Multi-Armed Bandits with Local Differential Privacy

Anonymous Authors[1]

Abstract

This paper investigates the problem of regret minimization for multi-armed bandit (MAB) problems with local differential privacy (LDP) guarantees. In stochastic bandit systems, the rewards may refer to the private activities that users do not want the agent to know, while the agent needs to know these activities to provide better services. LDP MAB is a model for handling this dilemma and we study its regret lower and upper bounds. For the case of homogeneous privacy levels, we provide the first known result with matching upper and lower bounds of the regret. In contrast to existing works, we also consider the LDP MAB problem where the users' privacy levels are heterogeneous, and the regret upper bounds are order-optimal under mild conditions. Numerical experiments also confirm our theoretical conclusions.

1. Introduction

1.1. Background and Motivation

The multi-armed bandit (MAB) (Berry and Fristedt, 1985) model is classic for abstracting sequential decision making under uncertainty and has attracted interests in various areas, such as communication networks, online advertising, clinical trials, etc. In an MAB model, there is a set of *arms*, and each *pull* of an arm generates a random *reward* according to some unknown latent distribution of this arm. The *agent* adaptively chooses arms to pull according to past observations to achieve some goal. A widely studied goal is *regret minimization*, where the regret is the expected gap between a proposed algorithm and an optimal algorithm that knows the latent distributions.

In recent years, users have become increasingly concerned about their private information and activities. However, many learning systems (e.g., medical experiments, recommender systems, advertisement allocators, and search engines, etc.) need such data to learn critical matters and provide better services. To handle this dilemma, we adopt the concept of differential privacy (DP), a widely accepted and applied metric to measure the privacy level. In theory, DP guarantees the difficulty for any party to determine whether or not an individual is listed in a private database. DP has been studied in many areas, such as data release (Mohammed et al., 2011), optimization (Huang et al., 2015), Q-learning (Wang and Hegde, 2019), just to name a few. Yet for DP in MAB, there are still many significant and open problems.

We take clinical trials as an example of DP in MAB. In an experiment of an illness with multiple treatments (aka arms), the experimenter (aka agent) wants to sequentially choose treatments for patients (aka individuals) based on past observations on treatment effects. This problem can be viewed as a regret minimization problem. However, the patients may not be willing to share the actual effects of the treatments with the experimenter due to privacy concerns. With the DP bandit algorithms, the actual effects will be concealed from the experimenter, which provides a certain level of privacy guarantee to all patients, while allowing the experimenter to learn from the observations efficiently.

The above example fits the *locally differentially private (LDP)* bandit model (Basu et al., 2019), where there is no trusted centroid curator. In this paper, we assume that each user has its own curator that can do randomized mapping on its data to provide privacy guarantee. These curators can be software or plugins embedded in the user's devices or terminals, and the non-private data will not leave users' control unless they have been processed by the curators. This differs from the DP bandits (e.g., (Mishra and Thakurta, 2015)) where there is a centroid curator with access to all the users' data. In machine learning systems like social media and search engines, it is unlikely that there is a third-party centroid curator that can gain access to the users' data, as these data are commonly viewed as valuable assets.

Existing works of LDP bandits (Gajane et al., 2018; Basu et al., 2019; Chen et al., 2020; Zheng et al., 2020) all focus on *homogeneous* privacy levels. This paper is the first to study the significant but missing case where the individuals' privacy levels are *heterogeneous*. In practice, some people

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

may be more willing to provide their data and the system can utilize these data and provide better services to all users. For instance, in the sequential medical experiments, some users take the medicine and are willing to share the results with the experimenters and the others only take the medicine while not revealing the results due to privacy concerns. This implies two types of privacy levels. The experimenters can utilize the information provided by individuals of type one to provide better treatments to all individuals. LDP MAB with heterogeneous privacy levels also covers the homogeneous case, and thus, is more general and can model the reality better.

1.2. Problem Statement

Bandit model. In this paper, there are n arms indexed by $1, 2, 3, ..., n$, and we use $[n]$[1] to denote the set of all arms. Each arm a is associated with an *unknown* latent distribution, and each pull of arm a returns a random reward according to its latent distribution. Let $R_{a,t}$ be the reward of the t-th pull of arm a and $\mu_a := \mathbb{E}[R_{a,t}]$ be the *mean reward* of arm a. We assume that the rewards are *independent* across arms and time, i.e., $(R_{a,t}, a \in [n], t \in \mathbb{Z}^+)$ are independent. Define $\mu^* = \max_{a\in[n]} \mu_a$. For any arm a, we define the mean reward *gap* as $\Delta_a := \mu^* - \mu_a$. An arm a is *optimal* iff $\Delta_a = 0$, and *sub-optimal* iff $\Delta_a > 0$.

Regret minimization. Given a time horizon $T > n$ (T may be unknown in this paper, fitting the online learning setting), the agent pulls the arms for at most T times. Let A_t be the t-th pulled arm and $N_{a,t} := \sum_{\tau=1}^{t} \mathbb{1}\{A^\tau = a\}$ be the number of pulls on arm a till time t. After T pulls, the (pseudo) regret is defined as

$$R(T) := T\mu^* - \sum_{t=1}^{T} \mathbb{E}[\mu_{A_t}] = \mathbb{E}[\sum_{a\in[n]} N_{a,T}\Delta_a].$$

The goal of the agent is to minimize the regret.

Local differential privacy (LDP). We first define ϵ-LDP (Duchi et al., 2013) in Definition 1.

Definition 1. *For $\epsilon > 0$, a randomized mapping $M : \mathcal{D} \to \mathbb{R}$ is said to be ϵ-LDP on $\mathcal{D} \subset \mathbb{R}$ if for any x, x' in $\mathcal{D}$ and a measurable subset E of $\mathbb{R}$, we have*

$$\mathbb{P}\{M(x) \in E\} \le e^\epsilon \cdot \mathbb{P}\{M(x') \in E\}.$$

The above inequality must hold if we switch x and x'.

This definition implies that for any two records in $\mathcal{D}$, under an ϵ-LDP mechanism, their statistical behaviors are similar. Hence, it is difficult for any party to determine which record is the source of the given output. Smaller values of ϵ implies higher levels of privacy. When $\epsilon = \infty$, there is no privacy.

[1]We define $[m] := \{1, 2, 3, ..., m\}$ for all $m \in \mathbb{Z}^+$.

We note that ϵ-LDP is also denoted as $(\epsilon, 0)$-DP, a stronger privacy level than (ϵ, δ)-LDP.

LDP bandit model. In the LDP bandit model, the users do not trust the agent and do not share the rewards to the agent. In each iteration, the agent makes a decision on which arm to pull according to past private responses and sends a request to the incoming user. Then the user sends the reward to its curator, and the curator returns a private response to the agent.

We first define HoLDP (LDP with Homogeneous privacy levels) aka ϵ-LDP. For random vectors X and Y, we use $X \in \sigma(Y)$ to represent that X is determined by Y plus some random factors independent of Y and the bandit instance. In the following definition, it implies that the agent does not know the actual rewards.

Definition 2 (HoLDP bandit model)**.** *Let A_t be the t-th pulled arm, R_t be the corresponding reward. For $\epsilon > 0$, the bandit model is said to be ϵ-LDP (HoLDP) if i) there is an ϵ-LDP mechanism $M : \mathcal{D} \to \mathbb{R}$ and ii) $A_{t+1} \in \sigma(A_s, M(R_s) : 1 \le s \le t)$ for any time t.*

Heterogeneous privacy levels. We assume that at each time, a new user arrives and the agent chooses an action (arm) for this user. Thus, the privacy provided to a user can be viewed as provided to the corresponding reward. In this paper, we allow the privacy levels of the users to be different. We define this type of LDP as Heterogeneous-LDP (HeLDP) in Definition 3. We use ϵ_t to denote the privacy level provided to the t-th user (and the corresponding reward), and $\epsilon_{a,t}$ to denote that provided to the t-th reward of arm a (and the corresponding user). If $\epsilon_t \equiv \epsilon$ for any t, then HeLDP reduces to HoLDP as a special case. For HeLDP MAB, we assume that the agent knows or is informed of the privacy level of each user, which is often reasonable in practice and vital for developing the algorithms.

Definition 3 (HeLDP bandit model)**.** *The bandit model is said to be HeLDP if there is a sequence of mechanisms $\{M_t\}_{t=1}^T$ such that $M_t : \mathcal{D} \to \mathbb{R}^l$ is ϵ_t-DP and $A_{t+1} \in \sigma(A_s, M_s(R_s) : 1 \le s \le t)$ for any time t.*

We further assume that the values of ϵ_t's are independently randomly drawn from a distribution $\mathcal{F}$. This is reasonable in many cases as in our modeling, each reward of an arm is generated by doing the corresponding action to an incoming user, and the corresponding privacy level can be viewed as drawn from the user pool. In practice, the agent may also know this distribution a priori, e.g., the server of a website may know the users' options regarding the privacy.

1.3. Related Work

Non-private MAB have been studied for decades, and either frequentist methods like UCB (Upper Confidence Bound) (Auer et al., 2002) or Bayesian methods like Thompson

Sampling (Agrawal and Goyal, 2012) can achieve optimal regret performance (up to constant factors). For a literature review on MAB, we refer readers to (Lattimore and Szepesvári, 2018). We also refer readers to (Dwork et al., 2014) for introductions and foundations of DP.

To the best of our knowledge, the earliest work that studied LDP bandits was (Gajane et al., 2018), which proposed an LDP bandit algorithm that works for Bernoulli arms. In contrast, our algorithms work for a much more general set of instances. A later work that studied LDP bandits is (Basu et al., 2019), in which distribution-dependent and distribution-free regret lower bounds were proved. We note that the distribution-dependent regret lower bound in (Basu et al., 2019) is looser than the one proved in this paper.

(Chen et al., 2020) proposed algorithms for HoLDP combinatorial bandits and proved a corresponding lower bound. For standard HoLDP MAB, their regret is of the same order as our Laplace mechanism[2], but has a larger ϵ-factor compared to our Bernoulli mechanism. Moreover, this paper also studies HeLDP MAB, and the algorithms or analysis for HoLDP MAB cannot be straightforwardly generalized to HeLDP MAB. According to the proof of Theorem 5 in (Chen et al., 2020), their lower bound assumes that the mean rewards of the arms are bounded away from 0 and 1 and use the inequality $D_{\mathrm{KL}}(p||q) \leq \frac{(p-q)^2}{q(1-q)}$ to get its lower bound, which makes the lower bound not optimal when the mean rewards of the arms are approaching 0 or 1. In contrast, our lower bound is optimal for arbitrary mean rewards in $[0, 1]$. (Zheng et al., 2020) proposed HoLDP MAB algorithms based on the Gaussian mechanism. Their algorithms only have (ϵ, δ)-LDP guarantee, weaker than ϵ-LDP, and their regrets have an additional $\log \frac{1}{\delta}$ factor compared to ours.

The following works are related to DP bandits but not directly comparable to our results. (Mishra and Thakurta, 2014; 2015; Sajed, 2019) studied the DP bandits, where there is a centroid curator that can aggregate the users' rewards. In (Shariff and Sheffet, 2018; Tossou and Dimitrakakis, 2015; 2016), the agent is trusted and the model does not want third-party eavesdroppers to learn the individual rewards from the agents' actions. The model in (Hannun et al., 2019) is also protecting the context features in the contextual bandit setting. (Tossou and Dimitrakakis, 2017) studied privacy-preserving adversarial bandits.

1.4. Main Contributions

1. We prove a regret lower bound for HoLDP MAB, which is *optimal* up to a constant factor.
2. We propose HeLDP MAB algorithms for the Laplace and the Bernoulli mechanisms, whose regrets are both optimal under mild assumptions. They can also be directly applied to HoLDP MAB, and the regrets *match* our lower bound up to constant factors.
3. For HoLDP bandits with unbounded and i.i.d.[3] sub-Gaussian noises, we propose a Sigmoid preprocessing and obtain algorithms with *tight* regret upper bounds (up to constant factors).

[2]We note that the algorithms in (Chen et al., 2020) are not online, but ours are.

2. Lower Bound

In this section, we establish a fundamental regret lower bound for HoLDP MAB. Typically, the lower bound of regret minimization depends on the KL-divergence (Cover and Thomas, 2012) between the latent distributions of the optimal arm and sub-optimal arms (Lai and Robbins, 1985). Let f and g be the probability density function (PDF) of two distributions, and we allow point masses in PDFs. The KL-divergence between f and g is defined as $D_{\mathrm{KL}}(f||g) := \int_{\mathbb{R}} f(x) \log \frac{f(x)}{g(x)} \mathrm{d}x$, where we stipulate $0 \cdot \log 0 = 0$.[4] (Basu et al., 2019; Chen et al., 2020) proved lower bounds for LDP bandit algorithms that depend on the KL-divergence between latent distributions. We note that from the proof of the lower bound in Chen et al. (2020), we can see that its lower bounds also depends on the KL-divergence. However, in this paper, our Theorem 4 states a tighter lower bound that depends on the values of $\Delta_a = \mu^* - \mu_a$ but not the KL-divergence. Later, Theorems 8 and 11 will show that when ϵ approaches zero, this lower bound is tight up to a constant factor. In the proof of Theorem 4, we will use the results in (Arratia and Gordon, 1989; Dragomir and Gluščevic, 2001).

Theorem 4. [5]*Let $\epsilon > 0$ be given. Assume that the rewards of all arms follow Bernoulli distributions. The regret $R(T)$ of any ϵ-LDP policy satisfies*

$$\liminf_{T\to\infty} \frac{R(T)}{\log T} \geq \frac{1}{(e^\epsilon - e^{-\epsilon})^2} \sum_{a:\Delta_a>0} \frac{1}{\Delta_a}.$$

When ϵ approaches 0, since $e^\epsilon - e^{-\epsilon} \simeq 2\epsilon$, we have $\liminf_{T\to\infty} \frac{R(T)}{\log T} \gtrsim \frac{1}{4\epsilon^2} \sum_{a:\Delta_a>0} \frac{1}{\Delta_a}$.

Remark. Theorem 4 is tight (according to Theorem 8) when ϵ approaches zero even if the mean rewards of the arms are approaching 0 or 1. In contrast, the lower bounds of existing works (Basu et al., 2019; Chen et al., 2020) depend on the KL-divergence, i.e., $\sum_{a:\Delta_a>0} \frac{\Delta_a}{D_{\mathrm{KL}}(\mu_a||\mu^*)}$, where $D_{\mathrm{KL}}(\mu_a||\mu^*) = \mu_a \log \frac{\mu_a}{\mu^*} + (1-\mu_a)\frac{1-\mu_a}{1-\mu^*}$. When μ_a and μ^* are approaching 0 or 1, one has $\frac{\Delta_a}{D_{\mathrm{KL}}(\mu_a||\mu^*)} =$

[3]"i.i.d." means "identically independently distributed".

[4]All log is this paper are natural log. $0 \cdot \log 0 = 0$ holds because $\lim_{x\to 0^+} x \log x = 0$.

[5]Due to space limitation, all proofs in this paper are provided in Supplementary Material.

$o(\frac{1}{\Delta_a})$ which is looser than our lower bound that depends on $\sum_{a:\Delta_a>0} \frac{1}{\Delta_a}$.

3. Algorithms

We propose two ϵ-LDP mechanisms and corresponding HeLDP algorithms, which can also be directly applied to HoLDP MAB. One mechanism is to add Laplace noise and the other is to convert rewards to Bernoulli responses.

Due to the HeLDP setting, the distributions of the private responses of the same arm may not be the same, which means that we cannot directly apply existing algorithms. Furthermore, the variances of the private responses are also heterogeneous, random, and even unbounded since ϵ_t's are random, making it harder to find a good confidence bound. In the following parts, we will introduce how we resolve these issues and develop the algorithms.

3.1. Laplace Mechanism

The Laplace mechanism is a widely used DP mechanism, and the key idea is to add an independent zero-mean Laplace$(1/\epsilon)$ noise to each reward. For any $b > 0$, the PDF of the Laplace(b) distribution is:

$$\text{Laplace}(b) : l(x; b) = (2b)^{-1} \exp(-|x|/b).$$

The variance of the Laplace(b) distribution is $2b^2$. The Laplace mechanism is described in Curator 1 and Lemma 5 states that it is ϵ-LDP.

Curator 1 Convert-to-Laplace (CTL)

Upon receiving reward r and privacy level ϵ:

$M_L(r; \epsilon) := r + L$, where $L \sim \text{Laplace}(1/\epsilon)$;

return $\epsilon, M_L(r; \epsilon)$;

Lemma 5 (Proposition 3.3 in (Dwork et al., 2016)). *CTL (i.e., $M_L(\cdot; \epsilon)$) is ϵ-LDP on $[0, 1]$.*

We then introduce a Chernoff concentration for the summation of independent Laplace random variables in Lemma 6. In the literature, the most common concentrations require the reward distributions to be either bounded or sub-Gaussian, but Laplace distributions satisfy neither. (Bubeck et al., 2013) proposed algorithms that can obtain regrets of the same order as sub-Gaussian bandits for bandits with heavy tailed rewards. However, in their work, the reward distributions of an arm are identical, while in this work, the distributions of the private responses of an arm are different due to the HeLDP setting. Thus, it is not obvious how to apply their techniques to our setting. Moreover, even for HoLDP, by using the confidence bound stated in Lemma 6, we do not need to use the complicated estimators or truncating methods as in (Bubeck et al., 2013), but only need to change the confidence bounds and can get order-optimal regret.

Lemma 6 (Lemma 2.8 in (Chan et al., 2011)). *Let $X_1, X_2, ..., X_m$ be independent zero-mean Laplace random variables with parameters $b_1, b_2, ..., b_m$. Define $Y_m := X_1 + X_2 + \cdots + X_m$ and $b_M := \max_{i=1}^m b_i$. For $\nu \geq \sqrt{\sum_{i=1}^m b_i^2}$ and $0 < \lambda < \frac{2\sqrt{2}\nu^2}{b_M}$, we have $\mathbb{P}\{Y_m > \lambda\} \leq \exp(-\frac{\lambda^2}{8\nu^2})$.*

We can use Lemma 6 to bound the empirical means of the private responses. Specifically, when $\sum_{r=1}^t \epsilon_{a,t}^{-2}$ is sufficiently large, we have concentration:

$$\mathbb{P}\Big\{ \sum_{r=1}^t L_{a,r} \geq \sqrt{8 \log t^4 \sum_{r=1}^t \epsilon_{a,r}^{-2}} \Big| \epsilon_{a,r} : r \in [t] \Big\} \leq t^{-4}.$$

However, since $\epsilon_{a,t}$'s are random, there are three new issues that are unseen in bandits with i.i.d. rewards.

First, some privacy level $\epsilon_{a,t}$ may be arbitrarily close to zero, making $\epsilon_{a,t}^{-1}$ and the above upper confidence bound arbitrarily large. Our solution is to remove private responses (ϵ_t, x_t) with ϵ_t smaller than a chosen parameter ϵ_{min}. This can also help improve the empirical regrets, which will be discussed in Section 4.

Second, in order for Lemma 6 to hold, we need $\sum_{r=1}^t \epsilon_{a,r}^{-2} > \epsilon_{min}^{-2} \log t^4$, which is not guaranteed in practice. Our solution is to first pull the arms until the values of $\sum_{r=1}^t \epsilon_{a,r}^{-2}$ are large enough to satisfy this condition before we move to the UCB pulling phase. This operation contributes to the second line of Theorem 7's regret.

Lastly, the confidence bound $\sqrt{\frac{8}{t^2} \sum_{r=1}^t \epsilon_{a,r}^{-2} \log t^4}$ itself is random due to the randomness of $\epsilon_{a,r}$'s, different from HoLDP. Thus, we need to take the additional randomness of $\sum_{r=1}^t \epsilon_{a,r}^{-2}$ into consideration too. Specifically, in the analysis, we will use another confidence bound to bound the confidence bound, which also contributes to the leftmost term of the third line of Theorem 7's regret.

With the above solutions, we present the algorithm in Agent 2. Its theoretical performance is stated in Theorem 7. Here, we let $\alpha(\epsilon) = \epsilon^{-2}$ and ϵ_{max} be the supreme of ϵ_t's region, and define

$$Q_L := \alpha(\epsilon_{min}) - \alpha(\epsilon_{max}),\ \alpha_0 := \mathbb{E}[\alpha(\epsilon_t) \mid \epsilon_t \geq \epsilon_{min}],$$
$$\text{and } p_0 := \mathbb{P}\{\epsilon_t \geq \epsilon_{min}\}.$$

Theorem 7. *HeLDP-UCB-L is HeLDP. Its distribution-*

Agent 2 HeLDP-UCB-L (HeLDP UCB by Laplace)

1: Set the privacy threshold $\epsilon_{min} > 0$;
2: $N_{a,0}, s_{a,0}, A_{a,0} \leftarrow 0$, and $u_{a,0} \leftarrow \infty$ for all arms a;
3: **for** $t = 1$ to T **do**
4: **if** $\exists$arm a having $A_{a,t} \le \epsilon_{min}^{-2} \log t^4$ **then**
5: $a_t \leftarrow a$;
6: **else**
7: $a_t \leftarrow \arg\max_{a\in[n]} u_{a,t}$;
8: **end if**
9: Pull arm a_t with respect to the incoming user;
10: Collect the private response (ϵ_t, x_t) from CTL;
11: **if** $\epsilon_t \ge \epsilon_{min}$ **then**
12: $N_{a_t,t} \leftarrow N_{a_t,t-1} + 1$;
13: $s_{a_t,t} \leftarrow s_{a_t,t-1} + x_t$;
14: $A_{a_t,t} \leftarrow A_{a_t,t-1} + \epsilon_t^{-2}$;
15: $u_{a_t,t} \leftarrow \frac{s_{a_t,t}}{N_{a_t,t}} + \sqrt{\frac{\log t^4}{2N_{a_t,t}}} + \sqrt{\frac{8A_{a_t,t}\log t^4}{N_{a_t,t}^2}}$;
16: **else**
17: Discard the response (ϵ_t, x_t);
18: **end if**
19: **for** arm a in $[n]$ with $N_{a,t}$ not updated **do**
20: $N_{a,t} \leftarrow N_{a,t-1}$; $s_{a,t} \leftarrow s_{a,t-1}$;
21: $A_{a,t} \leftarrow A_{a,t-1}$; Update $u_{a,t}$ as Line 14;
22: **end for**
23: **end for**

dependent regret is at most

$$\inf_{\kappa>0}\Big\{\sum_{a:\Delta_a>0}\Big[\max\Big\{\frac{8}{p_0\Delta_a}(1+4\sqrt{\alpha_0+\kappa Q_L})^2\log T, \frac{2\Delta_a}{p_0\alpha_0^2}(4\alpha_0\epsilon_{min}^{-2}+\epsilon_{min}^{-4})\log T\Big\} + \frac{\Delta_a}{2p_0\kappa^2}\log(p_0T\exp\{2\kappa^2\}) + \frac{2\Delta_a}{p_0} + \frac{2\pi^2\Delta_a}{3}\Big]\Big\};$$

its distribution-free regret is $O(\sqrt{\frac{(1+\alpha_0)nT\log T}{p_0}})$.

Remark: i) When $\epsilon_{min} = \Omega(\epsilon_{max})$ and $p_0 = \Omega(1)$, by setting $\kappa = 1$, the regret upper bound of HeLDP-UCB-L reduces to $O(\log T)\sum_{a:\Delta_a>0}\frac{1}{\Delta_a}(1+\epsilon_{max}^{-1})^2$, matching the lower bound in Theorem 4 with $\epsilon = \epsilon_{max}$ up to a constant factor. In this case, the distribution-free regret will reduce to $O((1+\epsilon_{max}^{-1})\sqrt{nT\log T})$, matching the lower bound of Basu et al. (2019) up to a $\sqrt{\log T}$ factor.

For HoLDP MAB, we have $\epsilon_t \equiv \epsilon$, and by setting $\epsilon_{min} = \epsilon$, we have $Q_L = 0$, $\alpha_0 = \epsilon^{-2}$, and $p_0 = 1$. First, we set $\kappa = \infty$, reducing the $\frac{\Delta_a}{2p_0\kappa^2}\log(p_0T\exp\{2\kappa^2\})$ term to Δ_a. Second, the term $\frac{2\Delta_a}{p_0\alpha_0^2}(4\alpha_0\epsilon_{min}^{-2}+\epsilon_{min}^{-4})\log T$ becomes $10\Delta_a\log T$, which is dominated by $\frac{8}{p_0\Delta_a}(1+4\sqrt{\alpha_0+\kappa Q_L})^2\log T$ and can be deleted. Therefore, we can obtain the HoLDP regret achieved by HeLDP-UCB-L in Corollary 8.

Corollary 8. *When $\epsilon_t \equiv \epsilon$ and $\epsilon_{min} = \epsilon$, HeLDP-UCB-L's distribution-dependent regret is at most*

$$\sum_{a:\Delta_a>0}\Big[\frac{8(1+4/\epsilon)^2}{\Delta_a}\log T + \Big(3+\frac{2\pi^2}{3}\Big)\Delta_a\Big];$$

its distribution-free regret is $O((1+\epsilon^{-1})\sqrt{nT\log T})$.

Remark. i) Compared to non-private UCB (Auer et al., 2002), the regret of HeLDP-UCB-L is increased by a $(1+4/\epsilon)^2$ factor, which can be viewed as the cost for preserving privacy. When ϵ approaches infinity (i.e., no privacy protection), this factor approaches one, and the regret approaches that of the non-private version. ii) According to Theorem 4, the distribution-dependent regret of HeLDP-UCB-L is optimal (up to a constant factor). iii) The distribution-free regret of HeLDP-UCB-L is optimal up to a $\sqrt{\log T}$ factor according to Basu et al. (2019).

Note that if we change "for any x, x' in $\mathcal{D}$" to "for any x, x' with $|x - x'| \le 1$" in Definition 1, then CTL is ϵ-LDP on $\mathbb{R}$ (Dwork et al., 2014). Thus, by changing the terms $\sqrt{\frac{\log t^4}{2N_{a,t}}}$ to proper confidence bounds, HeLDP-UCB-L is still HoLDP or HeLDP for bandit instances without a bounded support.

3.2. Bernoulli Mechanism

In addition to the Laplace mechanism, we propose another mechanism called Convert-to-Bernoulli (CTB), which converts bounded rewards to Bernoulli responses. (Gajane et al., 2018) proposed a similar mechanism that only works for Bernoulli rewards with homogeneous privacy levels. By contrast, in this paper, we allow the rewards to be arbitrary values in $[0, 1]$ and the privacy levels to be heterogeneous. CTB is described in Curator 3.

Curator 3 Convert-to-Bernoulli (CTB)

Upon receiving reward r and privacy level ϵ:
$M_B(r;\epsilon) \sim$ Bernoulli$(\frac{re^\epsilon+1-r}{1+e^\epsilon})$;
return $\epsilon, M_B(r;\epsilon)$;

Lemma 9. *CTB is ϵ-LDP on $[0, 1]$, and the returned value follows the Bernoulli distribution with mean $\mu_{a,\epsilon} := \frac{1}{2} + (2\mu_a - 1)\cdot\frac{e^\epsilon-1}{2(e^\epsilon+1)}$.*

Unlike non-private or HoLDP MAB algorithms, for HeLDP MAB, the means of the private responses of the same arm a are no longer the same since the privacy levels ϵ of the rewards are different. Thus, we cannot directly use the empirical means of the private responses to estimate the mean rewards of the arms.

To handle this, we use the following mapping

$$g(x;\epsilon) := \begin{cases} \frac{1}{2}\left(1+\frac{e^\epsilon+1}{e^\epsilon-1}\right) & \text{if } x=1, \\ \frac{1}{2}\left(1-\frac{e^\epsilon+1}{e^\epsilon-1}\right) & \text{if } x=0, \end{cases}$$

such that $\mathbb{E}[g(M_B(R_t;\epsilon_t);\epsilon_t)] = \mathbb{E}[R_t]$. Therefore, we can use the empirical means of the g-values to estimate the mean rewards of the arms. Besides, after the mapping $g(\cdot)$, the mapped values still have bounded supports (although the supports may be different). Thus, we can use the general Hoeffding inequality to get the following concentration by setting $b = \sqrt{\frac{1}{2t^2}\sum_{r=1}^{t}(\frac{e^{\epsilon_{a,r}}+1}{e^{\epsilon_{a,r}}-1})^2 \log t^4}$ and $X_{a,r} = M_B(R_{a,r};\epsilon_{a,r})$:

$$\mathbb{P}\Big\{\mu_a \geq \sum_{r=1}^{t}\frac{g(R_{a,r};\epsilon_{a,r})}{t} + b \Big| \epsilon_{a,r} : r \in [t]\Big\} \leq t^{-4}.$$

Here, another problem arises: when some $\epsilon_{a,t}$ is close to zero, the corresponding UCB will approach infinity, making the upper confidence bound infinitely large. Similar to the algorithm for the Laplace mechanism, we set a number ϵ_{min} and drop all private responses with privacy parameter smaller than ϵ_{min}.

Similar to the Laplace mechanism, the term $\sqrt{\frac{1}{2t^2}\sum_{r=1}^{t}(\frac{e^{\epsilon}_{a,r}+1}{e^{\epsilon}_{a,r}-1})^2 \log t^4}$ added to the empirical mean is random. Thus, we need also take this additional randomness into consideration. In the analysis, we use another confidence bound to bound the value of $\sum_{r=1}^{t}(\frac{e^{\epsilon}_{a,r}+1}{e^{\epsilon}_{a,r}-1})^2$, which contributes to the leftmost term of the second line of Theorem 10.

The algorithm is described in Agent 4. Its theoretical guarantee is stated in Theorem 10. Here, we let $\beta(\epsilon) = (\frac{e^\epsilon+1}{e^\epsilon-1})^2$ and ϵ_{max} be supreme of ϵ_t's region, and define:

$$Q_B := \beta(\epsilon_{min}) - \beta(\epsilon_{max}),\ \beta_0 := \mathbb{E}[\beta(\epsilon_t) \mid \epsilon_t \geq \epsilon_{min}], \text{ and } p_0 := \mathbb{P}\{\epsilon_t \geq \epsilon_{min}\}.$$

Theorem 10. *HeLDP-UCB-B is HeLDP. Its distribution-dependent regret is at most*

$$\inf_{\kappa>0}\Big\{\sum_{a:\Delta_a>0}\Big[\frac{8}{p_0\Delta_a}\cdot(\beta_0+\kappa Q_B)\log T + \frac{\Delta_a}{2p_0\kappa^2}\log(p_0 T\exp\{2p_0\kappa^2\}) + \frac{\Delta_a}{p_0} + \frac{\pi^2\Delta_a}{3}\Big]\Big\};$$

its distribution-free regret is $O(\sqrt{\frac{\beta_0 nT\log T}{p_0}})$.

Remark: i) Since for the Bernoulli mechanism, we do not need $B_{a,t}$ to be larger than a value as the Laplace mechanism, in the regret, there is one less term compared to Theorem 7. ii) For the case where $\epsilon_{min} = \Omega(\epsilon_{max})$ and $p_0 = \Omega(1)$, by setting $\kappa = 1$, the regret upper bound of HeLDP-UCB-B reduces to $O(\log T)\sum_{a:\Delta_a>0}\frac{1}{\Delta_a}(\frac{e^{\epsilon_{max}}+1}{e^{\epsilon_{max}}-1})^2$, matching the lower bound in Theorem 4 with $\epsilon = \epsilon_{max}$ up to a constant factor. In this case, the distribution-free regret will reduce to $O(\sqrt{\beta(\epsilon_{max})nT\log T}) = O((1+\epsilon_{max}^{-1})\sqrt{nT\log T})$, matching the lower bound of Basu et al. (2019) up to a $\sqrt{\log T}$ factor.

Agent 4 HeLDP-UCB-B (HeLDP-UCB by Bernoulli)

1: Set the privacy threshold $\epsilon_{min} > 0$;
2: $N_{a,0}, s_{a,0}, B_{a,0} \leftarrow 0$, and $u_{a,0} \leftarrow \infty$ for all arms a;
3: **for** $t = 1$ to T **do**
4: $\quad a_t \leftarrow \arg\max_{a\in[n]} u_{a,t}$;
5: $\quad$ Pull arm a_t with respect to the incoming user;
6: $\quad$ Collect the private response (ϵ_t, x_t) from CTB;
7: $\quad$ **if** $\epsilon_t \geq \epsilon_{min}$ **then**
8: $\quad\quad N_{a_t,t} \leftarrow N_{a_t,t-1} + 1$;
9: $\quad\quad s_{a_t,t} \leftarrow s_{a_t,t-1} + g(x_t;\epsilon_t)$;
10: $\quad\quad B_{a_t,t} \leftarrow B_{a_t,t-1} + \beta(\epsilon_t)$;
11: $\quad\quad u_{a_t,t} \leftarrow \frac{s_{a_t,t}}{N_{a_t,t}} + \sqrt{\frac{B_{a_t,t}\log t^4}{(2N_{a,t}^2)}}$
12: $\quad$ **else**
13: $\quad\quad$ Discard the response (ϵ_t, x_t);
14: $\quad$ **end if**
15: $\quad$ **for** arm a in $[n]$ with $N_{a,t}$ not updated **do**
16: $\quad\quad N_{a,t} \leftarrow N_{a,t-1}$; $s_{a,t} \leftarrow s_{a,t-1}$;
17: $\quad\quad B_{a,t} \leftarrow B_{a,t-1}$; Update $u_{a,t}$ as Line 11;
18: $\quad$ **end for**
19: **end for**

For HoLDP MAB, we have $\epsilon_t \equiv \epsilon$, and by setting $\epsilon_{min} = \epsilon$, we have $Q_B = 0$, $\beta_0 = (\frac{e^\epsilon+1}{e^\epsilon-1})^2$, and $p_0 = 1$. By setting $\kappa = \infty$, we can reduce the $\frac{\Delta_a}{2p_0\kappa^2}\log(p_0 T\exp\{2p_0\kappa^2\})$ term to Δ_a. Therefore, we can obtain the HoLDP regret achieved by HeLDP-UCB-B in Corollary 11.

Corollary 11. *When $\epsilon_t \equiv \epsilon$ and $\epsilon_{min} = \epsilon$, HeLDP-UCB-B's distribution-dependent regret is at most*

$$\sum_{a:\Delta_a>0}\Big[\frac{8}{\Delta_a}\Big(\frac{e^\epsilon+1}{e^\epsilon-1}\Big)^2\log T + \Big(2+\frac{\pi^2}{3}\Big)\Delta_a\Big];$$

its distribution-free regret is $O((1+\epsilon^{-1})\sqrt{nT\log T})$.

Remark. i) Compared to non-private UCB algorithms (Auer et al., 2002), the regret of HeLDP-UCB-B is increased by a $(\frac{e^\epsilon+1}{e^\epsilon-1})^2$ factor, which can be viewed as the cost for preserving privacy. When ϵ approaches infinity (i.e., no privacy protection), this factor approaches one, and the regret approaches that of the non-private version. ii) According to Theorem 4, the distribution-dependent regret of HeLDP-UCB-B is optimal (in order sense) since $(\frac{e^\epsilon+1}{e^\epsilon-1})^2 = \Theta(\epsilon^{-2})$

when $\epsilon \to 0$. iii) The distribution-free regret of HeLDP-UCB-B is optimal up to a $\sqrt{\log T}$ factor according to Basu et al. (2019). iv) The ϵ-factor $(\frac{e^\epsilon+1}{e^\epsilon-1})^2$ of HeLDP-UCB-B is always smaller than $(1+\frac{4}{\epsilon})^2$, that of HeLDP-UCB-L. Later, the simulations will also indicate that HeLDP-UCB-B tends to perform better than HeLDP-UCB-L for HoLDP.

3.3. LDP with Unbounded Reward Supports

In some cases, rewards may not have bounded supports, and the mechanisms studied in the last subsections do not work. Adding Laplace(s/ϵ) noise to the rewards is not ϵ-LDP if the difference between two rewards is larger than s. For the Bernoulli mechanism, when $r < 0$ or $r > 1$, the value $(re^\epsilon + 1 - r)/(e^\epsilon + 1)$ is outside $[0, 1]$, making it ill-defined.

We find that for bandits with i.i.d. sub-Gaussian noise, under a Sigmoid mapping, the gaps between arms a and b are $\Omega(|\mu_a - \mu_b|)$. The Sigmoid function is defined as

$$s(x) := \frac{1}{1+e^{-x}} \; \forall x \in \mathbb{R}.$$

This property is stated in Lemma 12, and we can use it to handle bandits with unbounded rewards supports and still achieve optimal regret upper bounds with larger constant factors, which are stated in Corollary 13. To avoid repetitiveness, the details of this part is relegated to Supplementary Material due to space limitation.

Lemma 12. *Let $0 \le \mu \le \lambda \le 1$. For $s(r) := (1+e^{-r})^{-1}$, $X = \lambda + Z_1$ and $Y = \mu + Z_2$, where Z_1, Z_2 are i.i.d. sub-Gaussian$(0, 1)$, we have $\mathbb{E}[s(X) - s(Y)] \ge c_s(\lambda - \mu)$ for a universal constant $c_s > 0$.*

Corollary 13. *Replacing CTL by CTL-S in HeLDP-UCB-L or replacing CTB by CTB-S in HeLDP-UCB-B, we get algorithms HeLDP-UCB-LS and HeLDP-UCB-BS. HeLDP-UCB-LS and HeLDP-UCB-BS are HeLDP. The regrets of HeLDP-UCB-LS and HeLDP-UCB-BS are at most c_s^{-2} that of HeLDP-UCB-L and HeLDP-UCB-LB, respectively.*

Remark. For MAB with unbounded reward supports, by using the Sigmoid mapping, we are still able to get optimal regrets (up to constant factors). We conjecture that other logistic functions or using truncating may have similar effects, which remain open for future studies.

4. Numerical Results

In this section, we provide numerical results. Due to space limitation, the results for bandits with unbounded supports are relegated to Supplementary Material. We do not compare the algorithms in (Zheng et al., 2020) since it only considers (ϵ, δ)-LDP, which is weaker than ϵ-LDP in this paper. We do not compare the algorithm in (Chen et al., 2020) as its regret is of the same order as HeLDP-UCB-L for HoLDP MAB.

In the experiments, we have $n = 20$ arms. The optimal arm has mean reward 0.9; five arms have mean rewards 0.8; five have mean rewards 0.7; five have mean rewards 0.6; and four have mean rewards 0.5. In Figure 1 (a,b,e-h), all rewards follow Bernoulli distributions. In Figure 1 (c,d), the rewards follow different types of distributions to show that our algorithms work beyond Bernoulli arms. Specifically, arms with mean rewards 0.9 or 0.6 generate rewards from Bernoulli distributions; arms with mean rewards 0.8 generate rewards from Beta$(4, 1)$ distribution; arms with mean rewards 0.7 generate rewards from $\{0.4, 1\}$ uniformly at random; and arms with mean rewards 0.5 generate rewards from $[0, 1]$ uniformly at random. Each line in each graph is averaged over 50 independent trials.

4.1. Homogeneous Privacy Levels

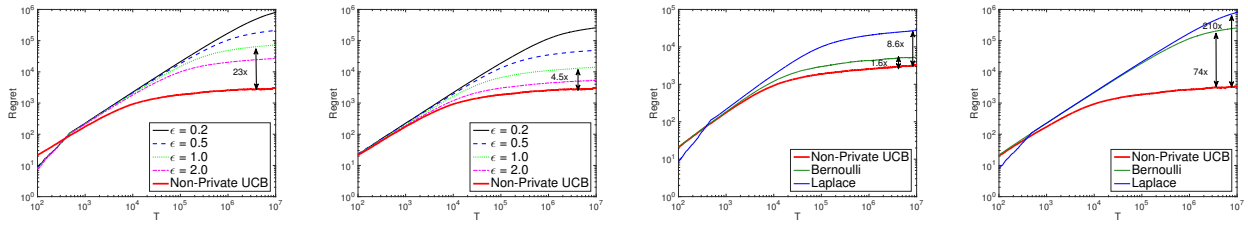
We first conduct experiments for HoLDP MAB, and numerical results are presented in Figure 1 (a-d). Specifically, we let the privacy levels ϵ_t be fixed at ϵ in the experiments and run our two algorithms.

In Figure 1 (a), we evaluate HeLDP-UCB-L with different ϵ-values. In Figure 1 (b), we do the same for HeLDP-UCB-B. According to Figure (a,b), the regrets of HeLDP-UCB-L and HeLDP-UCB-B both increase with $\frac{1}{\epsilon}$ and the convergence speed both decrease as ϵ decreases. In Figure 1 (c), we fix $\epsilon = 2.0$. We can see that the regret of HeLDP-UCB-B is slightly larger than the non-private UCB and smaller than HeLDP-UCB-L. The ratio of the regrets of HeLDP-UCB-B to non-private UCB is 1.6, and that of HeLDP-UCB-L is 8.6, which is close to theoretical values $(\frac{e^\epsilon+1}{e^\epsilon-1})^2 = 1.7$ for HeLDP-UCB-B and $(1+\frac{4}{\epsilon})^2 = 9.0$ for HeLDP-UCB-L. Thus, the numerical results in Figure 1 (c) are consistent with our theoretical predictions. In Figure 1 (d), we fix $\epsilon = 0.2$. The ratio of the regret of HeLDP-UCB-L (HeLDP-UCB-B) to non-private UCB becomes much larger, which is also consistent with the theory. The theoretical ratios are 441 and 101, respectively, which are larger and not substantially larger than the empirical results 210 and 74.

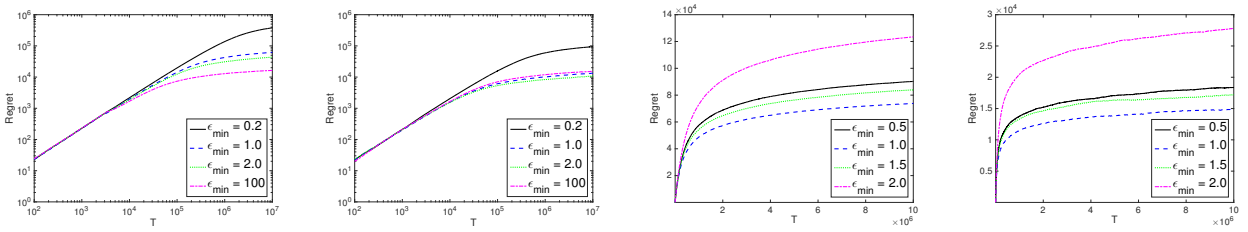
4.2. Heterogeneous Privacy Levels

We perform experiments on two privacy level distributions, where one is discrete and the other is continuous. The discrete distribution is the uniform distribution on $\{0, 0.2, 1, 2, 100\}$, and we set $\epsilon_{min} = \{0.2, 1, 2, 100\}$. The continuous distribution is Gaussian$(1, 1)$ distribution with negative values truncated to 0 and values larger than 100 truncated to 100. We set $\epsilon_{min} = \{0.5, 1.0, 1.5, 2.0\}$. In both distributions, there are approximately 20% users that do not cooperate (i.e., $\epsilon_t = 0$). The results are presented in Figure 1 (e-h).

From the results in Figure 1 (e-f), we can see that for all our



(a) HeLDP-UCB-L for HoLDP MAB, Vary ϵ. (b) HeLDP-UCB-B for HoLDP MAB, Vary ϵ. (c) Mixed Types of Arms, HoLDP, $\epsilon = 2.0$. (d) Mixed Types of Arms, HoLDP, $\epsilon = 0.2$.



(e) Discrete Distribution of ϵ_t, HeLDP-UCB-L. (f) Discrete Distribution of ϵ_t, HeLDP-UCB-B. (g) Continuous Distribution of ϵ_t, HeLDP-UCB-L. (h) Continuous Distribution of ϵ_t, HeLDP-UCB-B.

Figure 1. Numerical Results for HeLDP-UCB-L and HeLDP-UCB-B.

choices of ϵ_{min} under both ϵ_t-distributions, the regrets of both algorithms tend to converge, which indicates that these two algorithms are learning the latent distributions of the arms even if they only use part of the private responses.

Choice of ϵ_{min}. When ϵ_{min} is too large, p_0 may be too small, potentially increasing the regrets, e.g., the lines with $\epsilon_{min} = 2.0$ in Figure 1 (g,h). When ϵ_{min} is too small, then the terms α_0 and β_0 may be too large, making the regrets large, e.g., the lines with $\epsilon_{min} = 0.2$ in Figure 1 (e,f). Thus, choosing a proper value of ϵ_{min} is critical for getting small regrets. We provide an approach below to heuristically choose ϵ_{min} to get small regrets.

Define $V_L := \mathbb{E}[(1 + 4\sqrt{\alpha(\epsilon_t)})^2/p_0]$ and $V_B := \mathbb{E}[\beta(\epsilon_t)/p_0]$. Since the leading terms of the two algorithms' regrets depend on $(1+4\sqrt{\alpha(\epsilon_t)})^2/p_0$ (for HeLDP-UCB-L) or $\beta(\epsilon_t)/p_0$ (for HeLDP-UCB-B), we hypothesize that the empirical regrets of the algorithms linearly increase with V_L or V_B. We numerically evaluate how well the values of V_L and V_B can help in choosing ϵ_{min}. To show this, we linearly fit the empirical regrets and the (numerically computed) values of V_L or V_B, and summarize the corresponding R^2-value in Figure 2. We note that the remaining factor $\sum_{a:\Delta_a>0} \frac{8\log T}{\Delta_a}$ of the regrets for both algorithms under both distributions are the same. Due to space limitation, the values of the empirical regrets and V_L or V_B are relegated to Supplementary Material.

From Figure 2, we can see that our linear fitting is universal for different algorithms and ϵ_t-distributions. Moreover, the R^2-value is 0.9977, which is very close to 1 and indicates

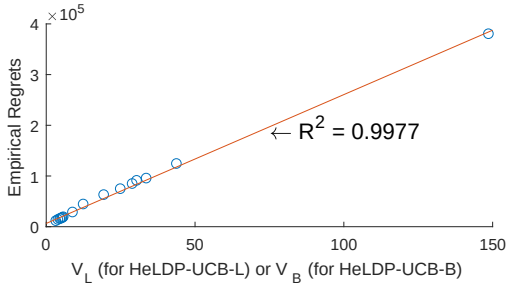


Figure 2. Linear Fitting of the Empirical Regrets and V_L (for HeLDP-UCB-L) or V_B (HeLDP-UCB-B).

that the empirical regrets of HeLDP-UCB-L (HeLDP-UCB-B) increase with V_L (V_B) almost linearly. This is consistent with our theoretical result in Theorem 7 (Theorem 10). Therefore, instead of using experiments, we may choose the best value of ϵ_{min} by computing V_L or V_B, which needs no pulling of arms and much little computation.

5. Conclusion

This paper studied the multi-armed bandit problem with local differential privacy guarantees. We proved the tight regret lower bound for HoLDP MAB and proposed HeLDP MAB algorithms with nearly optimal regrets, which can also be applied to HoLDP MAB and get order-optimal regrets. Numerical results also confirmed our theoretical results.

References

Agrawal, S. and Goyal, N. (2012). Analysis of Thompson sampling for the multi-armed bandit problem. In *Conference on learning theory*, pages 39–1.

Arratia, R. and Gordon, L. (1989). Tutorial on large deviations for the binomial distribution. *Bulletin of Mathematical Biology*, 51(1):125–131.

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

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

Berry, D. A. and Fristedt, B. (1985). Bandit problems: Sequential allocation of experiments (Monographs on statistics and applied probability). *London: Chapman and Hall*, 5:71–87.

Bubeck, S., Cesa-Bianchi, N., and Lugosi, G. (2013). Bandits with heavy tail. *IEEE Transactions on Information Theory*, 59(11):7711–7717.

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

Chen, X., Zheng, K., Zhou, Z., Yang, Y., Chen, W., and Wang, L. (2020). (Locally) differentially private combinatorial semi-bandits. *arXiv preprint arXiv:2006.00706*.

Cover, T. M. and Thomas, J. A. (2012). *Elements of information theory*. John Wiley & Sons.

Dragomir, S. S. and Gluščevic, V. (2001). Some inequalities for the kullback-leibler and χ2-distances in information theory and applications. *Tamsui Oxford Journal of Mathematical Sciences*, 17(2):97–111.

Duchi, J. C., Jordan, M. I., and Wainwright, M. J. (2013). Local privacy and statistical minimax rates. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pages 429–438. IEEE.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. (2016). Calibrating noise to sensitivity in private data analysis. *Journal of Privacy and Confidentiality*, 7(3):17–51.

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

Gajane, P., Urvoy, T., and Kaufmann, E. (2018). Corrupt bandits for preserving local privacy. In *Algorithmic Learning Theory*, pages 387–412.

Hannun, A., Knott, B., Sengupta, S., and van der Maaten, L. (2019). Privacy-preserving multi-party contextual bandits. *arXiv preprint arXiv:1910.05299*.

Huang, Z., Mitra, S., and Vaidya, N. (2015). Differentially private distributed optimization. In *International Conference on Distributed Computing and Networking*, pages 1–10.

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

Lattimore, T. and Szepesvári, C. (2018). Bandit algorithms.

Mishra, N. and Thakurta, A. (2014). Private stochastic multi-arm bandits: From theory to practice. In *ICML Workshop on Learning, Security, and Privacy*.

Mishra, N. and Thakurta, A. (2015). (Nearly) optimal differentially private stochastic multi-arm bandits. In *Conference on Uncertainty in Artificial Intelligence*, pages 592–601.

Mohammed, N., Chen, R., Fung, B., and Yu, P. S. (2011). Differentially private data release for data mining. In *ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 493–501. ACM.

Sajed, T. (2019). Optimal differentially private finite armed stochastic bandit.

Shariff, R. and Sheffet, O. (2018). Differentially private contextual linear bandits. In *Advances in Neural Information Processing Systems*, pages 4296–4306.

Tossou, A. C. and Dimitrakakis, C. (2015). Differentially private, multi-agent multi-armed bandits. In *European Workshop on Reinforcement Learning (EWRL)*.

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

Tossou, A. C. Y. and Dimitrakakis, C. (2017). Achieving privacy in the adversarial multi-armed bandit. In *AAAI Conference on Artificial Intelligence*.

Wang, B. and Hegde, N. (2019). Privacy-preserving Q-learning with functional noise in continuous spaces. In *Advances in Neural Information Processing Systems*, pages 11323–11333.

Zheng, K., Cai, T., Huang, W., Li, Z., and Wang, L. (2020). Locally differentially private (contextual) bandits learning. *arXiv preprint arXiv:2006.00701*.