Near-Optimal Thompson Sampling-based Algorithms for Differentially Private Stochastic Bandits (Appendix)

Bingshan Hu[1,2] **Nidhi Hegde**[1,2]

[1]Department of Computing Science, University of Alberta, Edmonton, Alberta, Canada
[2]Amii (Alberta Machine Intelligence Institute)

Abstract

We address differentially private stochastic bandits. We present two (near)-optimal Thompson Sampling-based learning algorithms: DP-TS and Lazy-DP-TS. The core idea in achieving optimality is the principle of optimism in the face of uncertainty. We reshape the posterior distribution in an optimistic way as compared to the non-private Thompson Sampling of Agrawal and Goyal [2017]. Our DP-TS achieves a $\sum_{j\in\mathcal{A}:\Delta_j>0} O\left(\frac{\log(T)}{\min\{\epsilon,\Delta_j\})}\log\left(\frac{\log(T)}{\epsilon\cdot\Delta_j}\right)\right)$ regret bound, where $\mathcal{A}$ is the arm set, Δ_j is the sub-optimality gap of a sub-optimal arm j, and ϵ is the privacy parameter. Our Lazy-DP-TS gets rid of the extra $\log$ factor by using the idea of dropping observations. The regret of Lazy-DP-TS is $\sum_{j\in\mathcal{A}:\Delta_j>0} O\left(\frac{\log(T)}{\min\{\epsilon,\Delta_j\}}\right)$, which matches the regret lower bound of Shariff and Sheffet [2018]. Additionally, we conduct experiments to compare the empirical performance of our proposed algorithms with the existing optimal algorithms for differentially private stochastic bandits.

1 INTRODUCTION

We consider the setting of differentially private stochastic multi-armed bandits[Mishra and Thakurta, 2015, Tossou and Dimitrakakis, 2016, Shariff and Sheffet, 2018, Sajed and Sheffet, 2019, Hu et al., 2021]. In the classical stochastic multi-armed bandit problem, we have a fixed and finite set of K arms and a stochastic environment. In each round $t = 1, 2, \dots, T$, the environment generates a random reward $X_j(t)$ for arm j which is revealed and collected if arm j is pulled in that round. For each arm j, the rewards $X_j(t) \in [0, 1]$ are i.i.d. over time according to a fixed but unknown probability distribution with mean μ_j. The goal of the learning algorithm is to pull arms sequentially to maximize the accumulated reward. The performance metric has traditionally been pseudo-*regret* [Bubeck and Cesa-Bianchi, 2012] which is a measure of the difference of the expected accumulated rewards compared to a given benchmark.

In the classical setting, the learning algorithm uses the true revealed rewards from previous rounds to make decisions on arms in future rounds. However, in many settings rewards may be private information that should be protected. For instance, consider an online search advertisement system where the objective is to display relevant ads for web queries. In such a setting the system would display a few advertisements to the user. When the user clicks on an ad, a reward is collected by the system, which in accumulation would allow the system and any external observers to learn user preferences. Rewards that represent user preferences are private information and may further allow inference on the user's other private characteristics.

Motivated by such applications, previous work [Mishra and Thakurta, 2015, Tossou and Dimitrakakis, 2016, Sajed and Sheffet, 2019, Hu et al., 2021] have studied the design of bandit learning algorithms with differential privacy[Dwork et al., 2014] for keeping reward information private. Differential privacy has been used as a framework because it provides robust privacy guarantees and a controlled tradeoff with regret guarantees in the case of bandit learning.

Two of the most common algorithms in the stochastic bandit setting are Upper Confidence Bound (UCB) sampling [Auer et al., 2002] and Thompson Sampling [Agrawal and Goyal, 2017]. Mishra and Thakurta [2015] present the first differentially private versions of these two algorithms and provide regret bounds for their algorithms. This was followed by differentially private algorithms in the contextual linear bandit setting [Shariff and Sheffet, 2018] and optimal differentially private algorithms based on Successive Elimination (DP-SE) [Sajed and Sheffet, 2019] and UCB (Anytime-Lazy-UCB) [Hu et al., 2021] in the stochastic bandit setting.

Accepted for the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2022).

However, to the best of our knowledge, there is still no optimal Thompson Sampling-based algorithm for differentially private stochastic bandits. Thompson Sampling-based algorithms exhibit better performance than UCB or SE-based methods, are applicable to a wider range of information models, and are more widely implemented in practical scenarios [Chapelle and Li, 2011, Gopalan et al., 2014]. Given their widespread implementation in practice, we provide two (near)-optimal Thompson Sampling-based algorithms for private stochastic bandits. The core concept in our algorithms still relies on the principle of optimism in the face of uncertainty. More specifically, we shift the posterior distribution of an arm to the right as compared to the posterior distribution in the non-private Thompson Sampling.

Our first algorithm, Differentially Private Thompson Sampling (DP-TS), can be viewed as a differentially private version of the standard Thompson Sampling for Bernoulli bandits [Agrawal and Goyal, 2017]: the learning algorithm makes a decision based on all observations obtained from the beginning and updates the statistics of the pulled arm at the end of each round. The regret bound of DP-TS is $\widetilde{O}\left(\frac{K\log(T)}{\min\{\epsilon,\Delta\}}\right)$, where $\widetilde{O}(\cdot)$ hides an extra $\log(\log(T)/(\epsilon\Delta))$ factor. Our second algorithm, Lazy Differentially Private Thompson Sampling (Lazy-DP-TS), drops observations during learning and updates the statistics of the pulled arm in a delayed manner. With these modifications we achieve the optimal $O\left(\frac{K\log(T)}{\min\{\Delta,\epsilon\}}\right)$ regret bound. Interestingly, as discussed in Section 3 and confirmed in Section 4 with numerical experiments, DP-TS may perform better than Lazy-DP-TS under some circumstances.

Contribution. We make the following key contributions. (1) We present (near)-optimal Thompson Sampling-based learning algorithms for differentially private stochastic bandits: DP-TS and Lazy-DP-TS; (2) The regret bound for DP-TS is $\sum_{j\in\mathcal{A}:\Delta_j>0} O\left(\max\left\{\frac{\log(T)}{\Delta_j}, \frac{\log(T)}{\epsilon}\log\left(\frac{\log(T)}{\epsilon\cdot\Delta_j}\right)\right\}\right)$, which is optimal up to a $\log\log(T)$ factor (Theorem 2); (3) The regret bound for Lazy-DP-TS is $\sum_{j\in\mathcal{A}:\Delta_j>0} O\left(\frac{\log(T)}{\min\{\epsilon,\Delta_j\}}\right)$, which is optimal (Theorem 4); (4) We show through numerical experiments performance improvement of our proposed learning algorithms as compared to the existing two optimal algorithms, DP-SE and Anytime-Lazy-UCB.

2 PROBLEM DEFINITION AND BACKGROUND

2.1 STOCHASTIC MULTI-ARMED BANDITS

We consider the stochastic multi-armed bandit setting, with a fixed set $\mathcal{A}$ of K arms and a stochastic environment. At each round $t = 1, 2, \ldots, T$, the environment generates a reward vector $X_t := (X_1(t), X_2(t), \ldots, X_K(t))$ with each $X_j(t) \in \{0, 1\}$ independently drawn from a Bernoulli distribution[1] with parameter $\mu_j \in (0, 1)$. The learning algorithm pulls an arm $J_t \in \mathcal{A}$ and at the end of round t, observes and obtains the reward of the pulled arm, $X_{J_t}(t)$. The goal of the algorithm to select an arm in each round such that the accumulated reward over T rounds is maximized.

Without loss of generality, we assume that the optimal arm is unique and let arm 1 be the unique optimal arm, i.e., $\mu_1 > \mu_j$ for all $j \in \mathcal{A}\backslash\{1\}$. Let $\Delta_j := \mu_1 - \mu_j$ be the mean reward gap, which indicates the performance loss in a single round when a sub-optimal arm j is pulled instead of the best arm 1. We use (pseudo)-regret $\mathcal{R}(T)$ to measure the performance, expressed as

$$\begin{aligned}\mathcal{R}(T) &= T\cdot\mu_1 - \sum_{j\in\mathcal{A}}\mathbb{E}\left[\sum_{t=1}^{T}\mathbf{1}\left\{J_t = j\right\}\right]\cdot\mu_j \\ &= \sum_{j\in\mathcal{A}:\Delta_j>0}\mathbb{E}\left[\sum_{t=1}^{T}\mathbf{1}\left\{J_t = j\right\}\right]\cdot\Delta_j \quad .\end{aligned}$$

2.2 DIFFERENTIAL PRIVACY

Differential privacy is a widely accepted framework of privacy and is based on the notion of plausible deniability: an adversary should learn nearly the same thing if one element in the dataset is changed or missing. In the context of bandits, a dataset is the stream of reward vectors drawn throughout the algorithm, and a change would refer to one reward vector in the stream. More formally, let $\boldsymbol{X}_{1:t}$ be the sequence of reward vectors up to time t and let $\boldsymbol{X}'_{1:t}$ be a neighbouring sequence which differs in at most one reward vector, say, at any round $s, s \le t$. The output of a bandit learning algorithm is the sequence of arm selections at each round. In this context, differential privacy is defined as follows, omitting the subscript t for clarity.

Definition 1. *An online learning algorithm $\mathcal{M}$ is ϵ-differentially private if, at every round $t = 1, \ldots T$, for any two neighbouring reward sequences $\boldsymbol{X}$ and $\boldsymbol{X}'$, and for any set $\mathcal{D}$ of decisions made, it holds that $\mathbb{P}\left\{\mathcal{M}(\boldsymbol{X}) \in \mathcal{D}\right\} \le e^{\epsilon}\cdot\mathbb{P}\left\{\mathcal{M}(\boldsymbol{X}') \in \mathcal{D}\right\}$.*

Remark. Our definition of differential privacy follows the standard notion that was introduced in [Dwork et al., 2014]. It can also be interpreted as and is very related to the Max Divergence $D_\infty(Q, Q') := \max_{y\in\text{Support}(Q')}\ln\frac{\mathbb{P}(Q=y)}{\mathbb{P}(Q'=y)}$ between two probability distributions Q and Q'. If we view Q as the output distribution (the distribution of the sequentially pulled arms) when working over the true reward sequences $\boldsymbol{X}$ and view Q' as the output distribution when working over $\boldsymbol{X}'$, an ϵ-differentially private algorithm guarantees that the maximum divergence between Q and Q' is at most

[1] As shown in [Agrawal and Goyal, 2017], the Bernoulli reward setting can be generalized to any bounded reward setting.

ϵ, i.e., $\ln\left(\frac{\mathbb{P}\{Q=y\}}{\mathbb{P}\{Q'=y\}}\right) \leq \epsilon$ for all possible output y. Actually, the quantity $\ln\left(\frac{\mathbb{P}\{Q=y\}}{\mathbb{P}\{Q'=y\}}\right)$ is the privacy loss that is occurred when an adversary witnesses the outcome y.

2.3 RELATED WORK

In the classical stochastic bandit setting, the UCB-based, Thompson Sampling-based, and elimination-based algorithms all achieve good theoretical guarantees. Essentially, these algorithms rely on the empirical means to make decisions and the regret bounds take the $O(K\log(T)/\Delta)$ form.

As shown in Proposition 2.1 of [Dwork et al., 2014], differential privacy is invariant to post-processing, i.e., if a learning algorithm takes the output of an ϵ-differentially private algorithm as input, then the output of this learning algorithm itself is also ϵ-differentially private. In designing stochastic bandit algorithms with differential privacy, if the internal algorithm to compute the empirical mean is designed to be ϵ-differentially private, then following from the post-processing property, we can claim the bandit algorithm itself is ϵ-differentially private. This property has indeed been used in the design of private algorithms in previous work [Mishra and Thakurta, 2015, Tossou and Dimitrakakis, 2016, Sajed and Sheffet, 2019, Hu et al., 2021].

Mishra and Thakurta [2015] present the first differntially-private versions of the UCB and Thompson Sampling-based algorithms. However, the regret bounds they derive, $O\left(\frac{K\log^3(T)}{\epsilon\cdot\Delta}\right)$ and $O\left(\frac{K\log^3(T)}{\epsilon^2\cdot\Delta^2}\right)$, are far from the $\Omega\left(\frac{K\log(T)}{\Delta}+\frac{K\log(T)}{\epsilon}\right)$ regret lower bound that is derived later by Shariff and Sheffet [2018]. The key reason for the sub-optimality is using the T-bounded Binary Mechanism[2] [Dwork et al., 2010, Chan et al., 2011] to add random noise to mask the empirical mean for an arm. Furthermore, since their algorithms need to know the time horizon T in advance to calibrate the distribution of the random noise, they cannot be anytime learning algorithms. More importantly, their Thompson Sampling-based algorithm has some operational issues: in some rounds, the total reward $r_a(t)$ computed by the tree-based mechanism can take negative values, resulting in invalid parameters for the posterior distribution, the Beta distribution. Note that the parameters of Beta distributions must be non-negative. Our proposed learning algorithms carefully use clipping to address this issue.

Recently, two optimal algorithms have been proposed for differentially private stochastic bandits. Sajed and Sheffet [2019] propose DP-SE, an optimal elimination-style algorithm, and Hu et al. [2021] propose Anytime-Lazy-UCB, an optimal UCB-based algorithm. The key idea in achieving optimality is to use fresh observations to compute the differentially private empirical means, thus minimizing the number of noise variables needed. Although DP-SE, Anytime-Lazy-UCB, and our proposed Lazy-DP-TS are all optimal, as will be shown in Section 4, Lazy-DP-TS always outperforms the other two algorithms.

[2] Dwork et al. [2010], call it the Tree-based Mechanism, but the core idea is identical.

3 ALGORITHMS AND ANALYSIS

We now present our algorithms for achieving differential privacy in the stochastic bandit setting. The algorithms rely on two key ideas. The first is to use the differential privacy property of invariance to post-processing to make the arm selection algorithm differentially private due to the internal algorithm of computing empirical means being differentially private. The second is based on the principle of optimism in the face of uncertainty. Note that the decisions for the Thompson Sampling-based algorithms fully depend on the generated random samples from the posterior distributions. Operating under the optimism principle, we reshape the posterior distribution in an optimistic way: we shift the posterior distribution in the private algorithm towards the right as compared to the posterior distribution for the non-private Thompson Sampling. This shifting makes it more likely to draw a "good" posterior sample as compared to the draw in the non-private setting.

While both our algorithms rely on these fundamental concepts, the key difference between them lies in the design of the internal algorithm to compute the differentially private empirical means. Our first algorithm, Differentially Private Thompson Sampling (DP-TS), uses all the observations from the beginning in computing the differentially private empirical means, whereas our second algorithm, Lazy Differentially Private Thompson Sampling (Lazy-DP-TS) uses only a subsequence of observations.

Suppose at a given time step, arm j has a sequence of n observations $(x_1, x_2, \ldots, x_n)$. In DP-TS, all n observations will be used to compute the differentially private empirical mean for arm j. We partition $(x_1, x_2, \ldots, x_n)$ into $(x_1, x_2, \ldots, x_m)$ and $(x_{m+1}, x_{m+1}, \ldots, x_n)$, where $m = 2^{\lfloor\log(n+1)\rfloor} - 1$. This partition guarantees $n - m \leq m$, i.e., the length of the first subsequence is always no smaller than the length of the second subsequence. The internal algorithm composes two differentially private mechanisms, each being 0.5ϵ-differentially private and acting on each partition, respectively, to process these n observations.

The first mechanism is a modified version of the Logarithmic Mechanism [Chan et al., 2011] and works over $(x_1, x_2, \ldots, x_m)$: a differentially private aggregated reward of these m observations will be computed. According to the original mechanism by Chan et al. [2011], random noise would be added to the reward of an arm whenever the number of observations of that arm hits 2^r, for all $r \geq 0$, while in our modified version, random noise is added whenever the number of observations hits $2^{r+1}-1$, so that fresh noise is added at longer epochs, resulting in less overall noise. The second mechanism is the bounded Binary Mechanism [Chan et al., 2011] and works over $(x_{m+1}, x_{m+2}, \dots, x_n)$: random noise will be added based on the bounded Binary Mechanism and a differentially private aggregated reward of these $n-m$ observations will be output. The differentially private empirical mean is thus computed by aggregating the outputs of these two mechanisms.

Note that a given observation may be used more than once in the calculation of the empirical means over rounds, which means more noise is required to maintain the same degree of privacy. Based on this remark, we propose Lazy-DP-TS, where the internal algorithm only uses a subsequence of all the observations obtained so far to compute the differentially private empirical mean and no observation can be reused, i.e., once an observation has been used, it will be abandoned. The length of the subsequences double each time, i.e., the internal algorithm adds a random noise to every $2^r, r \geq 0$ observations and outputs a differentially private empirical mean. This restriction of using an observation only once in the calculation of the empirical mean minimizes added noise and is thus the key to the optimality of differentially private online learning algorithms.

Notation. Let $\text{Beta}(\alpha, \beta)$ be a Beta distribution with parameters α, β and $\text{Lap}(b)$ be a Laplace distribution centered at 0 with scale b. The pdfs of $\text{Beta}(\alpha, \beta)$ and $\text{Lap}(b)$ are shown in Facts 1 and 2 in Appendix A. $\log(x)$ is the base-2 logarithm of x and $\ln(x)$ is the base-e logarithm of x.

3.1 DP-TS

We now present DP-TS, followed by its guarantees.

3.1.1 Algorithm

We present some notation specific to this algorithm – used in this section and Appendix B. $O_j(t-1) := \sum_{s=1}^{t-1} \mathbf{1}\{J_s = j\}$ counts the number of pulls of arm j by the end of round $t-1$ and $\widehat{\mu}_{j,O_j(t-1)}$ is the empirical mean over these $O_j(t-1)$ observations. Let $\widetilde{\mu}_{j,O_j(t-1)}$ be the private empirical mean, i.e., $\widehat{\mu}_{j,O_j(t-1)}$ plus some noise.

DP-TS is presented in Algorithm 1. Lines 2 to 4 initialize the algorithm. We pull each arm once and set $\Psi_j = \{\}$ to hold future observations. Let C_j track the differentially private aggregated reward computed by the modified Logarithmic Mechanism and B_j track the private aggregated reward returned by the Binary Mechanism. Since for each arm the modified Logarithmic Mechanism processes observations in epochs, we use r_j to index the arm-specific epoch, i.e., the modified Logarithmic Mechanism will add a noise variable to mask the aggregated reward of 2^{r_j} observations at the end of epoch r_j. We initialize $r_j = 0$ and the initialization phase adds random noise to the first observation.

Algorithm 1 DP-TS

1: **Input:** Arm set $\mathcal{A}$ and privacy parameter ϵ
2: **for** $t = 1, 2, \dots, K$ **do**
3: Pull $J_t \leftarrow t$; Set $O_{J_t} \leftarrow 1$, $\Psi_{J_t} \leftarrow \{\}$, $C_{J_t} \leftarrow X_{J_t}(t) + \text{Lap}\left(\frac{1}{0.5\epsilon}\right)$, $r_{J_t} \leftarrow 0$, $B_{J_t} \leftarrow 0$, $\widetilde{\mu}_{J_t,O_{J_t}} \leftarrow \frac{C_{J_t}+B_{J_t}}{O_{J_t}}$
4: **end for**
5: **for** $t = K+1, K+2, \dots$ **do**
6: **for** $j \in \mathcal{A}$ **do**
7: Set $\overline{\mu}_{j,O_j} = \max\left\{0, \min\left\{\widetilde{\mu}_{j,O_j} + \frac{6\sqrt{8}\log(O_j+1)\log(t)}{\epsilon \cdot O_j}, 1\right\}\right\}$
8: Set $\widetilde{\alpha}_j \leftarrow \overline{\mu}_{j,O_j} \cdot O_j$, $\widetilde{\beta}_j \leftarrow (1-\overline{\mu}_{j,O_j}) \cdot O_j$
9: Sample $\theta_j(t) \sim \text{Beta}(\widetilde{\alpha}_j + 1, \widetilde{\beta}_j + 1)$
10: **end for**
11: Pull arm $J_t \in \arg\max_{j \in \mathcal{A}} \theta_j(t)$
12: Set $O_{J_t} \leftarrow O_{J_t} + 1$; Append $X_{J_t}(t)$ to Ψ_{J_t}
13: **if** $O_{J_t} = \sum_{s=0}^{r_{J_t}+1} 2^s$ **then**
14: Set $C_{J_t} \leftarrow C_{J_t} + \sum \Psi_{J_t} + \text{Lap}\left(\frac{1}{0.5\epsilon}\right)$
15: Set $\Psi_{J_t} \leftarrow \{\}$, $r_{J_t} \leftarrow r_{J_t} + 1$, $B_{J_t} \leftarrow 0$
16: **else**
17: Invoke $2^{r_{J_t}+1}$-bounded Binary Mechanism with Input $(0.5\epsilon, \Psi_{J_t})$ and Output B_{J_t}
18: **end if**
19: Set $\widetilde{\mu}_{J_t,O_{J_t}} \leftarrow \frac{C_{J_t}+B_{J_t}}{O_{J_t}}$.
20: **end for**

Let $v_{\epsilon,O_j(t-1),t} := \frac{6\sqrt{8}\log(O_j(t-1)+1)\log(t)}{\epsilon \cdot O_j(t-1)}$. For all the rounds $t \geq K+1$, we first compute $\overline{\mu}_{j,O_j(t-1)} = \max\left\{0, \min\left\{\widetilde{\mu}_{j,O_j(t-1)} + v_{\epsilon,O_j(t-1),t}, 1\right\}\right\}$. Note that the empirical means are clipped so that $\overline{\mu}_{j,O_j(t-1)} \in [0,1]$. We set $\widetilde{\alpha}_j(t) := \overline{\mu}_{j,O_j(t-1)} \cdot O_j(t-1)$ and $\widetilde{\beta}_j(t) := \left(1 - \overline{\mu}_{j,O_j(t-1)}\right) \cdot O_j(t-1)$. We then generate a random posterior sample $\theta_j(t) \sim \text{Beta}\left(\widetilde{\alpha}_j(t)+1, \widetilde{\beta}_j(t)+1\right)$ for each arm and pull the arm with the highest sample, i.e., $J_t \in \arg\max_{j \in \mathcal{A}} \theta_j(t)$. Since $\overline{\mu}_{j,O_j(t-1)} \in [0,1]$, the parameters of Beta distribution are valid.

To update the private empirical mean of the pulled arm, we append $X_{J_t}(t)$ to Ψ_{J_t}. If the number of observations in Ψ_{J_t} hits $2^{r_{J_t}+1}$, we add random noise drawn from $\text{Lap}\left(\frac{1}{0.5\epsilon}\right)$ and update C_{J_t}. Since now all observations in Ψ_{J_t} are used by the modified Logarithmic Mechanism, we reset Ψ_{J_t} and B_{J_t}, and increment r_{J_t} by one. If the number of observations in Ψ_{J_t} has not reached $2^{r_{J_t}+1}$, we invoke the $2^{r_{J_t}+1}$-bounded Binary Mechanism [Chan et al., 2011] taking Ψ_{J_t} as input and preserving 0.5ϵ-differential privacy. Note the number of observations in Ψ_{J_t} is at most $2^{r_{J_t}+1}$.

Remark. (a) r_j is determined by $O_j(t-1)$ as r_j will only increment by one whenever the number of observations in Ψ_j hits 2^{r_j+1}. Indeed, $r_j = \lfloor \log(O_j(t-1)+1) \rfloor - 1$. (b) Regarding the noise variables included in the differentially private empirical mean, there are exactly r_j+1 i.i.d. random variables that are drawn from Lap $\left(\frac{1}{0.5\epsilon}\right)$ and at most r_j+1 i.i.d. random variables that are drawn from Lap $\left(\frac{r_j+1}{0.5\epsilon}\right)$.

We now compare Algorithm 1 to the non-private Thompson Sampling by Agrawal and Goyal [2017]. Let $\alpha'_j(t) := \widehat{\mu}_{j,O_j(t-1)} \cdot O_j(t-1)$ be the number of successes and $\beta'_j(t) := \left(1-\widehat{\mu}_{j,O_j(t-1)}\right) \cdot O_j(t-1)$ be the number of failures among $O_j(t-1)$ Bernoulli trials. Recall that in the non-private Thompson Sampling, we draw $\theta'_j(t) \sim$ Beta $\left(\alpha'_j(t)+1, \beta'_j(t)+1\right)$. By adding $\upsilon_{\epsilon,O_j(t-1),t}$ to $\widetilde{\mu}_{j,O_j(t-1)}$, we have, with high probability, $\overline{\mu}_{j,O_j(t-1)} \geq \widehat{\mu}_{j,O_j(t-1)}$, i.e., the posterior distribution for the differentially private version is shifted towards the right as compared to the non-private version.

3.1.2 Analysis

We present privacy and regret guarantees for Algorithm 1.

Theorem 1. *Algorithm 1 is ϵ-differentially private.*

Proof. We first show the internal algorithm to compute the empirical mean, i.e., from Lines 12 to 19, is ϵ-differentially private. Then, from Proposition 2.1 of Dwork et al. [2014], we conclude that Algorithm 1 is ϵ-differentially private. Note that Lines 6 to 11 can be viewed as post-processing since in these steps, the learning algorithm does not touch any revealed observations. Suppose reward sequences $\boldsymbol{X}$ and $\boldsymbol{X}'$ differ in round h, i.e., the reward vectors $X_h = (X_1(h), \ldots, X_K(h))$ and $X'_h = (X'_1(h), \ldots, X'_K(h))$ are not the same. Note that changing from X_h to X'_h has no impact on other arms except arm J_h as only the reward of the pulled arm, J_h, is revealed in round h. Let $J_h = j$. At the end of round h, the differentially private empirical mean of arm j will be updated. According to Algorithm 1, changing from $X_j(h)$ to $X'_j(h)$ impacts C_j by at most 1. From Theorem 3.6 of Dwork et al. [2014], we know the internal algorithm to compute C_j (Line 14) is 0.5ϵ-differentially private. From Theorem 3.5 of Chan et al. [2011], we know the internal algorithm to compute B_j (Line 17) is 0.5ϵ-differentially private. Composing these two internal algorithms together, from Theorem 3.14 in [Dwork et al., 2014], we conclude that the internal algorithm (Line 19) to compute the differentially private empirical mean is ϵ-differentially private. □

Theorem 2. *The regret $\mathcal{R}_{DP\text{-}TS}(T)$ of Algorithm 1 is at most*

$$\sum_{j \in \mathcal{A}:\Delta_j>0} O\left(\max\left\{\frac{\log(T)}{\Delta_j}, \frac{\log(T)}{\epsilon}\log\left(\frac{\log(T)}{\epsilon\cdot\Delta_j}\right)\right\}\right) .$$

Remark. Several remarks are in order. (a): DP-TS is optimal up to a $\log\log(T)$ factor. (b): When setting $\epsilon \to \infty$, Algorithm 1 boils down to the same algorithm as the one by Agrawal and Goyal [2017]. However, our derived regret bound, Theorem 2, is only order-optimal instead of asymptotically optimal. Note that the regret bound of the non-private Thompson Sampling can be asymptotically optimal, i.e., a regret bound attaining the best possible coefficient for the leading term asymptotically. (c): Algorithm 1 also has an $O\left(\sqrt{KT\log(T)} + \frac{K\log(T)}{\epsilon}\log\left(\frac{\sqrt{T\log(T)}}{\sqrt{K}\epsilon}\right)\right)$ problem-independent regret bound. Note that it is known that Thompson Sampling is able to achieve the $\Omega\left(\sqrt{KT}\right)$ minimax lower bound for non-private stochastic bandits [Jin et al., 2021]. Therefore, the $O\left(\sqrt{KT\log(T)}\right)$ term in Theorem 2 is $\sqrt{\log(T)}$ far from being minimax optimal. Note that the price of introducing differential privacy is $\Omega\left(\frac{K\log(T)}{\epsilon}\right)$ [Shariff and Sheffet, 2018]. This lower bound implies DP-TS is $\log\left(\frac{\sqrt{T\log(T)}}{\sqrt{K}\epsilon}\right)$ far from being optimal in the private setting. The detailed proof for the problem-independent result is deferred to Appendix C.

We now provide a proof sketch for Theorem 2. The detailed proof is deferred to Appendix B. Let $\mathcal{F}_{t-1}$ collect all the history information containing the pulled arms, the rewards associated with the pulled arms, and the added noise. Define $\mathcal{F}_0 = \{\}$. Let $y_j := \mu_1 - \frac{\Delta_j}{6}$ and define $E_j^\theta(t)$ as the event that $\{\theta_j(t) \leq y_j\}$. Let $C_j(t-1)$ be the event that $\left\{\left|\mu_j - \widehat{\mu}_{j,O_j(t-1)}\right| \leq \sqrt{\frac{3\log(t)}{O_j(t-1)}}\right\}$. Let $G_j(t-1)$ be the event that $\left\{\left|\widehat{\mu}_{j,O_j(t-1)} - \widetilde{\mu}_{j,O_j(t-1)}\right| \leq \upsilon_{\epsilon,O_j(t-1),t}\right\}$.

Proof sketch of Theorem 2. We upper bound $\mathbb{E}[O_j(T)]$. Let $\mathcal{L}_j := \max\left\{\frac{108\log(T)}{\Delta_j^2}, \frac{72\log(T)}{\epsilon\cdot\Delta_j}\log\left(\frac{72\log(T)}{\epsilon\cdot\Delta_j}\right)\right\}$. We separate all T rounds into two regimes based on whether $O_j(t-1) \geq \mathcal{L}_j$. For all rounds t s.t. $O_j(t-1) < \mathcal{L}_j$, the total regret is at most $\mathcal{L}_j \cdot \Delta_j$. In a round when $O_j(t-1) \geq \mathcal{L}_j$, w.t.p., we have $\overline{\mu}_{j,O_j(t-1)} \leq \mu_j + \frac{4\Delta_j}{6}$, which implies $\overline{E_j^\theta(t)}$ is a low probability event. Meanwhile, w.h.p., we also have $\overline{\mu}_{1,O_1(t-1)} \geq \widehat{\mu}_{1,O_1(t-1)}$, which allows us to reduce the proof to the non-private setting.

With these ideas in hand, we have $\sum_{t=1}^T \mathbb{E}\left[\mathbf{1}\left\{J_t = j\right\}\right]$

$$\begin{aligned}
\leq \quad & \underbrace{\mathcal{L}_j + \sum_{t=1}^T \mathbb{P}\left\{\overline{C_j(t-1)}\right\} + \sum_{t=1}^T \mathbb{P}\left\{\overline{G_j(t-1)}\right\}}_{=:\omega_0} \\
+ \quad & \underbrace{\sum_{t=1}^T \mathbb{P}\left\{O_j(t-1) > \mathcal{L}_j, C_j(t-1), G_j(t-1), \overline{E_j^\theta(t)}\right\}}_{=:\omega_1} \\
+ \quad & \underbrace{\sum_{t=1}^T \mathbb{P}\left\{J_t = j, E_j^\theta(t)\right\}}_{=:\omega_2} \quad .
\end{aligned} \tag{1}$$

Via well-known concentration inequalities, we have $\omega_0 \leq O(1)$ (shown in Lemmas 8 and 9 in Appendix B). For ω_1, we use the argument that if events $C_j(t-1)$ and $G_j(t-1)$ are true simultaneously and arm j has been pulled at least $\mathcal{L}_j$ times, we have $\overline{\mu}_{j,O_j(t-1)} \leq \mu_j + \frac{4\Delta_j}{6}$. Since $\theta_j(t) \sim \text{Beta}\left(\widetilde{\alpha}_j(t)+1, \widetilde{\beta}_j(t)+1\right)$, from the properties of the Beta distribution, we know that it is very unlikely to draw $\theta_j(t) > \mu_j + \frac{5\Delta_j}{6}$. Lemma 11 in Appendix B shows that $\omega_1 \leq O(1)$.

The key challenge is to upper bound ω_2. We first reduce the proof to the non-private Thompson Sampling. Then, we reuse Lemmas 2.9 and 2.10 in [Agrawal and Goyal, 2017] to conclude the proof. Now, we show how to reduce the proof to the non-private setting. By introducing $G_1(t-1)$ and $\overline{G_1(t-1)}$, term ω_2 is at most

$$\sum_{t=1}^{T} \mathbb{P}\left\{J_t = j, G_1(t-1), E_j^{\theta}(t)\right\} \quad + \quad \sum_{t=1}^{T} \mathbb{P}\left\{\overline{G_1(t-1)}\right\}.$$

For the second term above, it is at most $O(1)$ (shown in Lemma 9). For the first term above, we have

$$\begin{aligned} &\sum_{t=1}^{T} \mathbb{P}\left\{J_t = j, G_1(t-1), E_j^{\theta}(t)\right\} \\ \leq \quad & \mathbb{E}\left[\sum_{t=1}^{T} \frac{\mathbb{P}\{\theta_1(t) \leq y_j \mid \mathcal{F}_{t-1}\}}{1-\mathbb{P}\{\theta_1(t) \leq y_j \mid \mathcal{F}_{t-1}\}} \{J_t = 1, G_1(t-1)\}\right] \\ \leq \quad & \mathbb{E}\left[\sum_{t=1}^{T} \frac{\mathbb{P}\left\{\theta_1'(t) \leq y_j \mid \mathcal{F}_{t-1}\right\}}{1-\mathbb{P}\left\{\theta_1'(t) \leq y_j \mid \mathcal{F}_{t-1}\right\}} \{J_t = 1\}\right], \end{aligned} \tag{2}$$

where $\theta_1'(t) \sim \text{Beta}\left(\alpha_j'(t)+1, \beta_j'(t)+1\right)$, the non-private posterior distribution for arm j conditioned on $\mathcal{F}_{t-1}$.

The first inequality in (2) links the probability of pulling a sub-optimal j to the probability of pulling the best arm by using Lemma 12 that we develop in Appendix B. The last inequality uses the fact that if $\overline{\mu}_{1,O_1(t-1)} \geq \widehat{\mu}_{1,O_1(t-1)}$, we have $\mathbb{P}\{\theta_1(t) \leq y_j \mid \mathcal{F}_{t-1}\} \leq \mathbb{P}\{\theta_1'(t) \leq y_j \mid \mathcal{F}_{t-1}\}$, i.e., $\text{Beta}\left(\widetilde{\alpha}_j(t)+1, \widetilde{\beta}_j(t)+1\right)$ stochastically dominates $\text{Beta}\left(\alpha_j'(t)+1, \beta_j'(t)+1\right)$. Since the proof now is reduced to the non-private setting, slightly modifying Lemmas 2.9 and 2.10 in [Agrawal and Goyal, 2017] concludes the proof. Lemma 10 in Appendix B shows $\omega_2 \leq O\left(\frac{\log(T)}{\Delta_j^2}\right)$. □

3.2 LAZY-DP-TS

We now present Lazy-DP-TS and its guarantees. The idea to achieve optimality is limiting the number of times an observation is used in computing the empirical mean to one.

3.2.1 Algorithm

We first present some notation specific to this algorithm – used in this section and Appendix D. Let $O_j(t-1)$ denote the number of observations that are used to compute the differentially private empirical mean and $\widehat{\mu}_{j,O_j(t-1)}$ denote the empirical mean of these $O_j(t-1)$ observations. Let $\widetilde{\mu}_{j,O_j(t-1)}$ be the differentially private empirical mean.

Algorithm 2 Lazy-DP-TS

1: **Input:** Arm set $\mathcal{A}$ and privacy parameter ϵ
2: **for** $t = 1, 2, \ldots, K$ **do**
3: Pull $J_t \leftarrow t$; Set $O_{J_t} \leftarrow 1$, $\widetilde{\mu}_{J_t,O_{J_t}} \leftarrow X_{J_t}(t) + \text{Lap}\left(\frac{1}{\epsilon}\right)$, $r_{J_t} \leftarrow 0$, $\Psi_{J_t} \leftarrow \{\}$
4: **end for**
5: **for** $t = K+1, K+2, \ldots$ **do**
6: **for** $j \in \mathcal{A}$ **do**
7: Set $\overline{\mu}_{j,O_j} = \max\left\{0, \min\left\{\widetilde{\mu}_{j,O_j} + \frac{3\log(t)}{\epsilon \cdot O_j}, 1\right\}\right\}$
8: Set $\widetilde{\alpha}_j \leftarrow \overline{\mu}_{j,O_j} \cdot O_j$, $\widetilde{\beta}_j \leftarrow (1-\overline{\mu}_{j,O_j}) \cdot O_j$
9: Sample $\theta_j(t) \sim \text{Beta}(\widetilde{\alpha}_j + 1, \widetilde{\beta}_j + 1)$
10: **end for**
11: Pull $J_t \in \arg\max_{j \in \mathcal{A}} \theta_j(t)$
12: Append $X_{J_t}(t)$ to Ψ_{J_t}
13: **if** number of observations in Ψ_{J_t} hits $2^{r_{J_t}+1}$ **then**
14: Set $O_{J_t} \leftarrow 2^{r_{J_t}+1}$, $\widetilde{\mu}_{J_t,O_{J_t}} \leftarrow \frac{\sum \Psi_{J_t} + \text{Lap}\left(\frac{1}{\epsilon}\right)}{O_{J_t}}$
15: Set $r_{J_t} \leftarrow r_{J_t} + 1$, $\Psi_{J_t} \leftarrow \{\}$
16: **end if**
17: **end for**

Lazy-DP-TS is presented in Algorithm 2. Lines 2 to 4 are the initialization. We pull each arm once and add random noise that is drawn from $\text{Lap}\left(\frac{1}{\epsilon}\right)$ to the obtained observation to initialize the differentially private empirical mean. We still use 2^{r_j} to track the number of observations that have been used to compute the differentially private empirical mean for arm j. Initially, we set $r_j = 0$ and $\Psi_j = \{\}$ to hold future observations.

For all rounds $t \geq K+1$, we first compute $\overline{\mu}_{j,O_j(t-1)} = \max\left\{0, \min\left\{\widetilde{\mu}_{j,O_j(t-1)} + \frac{3\log(t)}{\epsilon \cdot O_j(t-1)}, 1\right\}\right\}$ and then compute $\widetilde{\alpha}_j(t) := \overline{\mu}_{j,O_j(t-1)} \cdot O_j(t-1)$ and $\widetilde{\beta}_j(t) := \left(1-\overline{\mu}_{j,O_j(t-1)}\right) \cdot O_j(t-1)$. Next, we generate a posterior sample $\theta_j(t) \sim \text{Beta}\left(\widetilde{\alpha}_j(t)+1, \widetilde{\beta}_j(t)+1\right)$ for each arm and pull the arm with the highest posterior sample, i.e., $J_t \in \arg\max_{j \in \mathcal{A}} \theta_j(t)$.

To process $X_{J_t}(t)$, we append it in Ψ_{J_t}. However, we may not update the differentially private empirical mean of the pulled arm in round t. We will only update it when the number of observations in Ψ_{J_t} hits $2^{r_{J_t}+1}$ and the updated differentially private empirical mean will be based on observations in Ψ_{J_t} only, i.e., the updated differentially private empirical mean is computed by adding a noise variable drawn from $\text{Lap}\left(\frac{1}{\epsilon}\right)$ to these fresh $2^{r_{J_t}+1}$ observations. Since now all observations in Ψ_{J_t} are used, we reset Ψ_{J_t} and increment r_{J_t} by one.

Remark. (a) The number of observations used to compute the differentially private empirical mean doubles each time, i.e., $O_j(t-1)$ takes values from $2^{r_j}, r_j \geq 0$. (b) The number of noise variables included in the private empirical mean of arm j is always 1 and it is drawn from Lap $\left(\frac{1}{\epsilon}\right)$.

3.2.2 Analysis

We now present privacy and regret guarantees for Algorithm 2.

Theorem 3. *Algorithm 2 is ϵ-differentially private.*

Proof. The internal algorithm to compute the differentially private empirical mean is shown in Lines 12 to 16 in Algorithm 2. Lines 5 to 11 can be viewed as post-processing. Now, we show that the internal algorithm is ϵ-differentially private. Suppose reward sequences $\boldsymbol{X}$ and $\boldsymbol{X}'$ differ in round h, i.e., the reward vectors $X_h = (X_1(h), \ldots, X_K(h))$ and $X'_h = (X'_1(h), \ldots, X'_K(h))$ are not the same. The changing from X_h to X'_h can only impact arm J_h. Let $J_h = j$. Since arm j's differentially private means are always based on fresh observations, the changing from X_h to X'_h can only impact the differentially private aggregated reward of arm j once and by at most 1. By adding a noise variable drawn from Lap $\left(\frac{1}{\epsilon}\right)$ to $\sum \Psi_j$, from Theorem 3.6 in [Dwork et al., 2014], we know that the internal algorithm to compute the differentially private empirical mean is ϵ-differentially private. ☐

Theorem 4. *The regret $\mathcal{R}_{\textit{Lazy-DP-TS}}(T)$ of Algorithm 2 is at most* $\sum_{j \in \mathcal{A}: \Delta_j > 0} O\left(\frac{\log(T)}{\min\{\epsilon, \Delta_j\}}\right)$.

Remark. Several remarks are in order. (a): Lazy-DP-TS is (order)-optimal as its regret upper bound matches the $\Omega\left(\sum_{j \in \mathcal{A}: \Delta_j > 0} \frac{\log(T)}{\Delta_j} + \frac{\log(T)}{\epsilon}\right)$ regret lower bound of Shariff and Sheffet [2018]. Our Lazy-DP-TS preserves the same regret guarantee as the one for Anytime-Lazy-UCB by Hu et al. [2021] and DP-SE by Sajed and Sheffet [2019]. However, as will be shown in Section 4, Lazy-DP-TS has better practical performance than Anytime-Lazy-UCB and DP-SE. Since Algorithm 2 drops observations as it learns, even if we set $\epsilon \to \infty$, the regret bound can never be asymptotically optimal. (b): Algorithm 2 also has an $O\left(\sqrt{KT\log(T)} + \frac{K\log(T)}{\epsilon}\right)$ problem-independent regret bound. Since the price of introducing differential privacy is $\Omega\left(\frac{K\log(T)}{\epsilon}\right)$, the $O\left(\frac{K\log(T)}{\epsilon}\right)$ term in Theorem 4 cannot be improved as it matches the lower bound of introducing differential privacy. Therefore, Lazy-DP-TS is minimax optimal up to a $\sqrt{\log(T)}$ factor in both private setting and non-private setting. The detailed proof for the problem-independent result is deferred to Appendix E.

We now present a proof sketch for Theorem 4. The detailed proof is deferred to Appendix D. We still define $C_j(t-1)$ as the event that the confidence interval of the empirical mean holds and $G_j(t-1)$ as the event that the noise injected is not too much. Let $\mathcal{F}_{t-1}$ collect all the history information and set $y_j := \mu_1 - \frac{\Delta_j}{6}$. Let event $E_j^\theta(t) := \{\theta_j(t) \leq y_j\}$.

Proof sketch of Theorem 4. We still upper bound the expected number of pulls of a sub-optimal arm j. However, we cannot separate all T rounds into two regimes since Algorithm 2 drops observations. Instead, we perform a decomposition as follows.

$$\begin{aligned} & \sum_{t=1}^{T} \mathbb{E}\left[\mathbf{1}\left\{J_t = j\right\}\right] \\ \leq \quad & \underbrace{\sum_{t=1}^{T} \mathbb{P}\left\{J_t = j, C_j(t-1), G_j(t-1), \overline{E_j^\theta(t)}\right\}}_{=:\omega_1} \\ + \quad & \underbrace{\sum_{t=1}^{T} \mathbb{P}\left\{J_t = j, E_j^\theta(t), G_1(t-1)\right\}}_{=:\omega_2} + O(1) \quad . \end{aligned} \tag{3}$$

Lemma 13 and Lemma 14 in Appendix D together show that the $O(1)$ term in (3) is an upper bound on $\sum_{t=1}^{T} \mathbb{P}\left\{\overline{C_j(t-1)}\right\} + \mathbb{P}\left\{\overline{G_j(t-1)}\right\} + \mathbb{P}\left\{\overline{G_1(t-1)}\right\}$.

To upper bound ω_1, we let $\mathcal{L}_j := \frac{72 \cdot \log(T)}{\Delta_j \cdot \min\{\epsilon, \Delta_j\}}$ and $d_j := \log\left(\mathcal{L}_j\right)$. Recall that for arm j, the numbers of observations that are used to compute the differentially private empirical means are 2^{r_j} for $0 \leq r_j \leq \log(T)$. Let τ_{r_j} be the round such that at the end of round τ_{r_j}, the learning algorithm will use 2^{r_j} observations to update the differentially private empirical mean for arm j. We separate $0 \leq r_j \leq \log(T)$ into two parts. The first part is when $0 \leq r_j \leq d_j$. Based on the definition of τ_{r_j}, we know that the total number of pulls of arm j is at most $\sum_{s=0}^{d_j} 2^s \leq O\left(\frac{\log(T)}{\Delta_j \cdot \min\{\epsilon, \Delta_j\}}\right)$ in all rounds up to (and including) τ_{d_j}. When $d_j < r_j \leq \log(T)$, we have $2^{r_j} > \mathcal{L}_j$, i.e., we have accumulated "enough" observations for arm j. For a fixed r_j, with high probability, the expected number of pulls of arm j is at most $O(1)$ in all rounds $t \in \left\{\tau_{r_j} + 1, \ldots, \tau_{r_j+1}\right\}$. Then, we know that the total expected number of pulls is at most $O(\log(T))$ in all rounds from $\tau_{d_j} + 1$ up to T. Lemma 16 in Appendix D shows $\omega_1 \leq O\left(\frac{\log(T)}{\Delta_j \cdot \min\{\epsilon, \Delta_j\}}\right)$.

The challenge still lies in upper bounding ω_2. We again use the ideas shown in (2) to reduce the proof to the non-private setting. We have

$$\omega_2 \leq \mathbb{E}\left[\sum_{t=1}^{T} \frac{\mathbb{P}\left\{\theta'_1(t) \leq y_j \mid \mathcal{F}_{t-1}\right\}}{1 - \mathbb{P}\left\{\theta'_1(t) \leq y_j \mid \mathcal{F}_{t-1}\right\}} \left\{J_t = 1\right\}\right] \quad . \tag{4}$$

However, now we cannot reuse Lemmas 2.9 and 2.10 from [Agrawal and Goyal, 2017] directly due to the fact that the observations for arm 1 are also dropped during the learning. To tackle this challenge, we separate all T rounds into multiple intervals based on whether arm 1's empirical mean is updated or not. Let τ_r be the round such that at the end of round τ_r, the learning algorithm will use 2^r observations for arm 1 to update arm 1's empirical mean, i.e., in all rounds $t \in \{\tau_r + 1, \ldots, \tau_{r+1}\}$, the posterior distribution for $\theta_1'(t)$ stays the same. Then, we have

$$\begin{aligned} \omega_2 &\leq \mathbb{E}\left[\sum_{r=0}^{\log(T)} \sum_{t=\tau_r+1}^{\tau_{r+1}} \frac{\mathbb{P}\{\theta_1'(t) \leq y_j | \mathcal{F}_{t-1}\}}{\mathbb{P}\{\theta_1'(t) > y_j | \mathcal{F}_{t-1}\}} \{J_t = 1\}\right] \\ &= \sum_{r=0}^{\log(T)} \mathbb{E}\left[\frac{\mathbb{P}\{\theta_1'(\tau_r+1) \leq y_j | \mathcal{F}_{\tau_r}\}}{\mathbb{P}\{\theta_1'(\tau_r+1) > y_j | \mathcal{F}_{\tau_r}\}} \sum_{t=\tau_r+1}^{\tau_{r+1}} \{J_t = 1\}\right] \\ &\leq \sum_{r=0}^{\log(T)} 2^{r+1} \cdot \mathbb{E}\left[\frac{\mathbb{P}\{\theta_1'(\tau_r+1) \leq y_j | \mathcal{F}_{\tau_r}\}}{\mathbb{P}\{\theta_1'(\tau_r+1) > y_j | \mathcal{F}_{\tau_r}\}}\right] . \end{aligned} \tag{5}$$

The last inequality uses the fact that the number of pulls for arm 1 in all rounds $t \in \{\tau_r + 1, \ldots, \tau_{r+1}\}$ is at most 2^{r+1} based on the definition of τ_{r+1}. Let $d_1 := \log\left(\frac{8}{\mu_1 - y_j}\right)$. We now analyze two cases separately based on whether $0 \leq r \leq \lfloor d_1 \rfloor$ or $r \geq \lceil d_1 \rceil$. By using Lemma 2.9 of Agrawal and Goyal [2017] and other analysis, we have $\sum_{r=0}^{\lfloor d_1 \rfloor} 2^{r+1} \mathbb{E}\left[\frac{\mathbb{P}\{\theta_1'(\tau_r+1) \leq y_j | \mathcal{F}_{\tau_r}\}}{\mathbb{P}\{\theta_1'(\tau_r+1) > y_j | \mathcal{F}_{\tau_r}\}}\right] \leq O\left(\frac{1}{\Delta_j^2}\right)$ and $\sum_{r=\lceil d_1 \rceil}^{\log(T)} 2^{r+1} \mathbb{E}\left[\frac{\mathbb{P}\{\theta_1'(\tau_r+1) \leq y_j | \mathcal{F}_{\tau_r}\}}{\mathbb{P}\{\theta_1'(\tau_r+1) > y_j | \mathcal{F}_{\tau_r}\}}\right] \leq O\left(\frac{\log(T)}{\Delta_j^2}\right)$. Lemma 15 in Appendix D shows $\omega_2 \leq O\left(\frac{\log(T)}{\Delta_j^2}\right)$. $\square$

4 EXPERIMENTAL RESULTS

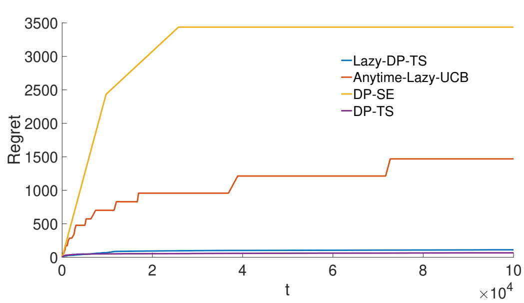

Figure 1: $\epsilon = 500$

We compare the practical performance among DP-TS, Lazy-DP-TS, DP-SE, and Anytime-Lazy-UCB under the experimental setting that has been used in [Sajed and Sheffet, 2019], i.e., we have $K = 5$ arms with mean rewards setting to $0.75, 0.625, 0.5, 0.375, 0.25$ and the privacy parameter ϵ setting to $0.1, 0.25, 0.5, 1.0, 500$. We set $T = 10^5$. Figure 1 shows the results of the setting where $\epsilon = 500$. It is not surprising that DP-TS outperforms Lazy-DP-TS as when ϵ is very large, DP-TS is asymptotically optimal while Lazy-DP-TS can only be order-optimal. Also, just as expected,

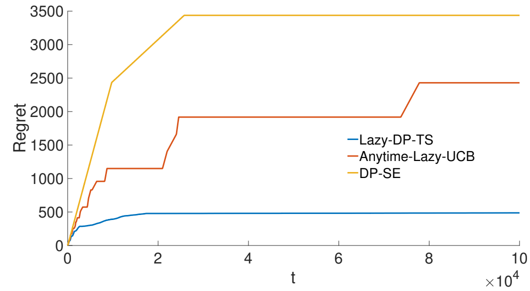

Figure 2: $\epsilon = 1.0$

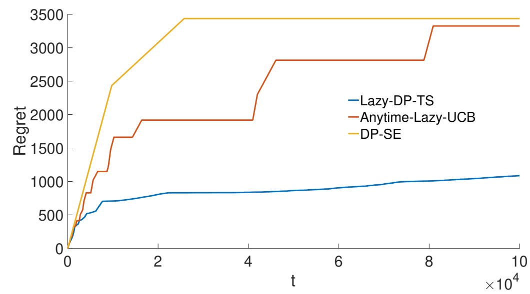

Figure 3: $\epsilon = 0.5$

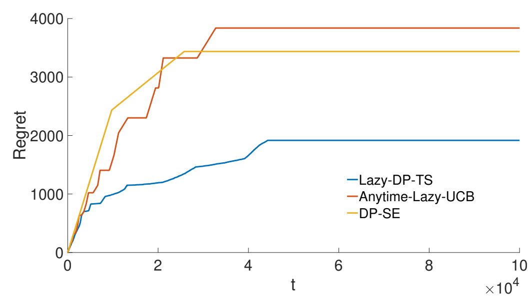

Figure 4: $\epsilon = 0.25$

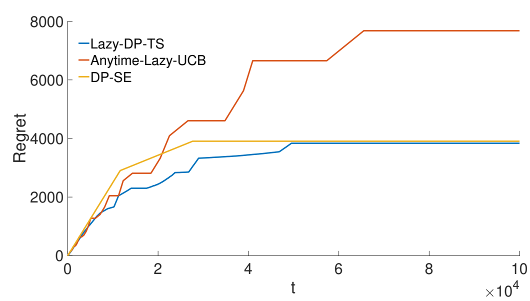

Figure 5: $\epsilon = 0.10$

Thompson Sampling-based algorithms outperform the UCB-based and elimination-style algorithms. Figures 2 to 5 show the results of the settings where $\epsilon = 1.0, 0.5, 0.25, 0.1$, we skip the plots of DP-TS as the practical performance of DP-TS is inferior to the remaining three optimal algorithms when ϵ is very small. From the experimental results we can see that Lazy-DP-TS always outperforms DP-SE and

Anytime-Lazy-UCB. More experimental results, including comparison of private and non-private algorithms, can be found in Appendix F.

5 CONCLUSION

We have presented optimal Thompson Sampling-based algorithms for differentially private stochastic bandits, filling a gap in the literature for differentially private online learning. The ideas used in this paper also contribute to developing optimal algorithms for other settings such as differentially private combinatorial multi-armed bandits [Chen et al., 2020]. Note that both the UCB and elimination-based algorithms are deterministic. So far, our proposed algorithms have not used the unique feature that only Thompson Sampling-based algorithms have, the randomness inherent in the learning algorithms. An interesting future direction is the design of optimal private Thompson Sampling-based algorithms using the fact that a random posterior sample may provide a degree of differential privacy for free [Wang et al., 2015, Foulds et al., 2016].

ACKNOWLEDGEMENTS

This work is supported by Amii Post-Doctoral Fellowships.

References

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

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

Sébastien Bubeck and Nicolo Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *arXiv preprint arXiv:1204.5721*, 2012.

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

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

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

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

Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3-4):211–407, 2014.

James Foulds, Joseph Geumlek, Max Welling, and Kamalika Chaudhuri. On the theory and practice of privacy-preserving bayesian data analysis. In *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence*, pages 192–201, 2016.

Aditya Gopalan, Shie Mannor, and Yishay Mansour. Thompson sampling for complex online problems. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, page I–100–I–108. JMLR.org, 2014.

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

Tianyuan Jin, Pan Xu, Jieming Shi, Xiaokui Xiao, and Quanquan Gu. Mots: Minimax optimal thompson sampling. In *International Conference on Machine Learning*, pages 5074–5083. PMLR, 2021.

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

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.

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

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

Yu-Xiang Wang, Stephen Fienberg, and Alex Smola. Privacy for free: Posterior sampling and stochastic gradient monte carlo. In *International Conference on Machine Learning*, pages 2493–2502. PMLR, 2015.

APPENDIX

The organization of the appendix is as follows.

1. Appendix A: Useful facts and inequalities ;
2. Appendix B: Proofs for Theorem 2 ;
3. Appendix C: Proofs for Problem-independent regret bound for Algorithm 1 ;
4. Appendix D: Proofs for Theorem 4 ;
5. Appendix E: Proofs for Problem-independent regret bound for Algorithm 2 ;
6. Appendix F: Additional experimental results .

A USEFUL FACTS AND INEQUALITIES

Fact 1. *(Beta distribution). The probability density function of a Beta distribution with parameters $\alpha, \beta > 0$, i.e., Beta(α, β), is*

$$f(x; \alpha, \beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\int_0^1 x^{\alpha-1}(1-x)^{\beta-1} dx} \quad . \tag{6}$$

Fact 2. *(Laplace distribution). The probability density function of a Laplace distribution (centered at 0) with scale b, i.e., Lap(b), is*

$$f(x; b) = \frac{1}{2b} e^{-\frac{|x|}{b}} \quad . \tag{7}$$

Fact 3. *(Hoeffding's inequality). Let $X_1, X_2, \ldots, X_n$ be n independent random variables with common range $[0, 1]$. Let $X = \sum_{j=1}^{n} X_j$. Then, for all $a > 0$, we have*

$$\mathbb{P}\{|X - \mathbb{E}[X]| \geq a\} \leq 2e^{-2a^2/n} \quad . \tag{8}$$

Fact 4. *(Chernoff-Hoeffding bound). Let $X_1, X_2, \ldots, X_n$ be n independent random variables with each $X_j \in \{0, 1\}$. Let $X = \frac{1}{n} \sum_{j=1}^{n} X_j$ and $\mu = \mathbb{E}[X]$. Then, for any $0 < \lambda < 1 - \mu$, we have*

$$\mathbb{P}\{X \geq \mu + \lambda\} \leq e^{-n \cdot d_{KL}(\mu+\lambda, \mu)} \quad , \tag{9}$$

and for any $0 < \lambda < \mu$, we have

$$\mathbb{P}\{X \geq \mu - \lambda\} \leq e^{-n \cdot d_{KL}(\mu-\lambda, \mu)} \quad , \tag{10}$$

where $d_{KL}(x, y) := x \ln\left(\frac{x}{y}\right) + (1-x)\ln\left(\frac{1-x}{1-y}\right)$ is the KL-divergence between two Bernoulli distributions with parameters x and y.

Fact 5. *For all positive integers α, β, we have*

$$F^{Beta}_{\alpha,\beta}(y) = 1 - F^{B}_{\alpha+\beta-1,y}(\alpha - 1) \quad , \tag{11}$$

where $F^{Beta}_{\alpha,\beta}(\cdot)$ is the cdf of a Beta distribution with parameters α, β and $F^{B}_{n,p}(\cdot)$ is the cdf of a Binomial distribution with parameters n, p.

Fact 6. *Let $Y \sim$ Lap(b), for any $\delta > 0$, we have*

$$\mathbb{P}\left\{|Y| \geq b \ln\left(\frac{1}{\delta}\right)\right\} \leq \delta \quad . \tag{12}$$

Fact 7. *(Corollary 2.9 in [Chan et al., 2011]). Let $Y_1, Y_2, \ldots Y_N$ be N i.i.d. random variables that are drawn according to distribution Lap (b). Let $Y := \sum_{i=1}^{N} Y_i$. Suppose $0 < \delta < 1$ and $v > b \cdot \max\left\{\sqrt{N}, \sqrt{\ln\left(\frac{2}{\delta}\right)}\right\}$. Then, we have*

$$\mathbb{P}\left\{|Y| > v\sqrt{8\ln\left(\frac{2}{\delta}\right)}\right\} \leq \delta \quad . \tag{13}$$

B PROOFS FOR THEOREM 2

Recall that $O_j(t-1)$ is the number of observations for arm j obtained from the very beginning until the end of round $t-1$; $\widehat{\mu}_{j,O_j(t-1)}$ is the empirical mean for arm j by the end of round $t-1$; $\widetilde{\mu}_{j,O_j(t-1)}$ is the differentially private empirical mean. Also, recall that $C_j(t-1)$ is the event such that $\left\{\left|\mu_j - \widehat{\mu}_{j,O_j(t-1)}\right| \leq \sqrt{\frac{3\log(t)}{O_j(t-1)}}\right\}$; $G_j(t-1)$ is the event such that $\left\{\left|\widehat{\mu}_{j,O_j(t-1)} - \widetilde{\mu}_{j,O_j(t-1)}\right| \leq \frac{6\sqrt{8}\log(O_j(t-1)+1)\log(t)}{\epsilon \cdot O_j(t-1)}\right\}$.

We now present two lemmas to show that, for every arm including the best arm, the probability that the complementary event of $C_j(t-1)$ or $G_j(t-1)$ occurs is very low.

Lemma 8. *For any arm j, we have*

$$\mathbb{P}\left\{\overline{C_j(t-1)}\right\} \leq O\left(\frac{1}{t^2}\right) \quad . \tag{14}$$

Lemma 9. *For any arm j, we have*

$$\mathbb{P}\left\{\overline{G_j(t-1)}\right\} \leq O\left(\frac{1}{t^2}\right) \quad . \tag{15}$$

For a sub-optimal arm j, recall that $\mathcal{L}_j = \max\left\{\frac{108\log(T)}{\Delta_j^2}, \frac{72\log(T)}{\epsilon \cdot \Delta_j}\log\left(\frac{72\log(T)}{\epsilon \cdot \Delta_j}\right)\right\}$ and $y_j = \mu_1 - \frac{\Delta_j}{6}$. Also, recall that $E_j^\theta(t)$ is the good event that the posterior sample of a sub-optimal arm j in round t is not too close to the mean reward of the best arm, i.e., event $E_j^\theta(t) = \{\theta_j(t) \leq y_j\}$.

We now present two key lemmas in order to complete the proof of Theorem 2. Lemma 10 states that the learning algorithm will only make a small amount of mistakes, i.e., at most $O\left(\frac{\log(T)}{\Delta_j^2}\right)$ mistakes, for pulling a sub-optimal arm j instead of the optimal arm, if the posterior samples of a sub-optimal arm j are always far from the mean reward of the best arm. Lemma 11 states that once a sub-optimal arm j has been pulled enough, i.e., at least $\mathcal{L}_j$ times, it is very unlikely to draw the posterior sample of arm j close to the mean reward of the best arm.

Lemma 10. *For any sub-optimal arm j, we have*

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}\left\{J_t = j, E_j^\theta(t)\right\}\right] \leq O\left(\frac{\log(T)}{\Delta_j^2}\right) \quad . \tag{16}$$

Lemma 11. *For any sub-optimal arm j, we have*

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}\left\{J_t = j, O_j(t-1) > \mathcal{L}_j, C_j(t-1), G_j(t-1), \overline{E_j^\theta(t)}\right\}\right] \leq O(1) \quad . \tag{17}$$

After these preparations, we now prove Theorem 2.

Proof of Theorem 2. We do the regret decomposition first and have

$$
\begin{aligned}
\mathcal{R}_{\text{DP-TS}}(T) &= \sum_{j\in\mathcal{A}:\Delta_j>0} \mathbb{E}\left[\sum_{t=1}^{T}\mathbf{1}\left\{J_t=j\right\}\right]\cdot\Delta_j \\
&\leq \sum_{j\in\mathcal{A}:\Delta_j>0}\mathcal{L}_j\cdot\Delta_j + \sum_{j\in\mathcal{A}:\Delta_j>0}\mathbb{E}\left[\sum_{t=1}^{T}\mathbf{1}\left\{J_t=j,O_j(t-1)>\mathcal{L}_j\right\}\right]\cdot\Delta_j \\
&\leq \sum_{j\in\mathcal{A}:\Delta_j>0}\mathcal{L}_j\cdot\Delta_j + \sum_{j\in\mathcal{A}:\Delta_j>0}\underbrace{\mathbb{E}\left[\sum_{t=1}^{T}\mathbf{1}\left\{J_t=j,O_j(t-1)>\mathcal{L}_j,C_j(t-1),G_j(t-1)\right\}\right]}_{=:\eta}\cdot\Delta_j \\
&+ \sum_{j\in\mathcal{A}:\Delta_j>0}\sum_{t=1}^{T}\underbrace{\mathbb{P}\left\{\overline{C_j(t-1)}\right\}}_{\leq O\left(\frac{1}{t^2}\right),\text{ Lemma 8}}\cdot\Delta_j + \sum_{j\in\mathcal{A}:\Delta_j>0}\sum_{t=1}^{T}\underbrace{\mathbb{P}\left\{\overline{G_j(t-1)}\right\}}_{\leq O\left(\frac{1}{t^2}\right),\text{ Lemma 9}}\cdot\Delta_j \quad .
\end{aligned}
\tag{18}
$$

We simply upper bound the indicator function directly for the first term in (18). For the third term, we apply Lemma 8 to upper bound it and for the last term, we apply Lemma 9 to upper bound it.

We now further decompose the η term in (18) based on $E_j^\theta(t)$ and $\overline{E_j^\theta(t)}$. We have

$$
\begin{aligned}
\eta &= \mathbb{E}\left[\sum_{t=1}^{T}\mathbf{1}\left\{J_t=j,O_j(t-1)>\mathcal{L}_j,C_j(t-1),G_j(t-1)\right\}\right] \\
&\leq \underbrace{\mathbb{E}\left[\sum_{t=1}^{T}\mathbf{1}\left\{J_t=j,E_j^\theta(t)\right\}\right]}_{\leq O\left(\frac{\log(T)}{\Delta_j^2}\right),\text{ Lemma 10}} + \underbrace{\mathbb{E}\left[\sum_{t=1}^{T}\mathbf{1}\left\{J_t=j,O_j(t-1)>\mathcal{L}_j,\overline{E_j^\theta(t)},C_j(t-1),G_j(t-1)\right\}\right]}_{\leq O(1),\text{ Lemma 11}} \quad .
\end{aligned}
\tag{19}
$$

By using Lemma 10, we know the first term in (19) is at most $O\left(\frac{\log(T)}{\Delta_j^2}\right)$; by using Lemma 11, we know the second term in (19) is at most $O(1)$. Then, we have $\eta \leq O\left(\frac{\log(T)}{\Delta_j^2}\right)$. By plugging the upper bound of η into (18), we have

$$
\begin{aligned}
\mathcal{R}_{\text{DP-TS}} &\leq \sum_{j\in\mathcal{A}:\Delta_j>0}\mathcal{L}_j\cdot\Delta_j + O\left(\frac{\log(T)}{\Delta_j^2}\right)\cdot\Delta_j + \sum_{t=1}^{T}O\left(\frac{1}{t^2}\right) \\
&\leq \sum_{j\in\mathcal{A}:\Delta_j>0}O\left(\max\left\{\frac{\log(T)}{\Delta_j},\frac{\log(T)}{\epsilon}\log\left(\frac{\log(T)}{\epsilon\cdot\Delta_j}\right)\right\}\right)+O(1) \quad ,
\end{aligned}
\tag{20}
$$

which concludes the proof. □

We now present the proofs of Lemma 8 and Lemma 9.

Proof of Lemma 8. To prove this lemma, only Hoeffding's inequality (shown in Fact 3) is needed. We have

$$
\begin{aligned}
&\mathbb{P}\left\{\overline{C_j(t-1)}\right\} \\
\leq\ &\mathbb{P}\left\{\left|\mu_j-\widehat{\mu}_{j,O_j(t-1)}\right|>\sqrt{\frac{3\log(t)}{O_j(t-1)}}\right\} \\
\leq\ &\mathbb{P}\left\{\left|\mu_j-\widehat{\mu}_{j,O_j(t-1)}\right|>\sqrt{\frac{4\ln(t)}{O_j(t-1)}}\right\} \\
\leq\ &\sum_{h=1}^{t-1}\mathbb{P}\left\{\left|\mu_j-\widehat{\mu}_{j,h}\right|>\sqrt{\frac{4\ln(t)}{h}}\right\} \\
\leq\ &\sum_{h=1}^{t-1}2e^{-2h\frac{4\ln(t)}{h}} \\
\leq\ &t\cdot\frac{2}{t^3} \\
\leq\ &O\left(\frac{1}{t^2}\right) \quad .
\end{aligned}
\tag{21}
$$

□

Proof of Lemma 9. We use Corollary 2.9 in [Chan et al., 2011] to prove this lemma. Corollary 2.9 of Chan et al. [2011] itself is shown in Fact 7. Let $\widehat{A}_{j,O_j(t-1)} := \widehat{\mu}_{j,O_j(t-1)} \cdot O_j(t-1)$ be the true aggregated reward over $O_j(t-1)$ observations. Let $\widetilde{A}_{j,O_j(t-1)} := \widetilde{\mu}_{j,O_j(t-1)} \cdot O_j(t-1)$ be the differentially private aggregated reward.

We have

$$\begin{aligned}
& \mathbb{P}\left\{\overline{G_j(t-1)}\right\} \\
= \;& \mathbb{P}\left\{\left|\widetilde{\mu}_{j,O_j(t-1)} - \widehat{\mu}_{j,O_j(t-1)}\right| > \tfrac{6\sqrt{8}\log(O_j(t-1)+1)\cdot\log(t)}{\epsilon\cdot O_j(t-1)}\right\} \\
= \;& \mathbb{P}\left\{\left|\widetilde{A}_{j,O_j(t-1)} - \widehat{A}_{j,O_j(t-1)}\right| > \tfrac{6\sqrt{8}\log(O_j(t-1)+1)\cdot\log(t)}{\epsilon}\right\} \\
\leq \;& \sum_{h=1}^{t-1} \mathbb{P}\left\{\left|\widetilde{A}_{j,h} - \widehat{A}_{j,h}\right| > \tfrac{6\sqrt{8}\log(h+1)\cdot\log(t)}{\epsilon}\right\} \quad .
\end{aligned} \tag{22}$$

For a fixed h, according to Algorithm 1, we know $r_j = \lfloor \log(h+1) \rfloor - 1 \leq \lfloor \log(t) \rfloor - 1$.

The first $\sum_{s=0}^{r_j} 2^s$ observations will use the modified Logarithmic Mechanism to inject noise. That is also to say, we add $r_j + 1$ random variables and each is independently drawn from Lap $\left(\frac{1}{0.5\epsilon}\right)$. Let $\mathcal{Y}_{j,1} := \sum_{n=0}^{r_j} Y_n$ be the aggregated noise injected by the modified Logarithmic Mechanism, where each $Y_n \sim$ Lap $\left(\frac{1}{0.5\epsilon}\right)$. Let $\widehat{C}_{j,0:r_j}$ be the true aggregated reward of the first $\sum_{s=0}^{r_j} 2^s$ observations. Let $\widetilde{C}_{j,0:r_j}$ be the differentially private aggregated reward of the first $\sum_{s=0}^{r_j} 2^s$ observations, i.e., $\widetilde{C}_{j,0:r_j} := \widehat{C}_{j,0:r_j} + \mathcal{Y}_{j,1}$.

The remaining $h - \sum_{s=0}^{r_j} 2^s \leq 2^{r_j+1}$ observations will use the 2^{r_j+1}-bounded Binary Mechanism to inject noise. Let $\mathcal{Y}_{j,2} := \sum_{n \in \mathcal{B}_{j,s}} Z_n$ be the aggregated noise injected by the 2^{r_j+1}-bounded Binary Mechanism, where $\mathcal{B}_{j,s}$ is the set of p-sums involved and each Z_n is i.i.d. from Lap $\left(\frac{r_j+1}{0.5\epsilon}\right)$. Note that the size of $\mathcal{B}_{j,s}$ is at most $r_j + 1$. Let $\widehat{B}_{j,r_j+1}$ be the true aggregated reward of these observations. Let $\widetilde{B}_{j,r_j+1}$ be the differentially private aggregated reward of these observations, i.e., $\widetilde{B}_{j,r_j+1} := \widehat{B}_{j,r_j+1} + \mathcal{Y}_{j,2}$.

Then, we have

$$\begin{aligned}
(22) \;\leq\; & \sum_{h=1}^{t-1} \mathbb{P}\left\{\left|\widetilde{A}_{j,h} - \widehat{A}_{j,h}\right| > \tfrac{6\sqrt{8}\log(h+1)\cdot\log(t)}{\epsilon}\right\} \\
\leq\; & \sum_{h=1}^{t-1} \mathbb{P}\left\{\left|\widetilde{C}_{j,0:r_j} - \widehat{C}_{j,0:r_j}\right| + \left|\widetilde{B}_{j,r_j+1} - \widehat{B}_{j,r_j+1}\right| > \tfrac{6\sqrt{8}\log(h+1)\cdot\log(t)}{\epsilon}\right\} \\
\leq\; & \sum_{h=1}^{t-1} \left(\mathbb{P}\left\{\left|\widetilde{C}_{j,0:r_j} - \widehat{C}_{j,0:r_j}\right| > \tfrac{6\sqrt{8}\log(h+1)\cdot\log(t)}{\epsilon}\right\} + \mathbb{P}\left\{\left|\widetilde{B}_{j,r_j+1} - \widehat{B}_{j,r_j+1}\right| > \tfrac{6\sqrt{8}\log(h+1)\cdot\log(t)}{\epsilon}\right\}\right) \quad .
\end{aligned} \tag{23}$$

We upper bound each term in (23) separately and have

$$\begin{aligned}
& \mathbb{P}\left\{\left|\widetilde{C}_{j,0:r_j} - \widehat{C}_{j,0:r_j}\right| > \tfrac{6\sqrt{8}\log(h+1)\cdot\log(t)}{\epsilon}\right\} \\
= \;& \mathbb{P}\left\{|\mathcal{Y}_{j,1}| > \tfrac{6\sqrt{8}\log(h+1)\cdot\log(t)}{\epsilon}\right\} \\
\leq \;& \mathbb{P}\left\{|\mathcal{Y}_{j,1}| > \tfrac{6\sqrt{8}\cdot\log(t)}{\epsilon}\right\} \\
= \;& \underbrace{\mathbb{P}\left\{|\mathcal{Y}_{j,1}| > \log(e)\frac{1}{0.5\epsilon}\sqrt{\ln(t^3)}\sqrt{8\ln(t^3)}\right\}}_{\text{Fact 7}} \\
\leq \;& \tfrac{2}{t^3} \quad ,
\end{aligned} \tag{24}$$

and

$$
\begin{array}{rl}
& \mathbb{P}\left\{\left|\widetilde{B}_{j,r_j+1}-\widehat{B}_{j,r_j+1}\right|>\frac{6\sqrt{8}\log(h+1)\cdot\log(t)}{\epsilon}\right\} \\
= & \mathbb{P}\left\{|\mathcal{Y}_{j,2}|>\frac{6\sqrt{8}\cdot\log(h+1)\cdot\log(t)}{\epsilon}\right\} \\
\leq & \underbrace{\mathbb{P}\left\{|\mathcal{Y}_{j,2}|>\log(e)\frac{\lfloor\log(h+1)\rfloor}{0.5\epsilon}\sqrt{\ln(t^3)}\sqrt{8\ln(t^3)}\right\}}_{\text{Fact 7}} \\
\leq & \frac{2}{t^3} \quad .
\end{array}
\tag{25}
$$

Regarding how to apply Fact 7, we have $r_j+1=\lfloor\log(h+1)\rfloor$ and $\max\left\{\sqrt{r_j+1},\sqrt{\ln(t^3)}\right\}=\max\left\{\sqrt{\lfloor\log(h+1)\rfloor},\sqrt{\ln(t^3)}\right\}=\sqrt{\ln(t^3)}$. Note that $\sqrt{\lfloor\log(h+1)\rfloor}\leq\sqrt{\lfloor\log(t)\rfloor}\leq\sqrt{\ln(t^3)}$ when $t\geq 2$.

By plugging (24) and (25) into (23), we conclude the proof. □

Before presenting the proof for Lemma 10, we first present Lemma 12, a lemma that will be used for the proof of Lemma 10. Lemma 12 itself is very similar to Lemma 2.8 in [Agrawal and Goyal, 2017].

Recall that $\mathcal{F}_{t-1}$ collects all the history information until the end of round $t-1$, which contains the pulled arms, the revealed rewards, and the injected noise. Define $\mathcal{F}_0=\{\}$.

Lemma 12. *For any sub-optimal j and any instantiation F_{t-1} of $\mathcal{F}_{t-1}$, we have*

$$
\mathbb{P}\left\{J_t=j,E_j^{\theta}(t),G_1(t-1)\mid\mathcal{F}_{t-1}=F_{t-1}\right\}\leq\frac{\mathbb{P}\{\theta_1(t)\leq y_j\mid\mathcal{F}_{t-1}=F_{t-1}\}}{\mathbb{P}\{\theta_1(t)>y_j\mid\mathcal{F}_{t-1}=F_{t-1}\}}\mathbb{P}\left\{J_t=1,E_j^{\theta}(t),G_1(t-1)\mid\mathcal{F}_{t-1}=F_{t-1}\right\}. \tag{26}
$$

We defer the proof of Lemma 12 to the end of this section.

We now start the proof for Lemma 10.

Proof of Lemma 10. We have

$$
\begin{array}{rl}
\text{LHS in (16)} = & \mathbb{E}\left[\sum_{t=1}^{T}\mathbf{1}\left\{J_t=j,E_j^{\theta}(t)\right\}\right] \\
\leq & \mathbb{E}\left[\sum_{t=1}^{T}\mathbf{1}\left\{J_t=j,E_j^{\theta}(t),G_1(t-1)\right\}\right]+\underbrace{\mathbb{E}\left[\sum_{t=1}^{T}\mathbf{1}\left\{\overline{G_1(t-1)}\right\}\right]}_{\leq O(1),\ \text{Lemma 9}} \\
\leq & \mathbb{E}\left[\sum_{t=1}^{T}\mathbf{1}\left\{J_t=j,E_j^{\theta}(t),G_1(t-1)\right\}\right]+O(1) \\
= & \mathbb{E}\left[\sum_{t=1}^{T}\underbrace{\mathbb{P}\left\{J_t=j,E_j^{\theta}(t),G_1(t-1)\mid\mathcal{F}_{t-1}\right\}}_{\text{Lemma 12}}\right]+O(1) \\
\leq & \mathbb{E}\left[\sum_{t=1}^{T}\frac{\mathbb{P}\{\theta_1(t)\leq y_j\mid\mathcal{F}_{t-1}\}}{\mathbb{P}\{\theta_1(t)>y_j\mid\mathcal{F}_{t-1}\}}\cdot\mathbb{P}\left\{J_t=1,E_j^{\theta}(t),G_1(t-1)\mid\mathcal{F}_{t-1}\right\}\right]+O(1) \\
\leq & \mathbb{E}\left[\sum_{t=1}^{T}\frac{\mathbb{P}\{\theta_1(t)\leq y_j\mid\mathcal{F}_{t-1}\}}{\mathbb{P}\{\theta_1(t)>y_j\mid\mathcal{F}_{t-1}\}}\cdot\mathbb{P}\left\{J_t=1,G_1(t-1)\mid\mathcal{F}_{t-1}\right\}\right]+O(1) \\
= & \underbrace{\mathbb{E}\left[\sum_{t=1}^{T}\frac{\mathbb{P}\left\{\theta_1(t)\leq y_j\mid\mathcal{F}_{t-1}\right\}}{\mathbb{P}\left\{\theta_1(t)>y_j\mid\mathcal{F}_{t-1}\right\}}\cdot\mathbf{1}\left\{J_t=1,G_1(t-1)\right\}\right]}_{=:\lambda}+O(1) \quad .
\end{array}
\tag{27}
$$

Let τ_k be the round such that in round τ_k, the k-th pull for arm 1 happens. That is also to say, in all rounds $t=\tau_k+1,\tau_k+2,\ldots,\tau_{k+1}$, there is no new observation for arm 1 and $O_1(t-1)=k$. Hence, arm 1's differentially private empirical mean cannot be updated. It is important to note that although arm 1's differentially private empirical mean cannot be updated during all the rounds in $\{\tau_k+1, \tau_k+2, \ldots, \tau_{k+1}\}$, it does not mean the posterior distributions $\theta_1(t)$ stay the same for all $t \in \{\tau_k+1, \tau_k+2, \ldots, \tau_{k+1}\}$.

We separate all T rounds into multiple intervals based on whether arm 1 is pulled or not. According to Algorithm 1, we know $\lambda_1 = 1$.

Now, we start to reduce the proof to the non-private setting.

We have

$$\begin{aligned} \lambda &= \mathbb{E}\left[\sum_{t=1}^{T} \frac{\mathbb{P}\{\theta_1(t)\le y_j|\mathcal{F}_{t-1}\}}{\mathbb{P}\{\theta_1(t)> y_j|\mathcal{F}_{t-1}\}} \cdot \mathbf{1}\{J_t=1, G_1(t-1)\}\right] \\ &\le \sum_{k=1}^{T} \mathbb{E}\left[\sum_{t=\tau_k+1}^{\tau_{k+1}} \frac{\mathbb{P}\{\theta_1(t)\le y_j|\mathcal{F}_{t-1}\}}{\mathbb{P}\{\theta_1(t)> y_j|\mathcal{F}_{t-1}\}} \cdot \mathbf{1}\{J_t=1, G_1(t-1)\}\right] \\ &\le \sum_{k=1}^{T} \mathbb{E}\left[\underbrace{\frac{\mathbb{P}\left\{\theta_1(\tau_{k+1})\le y_j \mid \mathcal{F}_{\tau_{k+1}-1}\right\}}{\mathbb{P}\left\{\theta_1(\tau_{k+1})> y_j \mid \mathcal{F}_{\tau_{k+1}-1}\right\}} \cdot \mathbf{1}\{G_1(\tau_{k+1}-1)\}}_{=:\rho}\right] \\ &\le \sum_{k=1}^{T} \mathbb{E}\left[\frac{\mathbb{P}\left\{\theta_1'(\tau_{k+1})\le y_j|\mathcal{F}_{\tau_{k+1}-1}\right\}}{\mathbb{P}\left\{\theta_1'(\tau_{k+1})> y_j|\mathcal{F}_{\tau_{k+1}-1}\right\}}\right] \quad , \end{aligned} \tag{28}$$

where $\theta_1'(\tau_{k+1}) \sim \text{Beta}\left(\widehat{\mu}_{j,k}\cdot k+1, (1-\widehat{\mu}_{j,k})\cdot k+1\right)$.

The last inequality in (28) uses the following arguments. We categorize all the instantiations $F_{\tau_{k+1}-1}$ of $\mathcal{F}_{\tau_{k+1}-1}$ into two types.

For the instantiation $F_{\tau_{k+1}-1}$ of $\mathcal{F}_{\tau_{k+1}-1}$ such that $\mathbf{1}\{G_1(\tau_{k+1}-1)\}=0$, we have $\rho=0$.

For the instantiation $F_{\tau_{k+1}-1}$ of $\mathcal{F}_{\tau_{k+1}-1}$ such that $\mathbf{1}\{G_1(\tau_{k+1}-1)\}=1$, we have

$$\begin{aligned} \overline{\mu}_{1,O_1(\tau_{k+1}-1)} &= \max\left\{0, \min\left\{\widetilde{\mu}_{1,O_1(\tau_{k+1}-1)} + \frac{6\sqrt{8}\log(O_1(\tau_{k+1}-1)+1)\log(\tau_{k+1})}{\epsilon\cdot O_1(\tau_{k+1}-1)}, 1\right\}\right\} \\ &= \max\left\{0, \min\left\{\widetilde{\mu}_{1,k} + \frac{6\sqrt{8}\log(k+1)\log(\tau_{k+1})}{\epsilon\cdot k}, 1\right\}\right\} \\ &\ge \max\left\{0, \min\left\{\left(\widehat{\mu}_{1,k} - \frac{6\sqrt{8}\log(k+1)\log(\tau_{k+1})}{\epsilon\cdot k}\right) + \frac{6\sqrt{8}\log(k+1)\log(\tau_{k+1})}{\epsilon\cdot k}, 1\right\}\right\} \\ &= \max\left\{0, \min\left\{\widehat{\mu}_{1,k}, 1\right\}\right\} \\ &= \widehat{\mu}_{1,k} \quad . \end{aligned} \tag{29}$$

From the fact that $\text{Beta}\left(\overline{\mu}_{1,O_1(\tau_{k+1}-1)}\cdot k+1, (1-\overline{\mu}_{1,O_1(\tau_{k+1}-1)})\cdot k+1\right)$ first-order stochastically dominates $\text{Beta}\left(\widehat{\mu}_{1,k}\cdot k+1, (1-\widehat{\mu}_{1,k})\cdot k+1\right)$, we have

$$\mathbb{P}\left\{\theta_1(\tau_{k+1})\le y_j \mid \mathcal{F}_{\tau_{k+1}-1}=F_{\tau_{k+1}-1}\right\} \le \mathbb{P}\left\{\theta_1'(\tau_{k+1})\le y_j \mid \mathcal{F}_{\tau_{k+1}-1}=F_{\tau_{k+1}-1}\right\} \quad , \tag{30}$$

which means

$$\frac{\mathbb{P}\left\{\theta_1(\tau_{k+1})\le y_j|\mathcal{F}_{\tau_{k+1}-1}=F_{\tau_{k+1}-1}\right\}}{\mathbb{P}\left\{\theta_1(\tau_{k+1})> y_j|\mathcal{F}_{\tau_{k+1}-1}=F_{\tau_{k+1}-1}\right\}} \le \frac{\mathbb{P}\left\{\theta_1'(\tau_{k+1})\le y_j|\mathcal{F}_{\tau_{k+1}-1}=F_{\tau_{k+1}-1}\right\}}{\mathbb{P}\left\{\theta_1'(\tau_{k+1})> y_j|\mathcal{F}_{\tau_{k+1}-1}=F_{\tau_{k+1}-1}\right\}} \quad . \tag{31}$$

Now, the proof is reduced to the non-private setting.

From Lemma 2.10 in [Agrawal and Goyal, 2017], we have

$$\begin{aligned} \lambda &\le \sum_{k=1}^{T} \mathbb{E}\left[\frac{\mathbb{P}\left\{\theta_1'(\tau_{k+1})\le y_j|\mathcal{F}_{\tau_{k+1}-1}\right\}}{\mathbb{P}\left\{\theta_1'(\tau_{k+1})> y_j|\mathcal{F}_{\tau_{k+1}-1}\right\}}\right] \\ &\le \frac{24}{(\mu_1-y_j)^2} + \sum_{k\ge \frac{8}{\mu_1-y_j}} \Theta\left(e^{-(\mu_1-y_j)^2\cdot k\cdot\frac{1}{2}} + \frac{e^{-2(\mu_1-y_j)^2\cdot k}}{(k+1)\cdot(\mu_1-y_j)^2} + \frac{1}{e^{\frac{(\mu_1-y_j)^2\cdot k}{4}}-1}\right) \\ &\le O\left(\frac{\log(T)}{(\mu_1-y_j)^2}\right) \\ &\le O\left(\frac{\log(T)}{\Delta_j^2}\right) \quad , \end{aligned} \tag{32}$$

which concludes the proof. $\square$

Proof of Lemma 11. Recall that $\mathcal{F}_{t-1}$ collect all the history information by the end of round $t-1$. We have

$$\begin{aligned}
& \mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}\left\{J_t = j, O_j(t-1) > \mathcal{L}_j, C_j(t-1), G_j(t-1), \overline{E_j^{\theta}(t)}\right\}\right] \\
= & \sum_{t=1}^{T} \mathbb{E}\left[\mathbf{1}\left\{O_j(t-1) > \mathcal{L}_j, C_j(t-1), G_j(t-1)\right\} \cdot \mathbf{1}\left\{J_t = j, \overline{E_j^{\theta}(t)}\right\}\right] \\
= & \sum_{t=1}^{T} \mathbb{E}\left[\mathbb{E}\left[\mathbf{1}\left\{O_j(t-1) > \mathcal{L}_j, C_j(t-1), G_j(t-1)\right\} \cdot \mathbf{1}\left\{J_t = j, \overline{E_j^{\theta}(t)}\right\} \mid \mathcal{F}_{t-1}\right]\right] \\
= & \sum_{t=1}^{T} \mathbb{E}\left[\underbrace{\underbrace{\mathbf{1}\left\{O_j(t-1) > \mathcal{L}_j, C_j(t-1), G_j(t-1)\right\}}_{=:\chi} \cdot \mathbb{E}\left[\mathbf{1}\left\{J_t = j, \overline{E_j^{\theta}(t)}\right\} \mid \mathcal{F}_{t-1}\right]}_{=:\Gamma}\right] .
\end{aligned} \tag{33}$$

Recall that $\mathcal{L}_j = \max\left\{\underbrace{108\frac{\log(T)}{\Delta_j^2}}_{=:\mathcal{H}_j}, \underbrace{72\sqrt{8}\frac{\log(T)}{\epsilon\cdot\Delta_j}\log\left(72\sqrt{8}\frac{\log(T)}{\epsilon\cdot\Delta_j}\right)}_{=:\mathcal{Q}_j}\right\}$ and let $\mathcal{H}_j := 108\frac{\log(T)}{\Delta_j^2}$ and $\mathcal{Q}_j := 72\sqrt{8}\frac{\log(T)}{\epsilon\cdot\Delta_j}\log\left(72\sqrt{8}\frac{\log(T)}{\epsilon\cdot\Delta_j}\right)$. Then, we have $\mathcal{L}_j \geq \mathcal{H}_j$ and $\mathcal{L}_j \geq \mathcal{Q}_j$.

We categorize all the instantiations F_{t-1} of $\mathcal{F}_{t-1}$ into two types based on whether the indicator function χ returns 1 or 0. Let $F^{\text{Beta}}_{\alpha,\beta}(\cdot)$ be the cdf of a Beta distribution with parameters α and β and let $F^{B}_{n,p}(\cdot)$ be the cdf of a Binomial distribution with parameters n and p.

For the F_{t-1} such that the indicator function χ returns 0, we have $\Gamma = 0$.

For the F_{t-1} such that the indicator function χ returns 1, we have

$$\begin{aligned}
\Gamma \quad = \quad & \mathbb{E}\left[\mathbf{1}\left\{J_t = j, \overline{E_j^{\theta}(t)}\right\} \mid \mathcal{F}_{t-1} = F_{t-1}\right] \\
= \quad & \mathbb{P}\left\{J_t = j, \overline{E_j^{\theta}(t)} \mid \mathcal{F}_{t-1} = F_{t-1}\right\} \\
\leq \quad & \mathbb{P}\left\{\theta_j(t) > y_j \mid \mathcal{F}_{t-1} = F_{t-1}\right\} \\
= \quad & 1 - F^{\text{Beta}}_{\overline{\mu}_{j,O_j(t-1)}\cdot O_j(t-1)+1, \left(1-\overline{\mu}_{j,O_j(t-1)}\right)\cdot O_j(t-1)+1}(y_j) \\
\leq^{(a)} \quad & 1 - F^{\text{Beta}}_{\left(\mu_j+\frac{4\Delta_j}{6}\right)\cdot O_j(t-1)+1, \left(1-\left(\mu_j+\frac{4\Delta_j}{6}\right)\right)\cdot O_j(t-1)+1}(y_j) \\
\leq \quad & 1 - F^{\text{Beta}}_{\left\lceil\left(\mu_j+\frac{4\Delta_j}{6}\right)\cdot O_j(t-1)\right\rceil+1, O_j(t-1)-\left\lceil\left(\mu_j+\frac{4\Delta_j}{6}\right)\cdot O_j(t-1)\right\rceil+1}(y_j) \\
=^{(b)} \quad & F^{B}_{O_j(t-1)+1, y_j}\left(\left\lceil\left(\mu_j + \frac{4\Delta_j}{6}\right)\cdot O_j(t-1)\right\rceil\right) \\
\leq \quad & F^{B}_{O_j(t-1)+1, y_j}\left(\left(\mu_j + \frac{4\Delta_j}{6} + \frac{1}{O_j(t-1)}\right)\cdot O_j(t-1)\right) \\
\leq^{(c)} \quad & F^{B}_{O_j(t-1)+1, y_j}\left(\left(\mu_j + \frac{4\Delta_j}{6} + \frac{\Delta_j}{12}\right)\cdot (O_j(t-1)+1)\right) \\
\leq^{(d)} \quad & e^{-(O_j(t-1)+1)\cdot d_{\text{KL}}\left(\mu_j+\frac{4\Delta_j}{6}+\frac{\Delta_j}{12}, y_j\right)} \\
\leq^{(f)} \quad & e^{-\mathcal{H}_j\cdot 2\frac{\Delta_j^2}{144}} \\
\leq \quad & \frac{1}{T} \quad .
\end{aligned} \tag{34}$$

By plugging the upper bound of Γ into (33), we conclude the proof.

We now present the detailed reasons about why each step holds in (34).

Let $\widetilde{\alpha}_j := \overline{\mu}_{j,O_j(t-1)}\cdot O_j(t-1)+1$ and $\widetilde{\beta}_j := \left(1-\overline{\mu}_{j,O_j(t-1)}\right)\cdot O_j(t-1)+1$. Let $\alpha'_j := \left(\mu_j + \frac{4\Delta_j}{6}\right)\cdot O_j(t-1)+1$ and $\beta'_j := \left(1-\mu_j-\frac{4\Delta_j}{6}\right)\cdot O_j(t-1)+1$. The inequality (a) in (34) uses the fact that $F^{\text{Beta}}_{\alpha'_j,\beta'_j}(y_j) \leq F^{\text{Beta}}_{\widetilde{\alpha}_j,\widetilde{\beta}_j}(y_j)$ if $\alpha'_j \geq \widetilde{\alpha}_j, \beta'_j \leq \widetilde{\beta}_j$. The equality (b) uses the relationship between the cdfs of Beta distribution and Binomial distribution, which is shown in Fact 5. The inequality (c) uses the fact that $\frac{1}{O_j(t-1)} \leq \frac{1}{\mathcal{H}_j} < \frac{\Delta_j}{12}$. The inequality (d) uses Chernoff-

Hoeffding bound that is shown in Fact 4. The inequality (f) uses the fact that $\mathcal{H}_j \leq O_j(t-1)+1$ and the Pinsker's inequality, i.e., $d_{\text{KL}}(x,y) \geq 2(x-y)^2$, where x, y are the parameters of Bernoulli distributions.

To show that $\alpha'_j \geq \widetilde{\alpha}_j$ and $\beta'_j \leq \widetilde{\beta}_j$, it suffices to show $\overline{\mu}_{j,O_j(t-1)} \leq \mu_j + \frac{4\Delta_j}{6}$.

Recall that instantiation F_{t-1} is the one such that the indicator function χ returns 1. Then, we know the following are true simultaneously:

1. $O_j(t-1) > \mathcal{L}_j \geq \mathcal{H}_j$;
2. $O_j(t-1) > \mathcal{L}_j \geq \mathcal{Q}_j$;
3. $\left|\widehat{\mu}_{j,O_j(t-1)} - \widetilde{\mu}_{j,O_j(t-1)}\right| \leq \frac{6\sqrt{8}\log(O_j(t-1)+1)\log(t)}{\epsilon\cdot O_j(t-1)}$;
4. $\left|\mu_j - \widehat{\mu}_{j,O_j(t-1)}\right| \leq \sqrt{\frac{3\log(t)}{O_j(t)}}$.

Let $f(x) = \sqrt{\frac{3\cdot\log(t)}{x}} + \frac{2\cdot 6\sqrt{8}\cdot\log(x+1)}{\epsilon}\cdot\frac{\log(t)}{x}$, where $x \geq e$. We now show $f(x) \leq \frac{4\Delta_j}{6}$ when $x > \mathcal{L}_j$.

When $x \geq \mathcal{L}_j$, it means $x \geq \mathcal{Q}_j$ and $x \geq \mathcal{H}_j$. Then, we have

$$\begin{aligned} f(x) &= \sqrt{\frac{3\cdot\log(t)}{x}} + \frac{2\cdot 6\sqrt{8}\cdot\log(x+1)}{\epsilon}\cdot\frac{\log(t)}{x} \\ &< \sqrt{\frac{3\cdot\log(t)}{x}} + \frac{3\cdot 6\sqrt{8}\cdot\log(x)}{\epsilon}\cdot\frac{\log(t)}{x} \\ &< \sqrt{\frac{3\cdot\log(t)}{\mathcal{H}_j}} + \frac{3\cdot 6\sqrt{8}\cdot\log(\mathcal{Q}_j)}{\epsilon}\cdot\frac{\log(t)}{\mathcal{Q}_j} \\ &\leq \sqrt{\frac{3\cdot\log(T)}{\frac{108\log(T)}{\Delta_j^2}}} + \frac{3\cdot 6\sqrt{8}\cdot\log(T)}{\epsilon}\frac{\log\left(72\sqrt{8}\frac{\log(T)}{\epsilon\cdot\Delta_j}\log\left(72\sqrt{8}\frac{\log(T)}{\epsilon\cdot\Delta_j}\right)\right)}{72\sqrt{8}\frac{\log(T)}{\epsilon\cdot\Delta_j}\log\left(72\sqrt{8}\frac{\log(T)}{\epsilon\cdot\Delta_j}\right)} \\ &\leq \frac{4\Delta_j}{6} \quad . \end{aligned} \tag{35}$$

Now, we have

$$\begin{aligned} \overline{\mu}_{j,O_j(t-1)} &= \max\left\{0, \min\left\{\widetilde{\mu}_{j,O_j(t-1)} + \frac{6\sqrt{8}\log(O_j(t-1)+1)\log(t)}{\epsilon\cdot O_j(t-1)}, 1\right\}\right\} \\ &\leq \max\left\{0, \min\left\{\widehat{\mu}_{j,O_j(t-1)} + \frac{2\cdot 6\sqrt{8}\log(O_j(t-1)+1)\log(t)}{\epsilon\cdot O_j(t-1)}, 1\right\}\right\} \\ &\leq \max\left\{0, \min\left\{\mu_j + \underbrace{\sqrt{\frac{3\log(t)}{O_j(t-1)}} + \frac{2\cdot 6\sqrt{8}\log\left(O_j(t-1)+1\right)\log(t)}{\epsilon\cdot O_j(t-1)}}_{f(O_j(t-1))\leq\frac{4\Delta_j}{6}}, 1\right\}\right\} \\ &\leq \mu_j + \frac{4\Delta_j}{6} \quad . \end{aligned} \tag{36}$$

□

Proof of Lemma 12. The LHS in (26) is

$$\begin{aligned} & \mathbb{P}\left\{J_t = j, E_j^\theta(t), G_1(t-1) \mid \mathcal{F}_{t-1} = F_{t-1}\right\} \\ \leq\ & \mathbb{P}\left\{G_1(t-1), \theta_j(t) \leq y_j, \forall j \in \mathcal{A} \mid \mathcal{F}_{t-1} = F_{t-1}\right\} \\ =\ & \mathbb{P}\left\{\theta_1(t) \leq y_j, G_1(t-1) \mid \mathcal{F}_{t-1} = F_{t-1}\right\}\cdot\mathbb{P}\left\{\theta_j(t) \leq y_j, \forall j \in \mathcal{A}\backslash\{1\} \mid \mathcal{F}_{t-1} = F_{t-1}\right\} \\ =\ & \mathbb{E}\left[\mathbf{1}\left\{\theta_1(t) \leq y_j, G_1(t-1)\right\} \mid \mathcal{F}_{t-1} = F_{t-1}\right]\cdot\mathbb{P}\left\{\theta_j(t) \leq y_j, \forall j \in \mathcal{A}\backslash\{1\} \mid \mathcal{F}_{t-1} = F_{t-1}\right\} \\ =\ & \mathbf{1}\left\{G_1(t-1)\right\}\mathbb{E}\left[\mathbf{1}\left\{\theta_1(t) \leq y_j\right\} \mid \mathcal{F}_{t-1} = F_{t-1}\right]\cdot\mathbb{P}\left\{\theta_j(t) \leq y_j, \forall j \in \mathcal{A}\backslash\{1\} \mid \mathcal{F}_{t-1} = F_{t-1}\right\} \\ =\ & \mathbf{1}\left\{G_1(t-1)\right\}\mathbb{P}\left\{\theta_1(t) \leq y_j \mid \mathcal{F}_{t-1} = F_{t-1}\right\}\cdot\mathbb{P}\left\{\theta_j(t) \leq y_j, \forall j \in \mathcal{A}\backslash\{1\} \mid \mathcal{F}_{t-1} = F_{t-1}\right\} \quad . \end{aligned} \tag{37}$$

We also have

$$\begin{aligned} & \mathbb{P}\left\{J_t = 1, E_j^\theta(t), G_1(t-1) \mid \mathcal{F}_{t-1} = F_{t-1}\right\} \\ \geq\ & \mathbb{P}\left\{G_1(t-1), \theta_1(t) > y_j \geq \theta_j(t), \forall j \in \mathcal{A}\backslash\{1\} \mid \mathcal{F}_{t-1} = F_{t-1}\right\} \\ =\ & \mathbb{P}\left\{\theta_1(t) > y_j, G_1(t-1) \mid \mathcal{F}_{t-1} = F_{t-1}\right\}\cdot\mathbb{P}\left\{y_j \geq \theta_j(t), \forall j \in \mathcal{A}\backslash\{1\} \mid \mathcal{F}_{t-1} = F_{t-1}\right\} \\ =\ & \mathbf{1}\left\{G_1(t-1)\right\}\mathbb{P}\left\{\theta_1(t) > y_j \mid \mathcal{F}_{t-1} = F_{t-1}\right\}\cdot\mathbb{P}\left\{\theta_j(t) \leq y_j, \forall j \in \mathcal{A}\backslash\{1\} \mid \mathcal{F}_{t-1} = F_{t-1}\right\} \quad . \end{aligned} \tag{38}$$

For the instantiation F_{t-1} such that $\mathbf{1}\left\{G_1(t-1)\right\}=0$, it is trivial to prove since both sides in (26) are 0. For the instantiation F_{t-1} such that $\mathbf{1}\left\{G_1(t-1)\right\}=1$, by combining (37) and (38), we conclude the proof. Note that for any $y_j \in (0,1)$, we have $\mathbb{P}\left\{\theta_1(t) > y_j \mid \mathcal{F}_{t-1}=F_{t-1}\right\} > 0$.

□

C PROOFS FOR PROBLEM-INDEPENDENT REGRET BOUND FOR ALGORITHM 1

Recall that the problem-dependent regret bound for Algorithm 1 is

$$\sum_{j\in\mathcal{A}:\Delta_j>0} O\left(\max\left\{\frac{\log(T)}{\Delta_j}, \frac{\log(T)}{\epsilon}\log\left(\frac{\log(T)}{\epsilon\cdot\Delta_j}\right)\right\}\right) \quad . \tag{39}$$

Now, let $\Delta := \sqrt{\frac{K\log(T)}{T}}$ be the critical gap. Then, we can express the regret as

$$\begin{aligned}
\mathcal{R}_{\text{DP-TS}}(T) &\leq T\cdot\Delta + \sum_{j\in\mathcal{A}:\Delta_j>\Delta} O\left(\max\left\{\frac{\log(T)}{\Delta_j}, \frac{\log(T)}{\epsilon}\log\left(\frac{\log(T)}{\epsilon\cdot\Delta_j}\right)\right\}\right)\\
&\leq \sqrt{KT\log(T)} + K\cdot O\left(\max\left\{\frac{\log(T)}{\Delta}, \frac{\log(T)}{\epsilon}\log\left(\frac{\log(T)}{\epsilon\cdot\Delta}\right)\right\}\right)\\
&= \sqrt{KT\log(T)} + O\left(\max\left\{\frac{K\log(T)}{\sqrt{\frac{K\log(T)}{T}}}, \frac{K\log(T)}{\epsilon}\log\left(\frac{\log(T)}{\epsilon\cdot\sqrt{\frac{K\log(T)}{T}}}\right)\right\}\right)\\
&\leq O(\sqrt{KT\log(T)}) + O\left(\frac{K\log(T)}{\epsilon}\log\left(\frac{\sqrt{T\log(T)}}{\epsilon\sqrt{K}}\right)\right) \quad .
\end{aligned} \tag{40}$$

D PROOFS FOR THEOREM 4

Recall that $O_j(t-1)$ is the number of "effective" observations for arm j at the end of round $t-1$, i.e., the number of observations that is used to compute the differentially private empirical mean for arm j. From Algorithm 2, we know that it can only take values of $2^r, r\geq 0$.

Recall $y_j = \mu_1 - \frac{\Delta_j}{6}$, and $E_j^\theta(t)$ is the event such that $\theta_j(t)\leq y_j$ and $\overline{E_j^\theta(t)}$ is the complementary event of $E_j^\theta(t)$.

Let $C_j(t-1)$ be the event such that $\left\{\left|\mu_j - \widehat{\mu}_{j,O_j(t-1)}\right| \leq \sqrt{\frac{3\log(t)}{O_j(t-1)}}\right\}$ and $\overline{C_j(t-1)}$ be the complementary event of $C_j(t-1)$. Let $G_j(t-1)$ be the event such that $\left\{\left|\widehat{\mu}_{j,O_j(t-1)} - \widetilde{\mu}_{j,O_j(t-1)}\right| \leq \frac{3\log(t)}{\epsilon\cdot O_j(t-1)}\right\}$ and $\overline{G_j(t-1)}$ be the complementary event of $G_j(t-1)$.

The following two lemmas show that $\overline{G_j(t-1)}$ and $\overline{C_j(t-1)}$ are low probability events for all $j\in\mathcal{A}$.

Lemma 13. *For any arm $j\in\mathcal{A}$, we have*

$$\sum_{t=1}^{T}\mathbb{E}\left[\mathbf{1}\left\{\overline{C_j(t-1)}\right\}\right] \leq O(1) \quad . \tag{41}$$

Lemma 14. *For any arm $j\in\mathcal{A}$, we have*

$$\sum_{t=1}^{T}\mathbb{E}\left[\mathbf{1}\left\{\overline{G_j(t-1)}\right\}\right] \leq O(1) \quad . \tag{42}$$

To prove Theorem 4, we prepare the following two lemmas which are similar to Lemma 10 and Lemma 11.

Lemma 15. *For any sub-optimal arm j, we have*

$$\sum_{t=1}^{T} \mathbb{E}\left[\mathbf{1}\left\{J_t = j, E_j^{\theta}(t)\right\}\right] \le O\left(\frac{\log(T)}{\Delta_j^2}\right) \quad . \tag{43}$$

Lemma 16. *For any sub-optimal arm j, we have*

$$\sum_{t=1}^{T} \mathbb{E}\left[\mathbf{1}\left\{J_t = j, C_j(t-1), G_j(t-1), \overline{E_j^{\theta}(t)}\right\}\right] \le O\left(\frac{\log(T)}{\Delta_j \cdot \min\left\{\epsilon, \Delta_j\right\}}\right) \quad . \tag{44}$$

After these preparations, we now prove Theorem 4.

Proof of Theorem 4. For each sub-optimal arm j, we upper bound its expected number of pulls.

We have

$$\begin{aligned}
& \sum_{t=1}^{T} \mathbb{E}\left[\mathbf{1}\left\{J_t = j\right\}\right] \\
\le \; & \underbrace{\sum_{t=1}^{T} \mathbb{E}\left[\mathbf{1}\left\{J_t = j, E_j^{\theta}(t)\right\}\right]}_{\le O\left(\frac{\log(T)}{\Delta_j^2}\right), \text{ Lemma 15}} + \underbrace{\sum_{t=1}^{T} \mathbb{E}\left[\mathbf{1}\left\{J_t = j, C_j(t-1), G_j(t-1), \overline{E_j^{\theta}(t)}\right\}\right]}_{\le O\left(\frac{\log(T)}{\Delta_j \cdot \min\left\{\epsilon, \Delta_j\right\}}\right), \text{ Lemma 16}} \\
+ \; & \underbrace{\sum_{t=1}^{T} \mathbb{E}\left[\mathbf{1}\left\{\overline{C_j(t-1)}\right\}\right]}_{\le O(1), \text{ Lemma 13}} + \underbrace{\sum_{t=1}^{T} \mathbb{E}\left[\mathbf{1}\left\{\overline{G_j(t-1)}\right\}\right]}_{\le O(1), \text{ Lemma 14}} \\
\le \; & O\left(\frac{\log(T)}{\Delta_j \cdot \min\left\{\epsilon, \Delta_j\right\}}\right) \quad ,
\end{aligned} \tag{45}$$

which concludes the proof. □

We now present the proofs for Lemma 13 to Lemma 16.

Proof of Lemma 13. We prove this lemma by using Hoeffding's inequality (shown in Fact 3). We have

$$\begin{aligned}
& \sum_{t=1}^{T} \mathbb{E}\left[\mathbf{1}\left\{\overline{C_j(t-1)}\right\}\right] \\
= \; & \sum_{t=1}^{T} \mathbb{P}\left\{\left|\widehat{\mu}_{j,O_j(t-1)} - \mu_j\right| > \sqrt{\frac{3\log(t)}{O_j(t-1)}}\right\} \\
\le \; & \sum_{t=1}^{T} \sum_{r=0}^{\log(t-1)} \mathbb{P}\left\{\left|\widehat{\mu}_{j,2^r} - \mu_j\right| > \sqrt{\frac{3\log(t)}{2^r}}\right\} \\
\le \; & \sum_{t=1}^{T} \sum_{r=0}^{\log(t-1)} \mathbb{P}\left\{\left|\widehat{\mu}_{j,2^r} - \mu_j\right| > \sqrt{\frac{4\ln(t)}{2^r}}\right\} \\
\le \; & \sum_{t=1}^{T} \sum_{r=0}^{\log(t-1)} 2e^{-2\cdot 2^r \frac{4\ln(t)}{2^r}} \\
\le \; & O(1) \quad ,
\end{aligned} \tag{46}$$

which concludes the proof. □

Proof of Lemma 14. We prove this lemma by using the measure concentration of Laplace random variables that is shown in

Fact 6. We have

$$
\begin{aligned}
& \sum_{t=1}^{T} \mathbb{E}\left[\mathbf{1}\left\{\overline{G_j(t-1)}\right\}\right] \\
= & \sum_{t=1}^{T} \mathbb{P}\left\{\left|\widehat{\mu}_{j,O_j(t-1)} - \widetilde{\mu}_{j,O_j(t-1)}\right| > \frac{3\log(t)}{\epsilon \cdot O_j(t-1)}\right\} \\
\leq & \sum_{t=1}^{T} \sum_{r=0}^{\log(t-1)} \mathbb{P}\left\{\left|\widehat{\mu}_{j,2^r} - \widetilde{\mu}_{j,2^r}\right| > \frac{3\log(t)}{\epsilon \cdot 2^r}\right\} \\
\leq & \sum_{t=1}^{T} \sum_{r=0}^{\log(t-1)} \mathbb{P}\left\{\left|2^r \cdot \widehat{\mu}_{j,2^r} - 2^r \cdot \widetilde{\mu}_{j,2^r}\right| > \frac{4\ln(t)}{\epsilon}\right\} \\
\leq & \sum_{t=1}^{T} \sum_{r=0}^{\log(t-1)} e^{-4\ln(t)} \\
\leq & \; O(1) \quad ,
\end{aligned}
\tag{47}
$$

which concludes the proof. □

Proof of Lemma 15. The key challenge is to reduce the proof to the non-private setting. Since Algorithm 2 drops observations along with the learning, we cannot reuse the proofs for Theorem 2. We need to use novel arguments to complete the proof.

We have

$$
\begin{aligned}
\text{LHS in (43)} \quad = & \quad \mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}\left\{J_t = j, E_j^{\theta}(t)\right\}\right] \\
& \text{(Add } G_1(t-1) \text{ and } \overline{G_1(t-1)} \text{ in the indicator function)} \\
\leq & \quad \underbrace{\mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}\left\{J_t = j, E_j^{\theta}(t), G_1(t-1)\right\}\right]}_{=:\zeta} + \underbrace{\mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}\left\{\overline{G_1(t-1)}\right\}\right]}_{\leq O(1),\ \text{Lemma 14}} \quad .
\end{aligned}
\tag{48}
$$

We now show that the first term ζ in (48) is upper bounded by $O\left(\frac{\log(T)}{\Delta_j^2}\right)$. From Lemma 14, we know that the second term is upper bounded by $O(1)$.

Recall $\mathcal{F}_{t-1}$ collects all the history information by the end of round $t-1$ containing the pulled arms, the rewards associated with the pulled arms, and the noise random variables and we set $\mathcal{F}_0 = \{\}$.

To upper bound ζ in (48), let λ_r be the round such that at the end of round λ_r, we use 2^r fresh observations to update arm 1's differentially private empirical mean. Note that based on the definition of λ_r, we know that in all rounds $t \in \{\lambda_r + 1, \lambda_r + 2, \ldots, \lambda_{r+1}\}$, the number of observations for arm 1 is exactly 2^r and the differentially private empirical mean for arm 1 stays the same. Note that $\theta_1(t) \sim \text{Beta}\left(\overline{\mu}_{1,O_1(t-1)} \cdot O_1(t-1) + 1, \left(1 - \overline{\mu}_{1,O_1(t-1)}\right) \cdot O_1(t-1) + 1\right)$. According to Algorithm 2, we know $\lambda_0 = 1$.

For the ζ term in (48), we have

$$\begin{aligned}
\zeta &= \mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}\left\{J_t = j, E_j^{\theta}(t), G_1(t-1)\right\}\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T} \underbrace{\mathbb{P}\left\{J_t = j, E_j^{\theta}(t), G_1(t-1) \mid \mathcal{F}_{t-1}\right\}}_{\text{Lemma 12}}\right] \\
&\leq \mathbb{E}\left[\sum_{t=1}^{T} \frac{\mathbb{P}\{\theta_1(t)\leq y_j|\mathcal{F}_{t-1}\}}{\mathbb{P}\{\theta_1(t)> y_j|\mathcal{F}_{t-1}\}} \cdot \mathbb{P}\left\{J_t = 1, E_j^{\theta}(t), G_1(t-1) \mid \mathcal{F}_{t-1}\right\}\right] \\
&\leq \mathbb{E}\left[\sum_{t=1}^{T} \frac{\mathbb{P}\{\theta_1(t)\leq y_j|\mathcal{F}_{t-1}\}}{\mathbb{P}\{\theta_1(t)> y_j|\mathcal{F}_{t-1}\}} \cdot \mathbf{1}\left\{J_t = 1, G_1(t-1)\right\}\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T} \frac{\mathbb{P}\{\theta_1(t)\leq y_j|\mathcal{F}_{t-1}\}}{\mathbb{P}\{\theta_1(t)> y_j|\mathcal{F}_{t-1}\}} \cdot \mathbf{1}\left\{J_t = 1\right\}\cdot \mathbf{1}\left\{G_1(t-1)\right\}\right] \\
&\quad \text{(Separate T rounds into multiple intervals based on whether the differentially private empirical mean for 1 changes)} \\
&\leq \underbrace{\mathbb{E}\left[\sum_{r=0}^{\log(T)} \sum_{t=\lambda_r+1}^{\lambda_{r+1}} \underbrace{\frac{\mathbb{P}\left\{\theta_1(t)\leq y_j \mid \mathcal{F}_{t-1}\right\}}{\mathbb{P}\left\{\theta_1(t)> y_j \mid \mathcal{F}_{t-1}\right\}} \cdot \mathbf{1}\left\{J_t = 1\right\}\cdot \mathbf{1}\left\{G_1(t-1)\right\}}_{=:\rho}\right]}_{=:\psi} .
\end{aligned} \tag{49}$$

We now start to reduce the proof to the non-private setting. Regarding term ρ in (49), we can categorize all the instantiations F_{t-1} of $\mathcal{F}_{t-1}$ into two types based on whether $\mathbf{1}\left\{G_1(t-1)\right\}$ returns 1 or not.

For the instantiation F_{t-1} such that $\mathbf{1}\left\{G_1(t-1)\right\} = 0$, we have $\rho = 0$.

For the instantiation F_{t-1} such that $\mathbf{1}\left\{G_1(t-1)\right\} = 1$, we have

$$\frac{\mathbb{P}\left\{\theta_1(t)\leq y_j \mid \mathcal{F}_{t-1}=F_{t-1}\right\}}{\mathbb{P}\left\{\theta_1(t)> y_j \mid \mathcal{F}_{t-1}=F_{t-1}\right\}} \leq \frac{\mathbb{P}\left\{\theta_1'(t)\leq y_j \mid \mathcal{F}_{t-1}=F_{t-1}\right\}}{\mathbb{P}\left\{\theta_1'(t)> y_j \mid \mathcal{F}_{t-1}=F_{t-1}\right\}} , \tag{50}$$

where $\theta_1'(t) \sim$ Beta $\left(\widehat{\mu}_{1,O_1(t-1)}\cdot O_1(t-1)+1, (1-\widehat{\mu}_{1,O_1(t-1)})\cdot O_1(t-1)+1\right)$.

The inequality in (50) uses the following arguments. If event $G_1(t-1)$ is true, we have $\widetilde{\mu}_{1,O_1(t-1)} \geq \widehat{\mu}_{1,O_1(t-1)} - \frac{3\log(t)}{\epsilon\cdot O_1(t-1)}$, which implies $\overline{\mu}_{1,O_1(t-1)} \geq \widehat{\mu}_{1,O_1(t-1)}$. Note that Beta $\left(\overline{\mu}_{1,O_1(t-1)}\cdot O_1(t-1)+1, (1-\overline{\mu}_{1,O_1(t-1)})\cdot O_1(t-1)+1\right)$ first-order stochastically dominates Beta $\left(\widehat{\mu}_{1,O_1(t-1)}\cdot O_1(t-1)+1, (1-\widehat{\mu}_{1,O_1(t-1)})\cdot O_1(t-1)+1\right)$.

Then, we have

$$\begin{aligned}
\psi &= \mathbb{E}\left[\sum_{r=0}^{\log(T)} \sum_{t=\lambda_r+1}^{\lambda_{r+1}} \frac{\mathbb{P}\{\theta_1(t)\leq y_j|\mathcal{F}_{t-1}\}}{\mathbb{P}\{\theta_1(t)> y_j|\mathcal{F}_{t-1}\}} \cdot \mathbf{1}\left\{J_t = 1\right\}\cdot \mathbf{1}\left\{G_1(t-1)\right\}\right] \\
&\leq \mathbb{E}\left[\sum_{r=0}^{\log(T)} \sum_{t=\lambda_r+1}^{\lambda_{r+1}} \frac{\mathbb{P}\left\{\theta_1'(t)\leq y_j|\mathcal{F}_{t-1}\right\}}{\mathbb{P}\left\{\theta_1'(t)> y_j|\mathcal{F}_{t-1}\right\}} \cdot \mathbf{1}\left\{J_t = 1\right\}\right] \\
&= \mathbb{E}\left[\sum_{r=0}^{\log(T)} \sum_{t=\lambda_r+1}^{\lambda_{r+1}} \frac{\mathbb{P}\left\{\theta_1'(\lambda_r+1)\leq y_j|\mathcal{F}_{\lambda_r}\right\}}{\mathbb{P}\left\{\theta_1'(\lambda_r+1)> y_j|\mathcal{F}_{\lambda_r}\right\}} \cdot \mathbf{1}\left\{J_t = 1\right\}\right] \\
&= \mathbb{E}\left[\sum_{r=0}^{\log(T)} \frac{\mathbb{P}\left\{\theta_1'(\lambda_r+1)\leq y_j|\mathcal{F}_{\lambda_r}\right\}}{\mathbb{P}\left\{\theta_1'(\lambda_r+1)> y_j|\mathcal{F}_{\lambda_r}\right\}} \cdot \underbrace{\sum_{t=\lambda_r+1}^{\lambda_{r+1}} \mathbf{1}\left\{J_t = 1\right\}}_{\leq 2^{r+1}}\right] \\
&\leq \sum_{r=0}^{\log(T)} 2^{r+1}\cdot \mathbb{E}\left[\frac{\mathbb{P}\left\{\theta_1'(\lambda_r+1)\leq y_j|\mathcal{F}_{\lambda_r}\right\}}{\mathbb{P}\left\{\theta_1'(\lambda_r+1)> y_j|\mathcal{F}_{\lambda_r}\right\}}\right] ,
\end{aligned} \tag{51}$$

which now the proof is reduced to the non-private setting.

The second equality in (51) uses the fact that the distribution of $\theta_1'(t)$ stays the same for all rounds $t = \lambda_r+1, \ldots, \lambda_{r+1}$. Recall that $\theta_1'(t) \sim$ Beta $\left(\widehat{\mu}_{1,O_1(t-1)}\cdot O_1(t-1)+1, (1-\widehat{\mu}_{1,O_1(t-1)})\cdot O_1(t-1)+1\right)$ and we know $O_1(t-1) = 2^r$ and

$\widehat{\mu}_{1,O_1(t-1)} = \widehat{\mu}_{1,2^r}$ for all $t \in \{\lambda_r+1,\ldots,\lambda_{r+1}\}$. The last inequality in (51) uses the fact that the number of pulls of arm 1 is at most 2^{r+1} during all the rounds in $\{\lambda_r+1,\ldots,\lambda_{r+1}\}$ based on the definition of λ_{r+1}.

Let $d_1 := \log\left(\frac{8}{\mu_1-y_j}\right)$. We now analyze the following two cases separately based on whether $r \le d_1$ or $r > d_1$. The intuitive understanding of this separation is when $r = 0,1,\ldots,\lfloor d_1 \rfloor$, we have $2^r \le 2^{d_1} \le \frac{8}{\mu_1-y_j}$. This means the number of observations for arm 1 is not "enough" for making arm 1's posterior distribution concentrated on its mean.

For all $0 \le r \le d_1$, from Lemma 2.9 in [Agrawal and Goyal, 2017], we have

$$\mathbb{E}\left[\frac{\mathbb{P}\left\{\theta_1'(\lambda_r+1) \le y_j \mid \mathcal{F}_{\lambda_r}\right\}}{\mathbb{P}\left\{\theta_1'(\lambda_r+1) > y_j \mid \mathcal{F}_{\lambda_r}\right\}}\right] \le \frac{3}{\mu_1-y_j} \quad . \tag{52}$$

When $\lceil d_1 \rceil \le r \le \log(T)$, we have $2^r \ge 2^{d_1} \ge \frac{8}{\mu_1-y_j}$. For all $\lceil d_1 \rceil \le r \le \log(T)$, from Lemma 2.9 in [Agrawal and Goyal, 2017], we have

$$\mathbb{E}\left[\frac{\mathbb{P}\left\{\theta_1'(\lambda_r+1)\le y_j|\mathcal{F}_{\lambda_r}\right\}}{\mathbb{P}\left\{\theta_1'(\lambda_r+1)> y_j|\mathcal{F}_{\lambda_r}\right\}}\right] \le \Theta\left(e^{-(\mu_1-y_j)^2\cdot 2^r\cdot\frac{1}{2}} + \frac{e^{-2(\mu_1-y_j)^2\cdot 2^r}}{(2^r+1)\cdot(\mu_1-y_j)^2} + \frac{1}{e^{\frac{(\mu_1-y_j)^2\cdot 2^r}{4}}-1}\right) \quad . \tag{53}$$

By applying (52) and (53) into (51), we have

$$\begin{aligned}
\psi &\le \sum_{r=0}^{\log(T)} 2^{r+1}\cdot\mathbb{E}\left[\frac{\mathbb{P}\left\{\theta_1'(\lambda_r+1)\le y_j|\mathcal{F}_{\lambda_r}\right\}}{\mathbb{P}\left\{\theta_1'(\lambda_r+1)> y_j|\mathcal{F}_{\lambda_r}\right\}}\right]\\
&\le \sum_{r=0}^{d_1} 2^{r+1}\cdot\frac{3}{\mu_1-y_j} + \sum_{r=\lceil d_1\rceil}^{\log(T)} 2^{r+1}\cdot\Theta\left(e^{-(\mu_1-y_j)^2\cdot 2^r\cdot\frac{1}{2}} + \frac{e^{-2(\mu_1-y_j)^2\cdot 2^r}}{(2^r+1)\cdot(\mu_1-y_j)^2} + \frac{1}{e^{\frac{(\mu_1-y_j)^2\cdot 2^r}{4}}-1}\right)\\
&\le O\left(\frac{1}{(\mu_1-y_j)^2}\right) + \int_{\lceil d_1\rceil}^{\log(T)}\Theta\left(2^r\cdot e^{-(\mu_1-y_j)^2\cdot 2^r\cdot\frac{1}{2}}\right)dr + O\left(\frac{\log(T)}{(\mu_1-y_j)^2}\right)\\
&\le O\left(\frac{\log(T)}{(\mu_1-y_j)^2}\right)\\
&\le O\left(\frac{\log(T)}{\Delta_j^2}\right) \quad ,
\end{aligned} \tag{54}$$

which concludes the proof. □

Proof of Lemma 16. Let λ_r be the round such that at the end of round λ_r, we use 2^r fresh observations to update arm j's differentially private empirical mean. Note that in all rounds $t \in \{\lambda_r+1,\lambda_r+2,\ldots,\lambda_{r+1}\}$, the number of observations for arm j is exactly 2^r and the differentially private empirical mean for arm j stays the same. According to Algorithm 2, we know that $\lambda_0 = j$. Recall $\mathcal{L}_j = \frac{72\cdot\log(T)}{\Delta_j\cdot\min\{\epsilon,\Delta_j\}}$ and let $d_j := \log\left(\mathcal{L}_j\right)$.

We have

$$
\begin{aligned}
& \text{LHS in } (44) \\
= & \sum_{t=1}^{T} \mathbb{E}\left[\mathbf{1}\left\{J_t = j, C_j(t-1), G_j(t-1), \overline{E_j^{\theta}(t)}\right\}\right] \\
& \text{(Separate all } T \text{ rounds into intervals based on whether the differentially private empirical mean for arm } j \text{ changes)} \\
\leq & \sum_{r=0}^{\log(T)} \mathbb{E}\left[\sum_{t=\lambda_r+1}^{\lambda_{r+1}} \mathbf{1}\left\{J_t = j, C_j(t-1), G_j(t-1), \overline{E_j^{\theta}(t)}\right\}\right] \\
& \text{(Break } \sum_{r=0}^{\log(T)} \text{ into } \sum_{r=0}^{\lfloor d_j \rfloor} + \sum_{r=\lceil d_j \rceil}^{\log(T)} \text{)} \\
\leq & \sum_{r=0}^{\lfloor d_j \rfloor} \mathbb{E}\left[\sum_{t=\lambda_r+1}^{\lambda_{r+1}} \mathbf{1}\left\{J_t = j, C_j(t-1), G_j(t-1), \overline{E_j^{\theta}(t)}\right\}\right] \sum_{r=\lceil d_j \rceil}^{\log(T)} \mathbb{E}\left[\sum_{t=\lambda_r+1}^{\lambda_{r+1}} \mathbf{1}\left\{J_t = j, C_j(t-1), G_j(t-1), \overline{E_j^{\theta}(t)}\right\}\right] \\
\leq & \sum_{r=0}^{\lfloor d_j \rfloor} \mathbb{E}\left[\sum_{t=\lambda_r+1}^{\lambda_{r+1}} \mathbf{1}\left\{J_t = j\right\}\right] + \sum_{r=\lceil d_j \rceil}^{\log(T)} \mathbb{E}\left[\sum_{t=\lambda_r+1}^{\lambda_{r+1}} \mathbf{1}\left\{J_t = j, C_j(t-1), G_j(t-1), \overline{E_j^{\theta}(t)}\right\}\right] \\
\leq & \sum_{r=0}^{\lfloor d_j \rfloor} 2^{r+1} + \sum_{r=\lceil d_j \rceil}^{\log(T)} \mathbb{E}\left[\sum_{t=\lambda_r+1}^{\lambda_{r+1}} \mathbf{1}\left\{J_t = j, C_j(t-1), G_j(t-1), \overline{E_j^{\theta}(t)}\right\}\right] \\
\leq & O\left(\frac{\log(T)}{\Delta_j \cdot \min\{\epsilon, \Delta_j\}}\right) + \underbrace{\sum_{r=\lceil d_j \rceil}^{\log(T)} \mathbb{E}\left[\sum_{t=\lambda_r+1}^{\lambda_{r+1}} \mathbf{1}\left\{J_t = j, C_j(t-1), G_j(t-1), \overline{E_j^{\theta}(t)}\right\}\right]}_{=:\eta} \quad .
\end{aligned}
\tag{55}
$$

We now show $\eta \leq O(\log(T))$. We have

$$
\begin{aligned}
\eta & = \sum_{r=\lceil d_j \rceil}^{\log(T)} \mathbb{E}\left[\sum_{t=\lambda_r+1}^{\lambda_{r+1}} \mathbf{1}\left\{J_t = j, C_j(t-1), G_j(t-1), \overline{E_j^{\theta}(t)}\right\}\right] \\
& \leq \sum_{r=\lceil d_j \rceil}^{\log(T)} \mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}\left\{J_t = j, C_j(t-1), G_j(t-1), O_j(t-1) = 2^r, \overline{E_j^{\theta}(t)}\right\}\right] \quad .
\end{aligned}
\tag{56}
$$

The inequality in (56) uses the fact that for all $t \in \{\lambda_r + 1, \ldots, \lambda_{r+1}\}$, the number of effective observations is exactly 2^r.

Recall $\mathcal{F}_{t-1}$ collects all the history information by the end of round $t-1$. Then, we have

$$
\begin{aligned}
\eta & \leq \sum_{r=\lceil d_j \rceil}^{\log(T)} \mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}\left\{J_t = j, C_j(t-1), G_j(t-1), O_j(t-1) = 2^r, \overline{E_j^{\theta}(t)}\right\}\right] \\
& = \sum_{r=\lceil d_j \rceil}^{\log(T)} \sum_{t=1}^{T} \mathbb{E}\left[\mathbb{E}\left[\mathbf{1}\left\{J_t = j, C_j(t-1), G_j(t-1), O_j(t-1) = 2^r, \overline{E_j^{\theta}(t)}\right\} \mid \mathcal{F}_{t-1}\right]\right] \\
& \leq \sum_{r=\lceil d_j \rceil}^{\log(T)} \sum_{t=1}^{T} \mathbb{E}\left[\underbrace{\underbrace{\mathbf{1}\left\{C_j(t-1), G_j(t-1), O_j(t-1) = 2^r\right\}}_{=:\chi} \mathbb{E}\left[\mathbf{1}\left\{\overline{E_j^{\theta}(t)}\right\} \mid \mathcal{F}_{t-1}\right]}_{=:\gamma}\right] \quad .
\end{aligned}
\tag{57}
$$

The last inequality in (57) uses the fact that $\mathbf{1}\left\{C_j(t-1), G_j(t-1), O_j(t-1) = 2^r\right\}$ is determined by $\mathcal{F}_{t-1}$.

We categorize all the instantiations F_{t-1} of $\mathcal{F}_{t-1}$ into two types based on whether the indicator function χ returns 1 or 0. Let $F^{\text{Beta}}_{\alpha,\beta}(\cdot)$ be the cdf of a Beta distribution with parameters α and β and let $F^{B}_{n,p}(\cdot)$ be the cdf of a Binomial distribution with parameters n and p.

For the F_{t-1} such that the indicator function χ returns 0, we have $\gamma = 0$.

For the F_{t-1} such that the indicator function χ returns 1, we have,

$$
\begin{aligned}
\gamma \quad &= \quad \mathbb{E}\left[\mathbf{1}\left\{\overline{E_j^{\theta}(t)}\right\} \mid \mathcal{F}_{t-1} = F_{t-1}\right] \\
&= \quad \mathbb{P}\left\{\overline{E_j^{\theta}(t)} \mid \mathcal{F}_{t-1} = F_{t-1}\right\} \\
&= \quad \mathbb{P}\left\{\theta_j(t) > y_j \mid \mathcal{F}_{t-1} = F_{t-1}\right\} \\
&= \quad 1 - F^{\text{Beta}}_{\overline{\mu}_{j,O_j(t-1)}\cdot O_j(t-1)+1,(1-\overline{\mu}_{j,O_j(t-1)})\cdot O_j(t-1)+1}(y_j) \\
&\leq^{(a)} \quad 1 - F^{\text{Beta}}_{\left(\mu_j+\frac{4\Delta_j}{6}\right)\cdot 2^r+1,\left(1-\left(\mu_j+\frac{4\Delta_j}{6}\right)\right)\cdot 2^r+1}(y_j) \\
&\leq \quad 1 - F^{\text{Beta}}_{\left\lceil\left(\mu_j+\frac{4\Delta_j}{6}\right)\cdot 2^r\right\rceil+1,2^r-\left\lceil\left(\mu_j+\frac{4\Delta_j}{6}\right)\cdot 2^r\right\rceil+1}(y_j) \\
&=^{(b)} \quad F^{B}_{2^r+1,y_j}\left(\left\lceil\left(\mu_j + \frac{4\Delta_j}{6}\right)\cdot 2^r\right\rceil\right) \\
&\leq \quad F^{B}_{2^r+1,y_j}\left(\left(\mu_j + \frac{4\Delta_j}{6}\right)\cdot 2^r + 1\right) \\
&= \quad F^{B}_{2^r+1,y_j}\left(\left(\mu_j + \frac{4\Delta_j}{6} + \frac{1}{2^r}\right)\cdot 2^r\right) \\
&\leq^{(c)} \quad F^{B}_{2^r+1,y_j}\left(\left(\mu_j + \frac{4\Delta_j}{6} + \frac{\Delta_j}{12}\right)\cdot (2^r+1)\right) \\
&\leq^{(d)} \quad e^{-(2^r+1)\cdot d_{\text{KL}}\left(\mu_j+\frac{4\Delta_j}{6}+\frac{\Delta_j}{12},y_j\right)} \\
&\leq^{(f)} \quad e^{-2^{d_j}\cdot 2\frac{\Delta_j^2}{144}} \\
&\leq \quad \frac{1}{T} \quad .
\end{aligned}
\tag{58}
$$

Since now we have shown $\gamma \leq \frac{1}{T}$, we know $\eta \leq O(\log(T))$. By plugging the upper bound of η into (55), we conclude the proof.

We now provide detailed explanations about why each key steps holds in (58). The inequality (a) uses the fact that $F^{\text{Beta}}_{\overline{\mu}_{j,O_j(t-1)}\cdot O_j(t-1)+1,(1-\overline{\mu}_{j,O_j(t-1)})\cdot O_j(t-1)+1}(y_j) \geq F^{\text{Beta}}_{\left(\mu_j+\frac{4\Delta_j}{6}\right)\cdot 2^r+1,\left(1-\left(\mu_j+\frac{4\Delta_j}{6}\right)\right)\cdot 2^r+1}(y_j)$ if $\overline{\mu}_{j,O_j(t-1)} \leq \mu_j + \frac{4\Delta_j}{6}$. The equality (b) uses the relationship between the cdfs of Beta distribution and Binomial distribution, which is shown in Fact 5. The inequality (c) uses the facts $\frac{1}{2^r} \leq \frac{1}{\mathcal{L}_j} < \frac{\Delta_j}{12}$ and $2^r < 2^r + 1$. Note that the cdf is non-decreasing. The inequality (d) uses the Chernoff-Hoeffding bound, which is shown in Fact 4. The inequality (f) uses the fact that $2^{d_j} \leq 2^r + 1$ and Pinsker's inequality, i.e., $d_{\text{KL}}(x,y) \geq 2(x-y)^2$.

We know prove our claim that $\overline{\mu}_{j,O_j(t-1)} \leq \mu_j + \frac{4\Delta_j}{6}$, when $r \geq d_j$. When $r \geq d_j$, we have $2^r \geq 2^{d_j} \geq \mathcal{L}_j \geq \frac{72\log(T)}{\Delta_j^2}$. Then, we have

$$
\begin{aligned}
\overline{\mu}_{j,O_j(t-1)} \quad &= \quad \max\left\{0, \min\left\{\widetilde{\mu}_{j,2^r} + \frac{3\log(t)}{\epsilon\cdot 2^r}, 1\right\}\right\} \\
&\leq \quad \max\left\{0, \min\left\{\widehat{\mu}_{j,2^r} + \frac{6\log(t)}{\epsilon\cdot 2^r}, 1\right\}\right\} \\
&\leq \quad \max\left\{0, \min\left\{\mu_j + \frac{3\log(t)}{\epsilon\cdot 2^r} + \sqrt{\frac{3\log(t)}{2^r}}, 1\right\}\right\} \\
&\leq \quad \max\left\{0, \min\left\{\mu_j + \frac{6\log(t)}{\epsilon\cdot\frac{72\log(t)}{\epsilon\cdot\Delta_j}} + \sqrt{\frac{3\log(t)}{\frac{72\log(t)}{\Delta_j^2}}}, 1\right\}\right\} \\
&\leq \quad \max\left\{0, \min\left\{\mu_j + \frac{6\Delta_j}{72} + \sqrt{\frac{3}{72}}\Delta_j, 1\right\}\right\} \\
&\leq \quad \mu_j + \frac{4\Delta_j}{6} \quad .
\end{aligned}
\tag{59}
$$

□

E PROOFS FOR PROBLEM-INDEPENDENT REGRET BOUND FOR ALGORITHM 2

Recall that the problem-dependent regret bound for Algorithm 2 is

$$
\sum_{j\in\mathcal{A}:\Delta_j>0} O\left(\frac{\log(T)}{\min\{\epsilon,\Delta_j\}}\right) \quad . \tag{60}
$$

Let $\Delta := \sqrt{\frac{K\log(T)}{T}}$ be the critical gap. Then, we can express the regret as

$$\begin{aligned} \mathcal{R}_{\text{Lazy-DP-TS}}(T) &\leq T\cdot\Delta + \sum_{j\in\mathcal{A}:\Delta_j>\Delta} O\left(\frac{\log(T)}{\min\{\epsilon,\Delta_j\}}\right) \\ &\leq T\cdot\sqrt{\frac{K\log(T)}{T}} + O\left(\frac{K\log(T)}{\Delta}+\frac{K\log(T)}{\epsilon}\right) \\ &\leq O\left(\sqrt{KT\log(T)}+\frac{K\log(T)}{\epsilon}\right) \quad . \end{aligned} \tag{61}$$

F ADDITIONAL EXPERIMENTAL RESULTS

F.1 ADDITIONAL EXPERIMENT 1

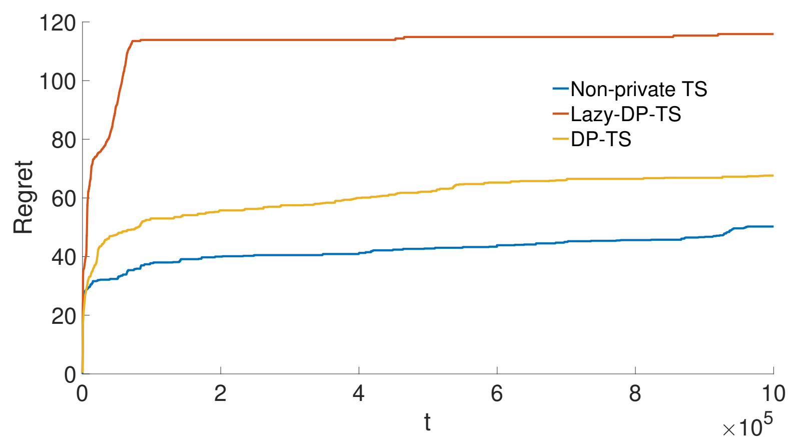

Figure 6: $\epsilon = 500$

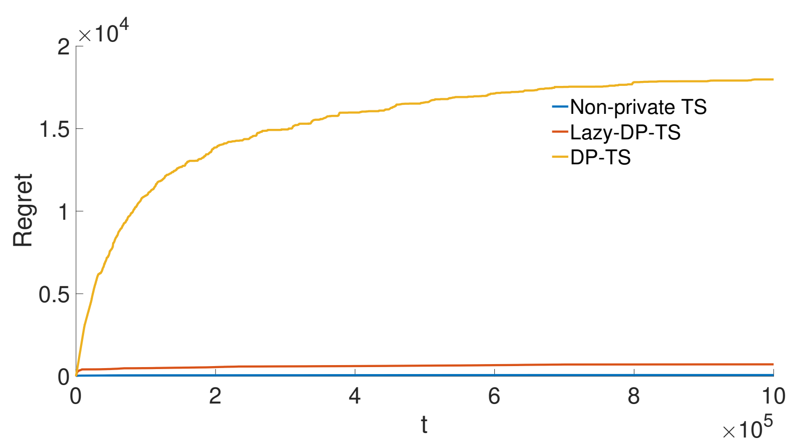

Figure 7: $\epsilon = 1.0$

Now, we provide more experimental results under the same setting that has been used in Section 4, i.e., $K = 5$ arms with mean rewards setting to $0.75, 0.625, 0.5, 0.375, 0.25$. We compare the practical performance of the non-private Thompson Sampling to differentially private Thompson Sampling with privacy parameter setting to $\epsilon = 0.1, 0.25, 0.5, 1, 500$. Figure 6 shows the experimental results for Non-private Thompson Sampling of Agrawal and Goyal [2017], DP-TS, and Lazy-DP-TS when setting $\epsilon = 500$. From the results we can see that even though we set a very large ϵ value, we still need to pay some cost for maintaining privacy. Figures 7 to 10 show the experimental results in the settings where $\epsilon = 1.0, 0.5, 0.25, 0.1$. From the experimental results we can see that the practical performance of Lazy-DP-TS is far better than DP-TS.

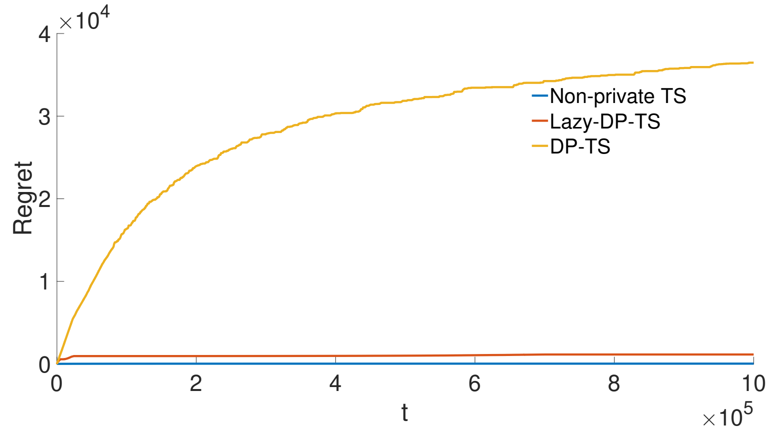

Figure 8: $\epsilon = 0.5$

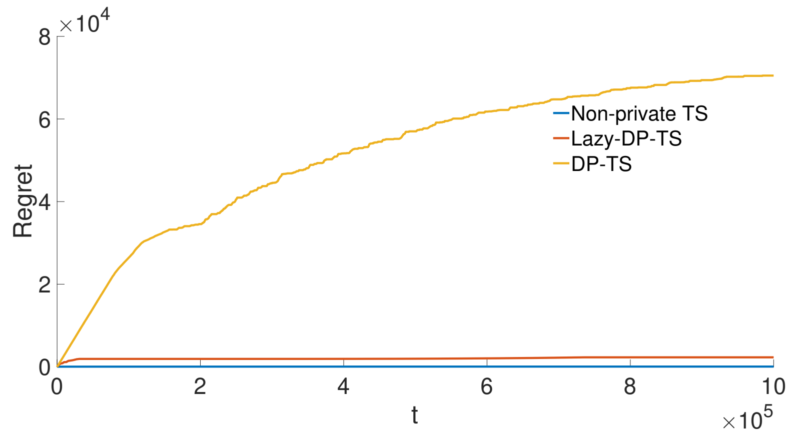

Figure 9: $\epsilon = 0.25$

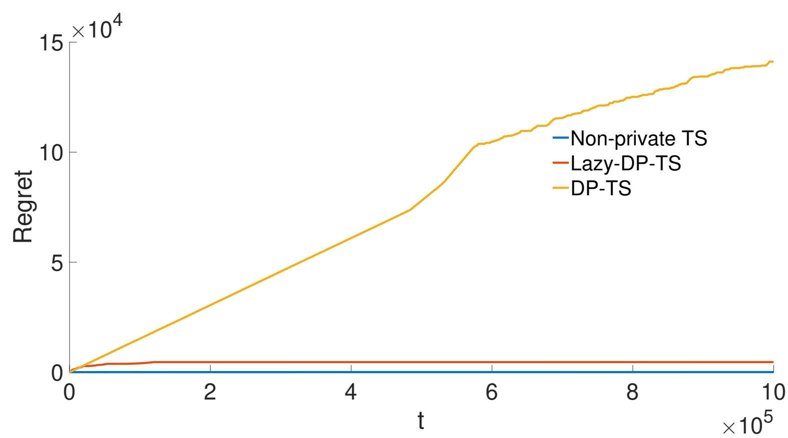

Figure 10: $\epsilon = 0.1$

F.2 ADDITIONAL EXPERIMENT 2

We set $T = 10^5$ and $K = 5$. The mean rewards are set to $0.8, 0.05, 0.05, 0.05, 0.05$, i.e., the mean reward gap is 0.75 for all sub-optimal arms. Now, only privacy parameter ϵ can impact the practical performance. We study the practical performance by setting $\epsilon = 0.5, 1.00, 500$. Figure 11 shows the results of the setting where $\epsilon = 500$. DP-TS still outperforms Lazy-DP-TS due to the fact that when ϵ is very large, DP-TS is asymptotically optimal while Lazy-DP-TS can only be order-optimal. We skip the plot of DP-SE as it is far inferior to the remaining three algorithms under this experimental setting. Figure 12 and Figure 13 show the results of the settings where $\epsilon = 1.0$ and $\epsilon = 0.5$, i.e., the values of ϵ are set to be close to the mean reward gap. From the experimental results we can see that the Lazy-DP-TS, Anytime-Lazy-UB, and DP-SE all perform better than the near-optimal DP-TS. However, Lazy-DP-TS still has better performance than the remaining two optimal algorithms.

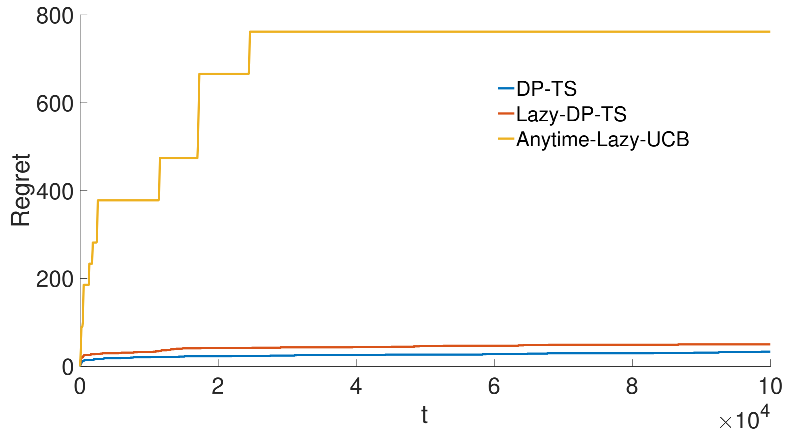

Figure 11: $\epsilon = 500$

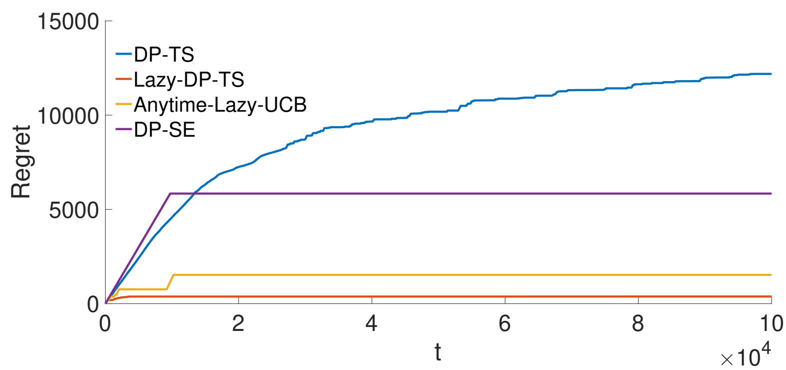

Figure 12: $\epsilon = 1.0$

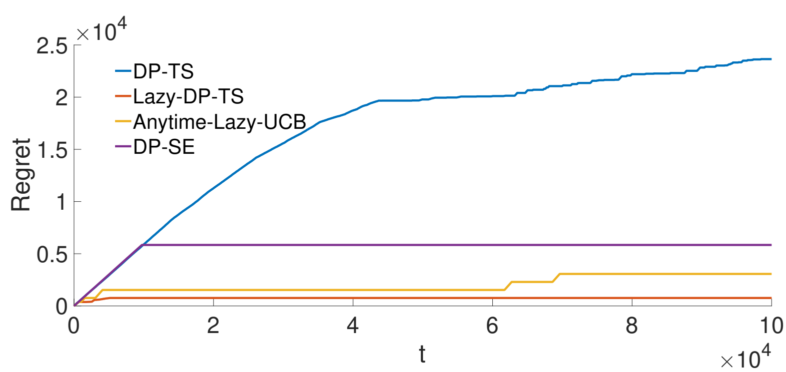

Figure 13: $\epsilon = 0.5$