 \cdot T_{a^*_k}(t-1)} \geq \hat{w}_{a^*_k, T_{a^*_k}(t-1)}(t-1)$. Since a random variable drawn from $\mathcal{N}(\mu, \sigma^2)$ first-order stochastically dominates a random variable drawn from $\mathcal{N}(\mu', \sigma^2)$ if $\mu \geq \mu'$, we have

$$\frac{1 - Y^\theta_{e,k}(t)}{Y^\theta_{e,k}(t)} = \frac{\mathbb{P}\left\{\theta_{a^*_k}(t) \leq y_{e,k} \,\middle|\, \mathcal{F}_{t-1} = F_{t-1}\right\}}{\mathbb{P}\left\{\theta_{a^*_k}(t) > y_{e,k} \,\middle|\, \mathcal{F}_{t-1} = F_{t-1}\right\}} \leq \frac{\mathbb{P}\left\{\hat{\theta}_{a^*_k}(t) \leq y_{e,k} \,\middle|\, \mathcal{F}_{t-1} = F_{t-1}\right\}}{\mathbb{P}\left\{\hat{\theta}_{a^*_k}(t) > y_{e,k} \,\middle|\, \mathcal{F}_{t-1} = F_{t-1}\right\}} ,$$

where $\hat{\theta}_{a_k^*}(t) \sim \mathcal{N}\left(\hat{w}_{a_k^*, T_{a_k^*}(t-1)}(t-1), \frac{1}{T_{a_k^*}(t-1)}\right)$.

From these two cases, for any F_{t-1}, we have

$$\eta \leq \frac{\mathbb{P}\left\{\hat{\theta}_{a_k^*}(t) \leq y_{e,k} \,\middle|\, \mathcal{F}_{t-1} = F_{t-1}\right\}}{\mathbb{P}\left\{\hat{\theta}_{a_k^*}(t) > y_{e,k} \,\middle|\, \mathcal{F}_{t-1} = F_{t-1}\right\}} \mathbb{1}\left\{a^t_{\pi_t^{-1}(k)} = a_k^*, G_{a_k^*}(t)\right\} \quad . \tag{34}$$

Now, the proof is reduced to the non-private setting. We divide all the n rounds depending on when the empirical mean of a_k^* changes, i.e., whether $\hat{w}_{a_k^*, T_{a_k^*}(t-1)}(t-1)$ changes. Let τ_s denote the round by the end of which we use fresh 2^s observations to update the empirical mean of a_k^*. Note that τ_s is random. Then, we have

$$\begin{aligned}
(33) &\leq \sum_{s=0}^{d^*_{e,k}} \mathbb{E}\left[\sum_{t=\tau_s+1}^{\tau_{s+1}} \frac{\mathbb{P}\left\{\hat{\theta}_{a_k^*}(t) \leq y_{e,k} \,\middle|\, \mathcal{F}_{t-1} = F_{t-1}\right\}}{\mathbb{P}\left\{\hat{\theta}_{a_k^*}(t) > y_{e,k} \,\middle|\, \mathcal{F}_{t-1} = F_{t-1}\right\}} \cdot \mathbb{1}\left\{a^t_{\pi_t^{-1}(k)} = a_k^*\right\} \cdot \mathbb{1}\left\{G_{a_k^*}(t)\right\}\right] \\
&\leq \sum_{s=0}^{d^*_{e,k}} \mathbb{E}\left[\sum_{t=\tau_s+1}^{\tau_{s+1}} \frac{\mathbb{P}\left\{\hat{\theta}_{a_k^*}(t) \leq y_{e,k} \,\middle|\, \mathcal{F}_{t-1} = F_{t-1}\right\}}{\mathbb{P}\left\{\hat{\theta}_{a_k^*}(t) > y_{e,k} \,\middle|\, \mathcal{F}_{t-1} = F_{t-1}\right\}} \cdot \mathbb{1}\left\{a^t_{\pi_t^{-1}(k)} = a_k^*\right\}\right] \\
&\leq \sum_{s=0}^{d^*_{e,k}} \mathbb{E}\left[2^{s+1} \cdot \frac{\mathbb{P}\left\{\hat{\theta}_{a_k^*}(\tau_s+1) \leq y_{e,k} \,\middle|\, \mathcal{F}_{\tau_s} = F_{\tau_s}\right\}}{\mathbb{P}\left\{\hat{\theta}_{a_k^*}(\tau_s+1) > y_{e,k} \,\middle|\, \mathcal{F}_{\tau_s} = F_{\tau_s}\right\}}\right] \\
&\leq O\left(\frac{\ln(Kn)}{\Delta^2_{e,k}}\right) \quad ,
\end{aligned} \tag{35}$$

where the last inequality uses Fact 4.

Similarly, we will have

$$\begin{aligned}
\Gamma_{2,k,e} &\leq \sum_{s=d^*_{e,k}+1}^{\log(n)} \mathbb{E}\left[2^{s+1} \cdot \frac{\mathbb{P}\left\{\hat{\theta}_{a_k^*}(\tau_s+1) \leq y_{e,k} \,\middle|\, \mathcal{F}_{\tau_s} = F_{\tau_s}\right\}}{\mathbb{P}\left\{\hat{\theta}_{a_k^*}(\tau_s+1) > y_{e,k} \,\middle|\, \mathcal{F}_{\tau_s} = F_{\tau_s}\right\}}\right] \cdot \Delta_{e,k} \\
&\leq O\left(\tfrac{1}{L^2}\right) \quad .
\end{aligned} \tag{36}$$

We use the following arguments to complete the proof. Now, as we are tracking the number of observations of the optimal base arm a_k^*, i.e., we are tracking $T_{a_k^*}(t-1)$. For a fixed k, we arrange all the mean reward gaps $\Delta_{e,k}$ for all $e \in \bar{A}^*$ in a descending order $\Delta_{e_1,k} \geq \Delta_{e_2,k} \geq \ldots \geq \Delta_{e_{L-K},k} =: \Delta_{e_{\min},k}$.

For all the rounds when the counter is in the range of $\left[0, l^*_{e_1,k}\right]$, the total regret among all these rounds is at most $\Delta_{e_1,k} \cdot O\left(l^*_{e_1,k}\right)$. When the counter is in the range of $\left[l^*_{e_1,k}+1, l^*_{e_2,k}\right]$, the total regret is at most $\Delta_{e_2,k} \cdot O\left(l^*_{e_2,k} - l^*_{e_1,k}\right) + \Gamma_{2,k,e_1} \leq \Delta_{e_2,k} \cdot O\left(l^*_{e_2,k} - l^*_{e_1,k}\right) + O(1/K)$. Finally, when the counter is in the range of $\left[l^*_{e_{L-K-1},k}+1, l^*_{e_{L-K},k}\right]$, the total regret among all these rounds is at most $\Delta_{e_{L-K-1},k} \cdot O\left(l^*_{e_{L-K+1},k} - l^*_{e_{L-K},k}\right) + \sum_{q=1}^{L-K-1} \Gamma_{2,k,e_q} \leq \Delta_{e_{L-K-1},k} \cdot O\left(l^*_{e_{L-K},k} - l^*_{e_{L-K-1},k}\right) + O(1/K)$. For all the rounds after the counter hits $l^*_{e_{L-K},k}$, the total regret is at most $\sum_{k=1}^{K} O(1/L^2) \leq O(1/K)$.

By combining all these pieces together and using Lemma 3, we have

$$U \leq \sum_{k=1}^{K} O\left(\frac{\ln(Kn)}{\Delta_{k,\min}}\right) \quad . \tag{37}$$

Lemma 4. We have

$$\sum_{t=1}^{n} \mathbb{E}\left[\mathbb{1}\left\{\overline{G_e(t)}\right\}\right] \leq O\left(\frac{1}{K^2 n^2}\right) ,$$

and

$$\sum_{t=1}^{n} \mathbb{E}\left[\mathbb{1}\left\{\overline{C_e(t)}\right\}\right] \leq O\left(\frac{1}{K^2 n^2}\right) .$$

Proof. Proof of Lemma 4:The proofs use concentration bounds of Laplace distributions and Hoeffding's inequality. We have

$$\begin{aligned}
\sum_{t=1}^{n}\mathbb{E}\left[\mathbb{1}\left\{\overline{G_e(t)}\right\}\right] &= \sum_{t=1}^{n}\mathbb{P}\left\{|\tilde{w}_{e,T_e(t-1)}(t-1)-\hat{w}_{e,T_e(t-1)}(t-1)|>\frac{3\ln(Kt)}{\varepsilon_0\cdot T_e(t-1)}\right\}\\
&\le \sum_{t=1}^{n}\sum_{s=0}^{\log(t-1)}\mathbb{P}\left\{|\tilde{w}_{e,2^s}(t-1)-\hat{w}_{e,2^s}(t-1)|>\frac{3\ln(Kt)}{\varepsilon_0\cdot 2^s}\right\}\\
&\le \sum_{t=1}^{n}\sum_{s=0}^{\log(t-1)}\mathbb{P}\left\{|2^s\cdot\tilde{w}_{e,2^s}(t-1)-2^s\cdot\hat{w}_{e,2^s}(t-1)|>\frac{3\ln(Kt)}{\varepsilon_0}\right\}\\
&= \sum_{t=1}^{n}\sum_{s=0}^{\log(t-1)}e^{-3\ln(Kt)}\\
&\le O\left(\frac{1}{K^2n^2}\right),
\end{aligned}$$

where the second Inequality used the concentration bound of a Laplace random variable (Fact 3).

Similarly, we have

$$\begin{aligned}
\sum_{t=1}^{n}\mathbb{E}\left[\mathbb{1}\left\{\overline{C_e(t)}\right\}\right] &= \sum_{t=1}^{n}\mathbb{P}\left\{|\hat{w}_{e,T_e(t-1)}(t-1)-\bar{w}(e)|>\sqrt{\frac{3\ln(Kt)}{T_e(t-1)}}\right\}\\
&\le \sum_{t=1}^{n}\sum_{s=0}^{\log(t-1)}\mathbb{P}\left\{|\hat{w}_{e,2^s}(t-1)-\bar{w}(e)|>\sqrt{\frac{3\ln(Kt)}{2^s}}\right\}\\
&\le \sum_{t=1}^{n}\sum_{s=0}^{\log(t-1)}2e^{-2\cdot 2^s\cdot\frac{3\ln(Kt)}{2^s}}\\
&\le O\left(\frac{1}{K^2n^2}\right),
\end{aligned}$$

where the second inequality uses Hoeffding's inequality (Fact 2). □

Lemma 5. For all t and for any instantiation F_{t-1} of $\mathcal{F}_{t-1}$, we have

$$\mathbb{P}\left\{a^t_{\pi_t^{-1}(k)}=e,\mathcal{E}^{\theta}_{e,k}(t),G_{a^*_k}(t)\,\middle|\,\mathcal{F}_{t-1}=F_{t-1}\right\}\le\frac{1-Y^{\theta}_{e,k}(t)}{Y^{\theta}_{e,k}(t)}\mathbb{P}\left\{a^t_{\pi_t^{-1}(k)}=a^*_k,\mathcal{E}^{\theta}_{e,k}(t),G_{a^*_k}(t)\,\middle|\,\mathcal{F}_{t-1}=F_{t-1}\right\}. \tag{38}$$

Proof of Lemma 5. We start by noting that the event $G_{a^*_k}(t)$ is determined by the history $\mathcal{F}_{t-1}$.

Case 1: If F_{t-1} is the one such that $G_{a^*_k}(t)$ is false, then both sides of the inequality shown in (38) are zero, and the inequality trivially holds.

Case 2: If F_{t-1} is the one such that $G_{a^*_k}(t)$ is true, we can omit $G_{a^*_k}$ in both sides in (38). Let $\pi_t^{-1}(k)=i$ and $A^t_{i-1}=\{a^t_1,\ldots,a^t_{i-1}\}$ be the set of the first $i-1$ base arms selected greedily by Algorithm 3. To complete the proof, it suffices to show

$$\mathbb{P}\left\{a^t_{\pi_t^{-1}(k)}=e,\mathcal{E}^{\theta}_{e,k}(t)\,\middle|\,\mathcal{F}_{t-1}=F_{t-1},A^t_{i-1}\right\}\le\frac{1-Y^{\theta}_{e,k}(t)}{Y^{\theta}_{e,k}(t)}\mathbb{P}\left\{a^t_{\pi_t^{-1}(k)}=a^*_k,\mathcal{E}^{\theta}_{e,k}(t)\,\middle|\,\mathcal{F}_{t-1}=F_{t-1},A^t_{i-1}\right\}. \tag{39}$$

Let $\sigma(A^t_{i-1})=\left\{e:e\in E\setminus A^t_{i-1},A^t_{i-1}\cup\{e\}\in\mathcal{I}\right\}$ be the set of base arms that can be added to the current solution set A^t_{i-1}. Note that $a^*_k\in\sigma(A^t_{i-1})$ and $a^*_k\notin A^t_{i-1}$.

We first construct an upper bound for the LHS of (39). We have

$$\begin{aligned}
&\mathbb{P}\left\{a^t_{\pi_t^{-1}(k)} = e, \mathcal{E}^\theta_{e,k}(t) \,\middle|\, \mathcal{F}_{t-1} = F_{t-1}, A^t_{i-1}\right\} \\
&\le \mathbb{P}\left\{\theta_j(t) \le y_{e,k}, \forall j \in \sigma(A^t_{i-1}) \,\middle|\, \mathcal{F}_{t-1} = F_{t-1}, A^t_{i-1}\right\} \\
&\overset{(\spadesuit)}{=} \mathbb{P}\left\{\theta_{a^*_k}(t) \le y_{e,k} \,\middle|\, A^t_{i-1}, \mathcal{F}_{t-1} = F_{t-1}\right\} \cdot \mathbb{P}\left\{\theta_j(t) \le y_{e,k}, \forall j \in \sigma(A^t_{i-1}) \setminus \{a^*_k\} \,\middle|\, \mathcal{F}_{t-1} = F_{t-1}, A^t_{i-1}\right\} \\
&= \mathbb{P}\left\{\theta_{a^*_k}(t) \le y_{e,k} \,\middle|\, \mathcal{F}_{t-1} = F_{t-1}\right\} \cdot \mathbb{P}\left\{\theta_j(t) \le y_{e,k}, \forall j \in \sigma(A^t_{i-1}) \setminus \{a^*_k\} \,\middle|\, \mathcal{F}_{t-1} = F_{t-1}, A^t_{i-1}\right\} \\
&= (1 - Y^\theta_{e,k}(t)) \cdot \mathbb{P}\left\{\theta_j(t) \le y_{e,k}, \forall j \in \sigma(A^t_{i-1}) \setminus \{a^*_k\} \,\middle|\, \mathcal{F}_{t-1} = F_{t-1}, A^t_{i-1}\right\},
\end{aligned} \tag{40}$$

where ($\spadesuit$) uses the fact that $\theta_{a^*_k}(t)$ and all base arms in $\sigma(A^t_{i-1})$ are independent.

Similarly, the RHS of (39) is lower bounded by

$$\begin{aligned}
&\mathbb{P}\left\{a^t_{\pi_t^{-1}(k)} = a^*_k, \mathcal{E}^\theta_{e,k}(t) \,\middle|\, \mathcal{F}_{t-1} = F_{t-1}, A^t_{i-1}\right\} \\
&\ge \mathbb{P}\left\{\theta_{a^*_k}(t) > y_{e,k} \ge \theta_j(t), \forall j \in \sigma(A^t_{i-1}) \setminus \{a^*_k\} \,\middle|\, \mathcal{F}_{t-1} = F_{t-1}, A^t_{i-1}\right\} \\
&= \mathbb{P}\left\{\theta_{a^*_k}(t) > y_{e,k} \,\middle|\, \mathcal{F}_{t-1} = F_{t-1} A^t_{i-1}\right\} \cdot \mathbb{P}\left\{\theta_j(t) \le y_{e,k}, \forall j \in \sigma(A^t_{i-1}) \setminus \{a^*_k\} \,\middle|\, \mathcal{F}_{t-1} = F_{t-1}, A^t_{i-1}\right\} \\
&= Y^\theta_{e,k}(t) \cdot \mathbb{P}\left\{\theta_j(t) \le y_{e,k}, \forall j \in \sigma(A^t_{i-1}) \setminus \{a^*_k\} \,\middle|\, \mathcal{F}_{t-1} = F_{t-1}, A^t_{i-1}\right\}.
\end{aligned} \tag{41}$$

Combining (40) and (41) gives

$$\mathbb{P}\left\{a^t_{\pi_t^{-1}(k)} = e, \mathcal{E}^\theta_{e,k}(t) \,\middle|\, \mathcal{F}_{t-1} = F_{t-1}, A^t_{i-1}\right\} \le \frac{1 - Y^\theta_{e,k}(t)}{Y^\theta_{e,k}(t)} \mathbb{P}\left\{a^t_{\pi_t^{-1}(k)} = a^*_k, \mathcal{E}^\theta_{e,k}(t) \,\middle|\, \mathcal{F}_{t-1} = F_{t-1}, A^t_{i-1}\right\}. \tag{42}$$

To get the stated result, we use law of total expectation and the fact that $Y^\theta_{e,k}$ is determined by $\mathcal{F}_{t-1}$. We have

$$\begin{aligned}
&\mathbb{P}\left\{a^t_{\pi_t^{-1}(k)} = e, \mathcal{E}^\theta_{e,k}(t) \,\middle|\, \mathcal{F}_{t-1} = F_{t-1}\right\} \\
=&\mathbb{E}\left[\mathbf{1}\left\{a^t_{\pi_t^{-1}(k)} = e, \mathcal{E}^\theta_{e,k}(t)\right\} \,\middle|\, \mathcal{F}_{t-1} = F_{t-1}\right] \\
=&\mathbb{E}\left[\mathbb{P}\left\{a^t_{\pi_t^{-1}(k)} = e, \mathcal{E}^\theta_{e,k}(t) \,\middle|\, A^t_{i-1}, \mathcal{F}_{t-1} = F_{t-1}\right\} \,\middle|\, \mathcal{F}_{t-1} = F_{t-1}\right] \\
\le&\mathbb{E}\left[\frac{1 - Y^\theta_{e,k}(t)}{Y^\theta_{e,k}(t)} \mathbb{P}\left\{a^t_{\pi_t^{-1}(k)} = a^*_k, \mathcal{E}^\theta_{e,k}(t) \,\middle|\, A^t_{i-1}, \mathcal{F}_{t-1} = F_{t-1}\right\} \,\middle|\, \mathcal{F}_{t-1} = F_{t-1}\right] \\
=&\frac{1 - Y^\theta_{e,k}(t)}{Y^\theta_{e,k}(t)} \mathbb{P}\left\{a^t_{\pi_t^{-1}(k)} = a^*_k, \mathcal{E}^\theta_{e,k}(t) \,\middle|\, \mathcal{F}_{t-1} = F_{t-1}\right\},
\end{aligned} \tag{43}$$

where the only inequality uses (42). □

Fact 2. (Hoeffding's inequality). Let $X_1, \ldots, X_n$ be independent random variables with each $X_i \in [a_i, b_i]$. Then, for any $\epsilon > 0$, we have

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n (X_i - \mathbb{E}X_i)\right| \ge \epsilon\right) \le 2\exp\left(\frac{-2n^2\epsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}\right). \tag{44}$$

Fact 3. ((Dwork et al., 2014, Fact 3.6); tail probability for Laplace distribution). If $Y \sim \mathrm{Lap}(b)$, for any $0 < \delta \le 1$, we have

$$\mathbb{P}\{|Y| \ge b\ln(1/\delta)\} = \delta. \tag{45}$$

Fact 4. (Agrawal & Goyal, 2017, Lemma 2.13). Let τ_s be the round by the end of which we use fresh 2^s observations to update the empirical mean of an optimal base arm a^*_k. Then, we have

$$\mathbb{E}\left[\frac{1 - Y^{\hat{\theta}}_{r,k}(\tau_s + 1)}{Y^{\hat{\theta}}_{r,k}(\tau_s + 1)}\right] \le \begin{cases} O(1) & \forall s, \\ O\left(\frac{1}{L^2 n}\right) & s \ge \log\left(l^*_{e,k}\right), \end{cases}$$

where $l^*_{e,k} := \left\lceil \log\left(\frac{288\ln\left(L^2\left(n + e^{32}\right)\right)}{\Delta^2_{e,k}}\right)\right\rceil$.

Fact 5. (Gaussian tail bound). Let X be a Gaussian distributed random variable with mean $\mathbb{E}[X]$ and variance σ^2, then for any t we have

$$\mathbb{P}\{X - \mathbb{E}[X] > t\} \le e^{-\frac{t^2}{2\sigma^2}}. \tag{46}$$