Journal of Machine Learning Research 25 (2024) 1-52 Submitted 10/21; Revised 11/23; Published 9/24

Optimal Learning Policies for Differential Privacy in Multi-armed Bandits

Siwei Wang[†] SIWEIWANG@MICROSOFT.COM
Microsoft Research Asia
Beijing, China

Jun Zhu[*] DCSZJ@TSINGHUA.EDU.CN
Department of Computer Science and Technology, BNRist Center, Tsinghua AI Institute, Tsinghua-Bosch Joint ML Center, Tsinghua University, Beijing, China
Pazhou Laboratory (Huangpu), Guangzhou, China

Editor: Pradeep Ravikumar

Abstract

This paper studies the multi-armed bandit problem with a requirement of differential privacy guarantee or global differential privacy guarantee. We first prove that, the lower bound for the extra regret to protect (ϵ, δ)-global differential privacy is $\Omega(\frac{N}{\epsilon}\log\frac{(e^\epsilon-1)T+\delta T}{(e^\epsilon-1)+\delta T})$ (N is the number of arms and T is the time horizon), which is independent with T for $\delta > 0$ and large enough T. Moreover, the lower bound for the extra regret to protect (ϵ, δ)-differential privacy can be no more than the above bound. This means that, different with the case $\delta = 0$, it is possible to design algorithms that protect privacy and achieve the same asymptotical regret upper bound as the non-private algorithms when $\delta > 0$. Then we adapt the Follow the Perturbed Leader (FTPL) framework, and propose learning policies with both Gaussian and Beta perturbed distributions (DP-FTPL-Gauss and DP-FTPL-Beta) to protect (ϵ, δ)-differential privacy. The analysis shows that they achieve an $O(\frac{N\log T}{\Delta_{\min}} + N\min\{\frac{1}{\delta^2}, \frac{1}{\epsilon^2}\log\frac{1}{\delta}\})$ regret upper bound, where $\Delta_{\min}$ is the minimum expected reward gap between the optimal arm and any other ones. We also design a unique perturbed distribution to protect (ϵ, δ)-differential privacy in the FTPL framework (DP-FTPL-New), which reduces the regret upper bound to $O(\frac{N\log T}{\Delta_{\min}} + \frac{N}{\epsilon}\log\frac{(e^\epsilon-1)T+\delta T}{(e^\epsilon-1)+\delta T})$. We further show that this perturbed distribution could also be used to protect (ϵ, δ)-global differential privacy, and design a corresponding algorithm GDP-Elim-New. We show that its regret upper bound is $O(\frac{\Delta_{\max}}{\Delta_{\min}}(\frac{N\log T}{\Delta_{\min}} + \frac{N}{\epsilon}\log\frac{(e^\epsilon-1)T+\delta T}{(e^\epsilon-1)+\delta T}))$. This shows that our $\Omega(\frac{N}{\epsilon}\log\frac{(e^\epsilon-1)T+\delta T}{(e^\epsilon-1)+\delta T})$ regret lower bound is tight (e.g. when $\frac{\Delta_{\max}}{\Delta_{\min}}$ is bounded).

Keywords: Multi-armed bandits, differential privacy, follow the perturbed leader, Thompson Sampling

1. Introduction

Multi-armed bandit (MAB) (Berry and Fristedt, 1985; Sutton and Barto, 1998) is an online learning model that captures the basic tradeoff between exploration and exploitation. In an MAB instance, there are totally N arms, and the player needs to choose one of them at

[*] The work was done when Siwei Wang[†] was doing postdoc at Tsinghua University. Jun Zhu is the corresponding author.

each time step. Then the player receives a random reward, which is drawn independently from an unknown distribution corresponding to the chosen arm. The goal of the player is to maximize the cumulative reward by choosing the arms properly. To evaluate a learning policy π, we use the metric *regret*, which is defined as the expected gap between the cumulative reward of π and the largest cumulative reward of any possible policies.

In recent years, due to the increasing scale of the Internet, online learning becomes a popular framework to model Internet applications. There are many prior works on applying the MAB model in reality, e.g., recommendation websites (Qin et al., 2014; Wang and Chen, 2017; Wang and Huang, 2018), online advertising (Schwartz et al., 2017; Chakrabarti et al., 2008), social networks (Buccapatnam et al., 2013; Liu and Zhao, 2010; Chen et al., 2013), search engines (Lu et al., 2010), etc. Along with the wide adoption, security problems in the MAB model are becoming increasingly more important. For example, in online search advertisement (Mishra and Thakurta, 2015), the system displays some advertisements to user t after he/she requests for a search query. Then the user will click the advertisements that he/she is interested in, and the search engine gets a reward of one if the user clicks an advertisement. In this case, the random reward is whether the user t is interested in the displayed advertisements and therefore it raises privacy concern for all the users, since their private data is used in the online learning policy. If the learning policy does not have any privacy guarantee, e.g., after one user who searches for keyword “a” clicks an advertisement “b”, it then displays advertisement “b” to the next user who searches for keyword “a” (to maximize the cumulative reward), then one can use such feedback to infer private information of the other users. Because of this, Internet users will concern about how their private information is protected, and refuse to use the systems that do not protect private information well. Therefore, except for how to achieve good regret behavior, people are also interested in privacy in MAB algorithms (Jain et al., 2012; Mishra and Thakurta, 2015; Debabrota et al., 2019). It is necessary to design algorithms for MAB problems that not only perform well (i.e., with low regret bound), but also protect private information properly.

In this paper, we mainly focus on the differential privacy (Dwork, 2008) in MAB problems. Specifically, we study the learning policies that protect (ϵ, δ)-differential privacy (Mishra and Thakurta, 2015) and (ϵ, δ)-global differential privacy (Debabrota et al., 2019) in MAB problems. Most of the existing MAB algorithms that guarantee differential privacy or global differential privacy suffer from an extra regret of $\Omega(N \log T)$ (here N is the number of arms and T is the time horizon), which may become the majority term when T is large enough. For example, to guarantee $(\epsilon, 0)$-differential privacy, Private-UCB (Mishra and Thakurta, 2015) has a regret upper bound of $O(N(\log T)^3)$; DP-UCB (Tossou and Dimitrakakis, 2016) has a regret upper bound of $O(N \log T (\log\log T)^2)$; DP-UCB-BOUND (Tossou and Dimitrakakis, 2016) has a regret upper bound of $O(N \log T \log\log T)$. To guarantee $(\epsilon, 0)$-global differential privacy, AdaP-UCB and AdaP-KLUCB (Azize and Basu, 2022) have an extra regret upper bound of $O(N \log T)$.

Then a natural question is whether we can design learning policies that guarantee differential privacy (or global differential privacy) but do not incur such large extra regrets, especially when T goes to infinity. And in this paper, we show that the answer is “yes” when $\delta > 0$, no matter we are considering differential privacy or global differential privacy.

When we are protecting differential privacy, we adopt the generic framework of Follow the Perturbed Leader (FTPL) (Kim and Tewari, 2019), which uses perturbations to choose arms and does not cause large extra regret, and design algorithms named DP-FTPL-Gauss, DP-FTPL-Beta and DP-FTPL-New. These algorithms use different kinds of perturbations, and their extra regret (to protect differential privacy) is independent of T when $\delta > 0$.

Specifically, we first adapt the idea of Thompson Sampling (TS) (Thompson, 1933; Agrawal and Goyal, 2013), the most famous learning policy under the generic FTPL framework, where the player uses observations to update the corresponding perturbed distributions via Bayes' rule. Compared with other policies under the FTPL framework, the perturbation in the vanilla TS often leads to a smaller cumulative regret in experiments. Therefore, it is worth to study how to use the perturbed distributions of TS policy in the differential privacy setting to reduce regret. Here, we consider to adapt two different kinds of perturbed distributions from TS, i.e., Beta distribution and Gaussian distribution, in our DP-FTPL framework. We show that they both guarantee (ϵ, δ)-differential privacy with $\delta > 0$, and their regrets are upper bounded by $O(\sum_i \frac{\log T}{\Delta_i} + \min\{\frac{\log \frac{1}{\delta}}{\epsilon^2}, \frac{1}{\delta^2}\})$, where Δ_i represents the expected reward gap between the best arm and arm i.

We then further explore the flexibility of FTPL and design perturbations that guarantee differential privacy with minimum extra regret. We show that using a specific perturbation (i.e., DP-FTPL-New) can guarantee (ϵ, δ)-differential privacy, and achieve a regret upper bound of $O(\sum_i(\frac{\log T}{\Delta_i} + \frac{1}{\epsilon}\log\frac{(e^\epsilon-1)T+\delta T}{(e^\epsilon-1)+\delta T}))$. We can see that for $\delta > 0$ and large enough T, the extra regret term is still independent with T. In fact, compared with Gaussian or Beta distribution, the new perturbation can significantly reduce the extra regret, and behave better when there is a high standard on the guarantee of differential privacy. On the other hand, different from DP-FTPL-Gauss and DP-FTPL-Beta, the perturbation in DP-FTPL-New could also deal with the case that $\delta = 0$. In this case, it has a regret upper bound $O(\sum_i(\frac{\log T}{\Delta_i} + \frac{1}{\epsilon}\log T))$, which matches with the best policy DP-SE before (Sajed and Sheffet, 2019). Because of the fact that DP-SE can only work in the case that $\delta = 0$, our DP-FTPL-New algorithm could be much more general.

As for the case of protecting global differential privacy, we merge the elimination framework (Maron and Moore, 1997; Evendar et al., 2006; Kaufmann and Kalyanakrishnan, 2013) and the perturbed distribution we used in DP-FTPL-New, and design an algorithm GDP-Elim-New to protect (ϵ, δ)-global differential privacy. We show that the regret upper bound of our algorithm is $O(\sum_i(\frac{\Delta_i \log T}{\Delta_{\min}^2} + \frac{\Delta_i}{\epsilon\Delta_{\min}}\log\frac{(e^\epsilon-1)T+\delta T}{(e^\epsilon-1)+\delta T}))$, and the extra regret term is also independent of T when $\delta > 0$. On the other hand, we also prove a matching regret lower bound, i.e., any algorithms that protect (ϵ, δ)-global differential privacy must suffer from regret of at least $\Omega(\frac{N}{\epsilon}\log\frac{(e^\epsilon-1)T+\delta T}{(e^\epsilon-1)+\delta T})$. This indicates that our regret lower bound is tight (e.g., when $\Delta_{\max}/\Delta_{\min}$ is bounded).

2. Related Work

Thompson Sampling (TS) is first proposed by Thompson (1933). It follows the Bayesian framework and fits the analysis in the Bayesian setting (i.e., the parameters of the game follow a known prior distribution) naturally, and people propose its sub-linear regret upper bound under the Bayesian setting as $O(\sqrt{NT\log T})$ (Russo and Van Roy, 2014). However, the analysis for using TS in the frequentist setting (i.e., the parameters of the game are fixed but unknown) is much more different. Agrawal and Goyal (2012) obtain the first sub-linear regret upper bound as $O(N^2 \log T)$. After that, Agrawal and Goyal (2013) and Kaufmann et al. (2012) obtain an $O(N \log T)$ regret upper bound, which is asymptotically optimal.

Follow the Perturbed Leader (FTPL) framework (Kim and Tewari, 2019) is a generation of Thompson Sampling policy. In each time slot, it draws random parameter samples from a series of perturbed distributions (which depend on the history information), and then chooses the arm with largest random parameter sample. The main difference from TS is that in FTPL the perturbed distributions are not updated by the Bayes' rule. Instead, FTPL chooses those perturbed distributions properly to fit the analysis in the frequentist setting, leading to a sub-linear regret upper bound. For example, Kim and Tewari (2019) show that using a uniform perturbed distribution (i.e., a uniform distribution between lower confidence bound and upper confidence bound) also leads to $O(N \log T)$ regret upper bound.

Mishra and Thakurta (2015) present a first attempt on the MAB problem with differential privacy. The authors derive a formal definition for the differential privacy in MAB algorithms, and propose a UCB-based algorithm to guarantee both $(\epsilon, 0)$-differential privacy and sub-linear regret upper bound of $O(\frac{N(\log T)^3}{\epsilon\Delta_{\min}})$. Similarly, Tossou and Dimitrakakis (2016) also provide UCB-based algorithms, such as DP-UCB and DP-UCB-BOUND, to guarantee $(\epsilon, 0)$-differential privacy. All these algorithms incur an extra regret of at least $O(\frac{N \log T \log\log T}{\epsilon})$, which is worse than the $O(\frac{N \log T}{\epsilon})$ extra regret term in our DP-FTPL-New policy. The DP-UCB-INT algorithm (Tossou and Dimitrakakis, 2016) achieves a T-independent additive term in its regret upper bound. However, its extra regret term is larger than ours when ϵ is smaller than δ, and it behaves much worse than our algorithms in experiments (see details in Section 8). Tossou and Dimitrakakis (2018) consider using a TS policy (with Gaussian prior) to protect (ϵ, δ)-differential privacy with $\delta > 0$, but their algorithm only works when $\delta = T^{-4}$, and has a much larger extra regret term than ours. Recently, Shariff and Sheffet (2018) show that the regret lower bound for the case $\delta = 0$ is $\Omega(\sum_i(\frac{\log T}{\Delta_i} + \frac{\log T}{\epsilon}))$. Based on their results, Sajed and Sheffet (2019) then design an optimal learning policy DP-SE for the case $\delta = 0$. Compared to the elimination-based algorithm DP-SE, our FTPL-based algorithm has a better behaviour in experiments (see details in Section 8), since the constant factor in the regret upper bounds of FTPL-based algorithms is usually smaller than the elimination-based algorithms.

Debabrota et al. (2019) first propose the definition of (ϵ, δ)-global differential privacy, and prove problem-independent regret lower bound for the case that $\delta = 0$ in both the Bayesian setting and the non-Bayesian setting. Following their works, Azize and Basu (2022) design AdaP-UCB and AdaP-KLUCB that protects $(\epsilon, 0)$-global differential privacy, and gives their corresponding problem-dependent regret upper bounds. Besides, they also propose problem-dependent and problem-independent regret lower bounds for the algorithms that protect $(\epsilon, 0)$-global differential privacy, and the regret lower bounds and regret upper bounds match with each other (in order). Compared to these results, our regret upper/lower bounds work in not only the case that $\delta = 0$, but also the case that $\delta > 0$, i.e., our analysis and algorithms are more general.

Chen et al. (2020), Zheng et al. (2020) and Gajane et al. (2018) go one step further and consider to protect local differential privacy (Duchi et al., 2014) in MAB problems. To guarantee local differential privacy, the learning policy needs to encrypt each user's data before collection, so that the attacker cannot deduce other users' data anyway. Therefore, a natural solution is to add noise on all the observations, which leads to an $O(\frac{N \log T}{\Delta_{\min}\epsilon^2})$ regret upper bound. Ren et al. (2020) show that this is indeed optimal, i.e., any algorithm that protects local differential privacy suffers from at least $\Omega(\frac{N \log T}{\Delta_{\min}\epsilon^2})$ regret. Except for regret minimization problems, the local differential privacy is also considered in other kinds of online learning models, e.g., the pure exploration problems (Féraud et al., 2019).

3. Model Setting

In this section, we present the basic setup of multi-armed bandits, as well as two definitions of differential privacy in multi-armed bandits, which are widely adopted in existing literature (Mishra and Thakurta, 2015; Tossou and Dimitrakakis, 2016; Sajed and Sheffet, 2019; Debabrota et al., 2019; Azize and Basu, 2022). These two differential privacy definitions can be applied in many real-world scenarios to protect users' privacy, such as search engines and advertising websites.

3.1 Multi-armed Bandit Problems

A multi-armed bandit instance is a tuple $\{A, \boldsymbol{D}, T\}$, where $A = \{1, 2, \cdots, N\}$ is the set of arms, $\boldsymbol{D} = \{D_1, \cdots, D_N\}$ are the corresponding distributions of the arms, and T is the number of time steps. At each time step t, the player can choose an arm $i(t) \in A$ to pull, and then observe a reward $x(t)$, which is drawn independently from $D_{i(t)}$, i.e., $x(t) = x_{i(t)}(t) \sim D_{i(t)}$. As assumed in many prior works (Mishra and Thakurta, 2015; Tossou and Dimitrakakis, 2016; Sajed and Sheffet, 2019), in this paper, we assume that D_i's are supported on $[0, 1]$. Let $\mu_i \triangleq \mathbb{E}_{x \sim D_i}[x]$ be the expected reward of distribution D_i, and assume that $\mu_1 > \mu_2 \geq \cdots \geq \mu_N$. Then we denote $\Delta_i \triangleq \mu_1 - \mu_i$, $\Delta_{\min} \triangleq \min_{i \geq 2} \Delta_i$. Also let $\mathcal{F}_{t-1} = \{(i(\tau), x(\tau))\}_{\tau=1}^{t-1}$ be the history of the game, then the player chooses the next arm $i(t) \sim \pi(\mathcal{F}_{t-1})$, where $\pi(\mathcal{F}_{t-1})$ denotes the probability distribution of the chosen arm by policy π given history $\mathcal{F}_{t-1}$. We use "regret" to evaluate the player's behaviour, which is defined as:

$$Reg(T) \triangleq T\mu_1 - \mathbb{E}\left[\sum_{t=1}^{T} \mu_{i(t)}\right].$$

The goal of the player is to design learning policies with as small regret as possible.

In this paper, we mainly consider the case that the distributions D_i's are all Bernoulli. If some of them are not, we can use the trick stated in the work of Agrawal and Goyal (2013), i.e., for any reward $x_i(t) \in [0, 1]$, we draw a random observation $y_i(t)$ independently from a Bernoulli distribution with parameter $x_i(t)$ and then use $y_i(t)$ as the observation in the algorithms. It is easy to check that $y_i(t)$'s are independent Bernoulli random variables with mean μ_i.

3.2 Differential Privacy in Multi-armed Bandits

In a multi-armed bandit instance, the definition of differential privacy is given as follows (Mishra and Thakurta, 2015):

Definition 1 *For learning policy π, it guarantees (ϵ, δ)-differential privacy if for any $t > 1$, $1 \leq \tau < t$, $x'(\tau) \in [0,1]$ and $A' \subseteq A$, we have that*

$$\Pr[i(t) \in A' | \mathcal{F}_{t-1}] \leq e^{\epsilon} \Pr[i(t) \in A' | \mathcal{F}'_{t-1}] + \delta, \tag{1}$$

where $\mathcal{F}'_{t-1}$ denotes the history of the game if we substitute one pair of arm-observation $(i(\tau), x(\tau))$ by $(i(\tau), x'(\tau))$.

Eq. (2) shows that when we arbitrarily change an observation $x(\tau)$ on arm i to be $x'(\tau)$, the probability distribution of the chosen arm $i(t)$ for any $t > \tau$ does not vary much. This means that an attacker cannot deduce other users' private information by collecting the data of single arm selections. Therefore, it can be applied in search engines or advertising websites to protect privacy (Mishra and Thakurta, 2015; Tossou and Dimitrakakis, 2016).

3.3 Global Differential Privacy in Multi-armed Bandits

For any bandit algorithm π, any action vector $\boldsymbol{a} \in A^T$ and reward vector $\boldsymbol{r} \in [0,1]^T$, we define $\pi(\boldsymbol{a}|\boldsymbol{r})$ as

$$\pi(\boldsymbol{a}|\boldsymbol{r}) \triangleq \prod_{t=1}^{T} \Pr_{a_t \sim \pi}[i(t) = a_t | (a_1, r_1), \cdots, (a_{t-1}, r_{t-1})].$$

That is, $\pi(\boldsymbol{a}|\boldsymbol{r})$ is the probability of π pulling arms according to $\boldsymbol{a}$, given the fact that its received observation vector is $\boldsymbol{r}$.

Then we state our definition of (ϵ, δ)-global differential privacy. To make it clear, we first define a data-revision rule as follows:

Definition 2 *A data-revision rule $\mathcal{R}$ on time horizon T is a set of TN conditional probabilities $p_t^{\mathcal{R}}(1|a)$ (and $p_t^{\mathcal{R}}(0|a)$) for all $t \in [T]$ and $a \in A$. When considering a revision of data vector $\boldsymbol{r} = [r_1, \cdots, r_T]$ corresponds to action vector $\boldsymbol{a} = [a_1, \cdots, a_T]$, for any step t, we draw a random variable $z_t \sim p_t^{\mathcal{R}}(\cdot|a_t)$, and only revise the data of r_t when $z_t = 1$.*

Definition 3 ((ϵ, δ)-global differential privacy) *For learning policy π, it guarantees (ϵ, δ)-global differential privacy if for any bandit instance I, any time horizon $T > 1$, any action vector set $S \subseteq A^T$ and any data-revision rule $\mathcal{R}$ such that for all $\boldsymbol{a} \in S$*

$$\sum_{t \in [T]} p_t^{\mathcal{R}}(1|a_t) \leq 1,$$

i.e., the expected number of revisions is less than 1, then

$$\sum_{\boldsymbol{a} \in S} \sum_{\boldsymbol{r}} P_I(\boldsymbol{r}|\boldsymbol{a}) \pi(\boldsymbol{a}|\boldsymbol{r}) \leq e^{\epsilon} \sum_{\boldsymbol{a} \in S} \sum_{\boldsymbol{r}} P_I(\boldsymbol{r}|\boldsymbol{a}) \pi(\boldsymbol{a}|\boldsymbol{r} \oplus \mathcal{R}) + \delta, \tag{2}$$

where $\boldsymbol{r} \oplus \mathcal{R}$ represents the new vector by applying data-revision rule $\mathcal{R}$ on $\boldsymbol{r}$, and $P_I(\boldsymbol{r}|\boldsymbol{a})$ is the probability of getting random reward vector $\boldsymbol{r}$ under bandit instance I and action vector $\boldsymbol{a}$.

Compared to the definition of $(\epsilon, 0)$-global differential privacy (Debabrota et al., 2019), i.e., for any $T > 1$, action vector $\boldsymbol{a}$ and adjacent reward vectors $\boldsymbol{r}, \boldsymbol{r'} \in [0,1]^T$, we have that $\pi(\boldsymbol{a}|\boldsymbol{r}) \le e^{\epsilon}\pi(\boldsymbol{a}|\boldsymbol{r'})$, we can see that there are three differences between these two definitions, and here we will explain them in detail.

The first difference is that instead of considering adjacent reward vectors $\boldsymbol{r}, \boldsymbol{r'}$, we introduce the definition of the data-revision rule, and consider such rules with an expected number of revisions at most 1. If we just let the data-revision rule $\mathcal{R}$ be: for some fixed $t \le T$, $p_t^{\mathcal{R}}(1|a_t) = 1$ for all $a_t \in A$. Then it is the same as the definition of adjacent reward vectors, and protects the privacy of the user comes in time step t. Except for this user-level privacy, our definition of the data-revision rule can also support action-level privacy. For example, if only the data of pulling some arm i is private (or more sensitive), then for a user who joins this system several times, as long as he/she chooses to pull this arm i for at most one time, his/her private information is also protected (by setting only $p_t^{\mathcal{R}}(1|i) = 1$ for this arm i and the steps t that he/she joins the system).

The second difference is that instead of any single action vector $\boldsymbol{a}$, we are considering any action vector set $S \subseteq A^T$. In fact, only if $\delta = 0$, i.e., the same as the work of Debabrota et al. (2019), considering single action vector $\boldsymbol{a}$ is the same as considering action vector set $S \subseteq A^T$. When $\delta > 0$, our definition of considering action vector set $S \subseteq A^T$ aligns with the classical differential privacy definition.

The third difference is that $P_I(\boldsymbol{r}|\boldsymbol{a})$ is involved in our definition. However, we want to emphasize that this is reasonable. Note that in bandit problems, not only the action vector $\boldsymbol{a}$ depends on the reward vector $\boldsymbol{r}$, but the reward vector $\boldsymbol{r}$ also depends on the action vector $\boldsymbol{a}$ (as a comparison, in classical definition of differential privacy in data sets, the data set is fixed, and the outcome depends on the data set). Hence, if we observe some reward vector $\boldsymbol{r}$, the posterior distribution of observing action vector $\boldsymbol{a}$ is not proportional to $\pi(\boldsymbol{a}|\boldsymbol{r})$, but proportional to $P_I(\boldsymbol{r}|\boldsymbol{a})\pi(\boldsymbol{a}|\boldsymbol{r})$. This implies that the classical way to define (ϵ, δ)-global differential privacy should be

$$\frac{1}{p(\boldsymbol{r})}\sum_{\boldsymbol{a}\in S} P_I(\boldsymbol{r}|\boldsymbol{a})\pi(\boldsymbol{a}|\boldsymbol{r}) \le e^{\epsilon}\frac{1}{p(\boldsymbol{r})}\sum_{\boldsymbol{a}\in S} P_I(\boldsymbol{r}|\boldsymbol{a})\pi(\boldsymbol{a}|\boldsymbol{r}\oplus\mathcal{R}) + \delta,$$

where $p(\boldsymbol{r}) = \sum_{\boldsymbol{a}\in A^T} P_I(\boldsymbol{r}|\boldsymbol{a})\pi(\boldsymbol{a}|\boldsymbol{r})$ is the probability of observing reward vector $\boldsymbol{r}$. However, the influence of the $\frac{1}{p(\boldsymbol{r})}$ factor on the above equation makes designing such algorithms much more complicated, e.g., the additive term between $\sum_{\boldsymbol{a}\in S} P_I(\boldsymbol{r}|\boldsymbol{a})\pi(\boldsymbol{a}|\boldsymbol{r})$ and $\sum_{\boldsymbol{a}\in S} P_I(\boldsymbol{r}|\boldsymbol{a})\pi(\boldsymbol{a}|\boldsymbol{r}\oplus\mathcal{R})$ should be upper bounded by $p(\boldsymbol{r})\delta$, which could be really small when $p(\boldsymbol{r})$ is small. Because of this, in this paper, we relax it a little bit, and consider the expectation gap (takes expectation over $\boldsymbol{r}$) rather than the gap for any fixed $\boldsymbol{r}$. This leads to our definition, i.e.,

$$\sum_{\boldsymbol{r}} p(\boldsymbol{r})\cdot\left(\frac{1}{p(\boldsymbol{r})}\sum_{\boldsymbol{a}\in S} P_I(\boldsymbol{r}|\boldsymbol{a})\pi(\boldsymbol{a}|\boldsymbol{r})\right) \le \sum_{\boldsymbol{r}} p(\boldsymbol{r})\cdot\left(e^{\epsilon}\frac{1}{p(\boldsymbol{r})}\sum_{\boldsymbol{a}\in S} P_I(\boldsymbol{r}|\boldsymbol{a})\pi(\boldsymbol{a}|\boldsymbol{r}\oplus\mathcal{R}) + \delta\right),$$

which is the same as Eq. (2). The meaning of taking expectation in this equation is that if all the other rewards are drawn according to a fixed bandit instance, then the attacker cannot deduce other users' private information by collecting the data of arm selection vectors.

4. Regret Lower Bound

Now we present our theoretical results in this paper, and we start with a regret lower bound for bandit algorithms that guarantee (ϵ, δ)-global differential privacy in this section (as stated in Theorem 4). This lower bound can be used as an evaluation criteria for the regret upper bounds in this paper.

Theorem 4 *For any bandit algorithm that guarantees (ϵ, δ)-global differential privacy, and achieves no more than $c_0 T^\alpha$ regret on any bandit instance for some constant $c_0 > 0, \alpha < 1$, then for any T satisfies that*

$$T \geq \frac{3}{\epsilon} \log \frac{(e^\epsilon - 1)T^{1-\alpha} + 5\delta T^{1-\alpha}}{40c_0(e^\epsilon - 1) + 5\delta T^{1-\alpha}},$$

there exists a two-arm bandit instance such that its regret is lower bounded by

$$Reg(T) \geq \frac{1}{16\epsilon} \log \frac{(e^\epsilon - 1)T^{1-\alpha} + 5\delta T^{1-\alpha}}{40c_0(e^\epsilon - 1) + 5\delta T^{1-\alpha}}. \tag{3}$$

Proof Let's consider the following two problem instances:

- In Instance 1 (I_1), there are 2 arms, where the first arm returns Bernoulli reward with mean $\frac{1}{2}$, and the second arm returns Bernoulli reward with mean $\frac{1}{2} - \Delta$.
- In Instance 2 (I_2), there are 2 arms, where the first arm returns Bernoulli reward with mean $\frac{1}{2}$, and the second arm returns Bernoulli reward with mean $\frac{1}{2} + \Delta$.

We will use $\Pr_1[\cdot], \Pr_2[\cdot]$ and $\mathbb{E}_1[\cdot], \mathbb{E}_2[\cdot]$ to denote the probability measures and expectation under Instance 1 and Instance 2, respectively. Let $Reg_1(T), Reg_2(T)$ denote the expected regret under Instance 1 and Instance 2, i.e., $Reg_1(T) = \Delta\mathbb{E}_1[N_2(T)], Reg_2(T) = \Delta\mathbb{E}_2[N_1(T)]$, where $N_1(T), N_2(T)$ are the number of times we pull arm 1 and arm 2.

We want to prove that any algorithm that guarantees (ϵ, δ)-global differential privacy, and achieves no more than $c_0 T^\alpha$ regret on Instance I_2 (i.e., $Reg_2(T) = \Delta\mathbb{E}_2[N_1(T)] \leq c_0 T^\alpha$), will suffer regret of at least $\frac{1}{4}N_0\Delta$ in Instance I_1 (i.e., $Reg_1(T) = \Delta\mathbb{E}_1[N_2(T)] > \frac{1}{4}N_0\Delta$), where

$$N_0 \triangleq \frac{1}{\epsilon p(\Delta)} \log \frac{(e^\epsilon - 1)T^{1-\alpha} + 5\delta T^{1-\alpha}}{5c_0'(e^\epsilon - 1) + 5\delta T^{1-\alpha}}. \tag{4}$$

Here $p(\Delta) = \frac{2\Delta}{\frac{1}{2}+\Delta}$, and $c_0' = \frac{2c_0}{\Delta}$, the detailed definition of them will be explained after a few lines.

We will prove this by contradiction.

Firstly, assume that algorithm π suffers regret at most $\frac{1}{4}N_0\Delta$ in Instance I_1, i.e., $\Delta\mathbb{E}_1[N_2(T)] \leq \frac{1}{4}N_0\Delta$, or $\mathbb{E}_1[N_2(T)] \leq \frac{1}{4}N_0$.

We use $\Pr_q[\boldsymbol{r}|\boldsymbol{a}] \triangleq \prod_{t=1}^{T} \Pr_q[x(t) = r_t | i(t) = a_t]$ (for $q = 1, 2$), then we know that $\Pr_q[\boldsymbol{r}|\boldsymbol{a}]\pi(\boldsymbol{a}|\boldsymbol{r})$ is the probability of observing action sequence $\boldsymbol{a}$ and reward sequence $\boldsymbol{r}$ when we run policy π on Instance I_q (for $q = 1, 2$).

Therefore, under the assumption that π suffers regret at most $\frac{1}{4}N_0\Delta$ on Instance I_1, by Markov's inequality, we must have that

$$\Pr_1[N_2(T) > N_0] \leq \frac{\mathbb{E}_1[N_2(T)]}{N_0} \leq \frac{\frac{1}{4}N_0}{N_0} = \frac{1}{4}.$$

This implies that

$$\mathrm{Pr}_1[N_2(T) \leq N_0] = \sum_{\boldsymbol{a}:N_2(\boldsymbol{a})<N_0} \sum_{\boldsymbol{r}} \mathrm{Pr}_1[\boldsymbol{r}|\boldsymbol{a}]\pi(\boldsymbol{a}|\boldsymbol{r}) \geq 1 - \frac{1}{4} = \frac{3}{4}, \tag{5}$$

where $N_2(\boldsymbol{a})$ represents the number of 2s in arm vector $\boldsymbol{a}$.

Similarly, under the assumption that π suffers regret at most $c_0 T^\alpha$ on Instance I_2, by Markov's inequality, we also have that

$$\mathrm{Pr}_2[N_1(T) \geq T - N_0] \leq \frac{\mathbb{E}_2[N_1(T)]}{(T-N_0)} \leq \frac{c_0 T^\alpha}{(T-N_0)\Delta}.$$

This implies that

$$\mathrm{Pr}_2[N_1(T) \geq T - N_0] = \mathrm{Pr}_2[N_2(T) \leq N_0] = \sum_{\boldsymbol{a}:N_2(\boldsymbol{a})\leq N_0} \sum_{\boldsymbol{r}} \mathrm{Pr}_2[\boldsymbol{r}|\boldsymbol{a}]\pi(\boldsymbol{a}|\boldsymbol{r}) \leq \frac{c_0 T^\alpha}{(T-N_0)\Delta}.$$

Note that we assume $T \geq \frac{3}{\epsilon}\log\frac{(e^\epsilon-1)T^{1-\alpha}+5\delta T^{1-\alpha}}{40c_0(e^\epsilon-1)+5\delta T^{1-\alpha}}$, then if we choose $\Delta = \frac{1}{4}$, we know that $N_0 = \frac{3}{2\epsilon}\log\frac{(e^\epsilon-1)T^{1-\alpha}+5\delta T^{1-\alpha}}{40c_0(e^\epsilon-1)+5\delta T^{1-\alpha}} \leq \frac{T}{2}$ (recall that $p(\Delta) = \frac{2\Delta}{\frac{1}{2}+\Delta}$ and $c_0' = \frac{2c_0}{\Delta}$). Therefore

$$\sum_{\boldsymbol{a}:N_2(\boldsymbol{a})\leq N_0} \sum_{\boldsymbol{r}} \mathrm{Pr}_2[\boldsymbol{r}|\boldsymbol{a}]\pi(\boldsymbol{a}|\boldsymbol{r}) \leq c_0' T^{\alpha-1}. \tag{6}$$

Now let's consider a data-revision rule $\mathcal{R}$, given by for all t, $p_t^{\mathcal{R}}(1|a_t=2) = p(\Delta)$ and $p_t^{\mathcal{R}}(1|a_t=1) = 0$. And when we do revision, we just change r_t to 0 (i.e., if r_t is 1, we change it to 0; and if r_t is already 0, then we do nothing).

Then, it is easy to check that the rule $\mathcal{R}$ changes a reward vector that is drawn from Instance 1 to a reward vector that is drawn from Instance 2 (recall that $p(\Delta) = \frac{2\Delta}{\frac{1}{2}+\Delta}$).

Therefore, we can rewrite Eq. (5) as

$$\sum_{\boldsymbol{a}:N_2(\boldsymbol{a})\leq N_0} \sum_{\boldsymbol{r}} \mathrm{Pr}_2[\boldsymbol{r}|\boldsymbol{a}]\pi(\boldsymbol{a}|\boldsymbol{r}\oplus\mathcal{R}) \geq \frac{3}{4} > \frac{1}{5}. \tag{7}$$

Then, since π can protect (ϵ,δ)-global differential privacy, it has the guarantee that for any $S \subseteq A^T$, $\sum_{\boldsymbol{a}\in S}\sum_{\boldsymbol{r}} \mathrm{Pr}_2(\boldsymbol{r}|\boldsymbol{a})\pi(\boldsymbol{a}|\boldsymbol{r}) \leq e^\epsilon \sum_{\boldsymbol{a}\in S}\sum_{\boldsymbol{r}} \mathrm{Pr}_2(\boldsymbol{r}|\boldsymbol{a})\pi(\boldsymbol{a}|\boldsymbol{r}\oplus\mathcal{R}) + \delta$, if in expectation, data revision rule $\mathcal{R}$ changes at most 1 reward in $\boldsymbol{r}$ under all $\boldsymbol{a}\in S$.

Note that for all $\boldsymbol{a}$ such that $N_2(\boldsymbol{a}) \leq N_0$, the expected number of changes on the reward vector $\boldsymbol{r}$ is $N_0 p(\Delta)$. Hence, by geometric progression analysis, we have that

$$\begin{aligned}
&\sum_{\boldsymbol{a}:N_2(\boldsymbol{a})\leq N_0} \sum_{\boldsymbol{r}} \mathrm{Pr}_2[\boldsymbol{r}|\boldsymbol{a}]\pi(\boldsymbol{a}|\boldsymbol{r}\oplus\mathcal{R}) \\
\leq \quad & e^{\epsilon N_0 p(\Delta)}\left(\sum_{\boldsymbol{a}:N_2(\boldsymbol{a})\leq N_0} \mathrm{Pr}_2[\boldsymbol{r}|\boldsymbol{a}]\pi(\boldsymbol{a}|\boldsymbol{r}) + \frac{\delta}{e^\epsilon-1}\right) - \frac{\delta}{e^\epsilon-1}.
\end{aligned}$$

Along with Eq. (7), we have

$$e^{\epsilon N_0 p(\Delta)}\left(\sum_{\boldsymbol{a}:N_2(\boldsymbol{a})\le N_0}\sum_{\boldsymbol{r}}\Pr{}_2[\boldsymbol{r}|\boldsymbol{a}]\pi(\boldsymbol{a}|\boldsymbol{r})+\frac{\delta}{e^{\epsilon}-1}\right)>\frac{1}{5}+\frac{\delta}{e^{\epsilon}-1},$$

which is the same as

$$e^{\epsilon N_0 p(\Delta)}\frac{\delta}{e^{\epsilon}-1}+e^{\epsilon N_0 p(\Delta)}\sum_{\boldsymbol{a}:N_2(\boldsymbol{a})\le N_0}\sum_{\boldsymbol{r}}\Pr{}_2[\boldsymbol{r}|\boldsymbol{a}]\pi(\boldsymbol{a}|\boldsymbol{r})>\frac{1}{5}+\frac{\delta}{e^{\epsilon}-1}. \tag{8}$$

By Eq. (6), and the fact that $N_0=\frac{1}{\epsilon p(\Delta)}\log\frac{(e^{\epsilon}-1)T^{1-\alpha}+5\delta T^{1-\alpha}}{5c_0'(e^{\epsilon}-1)+5\delta T^{1-\alpha}}$, we also have that

$$\begin{aligned}
& e^{\epsilon N_0 p(\Delta)}\frac{\delta}{e^{\epsilon}-1}+e^{\epsilon N_0 p(\Delta)}\sum_{\boldsymbol{a}:N_2(\boldsymbol{a})\le N_0}\sum_{\boldsymbol{r}}\Pr{}_2[\boldsymbol{r}|\boldsymbol{a}]\pi(\boldsymbol{a}|\boldsymbol{r})\\
\le\ & e^{\epsilon N_0 p(\Delta)}\frac{\delta}{e^{\epsilon}-1}+e^{\epsilon N_0 p(\Delta)}c_0'T^{\alpha-1}\\
=\ & e^{\epsilon N_0 p(\Delta)}\left(\frac{\delta}{e^{\epsilon}-1}+c_0'T^{\alpha-1}\right)\\
=\ & \frac{(e^{\epsilon}-1)T^{1-\alpha}+5\delta T^{1-\alpha}}{5c_0'(e^{\epsilon}-1)+5\delta T^{1-\alpha}}\left(\frac{\delta}{e^{\epsilon}-1}+c_0'T^{\alpha-1}\right)\\
=\ & \frac{1}{5}+\frac{\delta}{e^{\epsilon}-1}.
\end{aligned}$$

This contradicts with Eq. (8).

Therefore, the assumption cannot be true, i.e., algorithm π must suffer regret at least $\frac{1}{4}N_0\Delta$ in Instance I_1.

Note that

$$\begin{aligned}
\frac{1}{4}N_0\Delta &= \frac{\Delta}{4\epsilon p(\Delta)}\log\frac{(e^{\epsilon}-1)T^{1-\alpha}+5\delta T^{1-\alpha}}{5c_0'(e^{\epsilon}-1)+5\delta T^{1-\alpha}}\\
&= \frac{\Delta}{4\epsilon\frac{2\Delta}{\frac{1}{2}+\Delta}}\log\frac{(e^{\epsilon}-1)T^{1-\alpha}+5\delta T^{1-\alpha}}{5c_0'(e^{\epsilon}-1)+5\delta T^{1-\alpha}}\\
&= \frac{\frac{1}{2}+\Delta}{8\epsilon}\log\frac{(e^{\epsilon}-1)T^{1-\alpha}+5\delta T^{1-\alpha}}{5c_0'(e^{\epsilon}-1)+5\delta T^{1-\alpha}}\\
&\ge \frac{1}{16\epsilon}\log\frac{(e^{\epsilon}-1)T^{1-\alpha}+5\delta T^{1-\alpha}}{5c_0'(e^{\epsilon}-1)+5\delta T^{1-\alpha}}.
\end{aligned}$$

By choosing $\Delta=\frac{1}{4}$, we know that $c_0'=8c_0$, and this proves the regret lower bound stated in Theorem 4 (i.e., Eq. (3)). ■

Remark 5 *For large enough T, if $\delta=0$, then this regret lower bound becomes*

$$\Omega\left(\frac{1}{\epsilon}\log\frac{(e^{\epsilon}-1)T^{1-\alpha}+5\delta T^{1-\alpha}}{40c_0(e^{\epsilon}-1)}\right)\quad=\quad\Omega\left(\frac{1}{\epsilon}\log\frac{(e^{\epsilon}-1)T+\delta T}{(e^{\epsilon}-1)}\right)$$

$$= \Omega\left(\frac{1}{\epsilon}\log\frac{(e^\epsilon-1)T+\delta T}{(e^\epsilon-1)+\delta T}\right).$$

If $\delta \neq 0$, then this regret lower bound becomes

$$\begin{aligned}\Omega\left(\frac{1}{\epsilon}\log\frac{(e^\epsilon-1)T^{1-\alpha}+5\delta T^{1-\alpha}}{5\delta T^{1-\alpha}}\right) &= \Omega\left(\frac{1}{\epsilon}\log\frac{(e^\epsilon-1)T+\delta T}{\delta T}\right)\\ &= \Omega\left(\frac{1}{\epsilon}\log\frac{(e^\epsilon-1)T+\delta T}{(e^\epsilon-1)+\delta T}\right).\end{aligned}$$

Therefore, we can also write this regret lower bound as $\Omega(\frac{1}{\epsilon}\log\frac{(e^\epsilon-1)T+\delta T}{(e^\epsilon-1)+\delta T})$.

Remark 6 *We only consider the two-arm case in Theorem 4. In fact, it is easy to use the same analysis to show that when there are N arms, the regret lower bound is*

$$\Omega\left(\frac{N}{\epsilon}\log\frac{(e^\epsilon-1)T+\delta T}{(e^\epsilon-1)+\delta T}\right).$$

Compared with Theorem 4, the analysis for the regret lower bound of algorithms that protect (ϵ,δ)-differential privacy can be much more difficult. The reason is that the influence of changing one observation in algorithms that protect (ϵ,δ)-differential privacy is much larger than changing one observation in those who protect (ϵ,δ)-global differential privacy. As a simple example, assume that the observation in time step t is changed. For algorithms that protect (ϵ,δ)-differential privacy, we know that the distribution of the pulled arm in time step $t+1$ will not change too much (by Definition 1). As for time step $t+2$, we only know that if the pulled arm in time step $t+1$ is not changed, then the distribution of the pulled arm in time step $t+2$ will not change too much. However, if the pulled arm in time step $t+1$ is changed (which appears with small but non-zero probability), then the distribution of the pulled arm in time step $t+2$ can change more. In fact, this kind of double influence on the trajectory of pulling arms makes the analysis for the lower bound much more complex. As a comparison, for algorithms that protect (ϵ,δ)-global differential privacy, when the observation in time step t is changed, we can directly bound its influence on the whole trajectory of pulling arms by Definition 3. Because of this, we only give the regret lower bound for algorithms that protect (ϵ,δ)-global differential privacy in this paper, and leave the regret lower bound for algorithms that protect (ϵ,δ)-differential privacy as an open problem.

Remark 7 *Due to the fact that global differential privacy is stronger than differential privacy, the lower bound of the global differential privacy case (i.e., $\Omega(\frac{N}{\epsilon}\log\frac{(e^\epsilon-1)T+\delta T}{(e^\epsilon-1)+\delta T})$) is an* upper bound *for the lower bound of the differential privacy case.*

5. Sufficient Conditions for Algorithms that Guarantee Differential Privacy or Global Differential Privacy

After the lower bound analysis, we begin to design algorithms that can guarantee differential privacy or global differential privacy. In this section, we first state two sufficient conditions for those algorithms, one for the differential privacy guarantee and one for the global differential privacy guarantee. With these conditions, we can adapt the existing learning policies (such that these conditions hold) to protect differential privacy or global differential privacy in bandit problems.

5.1 Differential Privacy Case

From Definition 1, we can obtain a sufficient condition for MAB algorithms that follow a specific framework to guarantee (ϵ, δ)-differential privacy.

Lemma 8 *For any MAB algorithm that satisfies: i) it draws random sample $\theta_i(t)$ from a continuous probability distribution for each arm i and then pull the arm $i(t) = \operatorname{argmax}_i \theta_i(t)$, ii) the distribution of $\theta_i(t)$ remains the same under the same history of arm i. It guarantees (ϵ, δ)-differential privacy if we have that:*

$$\begin{aligned} F(x) &\leq e^{\epsilon} F'(x) + \delta, \\ 1 - F(x) &\leq e^{\epsilon}(1 - F'(x)) + \delta, \end{aligned}$$

where $F(x)$ and $F'(x)$ are the cumulative distribution functions (CDFs) of random variable $\theta_i(t)$ with the original history $\mathcal{F}_{t-1}$ and the modified history $\mathcal{F}'_{t-1}$ (as stated in Definition 1), respectively.

Proof Let's consider the probability distribution of $i(t)$ when one observation $x(\tau)$ changes to be $x'(\tau)$. To simplify the notations, we denote $i = i(\tau)$ in this proof.

Note that when we change the observed reward $x(\tau)$ to be $x'(\tau)$, the sample distributions of all the other arms $j \neq i$ at the beginning of time step t remain the same. Thus. it is sufficient to prove that for any fixed sample values $\boldsymbol{\theta}_{-i}(t) \triangleq \{\theta_j(t) : j \neq i\}$ and any set $A' \subseteq A$, we have that $\Pr_{\theta_i(t)\sim D}[i(t) \in A' | \boldsymbol{\theta}_{-i}(t)] \leq e^{\epsilon} \Pr_{\theta_i(t)\sim D'}[i(t) \in A' | \boldsymbol{\theta}_{-i}(t)] + \delta$, where D and D' denote the probability distributions for $\theta_i(t)$, given the τ-th observation to be $x(\tau)$ and $x'(\tau)$, respectively. If this holds, then we know that for any $A' \subseteq A$,

$$\begin{aligned} \Pr_{\theta_i(t)\sim D}[i(t) \in A'] &= \sum_{\boldsymbol{\theta}_{-i}(t)} \Pr[\boldsymbol{\theta}_{-i}(t)] \Pr_{\theta_i(t)\sim D}[i(t) \in A' | \boldsymbol{\theta}_{-i}(t)] \\ &\leq \sum_{\boldsymbol{\theta}_{-i}(t)} \Pr[\boldsymbol{\theta}_{-i}(t)] \left(e^{\epsilon} \Pr_{\theta_i(t)\sim D'}[i(t) \in A' | \boldsymbol{\theta}_{-i}(t)] + \delta \right) \\ &= \sum_{\boldsymbol{\theta}_{-i}(t)} \left(\Pr[\boldsymbol{\theta}_{-i}(t)] e^{\epsilon} \Pr_{\theta_i(t)\sim D'}[i(t) \in A' | \boldsymbol{\theta}_{-i}(t)] \right) + \delta \\ &= e^{\epsilon} \Pr_{\theta_i(t)\sim D'}[i(t) \in A'] + \delta. \end{aligned}$$

For fixed $\boldsymbol{\theta}_{-i}(t)$, denote $\theta_{\max,-i}(t) = \max_{j\neq i} \theta_j(t)$. Then there are only two possible events on the chosen arm $i(t)$ when $\boldsymbol{\theta}_{-i}(t)$ is fixed, i.e., either $\theta_i(t) > \theta_{\max,-i}(t)$ and we choose $i(t) = i$ or $\theta_i(t) < \theta_{\max,-i}(t)$ and we choose $i(t) = \operatorname{argmax}_{j\neq i} \theta_j(t)$. For fixed sample values $\boldsymbol{\theta}_{-i}(t)$, $\theta_{\max,-i}(t)$ is also a fixed value. Therefore it is sufficient to prove that for any fixed value x, we have $\Pr_{\theta_i(t)\sim D}[\theta_i(t) < x] \leq e^{\epsilon} \Pr_{\theta'_i(t)\sim D'}[\theta'_i(t) < x] + \delta$ (for the case that $\operatorname{argmax}_{j\neq i} \theta_j(t) \in A'$ but $i \notin A'$) and $\Pr_{\theta_i(t)\sim D}[\theta_i(t) > x] \leq e^{\epsilon} \Pr_{\theta'_i(t)\sim D'}[\theta'_i(t) > x] + \delta$ (for the case that $i \in A'$ but $\operatorname{argmax}_{j\neq i} \theta_j(t) \notin A'$), which means that we only need $F(x) \leq e^{\epsilon} F'(x) + \delta$ and $1 - F(x) \leq e^{\epsilon}(1 - F'(x)) + \delta$ to make sure that the algorithm guarantees (ϵ, δ)-differential privacy. ■

In our proposed policies, $\theta_i(t)$ is drawn independently from the corresponding perturbed distribution (which is always a continuous probability distribution) of arm i. Therefore, Lemma 8 can be used as a sufficient condition of differential privacy guarantee, i.e., if the perturbed distribution in our framework satisfies the inequalities in Lemma 8, then the algorithm must guarantee (ϵ, δ)-differential privacy.

5.2 Global Differential Privacy Case

To protect global differential privacy, we want the influence of every observation to be limited, i.e., instead of using the observation in time step t to compute all the empirical means after t (as the algorithm framework in Lemma 8), we choose to use the observation to compute an empirical mean for only *one* time. Motivated by this idea, we are going to apply the elimination framework, and obtain a sufficient condition for MAB algorithms that follow this specific framework to guarantee (ϵ, δ)-global differential privacy from Definition 3.

Lemma 9 *For any MAB algorithm that satisfies: i) it divides the learning procedure into several phases, and in each phase, it pulls each arm for $N(k)$ times; ii) at the end of each phase k, it draws random sample $\theta_i(k)$ for all the arms from a continuous distribution, and turn to exploitation if there exists some arm i such that $\theta_i(k) \geq \max_{j \neq i} \theta_j(k) + \Delta(k)$ for some constant $\Delta(k)$ (i.e., after this phase, we choose arm i forever); and iii) the distribution of $\theta_i(k)$ remains the same under the same history of arm i in phase k. It guarantees (ϵ, δ)-global differential privacy if for any $-\infty \leq x \leq y \leq +\infty$, we have that:*

$$\begin{aligned} F(y) - F(x) &\leq e^{\epsilon}(F'(y) - F'(x)) + \frac{\delta}{2} \\ 1 - F(y) + F(x) &\leq e^{\epsilon}(1 - F'(y) + F'(x)) + \frac{\delta}{2}, \end{aligned}$$

where $F(x)$ and $F'(x)$ are the CDFs of random variable $\theta_i(k)$ with two adjunct reward vectors of arm i in phase k, respectively.

Instead of eliminating the arms that are sub-optimal with high probability after each phase (like normal elimination-based algorithms), here we choose to eliminate all the sub-optimal arms together after one phase. The reason is that under this kind of elimination rule, if some step t is not an exploitation step, then the pulled arm at this step is fixed. This makes sure that the effect of any data-revision rule $\mathcal{R}$ on any non-exploitation step t is the same (regardless of the action vector $\boldsymbol{a}$), which plays a very important role in the following proof.

Proof Let's consider an arbitrary set of feasible action vectors S, and it is easy to see that for any element in S, we can use a phase-arm pair (k, i) to represent it, i.e., it turns to exploitation after phase k, and the chosen arm becomes arm i.

Let $k_{\max}$ be the maximum k in those (k, i) pairs in S. Then, for a data-revision rule $\mathcal{R}$, we know that: i) those t after phase $k_{\max}$ with $p_t^{\mathcal{R}}(1|a_t) \geq 0$ does not change the probability $\sum_{\boldsymbol{a} \in S} \sum_{\boldsymbol{r}} P_I(\boldsymbol{r}|\boldsymbol{a})\pi(\boldsymbol{a}|\boldsymbol{r})$ at all. Hence we do not need to consider the case that we allocate data-revision on those time steps; and ii) since in Definition 3, it is required that for all $\boldsymbol{a} \in S$, $\sum_{t \in [T]} p_t^{\mathcal{R}}(1|a_t) \leq 1$, and the pulled arm on any non-exploitation step t is the same, we must have that the budget of data-revision in the first $k_{\max}$ phases is at most 1, otherwise for those $\boldsymbol{a}$ that starts to exploit after phase $k_{\max}$, the expected number of data-revision is higher than 1.

Under the above two conditions, it suffices to show that when the data-revision rule $\mathcal{R}$ is to change the reward in only *one* step in the first $k_{\max}$ phases, $\sum_{\boldsymbol{a} \in S} \sum_{\boldsymbol{r}} P_I(\boldsymbol{r}|\boldsymbol{a}) \pi(\boldsymbol{a}|\boldsymbol{r}) \leq e^{\epsilon} \sum_{\boldsymbol{a} \in S} \sum_{\boldsymbol{r}} P_I(\boldsymbol{r}|\boldsymbol{a}) \pi(\boldsymbol{a}|\boldsymbol{r} \oplus \mathcal{R}) + \delta$.

Now let (here we write $P_I(\boldsymbol{r})$ instead of $P_I(\boldsymbol{r}|\boldsymbol{a})$ since there is only one feasible $\boldsymbol{a}$ in phase k, if we do not start to exploit after phase $k-1$ under this kind of algorithms)

$$\begin{aligned} q_{k,in} &= \sum_{r \in \{0,1\}^{N \cdot N(k)}} P_I(\boldsymbol{r}) \Pr[\text{Start exploiting } i \text{ after phase } k,\ (k,i) \text{ is in } S|\boldsymbol{r}] \\ q_{k,cont} &= \sum_{r \in \{0,1\}^{N \cdot N(k)}} P_I(\boldsymbol{r}) \Pr[\text{Do not start exploiting after phase } k|\boldsymbol{r}] \end{aligned}$$

Then it is easy to check that

$$\sum_{\boldsymbol{a} \in S} \sum_{\boldsymbol{r}} P_I(\boldsymbol{r}|\boldsymbol{a}) \pi(\boldsymbol{a}|\boldsymbol{r}) = \sum_k q_{k,in} \prod_{\ell=1}^{k-1} q_{\ell,cont}$$

Now consider our data-revision rule $\mathcal{R}$ is to change a reward in phase k^*. In this case, we can rewrite the above equation as

$$\sum_{\boldsymbol{a} \in S} \sum_{\boldsymbol{r}} P_I(\boldsymbol{r}|\boldsymbol{a}) \pi(\boldsymbol{a}|\boldsymbol{r}) = q_{k^*,in} p_{k^*,in} + q_{k^*,cont} p_{k^*,cont} + p_{k^*,others},$$

with

$$\begin{aligned} p_{k^*,in} &= \prod_{\ell=1}^{k^*-1} q_{\ell,cont}, \\ p_{k^*,cont} &= \sum_{k > k^*} q_{k,in} \prod_{\ell < k, \ell \neq k^*} q_{\ell,cont}, \\ p_{k^*,others} &= \sum_{k < k^*} q_{k,in} \prod_{\ell=1}^{k-1} q_{\ell,cont}. \end{aligned}$$

It is also easy to check that $p_{k^*,in}, p_{k^*,cont}, p_{k^*,others}$ are in $[0,1]$.

On the other hand, after applying the data-revision rule $\mathcal{R}$, we have another

$$\begin{aligned} q'_{k^*,in} &= \sum_{r \in \{0,1\}^{N \cdot N(k^*)}} P_I(\boldsymbol{r}) \Pr[\text{Start exploiting } i \text{ after phase } k^*,\ (k^*,i) \text{ is in } S|\boldsymbol{r} \oplus \mathcal{R}] \\ q'_{k^*,cont} &= \sum_{r \in \{0,1\}^{N \cdot N(k^*)}} P_I(\boldsymbol{r}) \Pr[\text{Do not start exploiting after phase } k^*|\boldsymbol{r} \oplus \mathcal{R}] \end{aligned}$$

for this phase k^*.

In this case,

$$\sum_{\boldsymbol{a}\in S}\sum_{\boldsymbol{r}} P_I(\boldsymbol{r}|\boldsymbol{a})\pi(\boldsymbol{a}|\boldsymbol{r}\oplus\mathcal{R}) = q'_{k^*,in}p_{k^*,in} + q'_{k^*,cont}p_{k^*,cont} + p_{k^*,others},$$

Let the arm with the changed reward be arm i. Then if we consider all the possible $\theta_i(k^*)$'s, there are three cases:

- If $\theta_i(k^*)$ is higher than $\max_{j\neq i}\theta_j(k^*)+\Delta$, we start exploitation after phase k^*, and the chosen arm is i;
- If $\theta_i(k^*)$ is lower than $\max_{j\neq i}\theta_j(k^*)+\Delta$, but higher than $\max_{j\neq i}\theta_j(k^*)-\Delta$, then we do not start exploitation after phase k^*;
- If $\theta_i(k^*)$ is lower than $\max_{j\neq i}\theta_j(k^*)-\Delta$, then either we start exploitation after phase k^*, and the chosen arm is $\arg\max_{j\neq i}\theta_j(k^*)$, or we do not start exploitation after phase k^*, depending on the other $\{\theta_j(k^*)\}_{j\neq i}$'s.

Similar to the proof of Lemma 8, it is easy to show that after if $\theta_i(k)$ satisfies the constraint in Lemma 9, we must have that $q_{k^*,cont}\le e^\epsilon q'_{k^*,cont}+\frac{\delta}{2}$ and $q_{k^*,in}\le e^\epsilon q'_{k^*,in}+\frac{\delta}{2}$. For example, if we start exploitation after phase k^* when $\theta_i(k^*)$ is lower than $\max_{j\neq i}\theta_j(k^*)-\Delta$, and both (k^*,i) and $(k^*,\arg\max_{j\neq i}\theta_j(k^*))$ are in S, then letting $x=\max_{j\neq i}\theta_j(k^*)-\Delta(k)$ and $y=\max_{j\neq i}\theta_j(k^*)+\Delta(k)$, we can get $q_{k^*,in}\le e^\epsilon q'_{k^*,in}+\frac{\delta}{2}$ by

$$1-F(y)+F(x)\le e^\epsilon(1-F'(y)+F'(x))+\frac{\delta}{2};$$

and we can get $q_{k^*,cont}\le e^\epsilon q'_{k^*,cont}+\frac{\delta}{2}$ by

$$F(y)-F(x)\le e^\epsilon(F'(y)-F'(x))+\frac{\delta}{2}.$$

As for the other cases, the analysis is almost the same.

Hence, we must have that

$$\begin{aligned}\sum_{\boldsymbol{a}\in S}\sum_{\boldsymbol{r}} P_I(\boldsymbol{r}|\boldsymbol{a})\pi(\boldsymbol{a}|\boldsymbol{r}) &= q_{k^*,in}p_{k^*,in}+q_{k^*,cont}p_{k^*,cont}+p_{k^*,others}\\ &\le \left(e^\epsilon q'_{k^*,in}+\frac{\delta}{2}\right)p_{k^*,in}+\left(e^\epsilon q'_{k^*,cont}+\frac{\delta}{2}\right)p_{k^*,cont}+p_{k^*,others}\\ &\le e^\epsilon q'_{k^*,in}p_{k^*,in}+e^\epsilon q'_{k^*,cont}p_{k^*,cont}+e^\epsilon p_{k^*,others}+\delta\\ &= e^\epsilon\sum_{\boldsymbol{a}\in S}\sum_{\boldsymbol{r}} P_I(\boldsymbol{r}|\boldsymbol{a})\pi(\boldsymbol{a}|\boldsymbol{r}\oplus\mathcal{R})+\delta\end{aligned}$$

This means this kind of algorithms can protect (ϵ,δ)-global differential privacy. ■

Similar to the differential privacy case, Lemma 9 can also be used as a sufficient condition of global differential privacy guarantee in an elimination framework. Specifically, in each phase k, we pull all the arms for a fixed number of time steps, and then draw random sample $\theta_i(k)$ independently from the corresponding perturbed distribution (which is always a continuous probability distribution) of arm i.

Algorithm 1 DP-FTPL-Gauss

```
1: Input: $\epsilon, \delta$, for each arm $i$, set $N_i = 0, R_i = 0$.
2: for $i = 1, 2, \cdots, N$ do
3:     while $N_i < N_G^* \triangleq \min\{\frac{1}{4\pi\delta^2}, \frac{1}{\epsilon^2}\log(\frac{e}{4\pi\delta^2})\}$ do
4:         Pull arm $i(t) = i$ and observe reward $x(t) = x_i(t) \in \{0, 1\}$.
5:         $R_i \leftarrow R_i + x_i(t)$, $N_i \leftarrow N_i + 1$.
6:     end while
7: end for
8: while true do
9:     For each arm $i$, draw sample $\theta_i(t)$ independently from Gaussian perturbed distribution
       $\mathcal{N}(\frac{R_i}{N_i}, \frac{2}{N_i})$.
10:    Choose arm $i(t) = \mathrm{argmax}_i\, \theta_i(t)$, and observe reward $x(t) = x_{i(t)}(t) \in \{0, 1\}$.
11:    $R_{i(t)} \leftarrow R_{i(t)} + x_{i(t)}(t)$, $N_{i(t)} \leftarrow N_{i(t)} + 1$.
12: end while
```

Remark 10 *Note that the condition on CDF in Lemma 8 is a special case of Lemma 9, i.e., either $x = -\infty$ or $y = \infty$ in Lemma 9. This is because that when the other random variables $\boldsymbol{\theta}_{-i}(t)$ (or $\boldsymbol{\theta}_{-i}(k)$) are fixed, the feasible region of $\theta_i(t)$ to pull any specific arm set (in Lemma 8) is either $(-\infty, y)$ or $(x, +\infty)$, while the feasible region of $\theta_i(k)$ to exploit on any specific arm set or continue to do explorations (in Lemma 9) could be (x, y) or $(-\infty, x) \cup (y, \infty)$ for some $-\infty \le x \le y \le +\infty$.*

6. Algorithms that Protect Differential Privacy in Multi-armed Bandits

We now introduce our learning algorithms to protect differential privacy in MAB problems, following the idea of Lemma 8. Inspired by TS policy (Thompson, 1933; Agrawal and Goyal, 2013), we first consider to use the common Gaussian distribution and Beta distribution as the perturbed distribution in Section 6.1 and Section 6.2. Then we design perturbations that fit the differential privacy setting better in Section 6.3.

6.1 DP-FTPL-Gauss

The DP-FTPL-Gauss algorithm is described in Algorithm 1. In the *start phase* (lines 2-7), the algorithm chooses each arm for $N_G^* \triangleq \min\{\frac{1}{4\pi\delta^2}, \frac{1}{\epsilon^2}\log(\frac{e}{4\pi\delta^2})\}$ times. After that, the algorithm starts to follow the standard TS (or FTPL) procedure, i.e., at each time step t, the algorithm first uses the perturbed distribution $\mathcal{N}(\frac{R_i(t)}{N_i(t)}, \frac{2}{N_i(t)})$ to draw random sample $\theta_i(t)$ for arm i, and then chooses the arm with the largest $\theta_i(t)$ to pull. Here $N_i(t)$ denotes the number of pulls on arm i until time t, and $R_i(t)$ denotes the cumulative reward on arm i until time t (in our setting, it equals to the number of times that we observe reward "1").

Compared with the classic TS (or FTPL) policy (Agrawal and Goyal, 2013; Kim and Tewari, 2019), here we add a start phase at the beginning of the game. Note that the differential privacy is naturally guaranteed in this start phase, since in this phase the next chosen arm is the same for any history $\mathcal{F}_{t-1}$. The start phase makes sure that each arm is pulled for a sufficient number of times. Then in the rest of the game, the perturbed distributions of all the arms cannot change too much if only one observation is modified, which means that Lemma 8 holds (with appropriate start phase size N_G^*). Therefore, our algorithm can guarantee differential privacy.

Theorem 11 *DP-FTPL-Gauss guarantees (ϵ, δ)-differential privacy, and its cumulative regret satisfies*

$$Reg(T) \leq \sum_{i=2}^{N} \max\left\{\frac{2\Delta_i \log T}{(\Delta_i - 2\lambda)^2}, N_G^* \Delta_i\right\} + \Theta\left(\frac{N}{\lambda^4}\right)$$

for any $\lambda < \frac{1}{2}\Delta_{\min}$.

Compared with the classic TS policy with Gaussian prior (Agrawal and Goyal, 2013; Kim and Tewari, 2019), the major difference of DP-FTPL-Gauss is a start phase with size $N \cdot N_G^*$. Therefore, it is not surprising that there is an extra regret term of $N \cdot N_G^*$. Because of this, here we only provide the proof of the privacy part in Theorem 11, and defer the proof of the regret part to Appendix A.

Proof (privacy part) Here we only prove the first inequality in Lemma 8, the second one could be proved by symmetry.

Denote i the arm with one observation changed, $N_i(t), R_i(t)$ the value of N_i, R_i at the beginning of time step t, and $R_i'(t)$ the total cumulative reward of arm i after we change one observation. Also let $\hat{\mu}_i(t) = \frac{R_i(t)}{N_i(t)}, \hat{\mu}_i'(t) = \frac{R_i'(t)}{N_i(t)}$ and $\Delta = \hat{\mu}_i'(t) - \hat{\mu}_i(t) = \frac{x_i'(\tau) - x_i(\tau)}{N_i(t)}$, then it is easy to check that $\Pr[\theta_i(t) \leq x] \leq \Pr[\theta_i'(t) \leq x]$ when $\Delta \leq 0$. Therefore we will then focus on the case that $\Delta > 0$. Let $f(x), f'(x)$ denote the probability density functions of $\theta_i(t), \theta_i'(t)$, respectively.

Then we have that

$$\begin{aligned}
\Pr[\theta_i(t) \leq x] - \Pr[\theta_i'(t) \leq x] &= \int_{-\infty}^{x} (f(y) - f'(y))dy \\
&= \int_{-\infty}^{x} f(y)dy - \int_{-\infty}^{x-\Delta} f(y)dy \\
&= \int_{x-\Delta}^{x} f(y)dy \\
&\leq \Delta\sqrt{\frac{N_i(t)}{4\pi}}.
\end{aligned}$$

Since $x_i(\tau) - x_i'(\tau) \leq 1$, we must have that $\Delta \leq \frac{1}{N_i(t)}$ and therefore $\Delta\sqrt{\frac{N_i(t)}{4\pi}} \leq \sqrt{\frac{1}{4\pi N_i(t)}}$. Thus, for $N_i(t) \geq \frac{1}{4\pi\delta^2}$, we must have that

$$\Pr[\theta_i(t) \leq x] \leq \Pr[\theta_i'(t) \leq x] + \delta \leq e^{\epsilon} \Pr[\theta_i'(t) \leq x] + \delta.$$

For $N_i(t) \leq \frac{1}{4\pi\delta^2}$, let's consider the value $x^* = \hat{\mu}_i(t) - \sqrt{\frac{2}{N_i(t)} \log \frac{1}{4\pi\delta^2 N_i(t)}}$, and we have that

$$\Delta \cdot f(x^*) \leq \sqrt{\frac{1}{4\pi N_i(t)}} \exp\left(-\frac{N_i(t)(x^* - \hat{\mu}_i(t))^2}{4}\right)$$

$$
\begin{aligned}
&= \sqrt{\frac{1}{4\pi N_i(t)}} \exp\left(-\frac{1}{2}\log\frac{1}{4\pi\delta^2 N_i(t)}\right) \\
&= \delta.
\end{aligned}
$$

For any $x > x^*$, we also have that

$$
\begin{aligned}
\frac{f(x)}{f'(x)} &= \exp\left(-\frac{N_i(t)(x-\hat{\mu}_i(t))^2}{4} + \frac{N_i(t)(x-\hat{\mu}_i'(t))^2}{4}\right) \\
&= \exp\left(\frac{2N_i(t)(\hat{\mu}_i(t)-x)\Delta + N_i(t)\Delta^2}{4}\right) \\
&\leq \exp\left(\sqrt{\frac{1}{2N_i(t)}\log\frac{1}{4\pi\delta^2 N_i(t)}} + \frac{1}{4N_i(t)}\right).
\end{aligned}
$$

Thus, if $N_i(t) \geq \frac{1}{\epsilon^2}\log(\frac{e}{4\pi\delta^2})$, then we have that:

$$
\frac{1}{4N_i(t)} + \sqrt{\frac{1}{2N_i(t)}\log\frac{1}{4\pi\delta^2 N_i(t)}} \leq \sqrt{\frac{1}{N_i(t)}\log\frac{e}{4\pi\delta^2 N_i(t)}} \leq \sqrt{\frac{1}{N_i(t)}\log\frac{e}{4\pi\delta^2}} \leq \epsilon.
$$

This implies that for any $x \leq x^*$, we must have that

$$
\Pr[\theta_i(t) \leq x] \leq \Pr[\theta_i'(t) \leq x] + \delta \leq e^\epsilon \Pr[\theta_i'(t) \leq x] + \delta,
$$

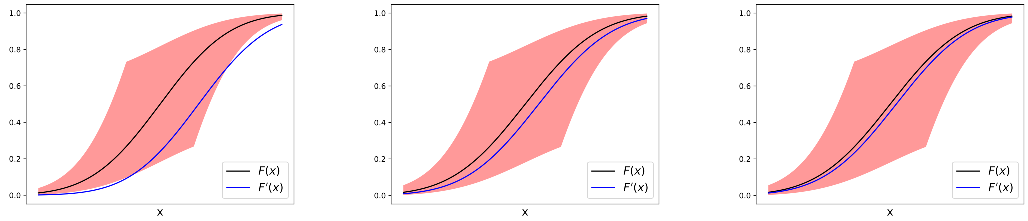
and for any $x > x^*$, we must have that

$$
\begin{aligned}
\Pr[\theta_i(t) \leq x] &= \Pr[\theta_i(t) \leq x^*] + \Pr[x^* \leq \theta_i(t) \leq x] \\
&\leq \Pr[\theta_i'(t) \leq x^*] + \delta + e^\epsilon \Pr[x^* \leq \theta_i'(t) \leq x] \\
&\leq e^\epsilon(\Pr[\theta_i'(t) \leq x^*] + \Pr[x^* \leq \theta_i'(t) \leq x]) + \delta \\
&= e^\epsilon \Pr[\theta_i'(t) \leq x] + \delta,
\end{aligned}
$$

i.e., the condition in Lemma 8 holds.

Therefore, for $N_i(t) \geq N_G^* = \min\{\frac{1}{4\pi\delta^2}, \frac{1}{\epsilon^2}\log(\frac{e}{4\pi\delta^2})\}$, i.e., after the start phase of DP-FTPL-Gauss, the (ϵ,δ)-differential privacy is always guaranteed. ■

Here we also use some figures (i.e., Fig. 1) to explain the correctness of the (ϵ,δ)-differential privacy guarantee in DP-FTPL-Gauss, as well as the necessity of the start phase. In Fig. 1, the black line is $F(x)$, the original CDF of $\theta_i(t)$, and the blue line is $F'(x)$, the CDF of $\theta_i(t)$ after one observation is changed. The red region is the feasible region of $F'(x)$ under constraints in Lemma 8 with $(\epsilon,\delta) = (1, 0.01)$, i.e., if $F'(x)$ lays in this region, then $(1, 0.01)$-differential privacy is protected. We can see that when $N_i(t) < N_G^*$ (Fig. 1(a)), $F'(x)$ is out of the feasible region, which means we cannot make sure that it has differential privacy guarantee. Hence the start phase that skips those small $N_i(t)$'s is necessary to protect differential privacy. On the other hand, when $N_i(t) = N_G^*$ (Fig. 1(b)), $F'(x)$ begins to lay in the feasible region. When $N_i(t)$ becomes larger (Fig. 1(c)), $F'(x)$ becomes closer to $F(x)$ and thus it must lay in the feasible region as well. Since after the start phase we always have $N_i(t) \geq N_G^*$, DP-FTPL-Gauss can guarantee differential privacy.



(a) $N_i(t) < N_G^*$

(b) $N_i(t) = N_G^*$

(c) $N_i(t) > N_G^*$

Figure 1: Explanation about how DP-FTPL-Gauss protects $(1, 0.01)$-differential privacy.

Algorithm 2 DP-FTPL-Beta

```
Input: $\epsilon, \delta$, for each arm $i$, set $N_i = 0, R_i = 0$.
for $i = 1, 2, \cdots, N$ do
    while $N_i < N_B^* \triangleq \max\left\{\min\left\{\frac{40e}{9\pi\delta^2}, \frac{8000\log\frac{e}{2\pi\delta^2}}{81\epsilon^2}\right\}, \frac{1000e}{9\pi}\right\}$ do
        Pull arm $i(t) = i$ and observe reward $x(t) = x_i(t) \in \{0, 1\}$.
        $R_i \leftarrow R_i + x_i(t)$, $N_i \leftarrow N_i + 1$.
    end while
end for
while true do
    For each arm $i$, draw sample $\theta_i(t)$ independently from Beta perturbed distribution
    $\mathcal{B}(a_i + k_i, b_i + k_i)$, where $a_i = R_i + 1$, $b_i = N_i - R_i + 1$ and $k_i = \lfloor N_i/8 \rfloor + 1$.
    Choose arm $i(t) = \text{argmax}_i\, \theta_i(t)$, and observe reward $x(t) = x_{i(t)}(t) \in \{0, 1\}$.
    $R_{i(t)} \leftarrow R_{i(t)} + x_{i(t)}(t)$, $N_{i(t)} \leftarrow N_{i(t)} + 1$.
end while
```

Remark 12 *The DP-UCB-INT algorithm (Tossou and Dimitrakakis, 2016) also achieves a T-independent additive term $O(\frac{1}{\epsilon^2}\log\frac{1}{\delta})$ in its regret upper bound. However, its additive regret term can be much larger than ours when ϵ is smaller than δ. Besides, DP-UCB-INT behaves much worse than our algorithms in experiments (see details in Section 8).*

6.2 DP-FTPL-Beta

The DP-FTPL-Beta algorithm is described in Algorithm 2. When we use Beta distribution as the perturbed distribution, a change on observations does not only cause a simple shift on the distribution of $\theta_i(t)$. In fact, it can change the entire shape of the perturbed distribution. Moreover, when the observations contain a large proportion of 0s (or 1s), a change on the observations can always lead to a large difference between perturbed distributions. To deal with these challenges, compared with the work of Agrawal and Goyal (2013), we modify the perturbed distributions in our algorithm, i.e., in DP-FTPL-Beta, we choose $N^*(\epsilon, \delta, T) = N_B^* \triangleq \max\left\{\min\left\{\frac{40e}{9\pi\delta^2}, \frac{8000\log\frac{e}{2\pi\delta^2}}{81\epsilon^2}\right\}, \frac{1000e}{9\pi}\right\}$, and the perturbed distribution to be the Beta distribution with parameters $(a_i + k_i, b_i + k_i)$, where $a_i = R_i + 1$, $b_i = N_i - R_i + 1$ and $k_i = \lfloor N_i/8 \rfloor + 1$. Compared with the traditional TS policy with the perturbed distribution to be the Beta distribution with parameters (a_i, b_i), here we add k_i 0s and k_i 1s to the observations on arm i, where k_i is approximately linear with the number of total pulls on arm i. This ensures that both the number of 0s and the number of 1s are larger than a constant (e.g., 0.1) proportion of all the observations on arm i, and makes sure that our algorithm guarantees (ϵ, δ)-differential privacy (as stated in the next theorem).

Theorem 13 *DP-FTPL-Beta guarantees (ϵ, δ)-differential privacy, and its cumulative regret satisfies*

$$Reg(T) \leq \sum_{i=2}^{N} \max\left\{\frac{5\Delta_i \log T}{2(\Delta_i - 5/2\lambda)^2}, N_B^*\Delta_i\right\} + \Theta\left(\frac{N}{\lambda^4}\right)$$

for any $\lambda < \frac{2}{5}\Delta_{\min}$.

Here we also provide the proof of the privacy part, and defer the proof of the regret part to Appendix B. Note that the proof of the privacy part in Theorem 13 is very different from that of Theorem 11, since they use totally different perturbed distributions. On the other hand, our perturbed distribution in DP-FTPL-Beta is also different from that in the work of Agrawal and Goyal (2013). Thus, the regret analysis also needs to be revised carefully to fit the new setting, and some of the theoretical results in our analysis can be of independent interest in other online learning models (please see Appendix B for details).

Proof (privacy part) We will use the following two facts in this proof.

Fact 1 *(Stirling's approximation, Olver et al. (2010))*

$$\sqrt{2\pi n}\left(\frac{n}{e}\right)^n \leq n! \leq \sqrt{2\pi e n}\left(\frac{n}{e}\right)^n.$$

Fact 2 *(Pinsker's Inequality, Csiszar and Körner (2011))*

$$KL(p, q) \geq \frac{1}{2}(p-q)^2.$$

Let's consider the distribution of $i(t)$, when one observation $x(\tau)$ changes to be $x'(\tau)$. We denote $i = i(\tau)$, and $a_i(t), b_i(t), k_i(t)$ are the value of a_i, b_i, k_i at the beginning of time step t.

Note that the Beta distribution of $\theta_i(t)$ (when no observation is changed) is $\mathcal{B}(a_i(t) + k_i(t), b_i(t) + k_i(t))$. To simplify the notations, we let $a = a_i(t) + k_i(t)$, and $b = b_i(t) + k_i(t) - 1$.

Similar as the proof of Theorem 11, we only need to prove that $\Pr[\theta_i(t) \leq x] \leq e^{\epsilon}\Pr[\theta_i'(t) \leq x] + \delta$, where $\theta_i(t)$ follows probability distribution $\mathcal{B}(a, b+1)$, and $\theta_i'(t)$ follows probability distribution $\mathcal{B}(a+1, b)$.

Then we denote f, f' the probability density functions (PDFs) of Beta distributions with parameters $(a, b+1)$ and $(a+1, b)$, respectively, and F, F' are the corresponding cumulative distribution functions (CDFs).

Existing results (Olver et al., 2010) show that $F(x) - F'(x) = \frac{(a+b)!x^a(1-x)^b}{a!b!}$. By Stirling's approximation (Fact 1), we have that

$$F(x) - F'(x) \quad = \quad \frac{(a+b)!x^a(1-x)^b}{a!b!}$$

$$
\begin{aligned}
&\leq \frac{\sqrt{2\pi e(a+b)}(\frac{a+b}{e})^{a+b}x^a(1-x)^b}{\sqrt{2\pi a}(\frac{a}{e})^a\sqrt{2\pi b}(\frac{b}{e})^b} \\
&= \sqrt{\frac{e(a+b)}{2\pi ab}}\frac{(a+b)^{a+b}x^a(1-x)^b}{a^ab^b}.
\end{aligned}
$$

Note that $x^a(1-x)^b \leq \frac{a^ab^b}{(a+b)^{a+b}}$ (when $x=\frac{a}{a+b}$, the equation holds). Thus if $\sqrt{\frac{e(a+b)}{2\pi ab}} \leq \delta$, then

$$
\Pr[\theta_i(t) \leq x] \leq \Pr[\theta_i'(t) \leq x] + \delta \leq e^{\epsilon}\Pr[\theta_i'(t) \leq x] + \delta.
$$

As for the case that $\sqrt{\frac{e(a+b)}{2\pi ab}} \geq \delta$, consider the point $x^* = \frac{a}{a+b} - \sqrt{\frac{1}{a+b}\log\frac{e(a+b)}{2\pi ab\delta^2}}$, we have that

$$
\begin{aligned}
\frac{1}{a+b}\log\frac{(\frac{a}{a+b})^a(\frac{b}{a+b})^b}{x^{*a}(1-x^*)^b} &= \frac{a}{a+b}\log\frac{\frac{a}{a+b}}{x^*} + \frac{b}{a+b}\log\frac{\frac{b}{a+b}}{(1-x^*)} \\
&= KL\left(\frac{a}{a+b}, x^*\right) \\
&\geq \frac{1}{2}\left(\frac{a}{a+b} - x^*\right)^2 \qquad (9) \\
&= \frac{1}{2(a+b)}\log\frac{e(a+b)}{2\pi ab\delta^2},
\end{aligned}
$$

where $KL(p,q) = p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q}$ denotes the KL divergency, and Eq. (9) comes from Pinsker's Inequality (Fact 2).

Therefore,

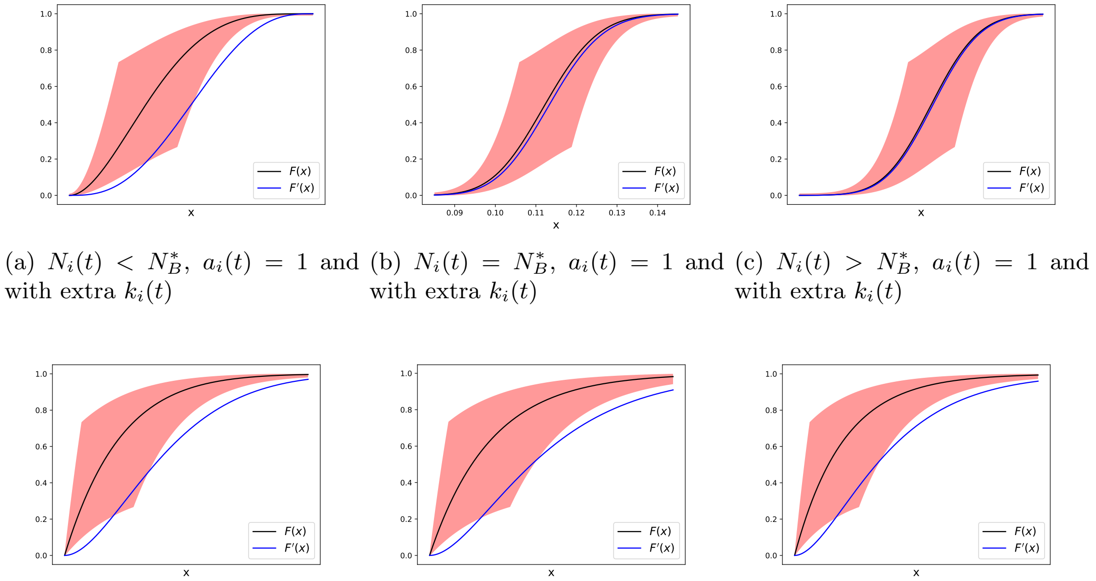
$$
\begin{aligned}
F(x^*) - F'(x^*) &\leq \sqrt{\frac{e(a+b)}{2\pi ab}}\frac{(a+b)^{a+b}x^{*a}(1-x^*)^b}{a^ab^b} \\
&= \sqrt{\frac{e(a+b)}{2\pi ab}}\frac{(a+b)^{a+b}(\frac{a}{a+b})^a(\frac{b}{a+b})^b}{a^ab^b}\frac{x^{*a}(1-x^*)^b}{(\frac{a}{a+b})^a(\frac{b}{a+b})^b} \\
&= \sqrt{\frac{e(a+b)}{2\pi ab}}\exp\left(-(a+b)\frac{1}{a+b}\log\frac{(\frac{a}{a+b})^a(\frac{b}{a+b})^b}{x^{*a}(1-x^*)^b}\right) \\
&\leq \sqrt{\frac{e(a+b)}{2\pi ab}}\exp\left(-(a+b)\frac{1}{2(a+b)}\log\frac{e(a+b)}{2\pi ab\delta^2}\right) \\
&= \delta.
\end{aligned}
$$

For any $x \geq x^*$, we have that

$$
\frac{f(x)}{f'(x)} = \frac{\frac{(1-x)}{b}}{\frac{x}{a}} = \frac{a(1-x)}{bx}.
$$

When $\sqrt{\frac{1}{a+b}\log\frac{e(a+b)}{2\pi ab\delta^2}} \leq \frac{\epsilon ab}{(a+b)(a+b+b\epsilon)}$, $\frac{f(x)}{f'(x)} \leq e^{\epsilon}$ holds for any $x \geq x^*$, which implies that

$$
\Pr[\theta_i(t) \leq x] = \Pr[\theta_i(t) \leq x^*] + \Pr[x^* \leq \theta_i(t) \leq x]
$$



(a) $N_i(t) < N_B^*$, $a_i(t) = 1$ and with extra $k_i(t)$

(b) $N_i(t) = N_B^*$, $a_i(t) = 1$ and with extra $k_i(t)$

(c) $N_i(t) > N_B^*$, $a_i(t) = 1$ and with extra $k_i(t)$

(d) $N_i(t) = 10$, $a_i(t) = 1$ and without extra $k_i(t)$

(e) $N_i(t) = 10^4$, $a_i(t) = 1$ and without extra $k_i(t)$

(f) $N_i(t) = 10^7$, $a_i(t) = 1$ and without extra $k_i(t)$

Figure 2: Explanation about how DP-FTPL-Beta protects $(1, 0.01)$-differential privacy and the necessity of the extra $k_i(t)$ in DP-FTPL-Beta.

$$\begin{aligned} &\leq \Pr[\theta'_i(t) \leq x^*] + \delta + e^\epsilon \Pr[x^* \leq \theta'_i(t) \leq x] \\ &\leq e^\epsilon (\Pr[\theta'_i(t) \leq x^*] + \Pr[x^* \leq \theta'_i(t) \leq x]) + \delta \\ &= e^\epsilon \Pr[\theta'_i(t) \leq x] + \delta. \end{aligned}$$

Therefore, we need $\frac{a^2b^2}{(a+b)^3} \geq \frac{\log \frac{e}{2\pi\delta^2}}{\epsilon^2}$ or $\frac{ab}{a+b} \geq \frac{e}{2\pi\delta^2}$ to guarantee (ϵ, δ)-differential privacy. Note that the inequality $(a_i(t)+k_i(t))(b_i(t)+k_i(t)) \geq 0.09(a_i(t)+b_i(t)+2k_i(t))^2$ always holds in DP-FTPL-Beta (since we choose $k_i(t) = \lfloor \frac{N_i(t)}{8} \rfloor + 1$). Thus when $a_i(t) + b_i(t) \geq \min\{\frac{40e}{9\pi\delta^2}, \frac{8000 \log \frac{e}{2\pi\delta^2}}{81\epsilon^2}\}$ (after the start phase), the (ϵ, δ)-differential privacy must be guaranteed. ■

Here we also use some figures to show the correctness of the differential privacy guarantee of DP-FTPL-Beta, and the necessity of the start phase and the extra $k_i(t)$ on parameters $(a_i(t), b_i(t))$ $(a_i(t), b_i(t), k_i(t)$ are the values of a_i, b_i, k_i at the beginning of time step $t)$. Similar as before, in Fig. 2, the black line is $F(x)$, the origin CDF of $\theta_i(t)$ and the blue line is $F'(x)$, the CDF of $\theta_i(t)$ after one observation is changed. The red region is the feasible region of $F'(x)$ under constraints in Lemma 8 with $(\epsilon, \delta) = (1, 0.01)$. In all these figures, we choose $a_i(t) = 1$.

Fig. 2(a), 2(b) and 2(c) show the correctness of the differential privacy guarantee of DP-FTPL-Beta, and the necessity of the start phase. We can see that when $N_i(t) < N_B^*$

(Fig. 2(a)), $F'(x)$ is out of the feasible region, which means the start phase is also necessary (similar to DP-FTPL-Gauss). On the other hand, when $N_i(t) = N_B^*$ (Fig. 2(b)), $F'(x)$ begins to lay in the feasible region. When $N_i(t)$ becomes larger (Fig. 2(c)), $F'(x)$ also becomes closer to $F(x)$ and thus it must lay in the feasible region as well. Since after the start phase we must have $N_i(t) \geq N_B^*$, DP-FTPL-Beta can guarantee differential privacy.

Fig. 2(d), 2(e) and 2(f) show the necessity of the extra $k_i(t)$ on parameters $(a_i(t), b_i(t))$. We can see that if we do not add extra $k_i(t)$ to parameters $(a_i(t), b_i(t))$, then no matter how large $N_i(t)$ is ($N_i(t) = 10$ in Fig. 2(d), $N_i(t) = 10^4$ in Fig. 2(e), and $N_i(t) = 10^7$ in Fig. 2(f)), the CDF $F'(x)$ is always out of the feasible region. This demonstrates that the usage of $k_i(t)$ is necessary.

Remark 14 *Our analysis shows that DP-FTPL-Beta achieves a similar behaviour as DP-FTPL-Gauss, i.e., after the start phase, the differential privacy is always protected. Similarly, the additive term of regret is still* $O(\min\{\frac{\log(1/\delta)}{\epsilon^2}, \frac{1}{\delta^2}\})$*. However, compared with the Gaussian case, here the start phase size contains a much larger constant factor. Therefore, for small value* T*, DP-FTPL-Beta can behave worse due to the long start phase (see details in Section 8).*

6.3 DP-FTPL-New

Theorems 11 and 13 show that the extra regret of DP-FTPL-Gauss and DP-FTPL-Beta are both $O(\min\{\frac{\log(1/\delta)}{\epsilon^2}, \frac{1}{\delta^2}\})$. They are much larger than the *upper bound* of the extra regret lower bound, i.e., $\Omega(\frac{1}{\epsilon}\log\frac{(e^\epsilon-1)T+\delta T}{(e^\epsilon-1)+\delta T})$ (as stated in Remark 7). For example, when $\delta \to 0$, $\min\{\frac{\log(1/\delta)}{\epsilon^2}, \frac{1}{\delta^2}\}$ becomes infinity, while $\frac{1}{\epsilon}\log\frac{(e^\epsilon-1)T+\delta T}{(e^\epsilon-1)+\delta T} = \frac{\log T}{\epsilon}$; and when $\epsilon \to 0$, $\min\{\frac{\log(1/\delta)}{\epsilon^2}, \frac{1}{\delta^2}\} = \frac{1}{\delta^2}$, while $\frac{1}{\epsilon}\log\frac{(e^\epsilon-1)T+\delta T}{(e^\epsilon-1)+\delta T} \approx \frac{1}{\delta}$. Therefore, the extra regret of DP-FTPL-Gauss and DP-FTPL-Beta is much larger than necessary. The reason is that both the Gaussian perturbed distribution and the Beta perturbed distribution do not suit the differential privacy setting well. For example, in both Fig 1(b) and Fig 2(b), there is still a large gap between the CDF $F'(x)$ and the boundary of the feasible region. Hence, if we want to design algorithms that achieve smaller extra regret, we need to make sure that the CDF $F'(x)$ is always the same as the boundary of the feasible region. Motivated by this idea, we design a new perturbed distribution, and design the DP-FTPL-New algorithm.

Similar with DP-FTPL-Gauss, we also let the perturbed distribution be symmetric and satisfy that a change on observations can only shift the distribution by at most $\frac{1}{N_i(t)}$. Then to ensure the differential privacy, it is sufficient to make sure that for any $-\infty \leq x \leq y \leq +\infty$ and any $z \in [-\frac{1}{N_i(t)}, \frac{1}{N_i(t)}]$, $(F(y) - F(x)) \leq e^\epsilon(F(y+z) - F(x+z)) + \delta$.

Hence, we come up with the following lemma.

Lemma 15 *Consider the random distribution with the following CDF:*

$$F(x) = \begin{cases} 0, & x < x_0 - \Delta_x \\ \frac{\left(\frac{e^\epsilon-1}{2\delta}+1\right)\exp(N_i(t)\epsilon(x-x_0))-1}{e^\epsilon-1}\delta, & x_0 - \Delta_x \leq x \leq x_0 \\ 1 - \frac{\left(\frac{e^\epsilon-1}{2\delta}+1\right)\exp(N_i(t)\epsilon(x_0-x))-1}{e^\epsilon-1}\delta, & x_0 < x \leq x_0 + \Delta_x \\ 1, & x > x_0 + \Delta_x \end{cases}$$

where $\Delta_x = \frac{1}{N_i(t)\epsilon}\log\left(\frac{e^\epsilon - 1}{2\delta} + 1\right)$. *This distribution satisfies that for any* $-\infty \le x \le y \le +\infty$ *and any* $z \in [-\frac{1}{N_i(t)}, \frac{1}{N_i(t)}]$,

$$\begin{aligned}(F(y) - F(x)) &\le e^\epsilon (F(y+z) - F(x+z)) + \delta; && (10)\\ (1 - F(y) + F(x)) &\le e^\epsilon (1 - F(y+z) + F(x+z)) + \delta. && (11)\end{aligned}$$

Specifically, when $\delta = 0$, *then the CDF can take limits for* $\delta \to 0$, *i.e.,*

$$F(x) = \begin{cases} \frac{1}{2}\exp(N_i(t)\epsilon(x - x_0)) & x \le x_0 \\ 1 - \frac{1}{2}\exp(N_i(t)\epsilon(x_0 - x)), & x > x_0 \end{cases}$$

On the other hand, if $\epsilon = 0$, *then the CDF can take limits for* $\epsilon \to 0$, *i.e.,*

$$F(x) = \begin{cases} 0, & x < x_0 - \Delta_x \\ \frac{1}{2} + N_i(t)\delta(x - x_0), & x_0 - \Delta_x \le x \le x_0 \\ \frac{1}{2} - N_i(t)\delta(x_0 - x), & x_0 < x \le x_0 + \Delta_x \\ 1, & x > x_0 + \Delta_x \end{cases}$$

where $\Delta_x = \frac{1}{2N_i(t)\delta}$.

Proof We only prove Eq. (10) when $\epsilon \neq 0$ and $\delta \neq 0$. The proofs for Eq. (11), as well as the case that one of (ϵ, δ) is 0 are almost the same.

Firstly, We divide the interval (x, y) to three parts, i.e., $I_1 = (x, y) \cap (-\infty, x_0 - \Delta_x]$, $I_2 = (x, y) \cap (x_0 - \Delta_x, x_0 + \Delta_x]$ and $I_3 = (x, y) \cap (x_0 + \Delta_x, +\infty)$. Then we denote F_1, F_2, F_3 (and F_1^z, F_2^z, F_3^z) as the probability mass of distribution with CDF $F(x)$ (and $F(x+z)$) in these three intervals, respectively.

For $z \ge 0$, it is easy to check that $F_1^z \le F_1 \le F(x_0 - \Delta_x) = \delta$, and $F_3 \le F_3^z$. As for the comparison between F_2 and F_2^z, one can use the corresponding PDF of F, i.e.,

$$f(x) = \begin{cases} 0, & x < x_0 - \Delta_x \\ \frac{\delta}{e^\epsilon - 1}(\frac{e^\epsilon - 1}{2\delta} + 1)N_i(t)\epsilon e^{N_i(t)\epsilon(x - x_0)}, & x_0 - \Delta_x \le x \le x_0 \\ \frac{\delta}{e^\epsilon - 1}(\frac{e^\epsilon - 1}{2\delta} + 1)N_i(t)\epsilon e^{N_i(t)\epsilon(x_0 - x)}, & x_0 < x \le x_0 + \Delta_x \\ 0, & x > x_0 + \Delta_x \end{cases}$$

If both $f(x)$ and $f(x+z)$ locates in the second part, then

$$\frac{f(x)}{f(x+z)} = \exp(-N_i(t)\epsilon z) \le 1. \tag{12}$$

If both $f(x)$ and $f(x+z)$ locates in the third part, then (recall that $z \le \frac{1}{N_i(t)}$)

$$\frac{f(x)}{f(x+z)} = \exp(N_i(t)\epsilon z) \le e^\epsilon \tag{13}$$

As for the case that $f(x)$ locates in the second part but $f(x+z)$ locates in the third part, then (recall that $z \le \frac{1}{N_i(t)}$ and in this case $x \le x_0$)

$$\frac{f(x)}{f(x+z)} = \exp(N_i(t)\epsilon(x - x_0 - x_0 + x + z)) \le e^\epsilon. \tag{14}$$

Algorithm 3 DP-FTPL-New

```
Input: $\epsilon, \delta$
for $t = 1, 2, \cdots, N$ do
    Pull arm $i(t) = t$ and observe reward $x(t) = x_t(t) \in \{0, 1\}$.
    $R_t \leftarrow x_t(t)$, $N_t \leftarrow 1$.
end for
while true do
    For each arm $i$, draw sample $\theta_i(t)$ independently from perturbed distribution $\mathcal{D}(N_i, R_i)$ (whose CDF is given as Lemma 15, and $x_0 = \hat{\mu}_i(t) + \sqrt{\frac{\log T}{N_i(t)}} + \frac{1}{N_i(t)\epsilon}\log\frac{T(e^\epsilon-1)+2T\delta}{2(e^\epsilon-1)+2T\delta}$).
    Choose arm $i(t) = \mathrm{argmax}_i\, \theta_i(t)$, and observe reward $x(t) = x_{i(t)}(t) \in \{0, 1\}$.
    $R_{i(t)} \leftarrow R_{i(t)} + x_{i(t)}(t)$, $N_{i(t)} \leftarrow N_{i(t)} + 1$.
end while
```

Eq. (12), (13) and (14) shows that $F_2 \le e^\epsilon F_2^z$. Therefore, we must have that

$$F_1 + F_2 + F_3 \le \delta + e^\epsilon F_2^z + F_3^z \le e^\epsilon(F_1^z + F_2^z + F_3^z) + \delta,$$

which implies that Eq. (10) holds.

Similarly, when $z \le 0$, one can also prove this equation by symmetry. ■

Now we introduce our DP-FTPL-New policy (as described in Algorithm 3), which uses the perturbed distribution in Lemma 15 to protect differential privacy. For an arm i, we can choose its x_0 (which is used in the CDF in Lemma 15) to be $\hat{\mu}_i(t) + \sqrt{\frac{\log T}{N_i(t)}} + \frac{1}{N_i(t)\epsilon}\log\frac{T(e^\epsilon-1)+2T\delta}{2(e^\epsilon-1)+2T\delta}$. In this case, if we change one observation, then $\hat{\mu}_i(t)$ can change for at most $\frac{1}{N_i(t)}$, while the other terms $\sqrt{\frac{\log T}{N_i(t)}} + \frac{1}{N_i(t)\epsilon}\log\frac{T(e^\epsilon-1)+2T\delta}{2(e^\epsilon-1)+2T\delta}$ remain the same. This means that the random distribution can shift for at most $\frac{1}{N_i(t)}$, and therefore based on Lemma 15, the algorithm can protect (ϵ, δ)-differential privacy.

Note that we do not require a start phase in DP-FTPL-New, since for any $N_i(t) > 0$, the conditions in Lemma 8 holds (which is different with DP-FTPL-Gauss and DP-FTPL-Beta). However, this does not mean that DP-FTPL-New incurs no extra regrets. The reason is that in DP-FTPL-New, the confidence radius of the perturbed distribution ($\frac{1}{N_i(t)\epsilon}\log\frac{T(e^\epsilon-1)+2T\delta}{2(e^\epsilon-1)+2T\delta}$) is not the same as the confidence radius of the expected reward ($\sqrt{\frac{\log T}{N_i}}$). In DP-FTPL-Gauss and DP-FTPL-Beta, we always choose the two confidence radiuses to have the same order, which means that these algorithms are normal bandit algorithms if they do not have the start phase (i.e., have regret upper bound $O(\sum_i \frac{\log T}{\Delta_i})$). Therefore, after we adding the start phase to guarantee differential privacy, the start phase size becomes the only extra additional term in their regret upper bounds. In DP-FTPL-New, the extra regret term comes from the different confidence radius $\frac{1}{N_i(t)\epsilon}\log\frac{T(e^\epsilon-1)+2T\delta}{2(e^\epsilon-1)+2T\delta}$. We also need this confidence radius to be less than $\Theta(\Delta_i)$ to make sure that with high probability, arm i will not be pulled. This means that $N_i(t) \ge \Theta(\frac{1}{\Delta_i\epsilon}\log\frac{T(e^\epsilon-1)+2T\delta}{2(e^\epsilon-1)+2T\delta})$, i.e., there is an extra additional term of $O(\frac{N}{\epsilon}\log\frac{T(e^\epsilon-1)+2T\delta}{2(e^\epsilon-1)+2T\delta})$ in the regret upper bound of DP-FTPL-New, which is shown in detail in the next theorem.

Theorem 16 *DP-FTPL-New can protect (ϵ, δ)-differential privacy, and its regret satisfies*

$$Reg(T) \le \sum_i \max\left\{\frac{16\log T}{\Delta_i}, \frac{4}{\epsilon}\log\frac{T(e^\epsilon-1)+2T\delta}{2(e^\epsilon-1)+2T\delta}\right\} + 4N$$

Proof The privacy part can be directly obtained by Lemma 8 and Lemma 15, therefore we only analyze the regret upper bound here, by showing that the used $\theta_i(t)$ is a kind of upper confidence bound. In this proof, we will use the following two facts.

Fact 3 *For any $t > 0$,*

$$\Pr\left[|\hat{\mu}_i(t) - \mu_i| \ge \sqrt{\frac{\log T}{N_i(t)}}\right] \le \frac{2}{T}.$$

Fact 4 *For any $t > 0$,*

$$\Pr\left[|\theta_i(t) - x_0| \ge \frac{1}{N_i(t)\epsilon}\log\frac{T(e^\epsilon-1)+2T\delta}{2(e^\epsilon-1)+2T\delta}\right] \le \frac{2}{T},$$

where $x_0 = \hat{\mu}_i(t) + \sqrt{\frac{\log T}{N_i(t)}} + \frac{1}{N_i(t)\epsilon}\log\frac{T(e^\epsilon-1)+2T\delta}{2(e^\epsilon-1)+2T\delta}$.

Fact 3 is based on Chernoff-Hoeffding inequality (which is shown in detail as Fact 5 in Appendix), i.e. by applying Chernoff-Hoeffding inequality to Eq. (15),

$$\begin{aligned}
\Pr\left[|\hat{\mu}_i(t) - \mu_i| \ge \sqrt{\frac{\log T}{N_i(t)}}\right] &= \sum_{n=1}^{T}\Pr\left[|\hat{\mu}_i(t) - \mu_i| \ge \sqrt{\frac{\log T}{N_i(t)}}, N_i(t) = n\right] \quad (15)\\
&\le \sum_{n=1}^{T}\frac{2}{T^2}\\
&\le \frac{2}{T}.
\end{aligned} \quad (16)$$

Fact 4 is based on the CDF described as in Lemma 15, i.e.,

$$\begin{aligned}
F\left(x_0 - \frac{1}{N_i(t)\epsilon}\log\frac{T(e^\epsilon-1)+2T\delta}{2(e^\epsilon-1)+2T\delta}\right) &= \frac{\left(\frac{e^\epsilon-1}{2\delta}+1\right)\frac{2(e^\epsilon-1)+2T\delta}{T(e^\epsilon-1)+2T\delta} - 1}{e^\epsilon-1}\delta\\
&= \frac{1}{T},
\end{aligned}$$

and similarly,

$$F\left(x_0 + \frac{1}{N_i(t)\epsilon}\log\frac{T(e^\epsilon-1)+2T\delta}{2(e^\epsilon-1)+2T\delta}\right) = 1 - \frac{1}{T}$$

Based on Facts 3 and 4, we know that with probability at least $1 - \frac{4}{T}$, we have

$$\theta_i(t) - \mu_i = (\theta_i(t) - x_0) + (x_0 - \hat{\mu}_i(t)) + (\hat{\mu}_i(t) - \mu_i) \geq 0,$$

and

$$\theta_i(t) - \mu_i = \theta_i(t) - x_0 + x_0 - \hat{\mu}_i(t) + \hat{\mu}_i(t) - \mu_i \leq 2\left(\sqrt{\frac{\log T}{N_i(t)}} + \frac{1}{N_i(t)\epsilon}\log\frac{T(e^\epsilon - 1) + 2T\delta}{2(e^\epsilon - 1) + 2T\delta}\right),$$

Hence, Let $\mathcal{E}_t$ be the event that $\forall i, 0 \leq \theta_i(t) - \mu_i \leq 2(\sqrt{\frac{\log T}{N_i(t)}} + \frac{1}{N_i(t)\epsilon}\log\frac{T(e^\epsilon-1)+2T\delta}{2(e^\epsilon-1)+2T\delta})$, then the expected regret at time step t when event $\neg\mathcal{E}_t$ happens is at most $\frac{4N}{T}$.

On the other hand, under event $\mathcal{E}_t$, we will pull an sub-optimal arm i only if

$$2\left(\sqrt{\frac{\log T}{N_i(t)}} + \frac{1}{N_i(t)\epsilon}\log\frac{T(e^\epsilon - 1) + 2T\delta}{2(e^\epsilon - 1) + 2T\delta}\right) \leq \Delta_i.$$

Otherwise

$$\theta_i(t) \leq \mu_i + 2\left(\sqrt{\frac{\log T}{N_i(t)}} + \frac{1}{N_i(t)\epsilon}\log\frac{T(e^\epsilon - 1) + 2T\delta}{2(e^\epsilon - 1) + 2T\delta}\right) \leq \mu_1 \leq \theta_1(t),$$

and we will not pull sub-optimal arm i.

Note that by basic calculations, if $N_i(t) \geq \max\{\frac{16\log T}{\Delta_i^2}, \frac{4}{\Delta_i\epsilon}\log\log\frac{T(e^\epsilon-1)+2T\delta}{2(e^\epsilon-1)+2T\delta}\}$, then $2(\sqrt{\frac{\log T}{N_i(t)}} + \frac{1}{N_i(t)\epsilon}\log\frac{T(e^\epsilon-1)+2T\delta}{2(e^\epsilon-1)+2T\delta})$ must be smaller than Δ_i.

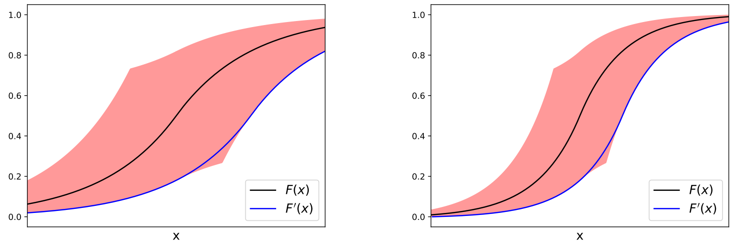
This means that the regret of DP-FTPL-New is upper bounded by

$$\begin{aligned} Reg(T) &\leq \sum_i \Delta_i \max\left\{\frac{16\log T}{\Delta_i^2}, \frac{4}{\Delta_i\epsilon}\log\frac{T(e^\epsilon - 1) + 2T\delta}{2(e^\epsilon - 1) + 2T\delta}\right\} + 4N \\ &= \sum_i \max\left\{\frac{16\log T}{\Delta_i}, \frac{4}{\epsilon}\log\frac{T(e^\epsilon - 1) + 2T\delta}{2(e^\epsilon - 1) + 2T\delta}\right\} + 4N. \end{aligned}$$

■

Here we use Fig. 3 to show the correctness of the differential privacy guarantee of DP-FTPL-New. We can see that in DP-FTPL-New, no matter how large $N_i(t)$ is, $F'(x)$ is always the same as the boundary of the feasible region. Therefore, it does not require the start phase, and always has differential privacy guarantee. Moreover, this is the best one can do (i.e., the lowest extra regret) to protect (ϵ, δ)-differential privacy based on Lemma 8.

Remark 17 *Note that Lemma 15 provides perturbed distributions not only for the case that both ϵ and δ are not zero, but also for the case that either $\delta = 0$ or $\epsilon = 0$. Therefore, Theorem 16 also works in the case that either $\delta = 0$ or $\epsilon = 0$ (and the extra regret term just takes limits of either $\delta \to 0$ or $\epsilon \to 0$, i.e., $O(\frac{N\log T}{\epsilon})$ and $O(\frac{N}{\delta})$). This means that DP-FTPL-New not only has a better regret upper bound, but is also more general than the existing algorithms, e.g., DP-UCB (only works when $\delta = 0$), DP-UCB-BOUND (only works when $\delta = 0$), DP-UCB-INT (only works when $\delta \neq 0$ and $\epsilon \neq 0$), DP-FTPL-Gauss (only works when $\delta \neq 0$) and DP-FTPL-Beta (only works when $\delta \neq 0$).*



(a) $N_i(t) = 10$

(b) $N_i(t) = 50$

Figure 3: Explanation about how DP-FTPL-New protects $(0.1, 0)$-differential privacy.

Algorithm 4 GDP-Elim-New

```
**Input:** $\epsilon, \delta$.
**for** Phase $k = 1, 2, \cdots$ **do**
    $\Delta(k) = 2^{-k}$.
    $N(k) = \max\{\frac{32 \log T}{\Delta^2(k)}, \frac{4}{\Delta(k)\epsilon} \log \frac{T^2(e^\epsilon - 1) + T^2\delta}{2(e^\epsilon - 1) + T^2\delta}\}$
    For each arm $i$, pull it for $N(k)$ times, and let its empirical mean in this $N(k)$ times of pull be $\hat{\mu}_i(k)$
    Let $\theta_i(k)$ be a random variable that is independently drawn from the distribution described in Lemma 15 with $\epsilon, \delta/2, x_0 = \hat{\mu}_i(k), N_i(t) = N(k)$.
    **if** there exists $i$ such that $\theta_i(k) \geq \max_{j \neq i} \theta_j(k) + \Delta(k)$ **then**
        Start to exploit on this arm $i$.
    **end if**
**end for**
```

7. Algorithm that Protect Global Differential Privacy in Multi-armed Bandits

Note that the distribution in Lemma 15 satisfies not only the condition in Lemma 8 (the differential privacy case), but also the condition in Lemma 9 (the global differential privacy case). Therefore, we can also design an algorithm GDP-Elim-New that protects (ϵ, δ)-global differential privacy based on this perturbed distribution.

The GDP-Elim-New is described as in Algorithm 4. Following Lemma 9, it divides the learning procedure into several phases. In each phase k, GDP-Elim-New will pull all the arms for $N(k)$ times, and then estimate their empirical means $\hat{\mu}_i(k)$'s in this phase. After that, to protect (ϵ, δ)-global differential privacy, it draws random samples $\theta_i(k)$'s from the perturbed distribution described in Lemma 15, and then start to exploit if it makes sure that some arm is the optimal arm with high probability.

Theorem 18 *Algorithm 4 can protect (ϵ, δ)-global differential privacy, and its regret is upper bounded by*

$$Reg(T) \leq \sum_i \left(\frac{1024 \Delta_i \log T}{\Delta_{\min}^2} + \frac{32 \Delta_i}{\epsilon \Delta_{\min}} \log \frac{T^2(e^\epsilon - 1) + T^2\delta}{2(e^\epsilon - 1) + T^2\delta} \right) + 4N.$$

Proof The privacy part can be directly obtained by Lemma 9 and Lemma 15, therefore we only analyze the regret upper bound here, and we first define the following two kinds of events:

- $\mathcal{A}(k) = \{\forall i, |\hat{\mu}_i(k) - \mu_i| < \sqrt{\frac{2\log T}{N(k)}}\}$;
- $\mathcal{B}(k) = \{\forall i, |\theta_i(k) - \hat{\mu}_i(k)| < \frac{1}{N(k)\epsilon}\log\frac{T^2(e^\epsilon-1)+T^2\delta}{2(e^\epsilon-1)+T^2\delta}\}$;

Similar with Fact 3 and Fact 4, we could prove that

$$\begin{aligned}\Pr[\neg\mathcal{A}(k)] &\leq \frac{2N}{T^2},\\ \Pr[\neg\mathcal{B}(k)] &\leq \frac{2N}{T^2}.\end{aligned}$$

Let $\mathcal{G}$ be the event that for any phase k, $\mathcal{A}(k)$ and $\mathcal{B}(k)$ happen. Then since there can be at most T phases, we have

$$\Pr[\neg\mathcal{G}] \leq \frac{4N}{T^2}\cdot T = \frac{4N}{T}.$$

Therefore the expected regret when $\neg\mathcal{G}$ happens is at most $4N$.

Note that if we choose $N(k) = \max\{\frac{32\log T}{\Delta^2(k)}, \frac{4}{\Delta(k)\epsilon}\log\frac{T^2(e^\epsilon-1)+T^2\delta}{2(e^\epsilon-1)+T^2\delta}\}$, then we must have

$$\sqrt{\frac{2\log T}{N(k)}} \leq \frac{1}{4}\Delta(k),$$

and

$$\frac{1}{N(k)\epsilon}\log\frac{T^2(e^\epsilon-1)+T^2\delta}{2(e^\epsilon-1)+T^2\delta} \leq \frac{1}{4}\Delta(k).$$

Thus, under event $\mathcal{A}(k)$ and $\mathcal{B}(k)$, for any sub-optimal arm $i \geq 2$, we must have that

$$\begin{aligned}\theta_1(k) &\geq \mu_1 - \sqrt{\frac{2\log T}{N(k)}} - \frac{1}{N(k)\epsilon}\log\frac{T^2(e^\epsilon-1)+T^2\delta}{2(e^\epsilon-1)+T^2\delta}\\ &\geq \mu_i - \frac{1}{4}\Delta(k) - \frac{1}{4}\Delta(k)\\ &\geq \theta_i(k) - \sqrt{\frac{2\log T}{N(k)}} - \frac{1}{N(k)\epsilon}\log\frac{T^2(e^\epsilon-1)+T^2\delta}{2(e^\epsilon-1)+T^2\delta} - \frac{1}{4}\Delta(k) - \frac{1}{4}\Delta(k)\\ &\geq \theta_i(k) - \Delta(k).\end{aligned}$$

This means that we can only exploit on arm 1 (the optimal arm).

As for sub-optimal arms $j \geq 2$, we must have that

$$\begin{aligned}\theta_j(k) &\leq \mu_j + \sqrt{\frac{2\log T}{N(k)}} + \frac{1}{N(k)\epsilon}\log\frac{T^2(e^\epsilon-1)+T^2\delta}{2(e^\epsilon-1)+T^2\delta}\\ &\leq \mu_j + \frac{1}{4}\Delta(k) + \frac{1}{4}\Delta(k)\end{aligned}$$

$$
\begin{aligned}
&= \mu_1 - \Delta_j + \frac{1}{4}\Delta(k) + \frac{1}{4}\Delta(k) \\
&\leq \theta_1(k) + \sqrt{\frac{2\log T}{N(k)}} + \frac{1}{N(k)\epsilon}\log\frac{T^2(e^\epsilon - 1) + T^2\delta}{2(e^\epsilon - 1) + T^2\delta} - \Delta_j + \frac{1}{4}\Delta(k) + \frac{1}{4}\Delta(k) \\
&\leq \theta_1(k) - \Delta_j + \Delta(k) \\
&\leq \theta_1(k) - \Delta(k) - (\Delta_j - 2\Delta(k)).
\end{aligned}
$$

This means that after phase k such that $2\Delta(k) \leq \Delta_j$, arm j should always satisfy $\theta_j(k) \leq \theta_1(k) - \Delta(k)$. Hence, if $2\Delta(k) \leq \Delta_{\min}$, we must start to exploit.

Denote this phase be k^*, i.e., $2\Delta(k^*) \leq \Delta_{\min}$ but $4\Delta(k^*) \geq \Delta_{\min}$.

Then, the expected regret when $\mathcal{G}$ happens is upper bounded by

$$
\begin{aligned}
\sum_i \sum_{k=1}^{k^*} N(k)\Delta_i &= \sum_i \sum_{k=1}^{k^*} \max\left\{\frac{32\log T}{\Delta^2(k)}, \frac{4}{\Delta(k)\epsilon}\log\frac{T^2(e^\epsilon - 1) + T^2\delta}{2(e^\epsilon - 1) + T^2\delta}\right\}\Delta_i \\
&\leq \sum_i \sum_{k=1}^{k^*} \left(\frac{32\log T}{\Delta^2(k)} + \frac{4}{\Delta(k)\epsilon}\log\frac{T^2(e^\epsilon - 1) + T^2\delta}{2(e^\epsilon - 1) + T^2\delta}\right)\Delta_i \\
&\leq \sum_i \left(\frac{64\Delta_i\log T}{\Delta^2(k^*)} + \frac{8\Delta_i}{\Delta(k^*)\epsilon}\log\frac{T^2(e^\epsilon - 1) + T^2\delta}{2(e^\epsilon - 1) + T^2\delta}\right) \\
&\leq \sum_i \left(\frac{1024\Delta_i\log T}{\Delta_{\min}^2} + \frac{32\Delta_i}{\epsilon\Delta_{\min}}\log\frac{T^2(e^\epsilon - 1) + T^2\delta}{2(e^\epsilon - 1) + T^2\delta}\right)
\end{aligned}
$$

Along with the fact that the expected regret when $\neg\mathcal{G}$ happens is at most $4N$, we can finally get the regret upper bound shown in Theorem 18. ■

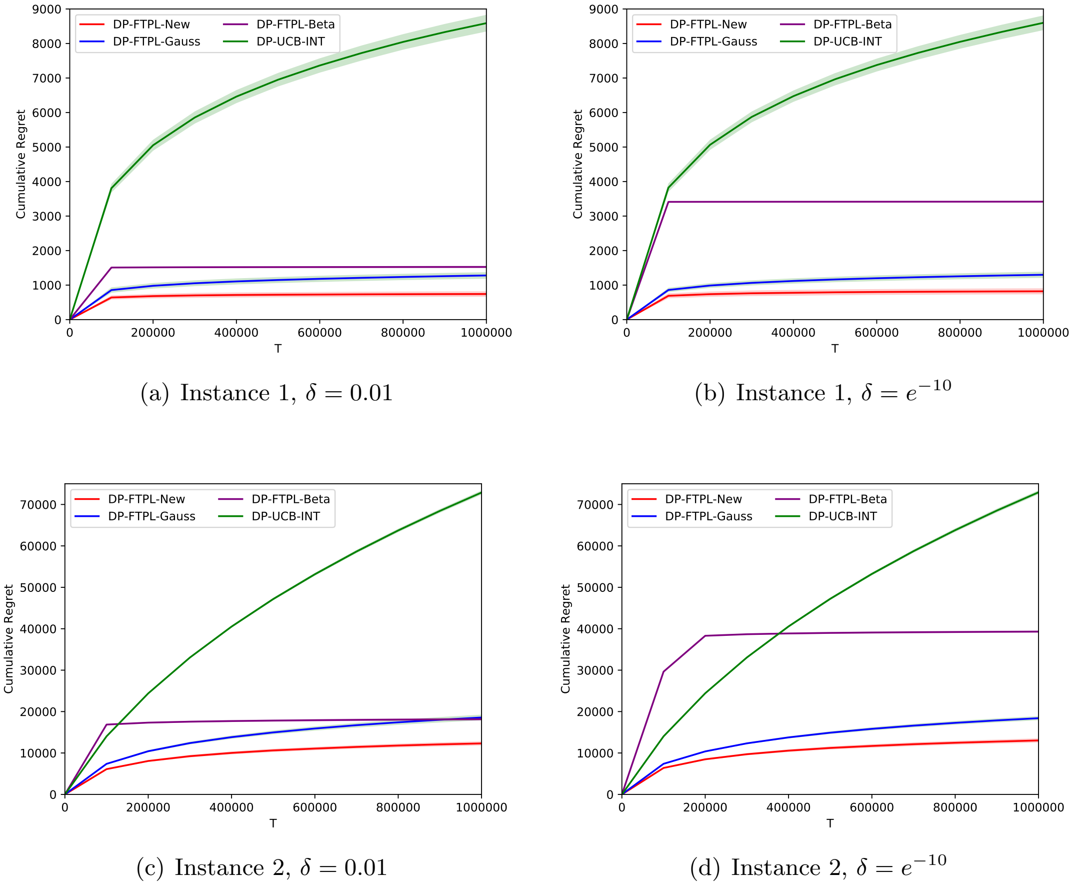
Theorem 18 shows that when $\frac{\Delta_{\max}}{\Delta_{\min}}$ is bounded, then the regret upper bound of GDP-Elim-New has the same order as DP-FTPL-New, and matches the regret lower bound in Theorem 4 (in order). This indicates that our regret lower bound is tight.

Remark 19 *Similar to DP-FTPL-New, GDP-Elim-New works in not only the case that $\delta \neq 0$ and $\epsilon \neq 0$, but also the case that either $\delta = 0$ or $\epsilon = 0$. Therefore, it is also more general than existing algorithms such as AdaP-UCB and AdaP-KLUCB (Azize and Basu, 2022), which only work in the case that $\delta = 0$.*

8. Experiments

In this section, we mainly consider the differential privacy case, and compare our algorithms with state-of-the-art baselines, including DP-UCB, DP-UCB-BOUND, DP-UCB-INT in the work of Tossou and Dimitrakakis (2016) and DP-SE in the work of Sajed and Sheffet (2019).

We consider two problem instances. In Instance 1 we set $N = 9$ and the expected reward vector is $[0.3, 0.35, 0.4, 0.45, 0.5, 0.55, 0.6, 0.65, 0.7]$, in Instance 2 we set $N = 101$ and the expected reward of arm $i \in [N]$ is $\mu_i = 0.3 + 0.004(i - 1)$. In both instances, we choose $T = 10^6$, and all the results (the average cumulative regrets and the standard deviations of cumulative regrets) take an average over 100 independent runs.



(a) Instance 1, $\delta = 0.01$

(b) Instance 1, $\delta = e^{-10}$

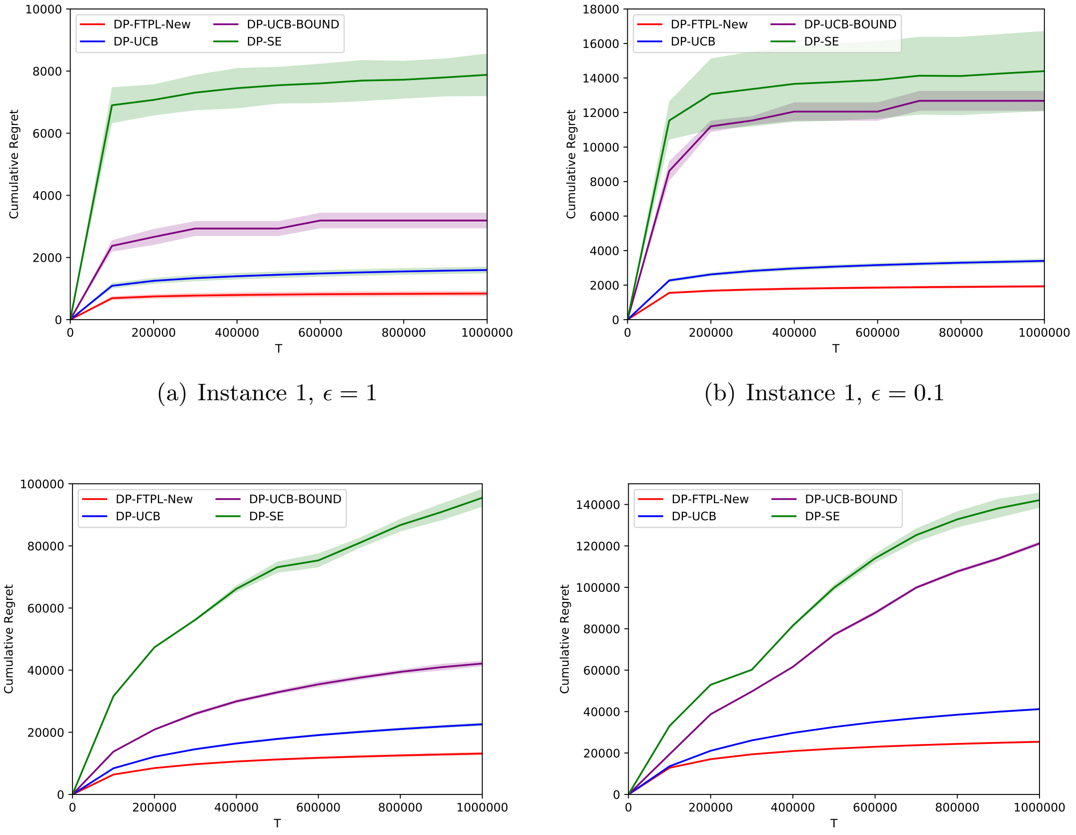
(c) Instance 2, $\delta = 0.01$

(d) Instance 2, $\delta = e^{-10}$

Figure 4: Experiments for (ϵ, δ)-differential privacy

We first consider the case that the learning policy needs to guarantee (ϵ, δ)-differential privacy. Here we compare the regret performances of our algorithms (DP-FTPL-Gauss, DP-FTPL-Beta and DP-FTPL-New) with DP-UCB-INT. In Fig. 4(a) and Fig. 4(b), we use Instance 1 and set (ϵ, δ) to be $(1, 0.01)$ or $(1, e^{-10})$. In Fig. 4(c) and Fig. 4(d), we use Instance 2 and set (ϵ, δ) to be $(1, 0.01)$ or $(1, e^{-10})$. From these experiments, we can see that our algorithms outperform DP-UCB-INT significantly. As we have explained, DP-FTPL-Beta suffers from a long start phase, and it behaves much worse when δ is small. However, we can see that after the start phase, its regret does not increase at all. This accords with our analysis, since its regret always equals to $\sum_i \Delta_i N_B^*$ when $\Delta_i N_B^*$ is larger than $\frac{\log T}{\Delta_i}$. As for DP-FTPL-Gauss, since it has a short start phase, its regret increases continuously as T grows up. On the other hand, DP-FTPL-New is always the optimal one in these four algorithms, since its extra regret term is the smallest.

Then we consider the case that the policy needs to guarantee $(\epsilon, 0)$-differential privacy. In this case, we compare the regret performance of our DP-FTPL-New policy with DP-UCB, DP-UCB-BOUND and DP-SE. In Fig. 5(a) and Fig. 5(b), we use Instance 1 and set ϵ to be 1 or 0.1. In Fig. 5(c) and Fig. 5(d), we use Instance 2 and set ϵ to be 1 or 0.1. From these experiments, we can see that DP-FTPL-New outperforms all the other algorithms



(a) Instance 1, $\epsilon = 1$

(b) Instance 1, $\epsilon = 0.1$

(c) Instance 2, $\epsilon = 1$

(d) Instance 2, $\epsilon = 0.1$

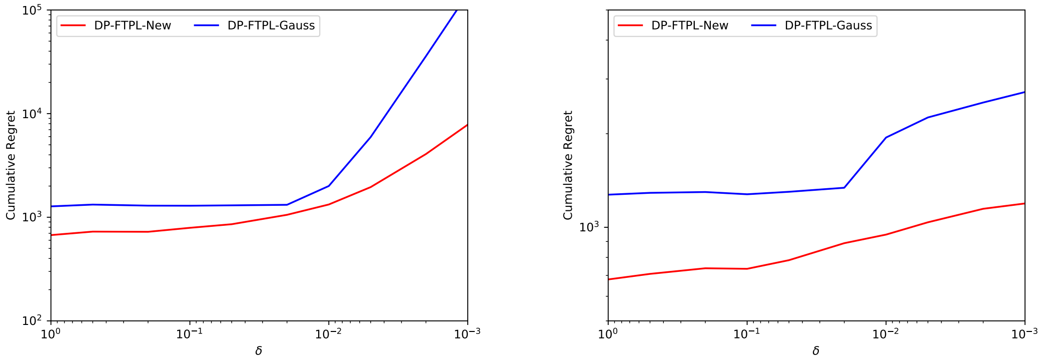
Figure 5: Experiments for $(\epsilon, 0)$-differential privacy

significantly. This accords with our analysis, i.e., the extra regret term of DP-FTPL-New is the smallest. As a contrast, another optimal algorithm DP-SE is an elimination-based algorithm and suffers from a large constant factor in its regret upper bound. Therefore it behaves even worse than the non-optimal UCB-based algorithms, e.g., DP-UCB and DP-UCB-BOUND.

Finally, we compare the performance of the same algorithm under different ϵ or δ to see how ϵ and δ influence the algorithms' regret. Here we use DP-FTPL-Gauss and DP-FTPL-New as examples, and consider their performances on Instance 1 with $T = 10^6$.

In Fig. 6(a), we fix $\epsilon = 0$ and in Fig. 6(b), we fix $\epsilon = 0.1$. We can see that when $\epsilon = 0$, the regret of both DP-FTPL-Gauss and DP-FTPL-New do not increase a lot when δ decreases from 1 to 0.01. However, when δ is less than 0.01, then the regret of DP-FTPL-Gauss grows much faster than DP-FTPL-New. In fact, from this log-log figure, one can see that the regret of DP-FTPL-Gauss is about $\frac{1}{\delta^2}$, while the regret of DP-FTPL-New is about $\frac{1}{\delta}$. This accords with our analysis, since when $\epsilon = 0$, the extra regret of DP-FTPL-Gauss is $O(\frac{N}{\delta^2})$, while the extra regret of DP-FTPL-New is $O(\frac{N}{\delta})$.

When $\epsilon = 0.1$, we can see that the curve of DP-FTPL-Gauss can be divided into three parts. This also accords with our analysis, i.e., when $\delta \in (0.02, 1)$, the major term in the



(a) Instance 1, $\epsilon = 0$

(b) Instance 1, $\epsilon = 0.1$

Figure 6: Experiments for fixed ϵ

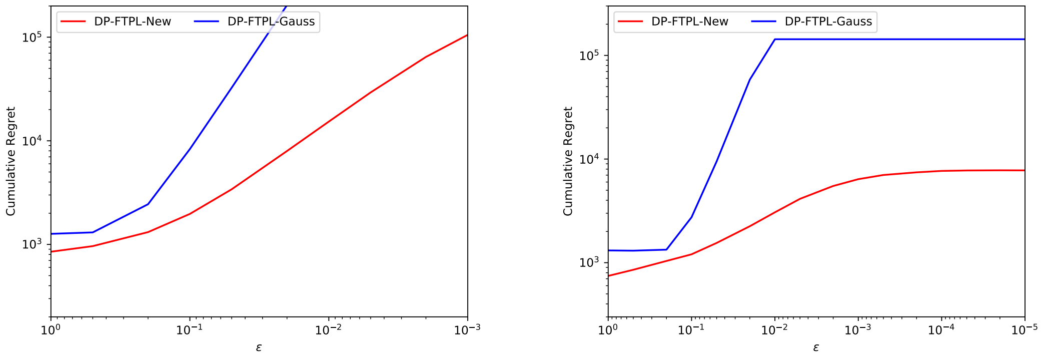
regret bound is $O(\sum_i \frac{\log T}{\Delta_i})$, hence the regret does not increase at all; when $\delta \in (0.01, 0.02)$, the major term in the regret bound is $O(\frac{N}{\delta^2})$, and the regret increases with rate about $\frac{1}{\delta^2}$; when $\delta < 0.01$, the major term in the regret bound becomes $O(\frac{N \log \frac{1}{\delta}}{\epsilon^2})$, and therefore the increasing rate of the regret becomes much slower than before. On the other hand, the extra regret of DP-FTPL-New is $O(\frac{N}{\epsilon} \log \frac{T(e^\epsilon - 1) + T\delta}{(e^\epsilon - 1) + T\delta})$, and this term becomes $O(N \log(1 + \frac{1}{\delta}))$ in this figure. Therefore, the regret of DP-FTPL-New is smaller than DP-FTPL-Gauss, and the curve of DP-FTPL-New is more smooth than DP-FTPL-Gauss.

Then we fix the value δ: in Fig. 7(a), we fix $\delta = 10^{-10}$ (since DP-FTPL-Gauss cannot deal with the case $\delta = 0$, we choose an extreme small δ instead of 0) and In Fig. 7(b), we fix $\delta = 0.001$.

We can see similar phenomena in Fig. 7(a) and Fig. 7(b) (as the case when we fix ϵ). When $\delta = 10^{-10}$, both the regret of DP-FTPL-Gauss and the regret of DP-FTPL-New do not increase a lot at the beginning. Then the regret of DP-FTPL-Gauss increases with rate about $\frac{1}{\epsilon^2}$, while the regret of DP-FTPL-Gauss increases with rate about $\frac{1}{\epsilon}$, since the extra regret in their corresponding regret upper bounds are $O(\frac{N \log \frac{1}{\delta}}{\epsilon^2})$ and $O(\frac{N \log T}{\epsilon})$, respectively. When $\delta = 0.001$, the curve of DP-FTPL-Gauss can also be divided into three parts, and the only difference here is that in the third part, the extra regret becomes $O(\frac{N}{\delta^2})$, and does not increase at all (as ϵ decreases). As for DP-FTPL-New, its curve is more smooth, and the regret almost stops increasing after $\epsilon < 10^{-4}$. This is because that its regret in this figure is about $O(\frac{N}{\epsilon} \log(1 + \frac{e^\epsilon - 1}{\delta}))$. When $\epsilon < 10^{-4}$, $\frac{N}{\epsilon} \log(1 + \frac{e^\epsilon - 1}{\delta}) \approx \frac{N}{\epsilon} \log(1 + \frac{\epsilon}{\delta}) \approx \frac{N}{\epsilon} \cdot \frac{\epsilon}{\delta} = \frac{N}{\delta}$, which does not increase as ϵ decreases.

9. Conclusions

In this paper, we study the algorithms that guarantee (ϵ, δ)-differential privacy or (ϵ, δ)-global differential privacy in MAB problems. We first adapt the famous Thompson Sampling policies to protect (ϵ, δ)-differential privacy (i.e., DP-FTPL-Gauss and DP-FTPL-Beta), and propose their regret upper bounds. Then we design a new perturbed distribution that suits the (ϵ, δ)-differential privacy setting well, and show that using this perturbed



(a) Instance 1, $\delta = 10^{-10}$

(b) Instance 1, $\delta = 0.001$

Figure 7: Experiments for fixed δ

distribution in FTPL framework (i.e., DP-FTPL-New) can significantly reduce the extra regret for privacy guarantee. This new kind of perturbed distribution, on the other hand, can also be used to protect (ϵ, δ)-global differential privacy. Based on this fact, we design the GDP-Elim-New algorithm, and give its regret upper bound. We also prove a regret lower bound for algorithms that protect (ϵ, δ)-global differential privacy, and this lower bound matches (in order) with the regret upper bound of GDP-Elim-New, indicating that the upper/lower bounds are tight. Compared to existing researches that only work in the case $\epsilon > 0$ and $\delta = 0$ (or only work in the case $\epsilon > 0$ and $\delta > 0$), our results work for any (ϵ, δ) as long as one of them is not zero. This means that our results are more general.

Acknowledgements

This work was supported by NSFC Projects (Nos. 62350080, 62106122, 92248303, 92370124, 62350080, 62276149, U2341228, 62076147), Tsinghua Institute for Guo Qiang, and the High Performance Computing Center, Tsinghua University. J. Zhu was also supported by the XPlorer Prize.

Appendix

Appendix A. Proof of Theorem 11

Theorem 11 *DP-FTPL-Gauss guarantees (ϵ, δ)-differential privacy, and its cumulative regret satisfies*

$$Reg(T) \leq \sum_{i=2}^{N} \max\left\{\frac{2\Delta_i \log T}{(\Delta_i - 2\lambda)^2}, N_G^* \Delta_i\right\} + \Theta\left(\frac{N}{\lambda^4}\right)$$

for any $\lambda < \frac{1}{2}\Delta_{\min}$.

Proof (regret part) In this proof, we need to use the following three facts.

Fact 5 *(Chernoff-Hoeffding Inequality, Hoeffding (1963))*

$$F^{Bino}_{n,p}((p-\lambda)n) \leq \exp\left(-2n\lambda^2\right),$$

where $F^{Bino}_{n,p}$ denotes the cumulative distribution function of Binomial distribution with parameter (n,p).

By Fact 5, one could get that: for any fixed n,

$$\Pr\left[\hat{\mu}_i(t) - \mu_i \leq -\sqrt{\frac{\log T}{N_i(t)}}, N_i(t) = n\right] \leq F^{Bino}_{n,\mu_i}\left(\left(\mu_i - \sqrt{\frac{\log T}{n}}\right)n\right) \leq \frac{1}{T^2}, \qquad (17)$$

and similarly,

$$\Pr\left[\hat{\mu}_i(t) - \mu_i \geq \sqrt{\frac{\log T}{N_i(t)}}, N_i(t) = n\right] \leq F^{Bino}_{n,1-\mu_i}\left(\left(1 - \mu_i - \sqrt{\frac{\log T}{n}}\right)n\right) \leq \frac{1}{T^2}. \qquad (18)$$

These equations are widely used in our analysis (e.g., to prove Fact 3).

Fact 6 *(Feller, 2008) For any fixed mean-variance pair $(\hat{\mu}_i(t), \frac{2}{N_i(t)})$, if $\theta_i(t)$ is drawn from Gaussian distribution $\mathcal{N}(\hat{\mu}_i(t), \frac{2}{N_i(t)})$, then*

$$\Pr\left[\theta_i(t) \geq \hat{\mu}_i(t) + \sqrt{\frac{2\log T}{N_i(t)}}\right] \leq \frac{1}{T}.$$

Fact 7 *(Birnbaum, 1942)*

$$\forall x > 0, \phi(x,0,1) \geq \frac{\sqrt{2/\pi}}{x + \sqrt{x^2+4}} \exp\left(\frac{-x^2}{2}\right).$$

Recall that $N_i(t), R_i(t)$ denote the value of N_i, R_i at the beginning of time step t, and $\hat{\mu}_i(t) = \frac{R_i(t)}{N_i(t)}$ is the empirical mean of arm i at time t. Then similar as prior works (Agrawal and Goyal, 2013), we can define the following events:

$$\begin{aligned}
\mathcal{A}_i(t) &= \{i(t) = i, \hat{\mu}_i(t) \geq \mu_i + \lambda\};\\
\mathcal{B}_i(t) &= \{i(t) = i, \hat{\mu}_i(t) < \mu_i + \lambda, \theta_i(t) \geq \mu_1 - \lambda\};\\
\mathcal{C}_i(t) &= \{i(t) = i, \theta_i(t) < \mu_1 - \lambda\}.
\end{aligned}$$

By definitions, we have that

$$\begin{aligned}
\mathbb{E}[N_i(T)] &= N^*_G + \sum_{t=T^*}^{T} \mathbb{E}[\mathbb{I}[i(t) = i]]\\
&= N^*_G + \sum_{t=T^*}^{T} \mathbb{E}[\mathbb{I}[\mathcal{A}_i(t) \cup \mathcal{B}_i(t) \cup \mathcal{C}_i(t)]]
\end{aligned}$$

$$
\begin{aligned}
&\leq N_G^* + \sum_{t=T^*}^{T} \left(\mathbb{E}[\mathbb{I}[\mathcal{A}_i(t)]] + \mathbb{E}[\mathbb{I}[\mathcal{B}_i(t)]] + \mathbb{E}[\mathbb{I}[\mathcal{C}_i(t)]]\right) \\
&\leq \sum_{t=T^*}^{T} \mathbb{E}[\mathbb{I}[\mathcal{A}_i(t)]] + \left(N_G^* + \sum_{t=T^*}^{T} \mathbb{E}[\mathbb{I}[\mathcal{B}_i(t)]]\right) + \sum_{t=T^*}^{T} \mathbb{E}[\mathbb{I}[\mathcal{C}_i(t)]],
\end{aligned}
$$

where $T^* = N \cdot N_G^* + 1$ is the first time step after the start phase.

For the term $\mathbb{E}[\mathbb{I}[\mathcal{A}_i(t)]]$, denote t_n as the time step that $N_i(t_n) = n-1$ and $i(t_n) = i$, i.e., the time step we choose arm i for the n-th time (also denote $t_0 = 0$), then we have that

$$
\begin{aligned}
\sum_{t=T^*}^{T} \mathbb{E}[\mathbb{I}[\mathcal{A}_i(t)]] &= \mathbb{E}\left[\sum_{t=T^*}^{T} \mathbb{I}[\mathcal{A}_i(t)]\right] \\
&\leq \mathbb{E}\left[\sum_{n=0}^{\infty}\sum_{t=t_n}^{t_{n+1}} \mathbb{I}[\mathcal{A}_i(t)]\right] \\
&= \mathbb{E}\left[\sum_{n=0}^{\infty}\sum_{t=t_n}^{t_{n+1}} \mathbb{I}[i(t) = i, \hat{\mu}_i(t) \geq \mu_i + \lambda]\right] \\
&= \mathbb{E}\left[\sum_{n=0}^{\infty} \mathbb{I}[\hat{\mu}_i(t_{n+1}) \geq \mu_i + \lambda]\right] \\
&= \sum_{n=0}^{\infty} \Pr[\hat{\mu}_i(t) \geq \mu_i + \lambda | N_i(t) = n] \\
&\leq \sum_{n=0}^{\infty} \exp\left(-2n\lambda^2\right) \qquad (19) \\
&\leq \frac{1}{1 - \exp(-2\lambda^2)} \\
&\leq \frac{1}{\frac{2}{e}\lambda^2} \\
&= \frac{e}{2\lambda^2},
\end{aligned}
$$

where Eq. (19) comes from Chernoff-Hoeffding Inequality (Fact 5).

As for the second term, when $N_i(t) \geq L_i(T) \triangleq \frac{2\log T}{(\mu_1 - \mu_i - 2\lambda)^2}$, we must have that (Eq. (20) is given by Fact 6)

$$
\begin{aligned}
&\Pr\left[\theta_i(t) \geq \mu_1 - \lambda; \hat{\mu}_i(t) \leq \mu_i + \lambda\right] \\
\leq\ &\Pr\left[\theta_i(t) \geq \hat{\mu}_i(t) + (\mu_1 - \mu_i) - 2\lambda\right] \\
=\ &\Pr\left[\theta_i(t) \geq \hat{\mu}_i(t) + \sqrt{\frac{2\log T}{L_i(t)}}\right] \\
\leq\ &\Pr\left[\theta_i(t) \geq \hat{\mu}_i(t) + \sqrt{\frac{2\log T}{N_i(t)}}\right]
\end{aligned}
$$

$$
\begin{aligned}
&= \sum_{n\geq L_i(T)} \sum_{\mu\in\{0/n,1/n,\cdots,n/n\}} \Pr[N_i(t)=n,\hat{\mu}_i(t)=\mu] \\
&\qquad\qquad \cdot \Pr\left[\theta_i(t)\geq\hat{\mu}_i(t)+\sqrt{\frac{2\log T}{N_i(t)}} \mid N_i(t)=n,\hat{\mu}_i(t)=\mu\right] \\
&\leq \sum_{n\geq L_i(T)} \sum_{\mu\in\{0/n,1/n,\cdots,n/n\}} \Pr[N_i(t)=n,\hat{\mu}_i(t)=\mu]\cdot\frac{1}{T} \qquad (20) \\
&\leq \frac{1}{T}.
\end{aligned}
$$

Therefore, if $L_i(T)\leq N_G^*$, then

$$
\begin{aligned}
\sum_{t=T^*}^{T}\mathbb{E}[\mathbb{I}[\mathcal{B}_i(t)]] &= \mathbb{E}\left[\sum_{t=T^*}^{T}\mathbb{I}[\mathcal{B}_i(t)]\right] \\
&\leq \mathbb{E}\left[\sum_{t=t_{L_i(T)+1}}^{T}\mathbb{I}[\mathcal{B}_i(t)]\right] \\
&= \mathbb{E}\left[\sum_{t=t_{L_i(T)+1}}^{T}\mathbb{I}[i(t)=i,\hat{\mu}_i(t)<\mu_i+\lambda,\theta_i(t)\geq\mu_1-\lambda]\right] \\
&\leq \mathbb{E}\left[\sum_{t=t_{L_i(T)+1}}^{T}\mathbb{I}[\theta_i(t)\geq\mu_1-\lambda;\hat{\mu}_i(t)\leq\mu_i+\lambda]\right] \\
&\leq \sum_{t=t_{L_i(T)+1}}^{T}\Pr[\theta_i(t)\geq\mu_1-\lambda;\hat{\mu}_i(t)\leq\mu_i+\lambda] \\
&\leq \sum_{t=t_{L_i(T)+1}}^{T}\frac{1}{T} \\
&\leq 1.
\end{aligned}
$$

Otherwise if $L_i(T)>N_G^*$, then

$$
\begin{aligned}
\sum_{t=T^*}^{T}\mathbb{E}[\mathbb{I}[\mathcal{B}_i(t)]] &= \mathbb{E}\left[\sum_{t=T^*}^{T}\mathbb{I}[\mathcal{B}_i(t)]\right] \\
&\leq \mathbb{E}\left[\sum_{t=T^*}^{t_{L_i(T)}}\mathbb{I}[\mathcal{B}_i(t)]+\sum_{t=t_{L_i(T)+1}}^{T}\mathbb{I}[\mathcal{B}_i(t)]\right] \\
&= \mathbb{E}\left[\sum_{n=N_G^*}^{L_i(T)}\sum_{t=t_n}^{t_{n+1}}\mathbb{I}[\mathcal{B}_i(t)]\right]+\mathbb{E}\left[\sum_{t=t_{L_i(T)+1}}^{T}\mathbb{I}[\mathcal{B}_i(t)]\right] \\
&\leq (L_i(T)-N_G^*)+\mathbb{E}\left[\sum_{t=t_{L_i(T)+1}}^{T}\mathbb{I}[\mathcal{B}_i(t)]\right]
\end{aligned}
$$

$$
\begin{aligned}
&= (L_i(T) - N_G^*) + \mathbb{E}\left[\sum_{t=t_{L_i(T)+1}}^{T} \mathbb{I}[i(t) = i, \hat{\mu}_i(t) < \mu_i + \lambda, \theta_i(t) \geq \mu_1 - \lambda]\right] \\
&\leq (L_i(T) - N_G^*) + \mathbb{E}\left[\sum_{t=t_{L_i(T)+1}}^{T} \mathbb{I}[\theta_i(t) \geq \mu_1 - \lambda; \hat{\mu}_i(t) \leq \mu_i + \lambda]\right] \\
&\leq (L_i(T) - N_G^*) + \sum_{t=t_{L_i(T)+1}}^{T} \Pr[\theta_i(t) \geq \mu_1 - \lambda; \hat{\mu}_i(t) \leq \mu_i + \lambda] \\
&\leq (L_i(T) - N_G^*) + \sum_{t=t_{L_i(T)+1}}^{T} \frac{1}{T} \\
&\leq (L_i(T) - N_G^*) + 1.
\end{aligned}
$$

These imply that

$$N_G^* + \sum_{t=T^*}^{T} \mathbb{E}[\mathbb{I}[\mathcal{B}_i(t)]] \leq \max\{L_i(T), N_G^*\} + 1.$$

Then we come to the third term, and we will use the following lemma, which is similar with Lemma 1 in the work of Agrawal and Goyal (2013). We defer the proof of Lemma 20 to Appendix A.1.

Lemma 20 *Denote $p_t = \Pr[\theta_1(t) \geq \mu_1 - \lambda | \mathcal{F}_{t-1}]$, then we have that*

$$\Pr[i(t) = i, \theta_i(t) \leq \mu_1 - \lambda | \mathcal{F}_{t-1}] \leq \frac{1 - p_t}{p_t} \Pr[i(t) = 1, \theta_i(t) \leq \mu_1 - \lambda | \mathcal{F}_{t-1}].$$

According to the analysis in the work of Agrawal and Goyal (2013), by Lemma 20, we have that

$$\sum_{t=T^*}^{T} \mathbb{E}[\mathbb{I}[\mathcal{C}_i(t)]] \leq \sum_{n=N_G^*}^{\infty} \mathbb{E}\left[\frac{1}{p_n} - 1\right],$$

where p_n denotes the random probability (due to the randomness on observations of arm 1) that $\theta_1(t) \geq \mu_1 - \lambda$ when there are totally n observations on arm 1, i.e., $N_i(t) = n$.

Note that the observations on arm 1 follow a Binomial distribution with parameters n, μ_1. Denote $f_{n,p}^{Bino}$ the probability mass function of Binomial distribution with parameters n, p, and $\phi(x, \mu, \sigma^2) = \Pr_{X \sim \mathcal{N}(\mu, \sigma^2)}[X \geq x]$, then the term $\mathbb{E}\left[\frac{1}{p_n}\right]$ can be written as

$$\mathbb{E}\left[\frac{1}{p_n}\right] = \sum_{s=0}^{n} \frac{f_{n,\mu_1}^{Bino}(s)}{\phi(\mu_1 - \lambda, \frac{s}{n}, \frac{2}{n})}.$$

Now we divide this sum into three parts: $S_H = \{s : s > (\mu_1 - \frac{\lambda}{2})n\}$, $S_M = \{s : (\mu_1 - \lambda)n \leq s \leq (\mu_1 - \frac{\lambda}{2})n\}$, and $S_L = \{s : s < (\mu_1 - \lambda)n\}$.

For the part S_H, when $n \leq \frac{8}{\lambda^2}$, we always have that $\phi(\mu_1 - \lambda, \frac{s}{n}, \frac{2}{n}) \geq \frac{1}{2}$, therefore

$$\sum_{s \in S_H} \frac{f^{Bino}_{n,\mu_1}(s)}{\phi(\mu_1 - \lambda, \frac{s}{n}, \frac{2}{n})} \leq 2.$$

When $n > \frac{8}{\lambda^2}$, we have that $\phi(\mu_1 - \lambda, \frac{s}{n}, \frac{2}{n}) \geq 1 - \exp(-n\frac{\lambda^2}{2})$ (for s in S_H), therefore

$$\sum_{s \in S_H} \frac{f^{Bino}_{n,\mu_1}(s)}{\phi(\mu_1 - \lambda, \frac{s}{n}, \frac{2}{n})} \leq \frac{1}{1 - \exp(-n\frac{\lambda^2}{2})}.$$

For the part S_M, we always have that $\sum_{s \in S_M} f^{Bino}_{n,\mu_1}(s) \leq \Pr[\hat{\mu}_1(t) \leq \mu_1 - \frac{\lambda}{2} | N_i(t) = n] \leq \exp(-n\frac{\lambda^2}{2})$ (by Fact 5), and $\phi(\mu_1 - \lambda, \frac{s}{n}, \frac{2}{n}) \geq \frac{1}{2}$, therefore

$$\sum_{s \in S_M} \frac{f^{Bino}_{n,\mu_1}(s)}{\phi(\mu_1 - \lambda, \frac{s}{n}, \frac{2}{n})} \leq 2 \exp\left(-n\frac{\lambda^2}{2}\right).$$

As for the part S_L, we have that

$$\begin{aligned}
\sum_{s \in S_L} \frac{f^{Bino}_{n,\mu_1}(s)}{\phi(\mu_1 - \lambda, \frac{s}{n}, \frac{2}{n})} &= \sum_{s \in S_L} \frac{\frac{n!}{s!(n-s)!}\mu_1^s(1-\mu_1)^{n-s}}{\phi(\mu_1 - \lambda, \frac{s}{n}, \frac{2}{n})} \\
&\leq \sum_{s \in S_L} \frac{\frac{\sqrt{2\pi en}(\frac{n}{e})^n \mu_1^s (1-\mu_1)^{n-s}}{\sqrt{2\pi s}(\frac{s}{e})^s \sqrt{2\pi(n-s)}(\frac{n-s}{e})^{n-s}}}{\phi(\mu_1 - \lambda, \frac{s}{n}, \frac{2}{n})} \qquad (21) \\
&= \sum_{s \in S_L} \frac{\sqrt{\frac{en}{2\pi s(n-s)}} \frac{n^n \mu_1^s (1-\mu_1)^{n-s}}{s^s (n-s)^{n-s}}}{\phi(\mu_1 - \lambda, \frac{s}{n}, \frac{2}{n})} \\
&\leq \sum_{s \in S_L} \sqrt{\frac{e}{2\pi}} \frac{\frac{(\mu_1)^s (1-\mu_1)^{n-s}}{(\frac{s}{n})^s (\frac{n-s}{n})^{n-s}}}{\phi(\mu_1 - \lambda, \frac{s}{n}, \frac{2}{n})} \\
&= \sum_{s \in S_L} \sqrt{\frac{e}{2\pi}} \frac{\left(\frac{(\mu_1)^{\frac{s}{n}} (1-\mu_1)^{\frac{n-s}{n}}}{(\frac{s}{n})^{\frac{s}{n}} (\frac{n-s}{n})^{\frac{n-s}{n}}}\right)^n}{\phi(\mu_1 - \lambda, \frac{s}{n}, \frac{2}{n})} \\
&= \sum_{s \in S_L} \sqrt{\frac{e}{2\pi}} \frac{\exp(-nKL(\frac{s}{n}, \mu_1))}{\phi(\mu_1 - \lambda, \frac{s}{n}, \frac{2}{n})} \\
&\leq \sum_{s \in S_L} \sqrt{\frac{e}{2\pi}} \frac{\exp(-\frac{n}{2}\left(\mu_1 - \frac{s}{n}\right)^2)}{\phi(\mu_1 - \lambda, \frac{s}{n}, \frac{2}{n})} \qquad (22) \\
&\leq \sum_{s \in S_L} \sqrt{\frac{e}{2\pi}} \frac{\exp(-\frac{n}{2}\left(\mu_1 - \frac{s}{n}\right)^2)}{\phi(\mu_1 - \lambda - \frac{s}{n}, 0, \frac{2}{n})} \\
&\leq \sum_{s \in S_L} \sqrt{\frac{e}{2\pi}} \frac{\exp(-\frac{n}{2}\left(\mu_1 - \frac{s}{n}\right)^2)}{\phi(\sqrt{\frac{n}{2}}\left(\mu_1 - \lambda - \frac{s}{n}\right), 0, 1)}
\end{aligned}$$

$$
\begin{aligned}
&\leq \sum_{s\in S_L} \sqrt{\frac{e}{2\pi}} \frac{\exp(-\frac{n}{2}\left(\mu_1 - \frac{s}{n}\right)^2)}{\frac{\sqrt{2/\pi}}{\sqrt{\frac{n}{2}}\left(\mu_1-\lambda-\frac{s}{n}\right)+\sqrt{\frac{n}{2}\left(\mu_1-\lambda-\frac{s}{n}\right)^2+4}} \exp(-\frac{n}{2}\left(\mu_1-\lambda-\frac{s}{n}\right)^2)} \qquad (23) \\
&\leq \sum_{s\in S_L} \sqrt{\frac{e}{2\pi}} \frac{\exp(-\frac{n}{2}\left(\mu_1 - \frac{s}{n}\right)^2)}{\frac{\sqrt{2/\pi}}{\sqrt{2n}\left(\mu_1-\lambda-\frac{s}{n}\right)+2} \exp(-\frac{n}{2}\left(\mu_1-\lambda-\frac{s}{n}\right)^2)} \\
&\leq \sum_{s\in S_L} \sqrt{e} \frac{\exp(-\frac{n}{2}\left(\mu_1 - \frac{s}{n}\right)^2)}{\frac{1}{\sqrt{n}+1} \exp(-\frac{n}{2}\left(\mu_1-\lambda-\frac{s}{n}\right)^2)} \\
&\leq \sum_{s\in S_L} \sqrt{e}(1+n) \frac{\exp(-\frac{n}{2}\left(\mu_1 - \frac{s}{n}\right)^2)}{\exp(-\frac{n}{2}\left(\mu_1-\lambda-\frac{s}{n}\right)^2)} \\
&\leq \sqrt{e}(1+n) \sum_{s\in S_L} \exp\left(-\frac{n}{2}(\mu_1 - \frac{s}{n})\lambda\right) \\
&= \sqrt{e}(1+n) \sum_{s\in S_L} \exp\left(-\frac{\lambda}{2}(n\mu_1 - s)\right) \\
&\leq \sqrt{e}(1+n)\left(\exp\left(-\frac{\lambda}{2}n\lambda\right) + \exp\left(-\frac{\lambda}{2}(n\lambda+1)\right) + \cdots\right) \\
&= \sqrt{e}(1+n)\exp\left(-\frac{\lambda}{2}n\lambda\right)\left(1 + \exp\left(-\frac{\lambda}{2}\right) + \cdots\right) \\
&= \sqrt{e}(1+n)\exp\left(-\frac{\lambda}{2}n\lambda\right)\frac{1}{1-\exp\left(-\frac{\lambda}{2}\right)} \\
&\leq 2\sqrt{e}(1+n)\exp\left(-n\frac{\lambda^2}{2}\right),
\end{aligned}
$$

where $KL(p,q) = p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q}$ is the KL-divergence between p and q. Eq. (21) comes from Stirling's approximation (Fact 1), Eq. (22) comes from Pinsker's Inequality (Fact 2) and Eq. (23) comes from the inequality on Millo's ratio (Fact 7).

Therefore, we have that:

$$
\begin{aligned}
\sum_{t=T^*}^{T} \mathbb{E}[\mathbb{I}[\mathcal{C}_i(t)]] &\leq \sum_{n=N_G^*}^{\infty} \mathbb{E}\left[\frac{1}{p_n} - 1\right] \\
&\leq \sum_{n=1}^{\infty} \mathbb{E}\left[\frac{1}{p_n} - 1\right] \\
&= \sum_{n=1}^{\infty} 2\sqrt{e}(1+n)\exp\left(-n\frac{\lambda^2}{2}\right) + \sum_{n=1}^{\infty} 2\exp\left(-n\frac{\lambda^2}{2}\right) \\
&\quad + \sum_{n=1}^{\frac{8}{\lambda^2}}(2-1) + \sum_{n=\frac{8}{\lambda^2}}^{\infty}\left(\frac{1}{1-\exp(-n\frac{\lambda^2}{2})} - 1\right) \\
&= \Theta\left(\frac{1}{\lambda^4}\right).
\end{aligned}
$$

Summing over the three terms, the cumulative regret of DP-FTPL-Gauss is upper bounded by

$$\begin{aligned}\sum_{i=2}^{N}\mathbb{E}[N_i(T)]\Delta_i &\leq \sum_{i}\left(\sum_{t=T^*}^{T}\mathbb{E}[\mathbb{I}[\mathcal{A}_i(t)]]+\left(N_G^*+\sum_{t=T^*}^{T}\mathbb{E}[\mathbb{I}[\mathcal{B}_i(t)]]\right)+\sum_{t=T^*}^{T}\mathbb{E}[\mathbb{I}[\mathcal{C}_i(t)]]\right)\Delta_i\\ &\leq \sum_{i=2}^{N}\max\left\{\frac{2\log T\Delta_i}{(\Delta_i-2\lambda)^2},N_G^*\Delta_i\right\}+\Theta\left(\frac{N}{\lambda^4}\right).\end{aligned}$$

■

A.1 Proof of Lemma 20

Lemma 19 *Denote* $p_t=\Pr[\theta_1(t)\geq\mu_1-\lambda|\mathcal{F}_{t-1}]$, *then we have that*

$$\Pr[i(t)=i,\theta_i(t)\leq\mu_1-\lambda|\mathcal{F}_{t-1}]\leq\frac{1-p_t}{p_t}\Pr[i(t)=1,\theta_i(t)\leq\mu_1-\lambda|\mathcal{F}_{t-1}].$$

Proof We first define event $\mathcal{M}_i(t)$ as follows:

$$\mathcal{M}_i(t)\triangleq\{\forall 2\leq j\leq N,\theta_j(t)\leq\theta_i(t)\}\cap\{\theta_i(t)\leq\mu_1-\lambda\}.$$

Then we have that

$$\begin{aligned}\{i(t)=i,\theta_i(t)\leq\mu_1-\lambda\} &\subseteq \mathcal{M}_i(t)\cap\{\theta_1(t)\leq\mu_1-\lambda\},\\ \{i(t)=1,\theta_i(t)\leq\mu_1-\lambda\} &\supseteq \mathcal{M}_i(t)\cap\{\theta_1(t)\geq\mu_1-\lambda\}.\end{aligned}$$

Also note that $\{\theta_1(t)\geq\mu_1-\lambda\}$ (or $\{\theta_1(t)\leq\mu_1-\lambda\}$) and $M_i(t)$ are independent events, therefore, we have that

$$\begin{aligned}\Pr[i(t)=i,\theta_i(t)\leq\mu_1-\lambda|\mathcal{F}_{t-1}] &\leq (1-p_t)\Pr[\mathcal{M}_i(t)|\mathcal{F}_{t-1}],\\ \Pr[i(t)=1,\theta_i(t)\leq\mu_1-\lambda|\mathcal{F}_{t-1}] &\geq p_t\Pr[\mathcal{M}_i(t)|\mathcal{F}_{t-1}].\end{aligned}$$

This implies that

$$\Pr[i(t)=i,\theta_i(t)\leq\mu_1-\lambda|\mathcal{F}_{t-1}]\leq\frac{1-p_t}{p_t}\Pr[i(t)=1,\theta_i(t)\leq\mu_1-\lambda|\mathcal{F}_{t-1}].$$

■

Appendix B. Proof of Theorem 13

Theorem 13 *DP-FTPL-Beta guarantees* (ϵ,δ)*-differential privacy, and its cumulative regret satisfies*

$$Reg(T)\leq\sum_{i=2}^{N}\max\left\{\frac{5\Delta_i\log T}{2(\Delta_i-5/2\lambda)^2},N_B^*\Delta_i\right\}+\Theta\left(\frac{N}{\lambda^4}\right)$$

for any $\lambda<\frac{2}{5}\Delta_{\min}$.

Proof (regret part) Recall that $a_i(t), b_i(t), k_i(t)$ are the value of a_i, b_i, k_i at the beginning of time step t, and also denote $N_i(t) = a_i(t) + b_i(t) - 2$, $M_i(t) = a_i(t) + b_i(t) + 2k_i(t) - 2$, $\mu'_i(t) = \frac{\mu_i N_i(t)+k_i(t)}{M_i(t)}$, $\hat{\mu}'_i(t) = \frac{a_i(t)+k_i(t)-1}{M_i(t)}$.

Then in this proof, the following fact holds (note that $\theta_i(t)$ is drawn from the Beta distribution $\mathcal{B}(a_i(t) + k_i(t), b_i(t) + k_i(t))$).

Fact 8 *(Agrawal and Goyal, 2013) For any fixed pair $(\hat{\mu}'_i(t), M_i(t))$, we have that*

$$\Pr\left[\theta_i(t) \geq \hat{\mu}'_i(t) + \sqrt{\frac{2\log T}{M_i(t)}}\right] \leq \frac{1}{T}.$$

We can also define the following events:

$$\begin{aligned}
\mathcal{A}_i(t) &= \{i(t) = i, \hat{\mu}'_i(t) \geq \mu'_i(t) + \lambda\};\\
\mathcal{B}_i(t) &= \{i(t) = i, \hat{\mu}'_i(t) < \mu'_i(t) + \lambda, \theta_i(t) \geq \mu'_1(t) - \lambda\};\\
\mathcal{C}_i(t) &= \{i(t) = i, \theta_i(t) < \mu'_1(t) - \lambda\}.
\end{aligned}$$

Similar as the proof of Theorem 11, by definitions, we have that

$$\mathbb{E}[N_i(T)] \leq \sum_{t=T^*}^{T} \mathbb{E}[\mathbb{I}[\mathcal{A}_i(t)]] + \left(N_B^* + \sum_{t=T^*}^{T} \mathbb{E}[\mathbb{I}[\mathcal{B}_i(t)]]\right) + \sum_{t=T^*}^{T} \mathbb{E}[\mathbb{I}[\mathcal{C}_i(t)]],$$

where $T^* = N \cdot N_B^* + 1$ is the first time step after the start phase.

For the term $\mathbb{E}[\mathbb{I}[\mathcal{A}_i(t)]]$, denote t_n as the time step that $N_i(t_n) = n - 1$ and $i(t_n) = i$, i.e., the time step we choose arm i for the n-th times (also denote $t_0 = 0$), then we have that (note that different with the proof of Theorem 11, here $\mathcal{A}_i(t) = \{i(t) = i, \hat{\mu}'_i(t) \geq \mu'_i(t) + \lambda\}$ but not $\{i(t) = i, \hat{\mu}_i(t) \geq \mu_i(t) + \lambda\}$)

$$\begin{aligned}
\sum_{t=T^*}^{T} \mathbb{E}[\mathbb{I}[\mathcal{A}_i(t)]] &= \mathbb{E}\left[\sum_{t=T^*}^{T} \mathbb{I}[\mathcal{A}_i(t)]\right]\\
&\leq \mathbb{E}\left[\sum_{n=0}^{\infty}\sum_{t=t_n}^{t_{n+1}} \mathbb{I}[\mathcal{A}_i(t)]\right]\\
&= \mathbb{E}\left[\sum_{n=0}^{\infty}\sum_{t=t_n}^{t_{n+1}} \mathbb{I}[i(t) = i, \hat{\mu}'_i(t) \geq \mu'_i(t) + \lambda]\right]\\
&= \mathbb{E}\left[\sum_{n=0}^{\infty} \mathbb{I}[\hat{\mu}'_i(t_{n+1}) \geq \mu'_i(t) + \lambda]\right]\\
&= \sum_{n=0}^{\infty} \Pr[\hat{\mu}'_i(t) \geq \mu'_i(t) + \lambda | N_i(t) = n]\\
&= \sum_{n=0}^{\infty} \Pr\left[\frac{a_i(t) + k_i(t) - 1}{M_i(t)} \geq \frac{\mu_i N_i(t) + k_i(t)}{M_i(t)} + \lambda | N_i(t) = n\right]\\
&= \sum_{n=0}^{\infty} \Pr\left[\frac{a_i(t) - 1}{N_i(t)} \geq \mu_i + \frac{M_i(t)}{N_i(t)}\lambda | N_i(t) = n\right]
\end{aligned}$$

$$\begin{aligned}
&\leq \sum_{n=0}^{\infty} \Pr\left[\frac{a_i(t)-1}{N_i(t)} \geq \mu_i + \frac{5}{4}\lambda | N_i(t) = n\right] \\
&\leq \sum_{n=0}^{\infty} \exp\left(-\frac{25}{8}n\lambda^2\right) \qquad (24) \\
&\leq \frac{1}{1-\exp(-\frac{25}{8}\lambda^2)} \\
&\leq \frac{1}{\frac{25}{8e}\lambda^2} \\
&= \frac{8e}{25\lambda^2},
\end{aligned}$$

where Eq. (24) is because of Chernoff-Hoeffding Inequality (Fact 5).

For the second term, by Fact 8, when $N_i(t) \geq L_i(T) = \frac{5\log T}{2(\mu_1-\mu_i-\frac{5}{2}\lambda)^2}$, we must have that

$$\begin{aligned}
& \Pr\left[\theta_i(t) \geq \mu_1'(t) - \lambda; \hat{\mu}_i'(t) \leq \mu_i'(t) + \lambda\right] \\
\leq\ & \Pr\left[\theta_i(t) \geq \hat{\mu}_i'(t) + (\mu_1'(t) - \mu_i'(t)) - 2\lambda\right] \\
\leq\ & \Pr\left[\theta_i(t) \geq \hat{\mu}_i'(t) + \frac{4}{5}(\mu_1 - \mu_i) - 2\lambda\right] \\
=\ & \Pr\left[\theta_i(t) \geq \hat{\mu}_i'(t) + \sqrt{\frac{2\log T}{\frac{5}{4}L_i(t)}}\right] \\
\leq\ & \Pr\left[\theta_i(t) \geq \hat{\mu}_i'(t) + \sqrt{\frac{2\log T}{\frac{5}{4}N_i(t)}}\right] \\
\leq\ & \Pr\left[\theta_i(t) \geq \hat{\mu}_i'(t) + \sqrt{\frac{2\log T}{M_i(t)}}\right] \\
=\ & \sum_{m,\mu'} \Pr[M_i(t) = m, \hat{\mu}_i'(t) = \mu'] \Pr\left[\theta_i(t) \geq \hat{\mu}_i'(t) + \sqrt{\frac{2\log T}{M_i(t)}} \mid M_i(t) = m, \hat{\mu}_i'(t) = \mu'\right] \\
\leq\ & \sum_{m,\mu'} \Pr[M_i(t) = m, \hat{\mu}_i'(t) = \mu'] \frac{1}{T} \\
\leq\ & \frac{1}{T}.
\end{aligned}$$

Therefore one could use a similar way (as bounding the second term in the proof of Theorem 11) to show that

$$N_B^* + \sum_{t=T^*}^{T} \mathbb{E}[\mathbb{I}[\mathcal{B}_i(t)]] \leq \max\{L_i(T), N_B^*\} + 1.$$

Then we come to the third term. The following lemma is also similar to Lemma 1 in the work of Agrawal and Goyal (2013), and its proof is almost the same as that for Lemma 20.

Lemma 21 *Denote* $p_t = \Pr[\theta_1(t) \geq \mu'_1 - \lambda | \mathcal{F}_{t-1}]$, *then we have that*

$$\Pr[i(t) = i, \theta_i(t) \leq \mu'_1(t) - \lambda | \mathcal{F}_{t-1}] \leq \frac{1 - p_t}{p_t} \Pr[i(t) = 1, \theta_i(t) \leq \mu'_1(t) - \lambda | \mathcal{F}_{t-1}].$$

Similarly, by Lemma 21, we also have that

$$\sum_{t=T^*}^{T} \mathbb{E}[\mathbb{I}[\mathcal{C}_i(t)]] \leq \sum_{n=N_B^*}^{\infty} \mathbb{E}\left[\frac{1}{p_n} - 1\right],$$

where p_n denotes the random probability (due to the randomness on observations of arm 1) that $\theta_1(t) \geq \mu'_1(t_n) - \lambda$ when there are totally n observations on arm 1, i.e., $N_i(t) = n$. Note that for fixed $N_i(t) = n$, the value of $k_i(t)$ is also fixed (and we denote it by k), also denote $m = n + 2k$.

Then we know that observations on arm 1 follow a Binomial distribution with parameters n, μ_1. For a history of observations with $a - 1$ number of 1s and $b - 1$ number of 0s (here $a + b - 2 = n$), the random probability p_n equals to $1 - F^{Beta}_{a+k,b+k}(\mu'_1(t_n) - \lambda)$, where $F^{Beta}_{a+k,b+k}$ denotes cumulative distribution function of $\mathcal{B}(a + k, b + k)$.

Therefore

$$\mathbb{E}\left[\frac{1}{p_n}\right] = \sum_{s=0}^{n} \frac{f^{Bino}_{n,\mu_1}(s)}{1 - F^{Beta}_{s+1+k,n-s+1+k}(\mu'_1(t_n) - \lambda)} = \sum_{s=0}^{n} \frac{f^{Bino}_{n,1-\mu_1}(s)}{F^{Beta}_{s+1+k,n-s+1+k}(1 - \mu'_1(t_n) + \lambda)}.$$

Prior analysis in Agrawal and Goyal (2013) also proves the following two inequalities:

Fact 9 *(Agrawal and Goyal, 2013) For any* $n \geq 1$, *we have that*

$$\sum_{s=0}^{n} \frac{f^{Bino}_{n,1-\mu_1}(s)}{F^{Beta}_{s+1,n-s+1}(1 - \mu_1 + \frac{5}{4}\lambda)} \leq \frac{12}{5\lambda}.$$

For $n \geq \frac{8}{\lambda^2}$, *we have that*

$$\sum_{s=(1-\mu_1+\lambda/2)n}^{n} \frac{f^{Bino}_{n,1-\mu_1}(s)}{F^{Beta}_{s+1,n-s+1}(1 - \mu_1 + \frac{5}{4}\lambda)} = \Theta\left(\frac{1}{\lambda^2} e^{-\frac{25n}{32}\lambda^2}\right).$$

The following lemma is the key lemma in our regret analysis of DP-FTPL-Beta, and its complete proof is stated in Appendix B.1.

Lemma 22 *For any* $y \in [0, 1]$, *and* $a, b, k \in \mathbb{N}_+$, *denote* $y' = \frac{a+b-2}{a+b-2+2k} y + \frac{k}{a+b-2+2k}$, *then if* $a + b + 2k - 2 \geq \frac{1250e}{9\pi}$ *and* $k \geq \frac{1}{8}(a + b - 2)$, *we always have that* $F^{Beta}_{a+k,b+k}(y') \geq \frac{F^{Beta}_{a,b}(y)}{10}$.

Using Lemma 22 with $y = 1 - \mu_1 + \frac{5}{4}\lambda$, $a = s + 1$, $b = n - s + 1$ and $k = \lfloor \frac{n}{8} \rfloor + 1$, we have that $F^{Beta}_{s+1+k,n-s+1+k}(1 - \mu'_1(t_n) + \lambda) \geq \frac{F^{Beta}_{s+1,n-s+1}(1 - \mu_1 + \frac{5}{4}\lambda)}{10}$. Note that our start phase size N_B^* is larger than $\frac{1000e}{9\pi}$, therefore $a_i(t) + b_i(t) + 2k_i(t) - 2 \geq \frac{1250e}{9\pi}$ always holds. Then by Fact 9, for any value n after the start phase, we have that

$$\sum_{s=0}^{n} \frac{f^{Bino}_{n,1-\mu_1}(s)}{F^{Beta}_{s+1+k,n-s+1+k}(1 - \mu'_1(t_n) + \lambda)} \leq \frac{24}{\lambda}.$$

For $n \geq \frac{8}{\lambda^2}$, we have that

$$\sum_{s=(1-\mu_1+\lambda/2)n}^{n} \frac{f^{Bino}_{n,1-\mu_1}(s)}{F^{Beta}_{s+1+k,n-s+1+k}(1-\mu_1'(t_n)+\lambda)} = \Theta\left(\frac{1}{\lambda^2}e^{-\frac{25n}{32}\lambda^2}\right).$$

For the case that $n \geq \frac{8}{\lambda^2}$ and $s < (1-\mu_1+\lambda/2)n$, we can use the following fact.

Fact 10 *(Beta-Binomial Trick, Agrawal and Goyal (2013)) Denote $F^{Bino}_{n,p}$ as the cumulative distribution function of Binomial distribution with parameters n, p, then*

$$F^{Beta}_{a,b}(x) = 1 - F^{Bino}_{a+b-1,x}(a-1).$$

By Fact 10 and Chernoff-Hoeffding inequality (Fact 5), for $s < (1-\mu_1+\lambda/2)n$, we have that

$$F^{Beta}_{s+1+k,n-s+1+k}(1-\mu_1'(t_n)+\lambda) = 1 - F^{Bino}_{n+2k+1,1-\mu_1'(t_n)+\lambda}(s+k+1) \geq 1 - \exp\left(\frac{-n\lambda^2}{2}\right).$$

Therefore,

$$\sum_{s=0}^{(1-\mu_1+\lambda/2)n} \frac{f^{Bino}_{n,1-\mu_1}(s)}{F^{Beta}_{s+1+k,n-s+1+k}(1-\mu_1'(t_n)+\lambda)} \leq \frac{1}{1-\exp\left(\frac{-n\lambda^2}{2}\right)} = 1 + \frac{1}{\exp\left(\frac{n\lambda^2}{2}\right)-1}.$$

Thus we have that

$$\begin{aligned}
\sum_{t=T^*}^{T} \mathbb{E}[\mathbb{I}[\mathcal{C}_i(t)]] &\leq \sum_{n=N_B^*}^{\infty} \mathbb{E}\left[\frac{1}{p_n} - 1\right] \\
&\leq \sum_{n=1}^{\frac{8}{\lambda^2}} \frac{24}{\lambda} + \sum_{n=\frac{8}{\lambda^2}+1}^{\infty} \left(\Theta\left(\frac{1}{\lambda^2}e^{-\frac{25n}{32}\lambda^2}\right) + 1 + \frac{1}{\exp\left(\frac{n\lambda^2}{2}\right)-1} - 1\right) \\
&= \Theta\left(\frac{1}{\lambda^3}\right) + \Theta\left(\frac{1}{\lambda^4}\right) + \Theta\left(\frac{1}{\lambda^2}\right) \\
&= \Theta\left(\frac{1}{\lambda^4}\right).
\end{aligned}$$

Summing over the three terms, the cumulative regret of DP-FTPL-Beta is upper bounded by

$$\begin{aligned}
\sum_{i=2}^{N} \mathbb{E}[N_i(T)]\Delta_i &\leq \sum_{i} \left(\sum_{t=T^*}^{T} \mathbb{E}[\mathbb{I}[\mathcal{A}_i(t)]] + \left(N_B^* + \sum_{t=T^*}^{T} \mathbb{E}[\mathbb{I}[\mathcal{B}_i(t)]]\right) + \sum_{t=T^*}^{T} \mathbb{E}[\mathbb{I}[\mathcal{C}_i(t)]]\right)\Delta_i \\
&\leq \sum_{i=2}^{N} \max\left\{\frac{5\log T\Delta_i}{2(\Delta_i - 5/2\lambda)^2}, N_B^*\Delta_i\right\} + \Theta\left(\frac{N}{\lambda^4}\right).
\end{aligned}$$

■

B.1 Proof of Lemma 22

Before the proof of Lemma 22, we first prove two other lemmas.

Lemma 23 *For any $y \in [0,1]$ and $a, b \in \mathbb{N}_+$, denote $y' = \frac{a+b-2}{a+b}y + \frac{1}{a+b}$ and $F_{a,b}^{Beta}(x)$ the cumulative distribution function of $\mathcal{B}(a,b)$. Then for any $y' \leq \frac{a-2}{a+b}$, we always have that $F_{a,b}^{Beta}(y) \leq F_{a+1,b+1}^{Beta}(y')$.*

Proof First we consider the case that $y \geq \frac{1}{2}$, in this case $y > y' > \frac{1}{2}$ and $y - y' = \frac{2y-1}{a+b}$.

For the cumulative distribution function $F_{a,b}^{Beta}(x)$, prior works (Olver et al., 2010) show that

$$F_{a+1,b+1}^{Beta}(y) = F_{a,b}^{Beta}(y) + \frac{(a+b-1)!y^a(1-y)^b}{(a-1)!b!} - \frac{(a+b)!y^a(1-y)^{b+1}}{a!b!}.$$

Since $y \geq y'$, then denote $f_{a,b}$ the probability density function of distribution $\mathcal{B}(a,b)$, i.e., $f_{a,b}(y) = \frac{y^{a-1}(1-y)^{b-1}(a+b-1)!}{(a-1)!(b-1)!}$, and we have that

$$\begin{aligned} F_{a+1,b+1}^{Beta}(y') &= F_{a+1,b+1}^{Beta}(y) - \int_{y'}^{y} f_{a+1,b+1}(x)dx \\ &= F_{a,b}^{Beta}(y) + \frac{(a+b-1)!y^a(1-y)^b}{(a-1)!b!} - \frac{(a+b)!y^a(1-y)^{b+1}}{a!b!} - \int_{y'}^{y} f_{a+1,b+1}(x)dx. \end{aligned}$$

Note that $y' \leq \frac{a-2}{a+b}$ and $y \geq \frac{1}{2}$ imply that $y' \leq y \leq \frac{a}{a+b}$, and $f_{a+1,b+1}$ is first increasing and then decreasing with maximum value at point $\frac{a}{a+b}$. Therefore, for all $x \in [y', y]$, $f_{a+1,b+1}(x) \leq f_{a+1,b+1}(y)$. This implies that

$$\begin{aligned} & F_{a+1,b+1}^{Beta}(y') \\ = \; & F_{a,b}^{Beta}(y) + \frac{(a+b-1)!y^a(1-y)^b}{(a-1)!b!} - \frac{(a+b)!y^a(1-y)^{b+1}}{a!b!} - \int_{y'}^{y} f_{a+1,b+1}(x)dx \\ \geq \; & F_{a,b}^{Beta}(y) + \frac{(a+b-1)!y^a(1-y)^b}{(a-1)!b!} - \frac{(a+b)!y^a(1-y)^{b+1}}{a!b!} - (y-y')f_{a+1,b+1}(y) \\ = \; & F_{a,b}^{Beta}(y) + \frac{(a+b-1)!y^a(1-y)^b}{(a-1)!b!} - \frac{(a+b)!y^a(1-y)^{b+1}}{a!b!} - (y-y')\frac{y^a(1-y)^b(a+b+1)!}{a!b!} \\ = \; & F_{a,b}^{Beta}(y) + \frac{(a+b-1)!y^a(1-y)^b}{a!b!}\big(a - (a+b)(1-y) - (y-y')(a+b+1)(a+b)\big) \\ = \; & F_{a,b}^{Beta}(y) + \frac{(a+b-1)!y^a(1-y)^b}{a!b!}\left(a - (a+b)(1-y) - \frac{2y-1}{a+b}(a+b+1)(a+b)\right) \\ = \; & F_{a,b}^{Beta}(y) + \frac{(a+b-1)!y^a(1-y)^b}{a!b!}\left(a - (a+b)(1-y) - (2y-1)(a+b+1)\right) \\ = \; & F_{a,b}^{Beta}(y) + \frac{(a+b-1)!y^a(1-y)^b}{a!b!}\left(a - (a+b)(1-y) - (2y-1)(a+b) - (2y-1)\right) \\ = \; & F_{a,b}^{Beta}(y) + \frac{(a+b-1)!y^a(1-y)^b}{a!b!}\left(a - y(a+b) - (2y-1)\right) \\ \geq \; & F_{a,b}^{Beta}(y) + \frac{(a+b-1)!y^a(1-y)^b}{a!b!}\left(a - 1 - y(a+b)\right). \end{aligned}$$

For $y' \leq \frac{a-2}{a+b}$, we must have that $y \leq \frac{a-1}{a+b}$, therefore $a-1-y(a+b) \geq 0$, which implies that $F^{Beta}_{a+1,b+1}(y') > F^{Beta}_{a,b}(y)$.

Then we consider the case that $y < \frac{1}{2}$, in this case $y < y' < \frac{1}{2}$ and $y' - y = \frac{1-2y}{a+b}$.

Similar as before, we have the following equations (note that now $y < y'$)

$$\begin{aligned} F^{Beta}_{a+1,b+1}(y') &= F^{Beta}_{a+1,b+1}(y) + \int_y^{y'} f_{a+1,b+1}(x)dx \\ &= F^{Beta}_{a,b}(y) + \frac{(a+b-1)!y^a(1-y)^b}{(a-1)!b!} - \frac{(a+b)!y^a(1-y)^{b+1}}{a!b!} + \int_y^{y'} f_{a+1,b+1}(x)dx. \end{aligned}$$

In this case, $y < y' \leq \frac{a-2}{a+b}$, which implies that for all $x \in [y, y']$, $f_{a+1,b+1}(x) \geq f_{a+1,b+1}(y)$. This means that

$$\begin{aligned} & F^{Beta}_{a+1,b+1}(y') \\ &= F^{Beta}_{a,b}(y) + \frac{(a+b-1)!y^a(1-y)^b}{(a-1)!b!} - \frac{(a+b)!y^a(1-y)^{b+1}}{a!b!} + \int_y^{y'} f_{a+1,b+1}(x)dx \\ &\geq F^{Beta}_{a,b}(y) + \frac{(a+b-1)!y^a(1-y)^b}{(a-1)!b!} - \frac{(a+b)!y^a(1-y)^{b+1}}{a!b!} + (y'-y)f_{a+1,b+1}(y) \\ &= F^{Beta}_{a,b}(y) + \frac{(a+b-1)!y^a(1-y)^b}{(a-1)!b!} - \frac{(a+b)!y^a(1-y)^{b+1}}{a!b!} + (y'-y)\frac{y^a(1-y)^b(a+b+1)!}{a!b!} \\ &= F^{Beta}_{a,b}(y) + \frac{(a+b-1)!y^a(1-y)^b}{a!b!}\big(a - (a+b)(1-y) + (y'-y)(a+b+1)(a+b)\big) \\ &= F^{Beta}_{a,b}(y) + \frac{(a+b-1)!y^a(1-y)^b}{a!b!}\left(a - (a+b)(1-y) + \frac{1-2y}{a+b}(a+b+1)(a+b)\right) \\ &= F^{Beta}_{a,b}(y) + \frac{(a+b-1)!y^a(1-y)^b}{a!b!}\left(a - (a+b)(1-y) + (1-2y)(a+b+1)\right) \\ &= F^{Beta}_{a,b}(y) + \frac{(a+b-1)!y^a(1-y)^b}{a!b!}\left(a - (a+b)(1-y) + (1-2y)(a+b) + (1-2y)\right) \\ &= F^{Beta}_{a,b}(y) + \frac{(a+b-1)!y^a(1-y)^b}{a!b!}\left(a - y(a+b) + (1-2y)\right) \\ &\geq F^{Beta}_{a,b}(y) + \frac{(a+b-1)!y^a(1-y)^b}{a!b!}\left(a - y(a+b)\right). \end{aligned}$$

For $y' \leq \frac{a-2}{a+b}$, we must have that $y \leq \frac{a}{a+b}$, therefore $a - y(a+b) \geq 0$, which implies that $F^{Beta}_{a+1,b+1}(y') > F^{Beta}_{a,b}(y)$ and we finish the proof of this lemma. ■

From Lemma 23, we can then prove the following Lemma 24.

Lemma 24 *For any $y \in [0,1]$, $a, b, k \in \mathbb{N}_+$, denote $y' = \frac{a+b-2}{a+b-2+2k}y + \frac{k}{a+b-2+2k}$ and $F^{Beta}_{a,b}(x)$ the cumulative distribution function of $\mathcal{B}(a,b)$, then if $y' \leq \frac{a-3+k}{a+b-2+2k}$, we always have that $F^{Beta}_{a+k,b+k}(y') \geq F^{Beta}_{a,b}(y)$.*

Proof Let $y = y_0$ and $y_m = \frac{a+b-2+2(m-1)}{a+b-2+2m} y_{m-1} + \frac{1}{a+b-2+2m}$. Then it is easy to check that

$$\begin{aligned} y_k &= \frac{a+b-2+2(k-1)}{a+b-2+2k} y_{k-1} + \frac{1}{a+b-2+2k} \\ &= \frac{a+b-2+2(k-2)}{a+b-2+2k} y_{k-2} + \frac{2}{a+b-2+2k} \\ &= \cdots \\ &= \frac{a+b-2+2}{a+b-2+2k} y_1 + \frac{k-1}{a+b-2+2k} \\ &= \frac{a+b-2}{a+b-2+2k} y_0 + \frac{k}{a+b-2+2k} \\ &= \frac{a+b-2}{a+b-2+2k} y + \frac{k}{a+b-2+2k} \\ &= y'. \end{aligned}$$

By the above equations, $y_k = y' \le \frac{a-3+k}{a+b-2+2k}$ implies that $y_m \le \frac{a-3+m}{a+b-2+2m}$ holds for any $1 \le m \le k$. Therefore we can use Lemma 23 for k times and get that

$$F^{Beta}_{a+k,b+k}(y') = F^{Beta}_{a+k,b+k}(y_k) \ge F^{Beta}_{a+k-1,b+k-1}(y_{k-1}) \ge \cdots \ge F^{Beta}_{a,b}(y_0) = F^{Beta}_{a,b}(y).$$

■

Based on Lemma 24, we can finally prove Lemma 22.

Lemma 21 *For any $y \in [0,1]$, and $a, b, k \in \mathbb{N}_+$, denote $y' = \frac{a+b-2}{a+b-2+2k} y + \frac{k}{a+b-2+2k}$, then if $a+b+2k-2 \ge \frac{1250e}{9\pi}$ and $k \ge \frac{1}{8}(a+b-2)$, we always have that $F^{Beta}_{a+k,b+k}(y') \ge \frac{F^{Beta}_{a,b}(y)}{10}$.*

Proof By Lemma 24, we know that for $y' \le \frac{a-3+k}{a+b-2+2k}$, $F^{Beta}_{a+k,b+k}(y') \ge F^{Beta}_{a,b}(y) \ge \frac{F^{Beta}_{a,b}(y)}{10}$ always holds.

For $y' > \frac{a-3+k}{a+b-2+2k}$, Fact 10 shows that that $F^{Beta}_{a+k,b+k}(y') = 1 - F^{Bino}_{a+b+2k-1,y'}(a+k-1) = F^{Bino}_{a+b+2k-1,1-y'}(b+k)$.

Note that $(a+b+2k-1)(1-y') < \frac{b+1+k}{a+b+2k-2}(a+b+2k-1) < b+2+k$, therefore $F^{Bino}_{a+b+2k-1,1-y'}(b+2+k) \ge \frac{1}{2}$. This implies that

$$F^{Bino}_{a+b+2k-1,1-y'}(b+k) \ge \frac{1}{2} - f^{Bino}_{a+b+2k-1,1-y'}(b+k+1) - f^{Bino}_{a+b+2k-1,1-y'}(b+k+2),$$

where $f^{Bino}_{n,p}$ denotes the probability mass function of Binomial distribution with parameters n, p.

Note that

$$\begin{aligned} f^{Bino}_{n,p}(s) &= \frac{n! p^s (1-p)^{n-s}}{s!(n-s)!} \\ &\le \frac{\sqrt{2\pi e n}(\frac{n}{e})^n p^s (1-p)^{n-s}}{\sqrt{2\pi s}(\frac{s}{e})^s \sqrt{2\pi(n-s)}(\frac{n-s}{e})^{n-s}} \\ &= \sqrt{\frac{en}{2\pi s(n-s)}} \frac{n^n x^s (1-x)^{n-s}}{s^s (n-s)^{n-s}} \end{aligned}$$

$$\leq \quad \sqrt{\frac{en}{2\pi s(n-s)}}.$$

Therefore

$$f^{Bino}_{a+b+2k-1,1-y'}(b+k+1) \leq \sqrt{\frac{e(a+b+2k-1)}{2\pi(b+k+1)(a+k-2)}} \leq \sqrt{\frac{e}{0.18\pi(a+b+2k-1)}}.$$

Similarly,

$$f^{Bino}_{a+b+2k-1,1-y'}(b+k+2) \leq \sqrt{\frac{e(a+b+2k-1)}{2\pi(b+k+1)(a+k-2)}} \leq \sqrt{\frac{e}{0.18\pi(a+b+2k-1)}}.$$

All these imply that

$$\begin{aligned} F^{Bino}_{a+b+2k-1,1-y'}(b+k) &\geq \frac{1}{2} - f^{Bino}_{a+b+2k-1,1-y'}(b+k+1) - f^{Bino}_{a+b+2k-1,1-y'}(b+k+2) \\ &\geq \frac{1}{2} - \sqrt{\frac{4e}{0.18\pi(a+b+2k-1)}}. \end{aligned}$$

If $a+b+2k-2 \geq \frac{1250e}{9\pi}$, then $F^{Beta}_{a+k,b+k}(y') = F^{Bino}_{a+b+2k-1,1-y'}(b+k) \geq \frac{1}{10}$ holds, and therefore $F^{Beta}_{a+k,b+k}(y') \geq \frac{F^{Beta}_{a,b}(y)}{10}$ must hold, and we finish the proof of this lemma. ■

References

Shipra Agrawal and Navin Goyal. Analysis of thompson sampling for the multi-armed bandit problem. In *Conference on learning theory*, pages 39–1, 2012.

Shipra Agrawal and Navin Goyal. Further optimal regret bounds for thompson sampling. In *International Conference on Artificial Intelligence and Statistics*, 2013.

Achraf Azize and Debabrota Basu. When privacy meets partial information: A refined analysis of differentially private bandits. In *Advances in Neural Information Processing Systems*, 2022.

Donald A Berry and Bert Fristedt. *Bandit problems: sequential allocation of experiments (Monographs on statistics and applied probability)*. Springer, 1985.

Zygmunt Wilhelm Birnbaum. An inequality for mill's ratio. *Ann. Math. Statist.*, 13:245–246, 1942.

Swapna Buccapatnam, Atilla Eryilmaz, and Ness B Shroff. Multi-armed bandits in the presence of side observations in social networks. In *52nd IEEE Conference on Decision and Control*, pages 7309–7314. IEEE, 2013.

Deepayan Chakrabarti, Ravi Kumar, Filip Radlinski, and Eli Upfal. Mortal multi-armed bandits. In *Advances in Neural Information Processing Systems*, volume 21, pages 273–280, 2008.

Wei Chen, Yajun Wang, and Yang Yuan. Combinatorial multi-armed bandit: General framework and applications. In *International Conference on Machine Learning*, pages 151–159, 2013.

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*, pages 1757–1767, 2020.

Imre Csiszar and János Körner. *Information theory: coding theorems for discrete memoryless systems.* Cambridge University Press, 2011.

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

John C Duchi, Michael I Jordan, and Martin J Wainwright. Privacy aware learning. *Journal of the ACM (JACM)*, 61(6):1–57, 2014.

Cynthia Dwork. Differential privacy: A survey of results. In *International conference on theory and applications of models of computation*, pages 1–19. Springer, 2008.

Eyal Evendar, Shie Mannor, and Yishay Mansour. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of Machine Learning Research*, 7:1079–1105, 2006.

Willliam Feller. *An introduction to probability theory and its applications, vol 2.* John Wiley & Sons, 2008.

Raphaël Féraud, Réda Alami, and Romain Laroche. Decentralized exploration in multi-armed bandits. In *International Conference on Machine Learning*, pages 1901–1909. PMLR, 2019.

Pratik Gajane, Tanguy Urvoy, and Emilie Kaufmann. Corrupt bandits for preserving local privacy. In *Algorithmic Learning Theory*, pages 387–412. PMLR, 2018.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

Prateek Jain, Pravesh Kothari, and Abhradeep Thakurta. Differentially private online learning. In *Conference on Learning Theory*, pages 24–1, 2012.

Emilie Kaufmann and Shivaram Kalyanakrishnan. Information complexity in bandit subset selection. In *Conference on Learning Theory*, pages 228–251, 2013.

Emilie Kaufmann, Nathaniel Korda, and Rémi Munos. Thompson sampling: An asymptotically optimal finite-time analysis. In *International conference on algorithmic learning theory*, pages 199–213. Springer, 2012.

Baekjin Kim and Ambuj Tewari. On the optimality of perturbations in stochastic and adversarial multi-armed bandit problems. In *Advances in Neural Information Processing Systems*, pages 2695–2704, 2019.

Keqin Liu and Qing Zhao. Distributed learning in multi-armed bandit with multiple players. *IEEE Transactions on Signal Processing*, 58(11):5667–5681, 2010.

Tyler Lu, Dávid Pál, and Martin Pál. Contextual multi-armed bandits. In *International conference on Artificial Intelligence and Statistics*, pages 485–492, 2010.

Oded Maron and Andrew W Moore. The racing algorithm: Model selection for lazy learners. *Artificial Intelligence Review*, 11(1):193–225, 1997.

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

Frank WJ Olver, Daniel W Lozier, Ronald F Boisvert, and Charles W Clark. *NIST handbook of mathematical functions hardback and CD-ROM*. Cambridge university press, 2010.

Lijing Qin, Shouyuan Chen, and Xiaoyan Zhu. Contextual combinatorial bandit and its application on diversified online recommendation. In *Proceedings of the 2014 SIAM International Conference on Data Mining*, pages 461–469. SIAM, 2014.

Wenbo Ren, Xingyu Zhou, Jia Liu, and Ness B Shroff. Multi-armed bandits with local differential privacy. *arXiv preprint arXiv:2007.03121*, 2020.

Daniel Russo and Benjamin Van Roy. Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39(4):1221–1243, 2014.

Touqir Sajed and Or Sheffet. An optimal private stochastic-mab algorithm based on optimal private stopping rule. In *International Conference on Machine Learning*, pages 5579–5588. PMLR, 2019.

Eric M Schwartz, Eric T Bradlow, and Peter S Fader. Customer acquisition via display advertising using multi-armed bandit experiments. *Marketing Science*, 36(4):500–522, 2017.

Roshan Shariff and Or Sheffet. Differentially private contextual linear bandits. In *Advances in Neural Information Processing Systems*, volume 31, pages 4296–4306, 2018.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.

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

Aristide CY Tossou and Christos Dimitrakakis. Algorithms for differentially private multi-armed bandits. In *AAAI Conference on Artificial Intelligence*, pages 2087–2093, 2016.

Aristide CY Tossou and Christos Dimitrakakis. On the differential privacy of thompson sampling with gaussian prior. *arXiv preprint arXiv:1806.09192*, 2018.

Qinshi Wang and Wei Chen. Improving regret bounds for combinatorial semi-bandits with probabilistically triggered arms and its applications. In *Advances in Neural Information Processing Systems*, pages 1161–1171, 2017.

Siwei Wang and Longbo Huang. Multi-armed bandits with compensation. In *Advances in Neural Information Processing Systems*, pages 5114–5122, 2018.

Kai Zheng, Tianle Cai, Weiran Huang, Zhenguo Li, and Liwei Wang. Locally differentially private (contextual) bandits learning. In *Advances in Neural Information Processing Systems*, 2020.