Thompson Sampling Itself is Differentially Private

Tingting Ou
Columbia University

Marco Avella Medina
Columbia University

Rachel Cummings
Columbia University

Abstract

In this work we first show that the classical Thompson sampling algorithm for multi-arm bandits is differentially private as-is, without any modification. We provide per-round privacy guarantees as a function of problem parameters and show composition over T rounds; since the algorithm is unchanged, existing $O(\sqrt{NT \log N})$ regret bounds still hold and there is no loss in performance due to privacy. We then show that simple modifications – such as pre-pulling all arms a fixed number of times, increasing the sampling variance – can provide tighter privacy guarantees. We again provide privacy guarantees that now depend on the new parameters introduced in the modification, which allows the analyst to tune the privacy guarantee as desired. We also provide a novel regret analysis for this new algorithm, and show how the new parameters also impact expected regret. Finally, we empirically validate and illustrate our theoretical findings in two parameter regimes and demonstrate that tuning the new parameters substantially improve the privacy-regret tradeoff.

1 INTRODUCTION

The Thompson Sampling algorithm is one of the earliest developed heuristics for the stochastic multi-arm bandits problem and has been proven to have good performance both empirically and theoretically. It is a Bayesian regret-minimization algorithm that initializes a prior distribution on the parameters of the reward distributions, plays the arm using the posterior probability of being the best arm, and updates the posterior distribution accordingly using the observations. It appeared for the first time by Thompson (1933) in a two-armed bandit problem motivated by clinical trials. After being largely ignored in the literature, strong empirical performance and theoretical guarantees led to a rapid surge of interest in this algorithm in the last 15 years (Russo et al., 2018). Indeed, the Thompson Sampling algorithm has since been widely studied and proven to be useful for solving a wide range of online learning problems (Agrawal and Goyal, 2012, 2017, 2013; Wang and Chen, 2018; Russo et al., 2018; Liu and Ročková, 2023; Huang et al., 2021; Hüyük and Tekin, 2020).

In this work, we first analyze the privacy guarantees of Thompson Sampling to show that the classical Thompson sampling algorithm for multi-arm bandits is differentially private (DP) as-is, without any modification. The algorithm determines the next arm to play at each timestep by first sampling an estimate of each arm's mean reward from the posterior, and then selecting the arm with the highest noisy posterior sample. When the algorithm is initialized with Gaussian priors on reward distributions, this step is equivalent to adding mean-zero Gaussian noise to the empirical mean of observed rewards, also known as the *Gaussian Mechanism* in the DP literature.

We provide per-round privacy guarantees as a function of problem parameters and show composition over T rounds for N arms. In the main body, we express per-round privacy in terms of Gaussian differential privacy (GDP) (Dong et al., 2022), compose GDP parameters across all rounds, and then translate the guarantees back to the standard DP parameters. GDP is known to be amenable to many rounds of composition and the addition of Gaussian noise, both of which occur in the Thompson Sampling algorithm, but it leads to less easily interpretable statements of DP guarantees. In Appendix C, we provide an alternative proof of the DP guarantees of Thompson Sampling, that relies on a direct analysis of the DP guarantees. Along the way, we also show that a more general version of the classic ReportNoisyMax algorithm (Dwork and Roth, 2014) with heterogeneous Gaussian noise still satisfies DP, which may be of independent interest. Since the algorithm

Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238. Copyright 2024 by the author(s).

is unchanged, existing regret bounds from non-private analysis of Thompson Sampling (Agrawal and Goyal, 2017) still hold, and there is no loss in performance due to privacy.

Next, we show that simple modifications – such as prepulling all arms b times or increasing the sampling variance by a factor of c – can provide even tighter privacy guarantees, and allow the analyst to tune the privacy parameters as desired. The analysis follows a similar proof structure as before – proving per-round GDP guarantees, composing across T rounds, and translating back to DP – but this time accounting for the impact of the new parameters b and c. Since we modify the Thompson Sampling algorithm, existing regret bounds no longer hold and must be re-derived. We provide a novel regret analysis for our algorithm that follows the same high-level structure of the original analysis by Agrawal and Goyal (2017), while tightening some intermediate steps and tracking the impact of the new parameters in the expected regret.

Finally, we empirically validate our theoretical findings for two different families of reward distributions: Bernoulli and truncated exponential. For both families, our experimental findings match our theoretical results: tuning b and c can lead to substantial improvements in the empirical regret, for the same fixed privacy guarantee. We also observe that the optimal tuning strategy in terms of the privacy-accuracy tradeoff appears to involve jointly tuning b and c, rather than tuning just one of the parameters.

1.1 Related Work

There is a large body of work that considers the setting of private stochastic multi-arm bandits (Mishra and Thakurta, 2015; Hu and Hegde, 2022; Tossou and Dimitrakakis, 2016; Sajed and Sheffet, 2019; Hu et al., 2021). These works have considered different algorithms for the problem of privately learning from bandits, adapting popular non-private procedures such as Successive Elimination, Upper Confidence Bound and Thompson Sampling.

We focus our attention in this work on the Thompson Sampling algorithm (Thompson, 1933) which has been shown to to be applicable in solving a wide range of online learning problems including the classic multi-arm bandits problem (Agrawal and Goyal, 2012, 2017), contextual bandits (Agrawal and Goyal, 2013), combinatorial semi-bandits (Wang and Chen, 2018) and other applications including online job scheduling, subset-selection, variable selection, opportunistic routing, combinatorial network optimization (Gopalan et al., 2014; Liu and Ročková, 2023; Huang et al., 2021; Hüyük and Tekin, 2020). The empirical efficacy of Thompson Sampling when applied to the stochastic multi-arm bandits problem was demonstrated by Chapelle and Li (2011) before the currently best known theoretical regret bounds were proven by Agrawal and Goyal (2012). The seminal work on Thompson Sampling by Agrawal and Goyal (2012) gave a regret bound of $O((\sum_{i=2}^{N} \frac{1}{\Delta_i})^2 \log T)$ for the stochastic multi-arm bandits problem with N arms over T timesteps, when the algorithm is instantiated with a Beta prior over rewards, where Δ_i is the reward gap between the best arm and arm i. When the Δ_i are all bounded away from 0, this gives a problem-independent regret bound of $O(N^2 \log T)$. Follow-up work (Agrawal and Goyal, 2017) gives an $O(\sqrt{NT \log T})$ regret bound for the Thompson Sampling algorithm for both a Beta prior and a Gaussian prior. Our work focuses on the Gaussian prior version of the Thompson Sampling algorithm and its privacy guarantees.

Some recent works on private stochastic bandits have specifically addressed the problem of privatizing the Thompson Sampling algorithm as a solution to the bandits problem. Mishra and Thakurta (2015) first designed an ϵ-differentially private variant of the Thompson sampling algorithm, with expected regret $O(N\frac{\log^3 T}{\Delta^2 \epsilon^2})$, where Δ is the reward gap between the best arm and the second best arm. Hu and Hegde (2022) presented two ϵ-differentially private (near)-optimal Thompson Sampling-based algorithms, DP-TS and Lazy-DP-TS, with regret bounds $\sum_j O(\frac{\log T}{\min\{\epsilon, \Delta_j\}} \log(\frac{\log T}{\epsilon \Delta_j}))$ and $\sum_j O(\frac{\log T}{\min\{\epsilon, \Delta_j\}})$, respectively. Both works involve significant modifications of the Thompson Sampling algorithm to guarantee privacy, while our work mainly analyzes the privacy guarantee of the *original* version of the Thompson Sampling algorithm.

This work builds upon works analyzing the privacy guarantees of existing well-studied randomized algorithms, most notably the paper by Blocki et al. (2012), which showed that the Johnson-Lindenstrauss transform itself preserves differential privacy, and the paper by Smith et al. (2020), which proved that the Flajolet-Martin Sketch itself is differentially private. Similar in spirit – although completely different in technical details – we show that the noise added in Thompson Sampling is sufficient to satisfy differential privacy.

2 MODEL AND PRELIMINARIES

We consider the classic stochastic multi-armed bandit (MAB) setting with N arms. At each time $t \in [T]$, an arm $i \in \mathcal{A} = [N]$ is chosen to be played based on the outcomes from the previous $(t-1)$ timesteps, and yields a random real-valued reward $r_t \in [0, 1]$ sampled from a fixed unknown distribution $\mathcal{D}_i$ with mean μ_i. The rewards obtained from playing any arm are sampled i.i.d. and are independent of time or the plays of other arms.

The analyst must specify a *policy* $\pi = \{\pi_t\}_{t\in[T]}$, where each π_t maps from the history $\mathcal{F}_{t-1} = \{(a_\tau, r_\tau)\}_{\tau<t}$ containing the sequence of arms played and rewards observed up to time $t-1$, to the new arm a_t played at time t.

We measure the analyst's success using the standard metric of expected total *regret*, which is the difference between the best possible expected cumulative reward (if the distributions $\mathcal{D}_i$ were known) and the expected total reward under policy π (Agrawal and Goyal, 2013). Letting $\mu^* = \max_i \mu_i$, the expected total regret is:

$$\mathbb{E}[\mathcal{R}(T,\pi)] = \mu^* T - \mathbb{E}_{a_t \leftarrow \pi}[\textstyle\sum_{t=1}^{T} r_t]. \qquad (1)$$

2.1 Thompson Sampling

The Thompson Sampling algorithm is a commonly used policy for minimizing regret in a MAB setting (Thompson, 1933; Agrawal and Goyal, 2012; Russo et al., 2018). One reason for the widespread use of Thompson Sampling is that it is known to achieve low regret. At a high-level, the algorithm operates as follows. It starts with a prior belief on the mean reward μ_i for each arm. After observing each reward $r \sim \mathcal{D}_i$, the algorithm performs a Bayesian update to compute a posterior distribution of $\mu_i|r$. At every timestep, the algorithm samples a value θ_i for each arm according to the posterior given all observed rewards, and then plays the arm with the highest sampled θ value.

The Thompson Sampling algorithm is parameterized only by its initial priors on $\mathcal{D}_i$. In this work we focus on the special case of Gaussian priors, where the algorithm is initialized with a Gaussian prior for each arm. We emphasize that the algorithm's priors on rewards need not match the true reward distributions, and thus this does not conflict with our modeling assumption of bounded rewards. These priors are a part of the algorithmic construction, and not a part of the underlying data model.

We focus on the single parameter setting where each reward distribution $\mathcal{D}_i$ has unknown mean μ_i and known variance $\sigma_i^2 = 1$. For each arm i, the algorithm is initialized with a prior belief that $\mu_i \sim \mathcal{N}(0,1)$; after observing a reward, the posterior distribution of each μ_i will remain Gaussian. Concretely, given the initial prior $\mathcal{N}(0,1)$ and a sequence of rewards $(r_1,\ldots,r_t)$, the posterior for arm i from which θ_i will be sampled is $\mathcal{N}(\hat{\mu}_{i,t}, \frac{1}{n_{i,t}+1})$, where $n_{i,t}$ is the number of times arm i is played up to time t, and $\hat{\mu}_{i,t} = \frac{1}{n_{i,t}+1}\sum_{\tau=1}^{t} 1_{a(\tau)=i} r_\tau$ is the empirical mean of observed rewards from arm i, with a slight offset to avoid degeneracy based on the initial prior.

The θ_i for each arm is sampled according to this posterior at each timestep, and the arm corresponding to the largest realized θ is selected, output, and the reward from the selected arm is observed internally by the algorithm. Algorithm 1 presents a formal description of this algorithm.

Algorithm 1 Thompson Sampling with Gaussian priors

```
Input: number of arms $N$, time horizon $T$
Initialize $\hat{\mu}_{i,0} = 0,\ n_{i,0} = 0$ for each $i = 1,\ldots,N$
for $t = 1,2,\ldots,T$ do
    For each arm $i = 1,\ldots,N$, sample independently $\theta_{i,t} \sim \mathcal{N}(\hat{\mu}_{i,t-1}, \frac{1}{n_{i,t-1}+1})$
    Play arm $a_t := \arg\max_i \theta_{i,t}$ and observe $r_t$
    Output $a_t$
    For $i = a_t$, update $\hat{\mu}_{i,t} = \frac{\hat{\mu}_{i,t-1}(n_{i,t-1}+1)+r_t}{n_{i,t-1}+2}$, and $n_{i,t} = n_{i,t-1}+1$
    For all $i \neq a_t$, update $\hat{\mu}_{i,t} = \hat{\mu}_{i,t-1}$, and $n_{i,t} = n_{i,t-1}$
end for
```

Regret as defined in Equation (1) provides a *problem-independent* definition, because it has no additional assumptions or dependence on the problem instance characterized by the true reward means $(\mu_1,\ldots,\mu_N)$. We will also consider an equivalent *problem-dependent* definition of regret, which provides guarantees based on the true gap between the arm means. Define the suboptimality gap of arm i to be $\Delta_i := \mu^* - \mu_i$. We can use this, and the random variable $n_{i,t}$, which denotes the number of times that arm i has been played up to time t, to re-write the expected regret in a problem-dependent manner:

$$\mathbb{E}[\mathcal{R}(T,\pi)] = \mathbb{E}_{a_t\leftarrow\pi}[\sum_{t=1}^{T}(\mu^* - \mu_{a_t})] = \sum_{i=1}^{N}\Delta_i \mathbb{E}_{a_t\leftarrow\pi}[n_{i,T}]. \qquad (2)$$

2.2 Differential Privacy

Differential privacy (DP) is a parameterized notion of database privacy, which ensures that changing a single element of the input database will lead to only small changes in the distribution over outputs. In our online setting, we consider database *streams*, where each data element (from a data universe $\mathcal{R}$) arrives one-by-one, and the algorithm must produce an output from some output space $\mathcal{O}$ at each timestep. Two length T streams $R, R' \in \mathcal{R}^T$ are said to be *neighboring* if they differ only in the data element received in a single timestep. We first recall the definition of differential privacy for streams, adapted from its standard presentation of Dwork et al. (2006) to the streaming setting as in Cummings et al. (2020).

Definition 1. *A streaming algorithm $\mathcal{M} : \mathcal{R}^T \to \mathcal{O}^T$ is (ϵ, δ)-differentially private if for any pair of neighboring streams $R, R' \in \mathcal{R}^T$ and for any set of outputs $S \subseteq \mathcal{O}^T$,*

$$\Pr[\mathcal{M}(R) \in S] \leq e^{\epsilon} \Pr[\mathcal{M}(R') \in S] + \delta.$$

In the context of Thompson Sampling, our *neighboring streams* correspond to sequences of rewards $R = \{r_t\}_{t\in[T]}$ and $R' = \{r'_t\}_{t\in[T]}$ that differ in a single reward value: there exists a single τ such that $r_\tau \neq r'_\tau$, and for all other $t \neq \tau$, $r_t = r'_t$. The output $\mathcal{O}^T$ is the sequence of arm pulls $\{a_t\}_{t\in[T]}$ output by the algorithm.

Note that streaming algorithms can be *adaptive*, where the chosen arm a_t at time t is a function of all previous rewards and arm pulls. We will also analyze the privacy of single-shot mechanisms, where the mechanisms output space is $\mathcal{R}$ instead of $\mathcal{R}^T$. This corresponds to analysis at a single fixed timestep, rather than across all time.

We will use Gaussian differential privacy (GDP) (Dong et al., 2022), which is a variant of DP that is more amenable to many rounds of composition – where private algorithms are applied many times to the same dataset – and the addition of Gaussian noise, both of which occur in the Thompson Sampling algorithm. First we see that it is easy to translate between GDP and the standard (ϵ, δ)-DP notion.

Definition 2 (Dong et al. (2022))**.** *A mechanism $\mathcal{M}$ is η-GDP if and only if it is $(\epsilon, \delta(\epsilon))$-DP for all $\epsilon \geq 0$, where $\delta(\epsilon) = \Phi(-\frac{\epsilon}{\eta} + \eta/2) - e^{\epsilon}\Phi(-\frac{\epsilon}{\eta} - \eta/2)$, where $\Phi(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} \exp(-u^2/2)du$ is the cumulative density function of the standard normal distribution.*

One notable advantage of GDP is its composition properties, meaning that the privacy parameter μ composes slowly as more computations are performed on the data.

Lemma 1 (Dong et al. (2022))**.** *Let $\mathcal{M}_t$ be an η_t-GDP mechanism, for $t = 1, \ldots, T$. Then the adaptive T-fold composition of all $\mathcal{M}_1, \ldots, \mathcal{M}_T$ is $\sqrt{\sum_{t=1}^{T} \eta_t^2}$-GDP.*

Our analysis will be based on the *Gaussian Mechanism*, which is a method for privately evaluating a real-valued function f, and is defined as:

$$\mathcal{M}(D, \sigma^2) = f(D) + Y, \quad \text{where } Y \sim \mathcal{N}(0, \sigma^2).$$

The Gaussian Mechanism satisfies (s_f/σ)-GDP (Dong et al., 2022), where $s_f = \max_{R,R'\text{neighbors}} |f(R) - f(R')|$ is the *sensitivity* of f, or the maximum change in the function's value between neighboring databases.

3 THOMPSON SAMPLING IS DP

In this section, we show that the original Thompson Sampling algorithm (with Gaussian priors) as presented in Algorithm 1 is differentially private. Intuitively, differential privacy requires that algorithms make randomized decisions based on the database. In the case of Thompson Sampling, the algorithm is inherently randomized, by selecting the next arm based on the (randomly generated) θ_i rather than the exact empirical mean of historical play.

To show this formally, we must show that the particular distributions of randomness used in Thompson Sampling satisfy the mathematical requirements of differential privacy. To prove this, we will first focus on the privacy guarantees of a single step at a fixed time t, and then show how the privacy guarantees composes across T timesteps. Since GDP is known to yield improved composition guarantees relative to regular DP composition and other competing DP-like guarantees such as Renyi-DP (Dong et al., 2022), the analysis will involve computing single-step privacy guarantees using GDP (Lemma 2), then applying GDP composition (Lemma 3), and finally converting back to a DP guarantee (Theorem 3). While the formal statement with all parameters is given later in Theorem 3, the main result can be stated informally as in Theorem 1 below.

Theorem 1 (Informal version of Theorem 3)**.** *Algorithm 1 satisfies (ϵ, δ)-differential privacy.*

Before formally stating and proving this privacy result, we remark that since differential privacy does not require any changes to the algorithm itself, then all existing regret bounds continue to hold under differential privacy without incurring additional loss. These regret bounds are stated below the Theorem 2, and are tight under the mild assumption that rewards are bounded in $[0, 1]$.

Theorem 2 (Agrawal and Goyal (2017))**.** *The Thompson Sampling algorithm with Gaussian priors (Algorithm 1) has regret at most $\sum_{i=1}^{N} O(\frac{\log T}{\Delta_i})$ (problem-dependent), or $O(\sqrt{NT \log N})$ (problem-independent).*

Returning to the privacy analysis, at time t, Algorithm 1 samples $\theta_{i,t} \sim \mathcal{N}(\hat{\mu}_{i,t-1}, \frac{1}{n_{i,t-1}})$ independently for each arm, and then selects $a_t = \arg\max_i \theta_{i,t}$.

First, consider the vector of sampled mean-estimates $\{\theta_{i,t+1}\}_{i\in[N]}$ that are generated by the Thompson Sampling algorithm at time $t+1$ given history $\mathcal{F}_t$. The single-step algorithm $\mathcal{M}_{TS}(\mathcal{F}_t)$ corresponding to one step of Thompson sampling can be described as follows:

$$\mathcal{M}_{TS}(\mathcal{F}_t) = \arg\max_i \theta_{i,t+1},$$

where $\theta_{i,t+1} \sim \mathcal{N}(\frac{1}{n_{i,t}+1} \sum_{\tau=1}^{t} 1_{a(\tau)=i} r_\tau, \frac{1}{n_{i,t}+1})$.

Our first result is that this single step mechanism satisfies Gaussian differential privacy.

Lemma 2. *The mechanism $\mathcal{M}_{TS}(\mathcal{F}_t)$ satisfies $\sqrt{\frac{1}{2}}$-GDP with respect to observed rewards.*

The full proof of Lemma 2 is deferred to Appendix A.1, but we give a brief proof sketch here for intuition. Each $\theta_{i,t+1}$ can be expressed as $\hat{\mu}_i(\mathcal{F}_t)$ plus an independent Gaussian noise term sampled from $\mathcal{N}(0, \frac{1}{n_{i,t}+1})$. Viewing $\hat{\mu}_i(\mathcal{F}_t)$ as the real-valued query on the data, this is simply an instantiation of the Gaussian Mechanism. The query has sensitivity $s \leq \frac{1}{n_{i,t}+1}$ for arm i, since rewards are bounded between 0 and 1, so the empirical means on neighboring reward vectors can differ by at most $\frac{1}{n_{i,t}+1}$. Since the variance of the Gaussian noise added is $\frac{1}{n_{i,t}+1}$, and assuming that $n_{i,t} \geq 1$ – meaning that at least one reward has been observed from arm i, this yields a $\sqrt{\frac{1}{2}}$-GDP guarantee for this $\theta_{i,t+1}$. Since neighboring reward sequences can only differ in one single reward, they will also differ in only one single arm pull, so we need not consider composition across all N arms. Finally, the outcome of $\mathcal{M}_{TS}(\mathcal{F}_t)$ is simply the argmax of the θ_i, which is post-processing on the private output.

In Appendix C, we give an alternative proof for the DP guarantees of $\mathcal{M}_{TS}(\mathcal{F}_t)$, by proving that ReportNoisyMax (Dwork and Roth, 2014) with heterogeneous Gaussian noise, rather than the standard Laplace noise with identifical variance, satisfies DP. This result may be of independent interest, but leads to looser overall privacy bounds for Thompson Sampling due to the composition over a large number of rounds.

Next, we apply the composition guarantees of GDP given in Lemma 1 to show that the repeated application of $\mathcal{M}_{TS}(\mathcal{F}_t)$ for T rounds – as in Thompson Sampling – also satisfies GDP.

Lemma 3. *The Thompson Sampling algorithm with Gaussian priors run for T total timesteps (Algorithm 1) satisfies $\sqrt{\frac{1}{2}T}$-GDP.*

The proof of Lemma 3 follows immediately from the fact that one round of Thompson Sampling is $\sqrt{\frac{1}{2}}$-GDP by Lemma 2, and then applying GDP composition (Lemma 1) to get that T rounds of Thompson Sampling is together $\sqrt{\frac{1}{2}T}$-GDP.

Finally, we use Definition 2 to convert the GDP guarantee of Lemma 3 back to the desired (ϵ, δ)-DP.

Theorem 3. *Thompson Sampling with Gaussian priors (Algorithm 1) run for T timesteps is $(\epsilon, \delta(\epsilon))$-DP for all $\epsilon \geq 0$, where $\delta(\epsilon) = \Phi(-\frac{\epsilon}{\sqrt{\frac{1}{2}T}} + \frac{1}{2}\sqrt{\frac{1}{2}T}) - e^{\epsilon}\Phi(-\frac{\epsilon}{\sqrt{\frac{1}{2}T}} - \frac{1}{2}\sqrt{\frac{1}{2}T})$.*

Remark 1. *Although GDP provides the tightest composition bounds, this comes at the cost of interpretability of the resulting differential privacy parameters, as observed in the statement of Theorem 3. In Appendix C, we show that the informal Theorem 1 can also be proved using the composition methods of standard DP (Theorem 1) and a variant known as Renyi DP (Theorem 8). We also show empirically in Appendix C.3 that although these other privacy methods provide more interpretable bounds in terms of the dependence on T, the privacy guarantees are often orders of magnitude weaker than those attained using GDP.*

4 IMPROVING THE PRIVACY-REGRET TRADE-OFF

In this section, we show how the privacy-regret trade-off of Thompson Sampling can be improved with a simple modification of the algorithm. Concretely, the modified algorithm first pulls each arm b times before beginning the Thompson Sampling procedure. This serves to give the algorithm a "warm start" with more accurate prior beliefs on rewards, rather than $\mathcal{N}(0, 1)$. It also decreases the sensitivity of the implicit Gaussian Mechanism that computes $\theta_{i,t}$ by ensuring that each $n_{i,t}$ is at least b, thus leading to smaller ϵ values; on the other hand, the algorithm does not improve its decisions during these bN rounds, and may incur maximum loss during these initial rounds.

The second modification is scaling up the variance used to sample $\theta_{i,t}$ by a factor of $c \geq 1$. This serves to improve the ϵ privacy guarantees of the algorithm since more noise is added, but it also adds higher levels of noise to the algorithm's estimated empirical reward of each arm, thus increasing regret. The complete Modified Thompson Sampling algorithm with both of these changes is presented in Algorithm 2.

We show that by tuning b and c, the Modified Thompson Sampling algorithm can achieve both tighter privacy guarantees and lower regret, than the values achieved under the existing settings of $b = 0$ and $c = 1$ that correspond to standard Thompson Sampling (Algorithm 1). The remainder of this section provides analysis of the privacy guarantee and the regret bound of Algorithm 2.

Algorithm 2 Modified Thompson Sampling with Gaussian priors

```
Input: number of arms $N$, time horizon $T$, number of pre-pulls $b$, variance multiplier $c \geq 1$
Initialize $\hat{\mu}_{i,0} = 0$, $n_{i,0} = 0$ for each $i = 1, \dots, N$
for $i = 1, 2, \dots, N$ do
    for $j = 1, 2, \dots, b$ do
        Play arm $i$ and observe reward $r_{i,j}$
        Output $i$
        Update $\hat{\mu}_{i,} = \frac{\hat{\mu}_{i,0}(n_{i,0}+1)+r_{i,j}}{n_{i,0}+2}$, and $n_{i,0} = n_{i,0} + 1$
    end for
end for
for $t = 1, 2, \dots, T - bN$ do
    For each arm $i = 1, \dots, N$, sample independently $\theta_{i,t} \sim \mathcal{N}(\hat{\mu}_{i,t-1}, \frac{c}{n_{i,t-1}+1})$
    Play arm $a_t := \arg\max_i \theta_{i,t}$ and observe $r_t$
    Output $a_t$
    For $i = a_t$, update $\hat{\mu}_{i,t} = \frac{\hat{\mu}_{i,t-1}(n_{i,t-1}+1)+r_t}{n_{i,t-1}+2}$, and $n_{i,t} = n_{i,t-1} + 1$
    For all $i \neq a_t$, update $\hat{\mu}_{i,t} = \hat{\mu}_{i,t-1}$, and $n_{i,t} = n_{i,t-1}$
end for
```

4.1 Privacy Guarantees

The privacy analysis of Algorithm 2 follows a similar structure as that of Algorithm 1 in Section 3. We start with Lemma 4, which gives the GDP guarantee.

Lemma 4. *Modified Thompson Sampling with Gaussian priors and input parameters (b, c) run for T timesteps satisfies $\sqrt{\frac{1}{c(b+1)}T}$-GDP.*

The full proof of Lemma 4 is presented in Appendix A.2, and we give a proof sketch here. Similar to the proof of Lemma 2 in Section 3, the proof begins with a privacy analysis of the single step of the mechanism at a fixed time t. The changes for this modified algorithm are in the sensitivity of $\hat{\mu}_{i,t}$ and in the noise that is added. Recall that the GDP parameter of the Gaussian Mechanism is s_f/σ when sensitivity of the function is s_f and the added Gaussian noise has variance σ^2. In Algorithm 2, this expression is:

$$\left|\frac{\frac{1}{n_{i,t}+1}}{\sqrt{\frac{c}{n_{i,t}+1}}}\right| = \frac{1}{\sqrt{c(n_{i,t}+1)}} \leq \frac{1}{\sqrt{c(b+1)}}.$$

This single shot GDP guarantee is then composed across T rounds using Lemma 1, to give $\sqrt{\frac{1}{c(b+1)}}T$-GDP over T rounds.

Finally, the GDP guarantee of Lemma 4 is converted to a differential privacy guarantee using Definition 2 to yield the final privacy guarantees of Algorithm 2, presented in Theorem 4.

Theorem 4. *Modified Thompson Sampling with Gaussian priors and input parameters (b, c) run for T timesteps satisfies $(\epsilon, \delta(\epsilon))$-DP for all $\epsilon \geq 0$, where $\delta(\epsilon) = \Phi(-\frac{\epsilon}{\sqrt{\frac{1}{c(b+1)}T}} + \frac{1}{2}\sqrt{\frac{1}{c(b+1)}T}) - e^{\epsilon}\Phi(-\frac{\epsilon}{\sqrt{\frac{1}{c(b+1)}T}} - \frac{1}{2}\sqrt{\frac{1}{c(b+1)}T})$.*

To illustrate the impact of b and c on the privacy guarantees of Theorem 4, Figure 1 visualizes the tradeoff between ϵ and δ under varying b and c. We observe that relative to the values $b = 0$ and $c = 1$ corresponding to the special case of standard Thompson Sampling, increasing these parameters can lead to substantial improvements in the privacy guarantee. This means that even a small amount of prepulling or increase in the variance of sampling θ can lead to dramatically tighter privacy guarantees, relative to the standard Thompson Sampling algorithm. Further empirical analysis, including the impact of b and c on regret, is deferred to our experimental results in Section 5.

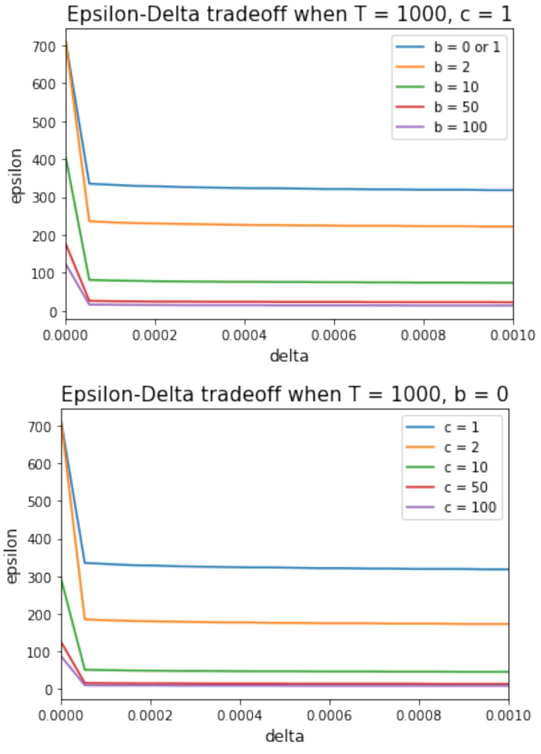

Figure 1: DP parameter ϵ as a function of δ when fixing $T = 1000$ and varying b (top) and c (bottom). Note that the per-round GDP parameter is $\sqrt{\frac{1}{c(b+1)}}$, so the role of c and b are nearly symmetric, leading to nearly identical plots on the top and bottom. Of course, these parameters can also be varied together.

4.2 Regret Guarantees

Since Algorithm 2 differs from the standard Thompson Sampling algorithm, new regret analysis is needed. Theorem 5 gives both problem-dependent and problem-independent regret bounds for Algorithm 2, based on both parameters b and c. Relative to the guarantees of Theorem 2 from the work by Agrawal and Goyal (2017) for standard Thompson Sampling, we see that regret increases by a factor of c.

Theorem 5. *Consider the Modified Thompson Sampling with Gaussian priors and input parameters* (b, c) *run for* $T > bN + \frac{4}{\min_i \Delta_i^2}$ *timesteps, where* $b \geq 0, c \geq 1$. *Then the algorithm has expected regret* $bN + O(c\sqrt{N(T - bN)\log N})$ *(problem-independent), or* $bN + \sum_{i=1}^{N} O(c\frac{\log(T-bN)}{\Delta_i})$ *(problem dependent).*

The full proof of Theorem 5 is deferred to Appendix B. It generalizes the analysis of Agrawal and Goyal (2017) to incorporate the new parameters b and c. Without loss of generality, we assume arm 1 is the unique optimal arm. The key idea is to note that $\mathbb{E}[\mathcal{R}(T, \pi)] = \sum_{i=1}^{N} \Delta_i \mathbb{E}[n_{i,T}]$ and $\mathbb{E}[n_{i,T}] = \sum_{t=1}^{T} \Pr[a_t = i]$. Therefore in order to control the expected regret it suffices to control the probability that arm i gets play at any time t. This can be done by intersecting this probability to events of the form $\{\hat{\mu}_{i,t-1} - \mu_i \leq \Delta_i/3\}$ and $\{\theta_{i,t} - \mu_1 \leq -\Delta_i/3\}$, where arm 1 is assumed WLOG to be the optimal arm. Since rewards are bounded in $[0, 1]$ and the $\theta_{i,t}$ are Gaussian, one can tightly control these probabilities by carefully conditioning on the history using Chernoff-type bounds tailored for this problem.

5 EXPERIMENTS

In this section, we evaluate the empirical performance of the modified Thompson Sampling algorithm under varying b and c parameter values. This both helps illustrate performance of the algorithm in terms of privacy and regret, and it also helps illustrate the role of the additional parameters. Recall that $b = 0$ and $c = 1$ is the special case corresponding to standard Thompson Sampling. We vary combinations of (b, c) that achieve the same fixed privacy budget, as measured by the GDP parameter, which we also vary. These experiments can also provide insight for identifying the optimal (b, c) values to minimize regret given a fixed privacy budget. We consider two families of true reward distributions: Bernoulli and exponentially distributed. In Appendix D, we also compare our algorithm against two recent non-TS-based DP algorithms for online bandit problems: DP-SE (Sajed and Sheffet, 2019) and Anytime-Lazy-UCB (Hu et al., 2021)). All experiments were run on a personal laptop with an M1 Pro chip in around 2 hours.

5.1 Bernoulli rewards

We start with the experimental setting of Hu and Hegde (2022), where $N = 5$ arms have Bernoulli rewards with means [0.75, 0.625, 0.5, 0.375, 0.25] respectively. We let $T = 10^5$ and vary the desired η-GDP privacy parameter to be $1, 2$ and 5. We vary the (b, c) parameters jointly to ensure that the desired GDP guarantee is obtained. Recall from Lemma 4 that to ensure μ-GDP for $T = 10^5$ rounds, b and c must satisfy $\eta = \sqrt{\frac{1}{c(b+1)}}10^{2.5}$.

Note that standard Thompson Sampling with $b = 0$ and $c = 1$ would yield $10^{2.5} \approx 316$-GDP, which is orders of magnitude higher than the privacy parameters considered here, so performance for these parameter values are not shown in the plots.

Figure 2 shows the *empirical regret* for each parameter combination over time, defined as,

$$\mathbb{E}[\mathcal{R}(T, \pi)] = \mu^* T - \textstyle\sum_{t=1}^{T} r_t. \qquad (3)$$

The empirical regret is averaged over 10 runs, for which we already observe convergence of the average regret empirically. For each parameter setting, we observe an initial period of high regret, corresponding to the pre-pulling phase; this is more pronounced for larger b values. Afterwards, there is a phase transition to much lower per-round regret, once the algorithm begins the traditional Thompson Sampling phase. Parameter regimes with larger b values perform extremely well in this phase, both because they have a more accurate warm-start from the pre-pulls, and because larger b corresponds to smaller c for the same fixed GDP guarantees, corresponding to lower variance in the sampling step. Smaller b values do not suffer this initial period of loss, but they do incur more per-round regret; at the extreme, $b = 0$ has so much noise that its estimates do not converge. The convex shape of the regret in the pre-pulling phase is an artifact of our specific setting, where the algorithm pre-pulls the arms with the highest average rewards first.

We also observe that the lowest regret at time T is achieved by parameter settings with intermediate values of b and c, rather than settings that only involve tuning each parameter alone. This suggests that an optimal tuning strategy would increase both b and c together. Finally, we observe that the values of b and c that lead to the lowest empirical regret depend on the privacy budget, meaning that parameter tuning to optimize the privacy-regret tradeoff must take into account the desired privacy level. The regret for all

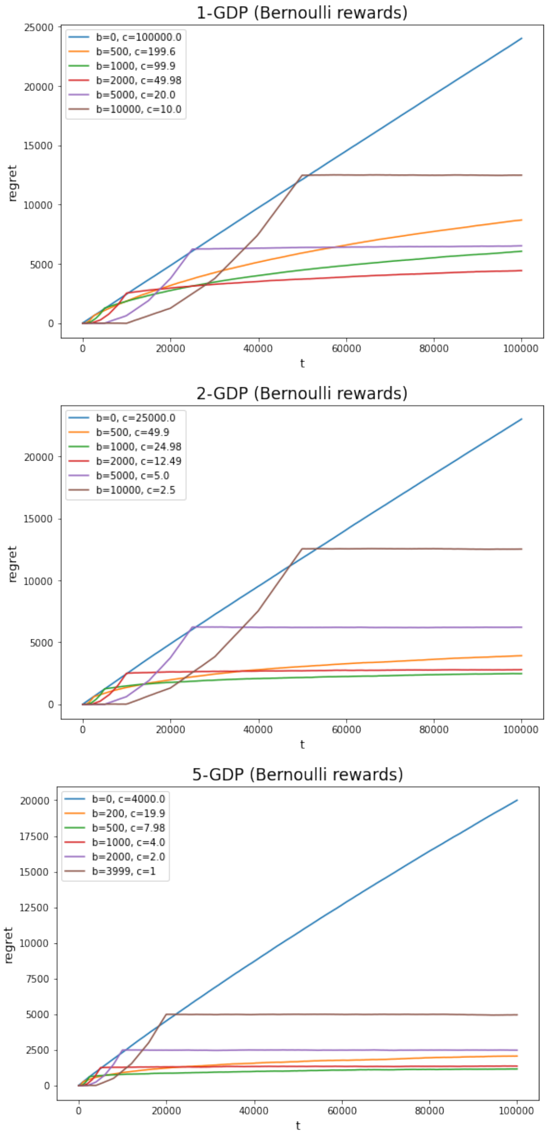

Figure 2: Empirical regret of Algorithm 2 under varying (b, c) when rewards are Bernoulli distributed.

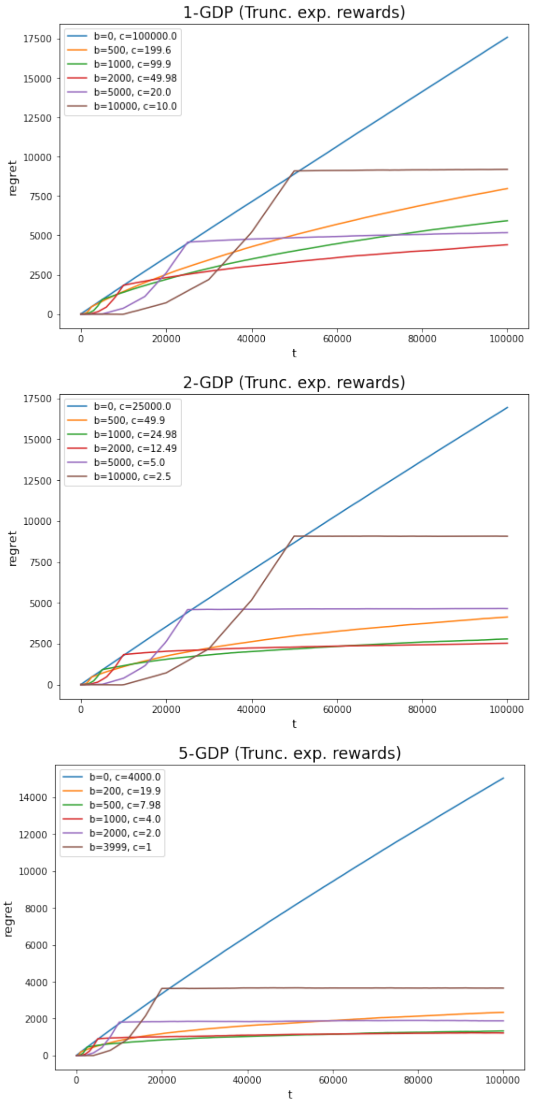

Figure 3: Empirical regret of Algorithm 2 under varying (b, c) when rewards are generated from a truncated exponential distribution.

parameters decreases with weaker privacy guarantees, as expected.

5.2 Truncated exponential rewards

Next, we consider rewards sampled from the truncate exponential distribution on [0,1] with varying rates. We again consider $N = 5$ arms respectively with exponential rates [0.1, 1, 2, 5, 10], corresponding to means of approximately [0.492, 0.418, 0.343, 0.193, 0.1]. We again fix $T = 10^5$, and vary the (b, c) parameters jointly to achieve desired GDP guarantees of $\eta = 1, 2, 5$.

Figure 3 shows the empirical regret as defined in Equation (3) for each parameter combination over time, averaged over 10 runs. We observe qualitatively similar findings as with the Bernoulli rewards: larger b corresponds to high initial regret and then low per-round regret; larger c corresponds to higher per-round regret, with the largest value of c (i.e., $b = 0$) having much higher regret; the optimal regret-minimizing parameter regimes involve intermediate values of b and c; and the optimal parameter values depend on the desired GDP parameter.

6 CONCLUSION

In this work, we analyze the privacy guarantees of the Thompson Sampling algorithm (Thompson, 1933; Russo et al., 2018) with Gaussian priors, which is commonly used for learning with bandit feedback. We show that the original Thompson Sampling algorithm satisfies differential privacy without any modifications, leveraging structural similarities between the algorithm's sampling procedure and existing DP tools, namely the Gaussian Mechanism. Importantly, this result means that there is *no* loss in performance from adding privacy, and known regret bounds (Agrawal and Goyal, 2017) still hold for the private algorithm.

Additionally, we show that two small modifications to the algorithm – namely pre-pulling each arm b times and scaling up the variance of sampling noise by a factor of c – enables tunable privacy guarantees. The resulting privacy and regret guarantees depend on the values of the new parameters b and c, which can be tuned to substantially improve the privacy guarantee at only a small increase in regret. We demonstrate our theoretical results empirically on two different reward distributions, and show substantial improvements in regret for a fixed privacy guarantee by properly tuning the parameters b and c.

Acknowledgements

T.O. and R.C. were supported in part by NSF grants CNS-2138834 (CAREER) and EEC-2133516. M.A.M. was supported in part by NSF grant DMS-2310973.

References

Shipra Agrawal and Navin Goyal. Analysis of Thompson Sampling for the multi-armed bandit problem. In *Proceedings of the 25th Annual Conference on Learning Theory*, volume 23 of *COLT '12*, pages 39.1–39.26, 2012.

Shipra Agrawal and Navin Goyal. Thompson Sampling for contextual bandits with linear payoffs. In *Proceedings of the 30th International Conference on Machine Learning - Volume 28*, ICML'13, 2013.

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

Jeremiah Blocki, Avrim Blum, Anupam Datta, and Or Sheffet. The Johnson-Lindenstrauss transform itself preserves differential privacy. In *2012 IEEE 53rd Annual Symposium on Foundations of Computer Science*, FOCS '12, pages 410–419, 2012.

Olivier Chapelle and Lihong Li. An empirical evaluation of Thompson Sampling. In *Advances in Neural Information Processing Systems*, volume 24 of *NeurIPS '11*, 2011.

Rachel Cummings, David M. Pennock, and Jennifer Wortman Vaughan. The possibilities and limitations of private prediction markets. *ACM Transactions on Economics and Computation*, 8(3):1–24, 2020.

Jinshuo Dong, Aaron Roth, and Weijie J. Su. Gaussian Differential Privacy. *Journal of the Royal Statistical Society, Series B*, 84(1):3–37, 2022.

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

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the Third Conference on Theory of Cryptography*, TCC '06, pages 265–284, 2006.

Cynthia Dwork, Guy N. Rothblum, and Salil Vadhan. Boosting and differential privacy. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, FOCS '10, pages 51–60, 2010.

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

Bingshan Hu and Nidhi Hegde. Near-optimal Thompson Sampling-based algorithms for differentially private stochastic bandits. In *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, volume 180 of *UAI '22*, pages 844–852, 2022.

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

Zhiming Huang, Yifan Xu, and Jianping Pan. TSOR: Thompson sampling-based opportunistic routing. *IEEE Transactions on Wireless Communications*, 20(11):7272–7285, 2021.

Alihan Hüyük and Cem Tekin. Thompson sampling for combinatorial network optimization in unknown environments. *IEEE/ACM Transactions on Networking*, 28(6):2836–2849, 2020.

Yi Liu and Veronika Ročková. Variable selection via Thompson sampling. *Journal of the American Statistical Association*, 118(541):287–304, 2023.

Ilya Mironov. Rényi differential privacy. In *IEEE Computer Security Foundations Symposium (CSF)*, 2017.

Nikita Mishra and Abhradeep Thakurta. (Nearly) optimal differentially private stochastic multi-arm bandits. In *Proceedings of the 31st Conference on Uncertainty in Artificial Intelligence*, UAI '15, pages 592–601, 2015.

Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. Smooth sensitivity and sampling in private data analysis. In *Proceedings of the Thirty-Ninth Annual ACM Symposium on Theory of Computing*, STOC '07, pages 75–84, 2007.

Daniel J Russo, Benjamin Van Roy, Abbas Kazerouni, Ian Osband, and Zheng Wen. A tutorial on Thompson sampling. *Foundations and Trends in Machine Learning*, 11(1):1–96, 2018.

Touqir Sajed and Or Sheffet. An optimal private stochastic-MAB algorithm based on optimal private stopping rule. In *International Conference on Machine Learning*, ICML '19, 2019.

Adam Smith, Shuang Song, and Abhradeep Guha Thakurta. The Flajolet-Martin sketch itself preserves differential privacy: Private counting with minimal space. In *Advances in Neural Information Processing Systems*, volume 33 of *NeurIPS '20*, pages 19561–19572, 2020.

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

Aristide C. Y. Tossou and Christos Dimitrakakis. Algorithms for differentially private multi-armed bandits. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI '16, pages 2087–2093, 2016.

Siwei Wang and Wei Chen. Thompson Sampling for combinatorial semi-bandits. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *ICML '18'*, pages 5114–5122, 2018.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. Yes
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. Not Applicable
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. Yes - in supplemental material

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. Yes
   (b) Complete proofs of all theoretical results. Yes - In supplemental material
   (c) Clear explanations of any assumptions. Yes

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). Yes - in supplemental material
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). Yes
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). Yes
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). Yes

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. Not Applicable
   (b) The license information of the assets, if applicable. Not Applicable
   (c) New assets either in the supplemental material or as a URL, if applicable. Not Applicable
   (d) Information about consent from data providers/curators. Not Applicable
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. Not Applicable

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. Not Applicable
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. Not Applicable
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. Not Applicable

Thompson Sampling Itself is Differentially Private
Supplementary Materials

A OMITTED PRIVACY PROOFS

A.1 Proof of Lemma 2

Lemma 2. *The mechanism $\mathcal{M}_{TS}(\mathcal{F}_t)$ satisfies $\sqrt{\frac{1}{2}}$-GDP with respect to observed rewards.*

Proof. We first note that Thompson Sampling itself is inherently an adaptive algorithm whose output at any time step t depends on its previous outputs at time steps $1, 2, \ldots, t-1$. In order to prove the privacy guarantee of Algorithm 1 composed for T time steps, it suffices to show that at every time step, conditioning on the outputs in the previous time steps, the algorithm is $\sqrt{1/2}$-GDP. To prove Lemma 2, consider two neighboring histories $\mathcal{F}_t$ and $\mathcal{F}'_t$ of length t that differ only in one observed reward. Consider the vector-valued N-dimensional query $\hat{\mu}(\mathcal{F}_t) = (\hat{\mu}_1(\mathcal{F}_t), \ldots, \hat{\mu}_N(\mathcal{F}_t))$ that computes the empirical mean of observed rewards from each arm, given reward history $\mathcal{F}_t$:

$$\hat{\mu}_i(\mathcal{F}_t) = \frac{1}{n_{i,t}+1} \sum_{\tau=1}^{t} 1_{a_\tau = i} r_\tau.$$

Then for each $i \in [N]$, $\theta_{i,t+1}$ can be expressed as $\hat{\mu}_i(\mathcal{F}_t)$ plus an independent Gassian noise term sampled from $\mathcal{N}(0, \frac{1}{n_{i,t}+1})$.

We then analyze the privacy guarantee of releasing $\{\theta_{i,t+1}\}_{i\in[N]}$, which is sufficient to determine the arm pulled at time $t+1$, as simply the argmax of all $\theta_{i,t+1}$. Let j denote this arm. Since $\theta_{j,t+1}$ is the only value that would depend on the reward observed at time t, then it would be the only value that differs between neighboring databases of reward histories $\mathcal{F}_t$ and $\mathcal{F}'_t$. Therefore the (single shot) N-dimensional empirical mean query has sensitivity $s = \max |\hat{\mu}_j(\mathcal{F}_t) - \hat{\mu}_j(\mathcal{F}'_t)| \leq \frac{1}{n_{j,t}+1}$. The standard deviation of noise added is $\sqrt{\frac{1}{n_{j,t}+1}}$, so the GDP parameter is at most:

$$\left| \frac{\frac{1}{n_{j,t}+1}}{\sqrt{\frac{1}{n_{j,t}+1}}} \right| = \frac{1}{\sqrt{n_{j,t}+1}} \leq \frac{1}{\sqrt{2}}. \tag{4}$$

The final inequality in Equation (4) comes from the assumption that $n_{j,t} \geq 1$, which is required to avoid the degenerate case of empty database $\hat{\mu}_{j,t}$ used in the Gaussian mechanism. Note that if this were not the case, then $\theta_{j,t+1} \sim \mathcal{N}(0,1)$, and there would be no privacy loss associated with arm j because there would be no data to protect.

□

A.2 Proof of Lemma 4

Lemma 4. *Modified Thompson Sampling with Gaussian priors and input parameters (b, c) run for T timesteps satisfies $\sqrt{\frac{1}{c(b+1)}T}$-GDP.*

Proof. We first show the per-round GDP guarantee of Algorithm 2 with respect to the observed rewards is $\sqrt{\frac{1}{(b+1)c}}$, and then compose across rounds using Lemma 1 to reach the final result. To prove the per-round privacy guarantee, consider two neighboring histories $\mathcal{F}_t$ and $\mathcal{F}'_t$ of length t that differ only in one observed reward. Consider the vector-valued N-dimensional query $\hat{\mu}(\mathcal{F}_t) = (\hat{\mu}_1(\mathcal{F}_t), \dots, \hat{\mu}_N(\mathcal{F}_t))$ that computes the empirical mean of observed rewards from each arm, given reward history $\mathcal{F}_t$:

$$\hat{\mu}_i(\mathcal{F}_t) = \frac{1}{n_{i,t}+1} \sum_{\tau=1}^{t} 1_{a_\tau = i} r_\tau .$$

For each $i \in [N]$, $\theta_{i,t+1}$ can be expressed as $\hat{\mu}_i(\mathcal{F}_t)$ plus an independent Gassian noise term sampled from $\mathcal{N}(0, \frac{c}{n_{i,t}+1})$.

Following the same argument as in the proof of Lemma 2, we analyze the privacy guarantee of releasing $\{\theta_{i,t+1}\}_{i\in[N]}$, which is sufficient to determine the arm pulled at time $t+1$ as $\arg\max_i \theta_{i,t+1}$. Denote this arm as j. As in the proof of Lemma 2, $\theta_{j,t+1}$ is the only value that would depend on the reward observed at time t, so it would be the only value that differs between neighboring histories $\mathcal{F}_t$ and $\mathcal{F}'_t$. Therefore the (single shot) N-dimensional empirical mean query is essentially a histogram query with sensitivity $s = \max |\hat{\mu}_j(\mathcal{F}_t) - \hat{\mu}_j(\mathcal{F}'_t)| \leq \frac{1}{n_{j,t}+1}$.

To calculate the GDP parameter, with a variance multiplier c, the standard deviation of the Gaussian noise added is now $\sqrt{\frac{c}{n_{j,t}+1}}$. Additionally, we have a new lower bound for $n_{j,t}$, of b, due to the b pre-pulls of each arm. Then the GDP parameter is at most:

$$\left| \frac{\frac{1}{n_{j,t}+1}}{\sqrt{\frac{c}{n_{j,t}+1}}} \right| = \frac{1}{\sqrt{c(n_{j,t}+1)}} \leq \frac{1}{\sqrt{c(b+1)}}.$$

Then, using Lemma 1 which composes the single-step privacy guarantee adaptively, the GDP parameter over T timesteps is

$$\sqrt{\sum_{t=1}^{T} \left(\frac{1}{\sqrt{c(b+1)}} \right)^2} = \sqrt{\frac{1}{c(b+1)}T}.$$

□

B OMITTED REGRET PROOFS

We adapt the arguments of Agrawal and Goyal (2017) to our modified Thompson sampling algorithm. We first provide the main argument of the proof in Section B.1, followed by proofs of three auxiliary lemmas in Section B.2.

B.1 Proof of Theorem 5

Theorem 5. *Consider the Modified Thompson Sampling with Gaussian priors and input parameters (b, c) run for $T > bN + \frac{4}{\min_i \Delta_i^2}$ timesteps, where $b \geq 0, c \geq 1$. Then the algorithm has expected regret $bN + O(c\sqrt{N(T-bN)\log N})$ (problem-independent), or $bN + \sum_{i=1}^{N} O(c\frac{\log(T-bN)}{\Delta_i})$ (problem dependent).*

Recall the notation of Algorithm 2, that a_t is the arm played and outputted at time t, $n_{i,t}$ is the number of times arm i is pulled after t time steps, and $\hat{\mu}_{i,t} = \frac{1}{n_{i,t}+1}(\sum_{j=1}^{b} r_{i,j} + \sum_{\tau=1:a_\tau=i}^{t} r_t)$ is the empirical mean of rewards for arm i after t time steps. At time step t, the sample $\theta_{i,t}$ is drawn from the posterior distribution from observations in the previous $t-1$ timesteps, i.e. $\theta_{i,t} \sim \mathcal{N}(\hat{\mu}_{i,t-1}, \frac{1}{n_{i,t-1}+1})$. We emphasize that at time step t, the sample $\theta_{i,t} \sim \mathcal{N}(\hat{\mu}_{i,t-1}, \frac{1}{n_{i,t-1}+1})$ is drawn from a distribution updated by data up to time $t-1$. That is, the decision a_t only depends on $\hat{\mu}_{i,t-1}$ and $\frac{1}{n_{i,t-1}+1}$.

Recall that the expected problem-dependent regret (Equation (2)) can be written as:

$$\mathbb{E}[\mathcal{R}(T,\pi)] = \sum_{i=1}^{N} \Delta_i \mathbb{E}[n_{i,T}].$$

We will therefore need to bound the expected number of suboptimal plays of each arm i during T time steps. Without loss of generality, we assume arm 1 is the optimal arm, and define $\Delta_i = \mu_1 - \mu_i$ to be the suboptimality gap of arm i; if there are multiple optimal arms, this will only improve regret.

For each i, define $x_i = \mu_i + \Delta_i/3$ and $y_i = \mu_1 - \Delta_i/3$. These will serve as two "mid-points" between μ_i and μ_1. Thus for $i \neq 1$, we have $\mu_i < x_i < y_i < \mu_1$, and for $i = 1$, we have $\mu_i = \mu_1 = x_1 = y_1$. We will also define the events $E_i^{\mu}(t) = \{\hat{\mu}_{i,t-1} \leq x_i\}$, and $E_i^{\theta}(t) = \{\theta_{i,t} \leq y_i\}$. Note that when the number of observed rewards of arm i increases, empirical mean $\hat{\mu}_{i,t-1}$ convergences to μ_i, making the probability of event $E_i^{\mu}(t)$ tend to 1 and that of $E_i^{\theta}(t)$ tend to 0. We denote the respective complements of these events by $\overline{E_i^{\mu}(t)}$ and $\overline{E_i^{\theta}(t)}$. Finally, we define $p_{i,t} := \Pr[\theta_{1,t} > y_i | \mathcal{F}_{t-1}]$. Note that this is a random variable that depends on $\mathcal{F}_{t-1}$, and has a fixed value if $\mathcal{F}_{t-1}$ is instantiated to be a particular history F_{t-1}.

Given b prepulls per arm and $T > bN$, this notation and the law of total probability leads to the following identity on the number of arm pulls.

$$\begin{aligned}\mathbb{E}[n_{i,T}] &= b + \sum_{t=1}^{T-bN} \Pr\left[a_t = i\right] \\ &= b + \sum_{t=1}^{T-bN} \Pr\left[a_t = i, \overline{E_i^{\mu}(t)}\right] + \sum_{t=1}^{T-bN} \Pr\left[a_t = i, E_i^{\mu}(t), \overline{E_i^{\theta}(t)}\right] + \sum_{t=1}^{T-bN} \Pr\left[a_t = i, E_i^{\mu}(t), E_i^{\theta}(t)\right]. \end{aligned} \tag{5}$$

We will upper bound each of these three terms separately. Lemmas 5 and 6 bound the first two terms in Equation (5), and are both proven in Section B.2.

Lemma 5. $\sum_{t=1}^{T-bN} \Pr\left[a_t = i, \overline{E_i^{\mu}(t)}\right] \leq \frac{9}{\Delta_i^2} e^{-2b\Delta_i^2/9}$.

Lemma 6. *Let* $T \geq bN + \frac{1}{\Delta_i^2} e^{\frac{1}{4\pi}}$*. Then,* $\sum_{t=1}^{T-bN} \Pr\left[a_t = i, E_i^{\mu}(t), \overline{E_i^{\theta}(t)}\right] \leq \max\{0, \frac{18c\log((T-bN)\Delta_i^2)}{\Delta_i^2} - b\} + \frac{1}{\Delta_i^2}$.

In order to bound the third term in Equation (5), we will need Lemma 7.

Lemma 7 (Agrawal and Goyal (2017)). *For all* t*,* $i \neq 1$ *and history* F_{t-1} *we have*

$$\Pr\left[a_t = i, E_i^{\mu}(t), E_i^{\theta}(t) | \mathcal{F}_{t-1} = F_{t-1}\right] \leq \frac{1-p_{i,t}}{p_{i,t}} \Pr\left[a_t = 1, E_i^{\mu}(t), E_i^{\theta}(t) | \mathcal{F}_{t-1} = F_{t-1}\right].$$

Taking conditional expectations with respect to $\mathcal{F}_{t-1}$, followed by Lemma 7 we see that

$$\begin{aligned}\sum_{t=1}^{T-bN} \Pr\left[a_t = i, E_i^{\mu}(t), E_i^{\theta}(t)\right] &= \sum_{t=1}^{T-bN} \mathbb{E}\left[\Pr(a_t = i, E_i^{\mu}(t), E_i^{\theta}(t) | \mathcal{F}_{t-1})\right] \\ &\leq \sum_{t=1}^{T-bN} \mathbb{E}\left[\frac{1-p_{i,t}}{p_{i,t}} \Pr(a_t = 1, E_i^{\mu}(t), E_i^{\theta}(t) | \mathcal{F}_{t-1})\right] \\ &= \sum_{t=1}^{T-bN} \mathbb{E}\left[\mathbb{E}\left[\frac{1-p_{i,t}}{p_{i,t}} \mathbb{1}\left(a_t = 1, E_i^{\mu}(t), E_i^{\theta}(t)\right) \middle| \mathcal{F}_{t-1}\right]\right] \\ &= \sum_{t=1}^{T-bN} \mathbb{E}\left[\frac{1-p_{i,t}}{p_{i,t}} \mathbb{1}(a_t = 1, E_i^{\mu}(t), E_i^{\theta}(t))\right]. \end{aligned} \tag{6}$$

The second step is an application of Lemma 7, and the third step uses the fact that $p_{i,t}$ is fixed given F_{t-1}.

Now, let τ_k be the time at which arm 1 is played for the k-th time, not counting the pre-pull stage, so that $n_{1,\tau_k} = b+k$. Note that $p_{i,t} = \Pr[\theta_{1,t} > y_i | \mathcal{F}_{t-1}]$ changes only if the distribution of $\theta_{1,t}$ changes. Thus, $p_{i,t}$ is the fixed for all $t \in \{\tau_k + 1, \dots, \tau_{k+1}\}$ for every k. Then,

$$\begin{aligned}
\sum_{t=1}^{T-bN} \mathbb{E}\left[\frac{1-p_{i,t}}{p_{i,t}} \mathbb{1}(a_t = 1, E_i^\mu(t), E_i^\theta(t))\right] &= \sum_{k=0}^{T-bN-1} \mathbb{E}\left[\frac{1-p_{i,\tau_k+1}}{p_{i,\tau_k+1}} \sum_{t=\tau_k+1}^{\tau_{k+1}} \mathbb{1}\left(a_t = 1, E_i^\mu(t), E_i^\theta(t)\right)\right] \\
&\leq \sum_{k=0}^{T-bN-1} \mathbb{E}\left[\frac{1-p_{i,\tau_k+1}}{p_{i,\tau_k+1}}\right] \\
&= \sum_{k=0}^{T-bN-1} \mathbb{E}\left[\frac{1}{p_{i,\tau_k+1}} - 1\right]. \qquad (7)
\end{aligned}$$

To continue bounding $\mathbb{E}[\frac{1}{p_{i,\tau_k+1}}] - 1$, we require Lemma 8, which is also proved in Section B.2.

Lemma 8. *Let τ_k be the first time when arm 1 is played for the k-th time excluding the pre-pulls, i.e. $n_{1,\tau_k} = b+k$ and $n_{1,t} < b+k$ for $t < \tau_k$. Then for $T \geq bN + \frac{4}{\Delta_i^2}$,*

$$\mathbb{E}\left[\frac{1}{p_{i,\tau_k+1}}\right] - 1 \leq \begin{cases} 71 & \textit{for all } k, \\ \frac{4}{(T-bN)\Delta_i^2} & \textit{for } k \geq \max\{1, \frac{72c}{\Delta_i^2}\log((T-bN)\Delta_i^2) - (b+1)\}. \end{cases}$$

For ease of notation, define $L = \lceil \frac{72c}{\Delta_i^2} \log((T-bN)\Delta_i^2) - (b+1) \rceil$. Combining (6), (7) and Lemma 8 we finally obtain the following upper bound on the third term in Equation (5):

$$\begin{aligned}
\sum_{t=1}^{T-bN} \Pr\left[a_t = i, E_i^\mu(t), E_i^\theta(t)\right] &\leq \sum_{k=0}^{T-bN-1} \mathbb{E}\left[\frac{1}{p_{i,\tau_k+1}} - 1\right] \\
&\leq 71L \cdot \mathbf{1}(L > 0) + \sum_{k=L}^{T-bN-1} \frac{4}{(T-bN)\Delta_i^2} \\
&\leq 71L \cdot \mathbf{1}(L > 0) + \frac{4}{\Delta_i^2}. \qquad (8)
\end{aligned}$$

The first step combines Equations (6) and (7), the second step applies Lemma 8, treating separately the terms in the sum with k larger or smaller than L, and conditioning on the case that $L > 0$, otherwise the first term in this expression will be upper bounded by 0. The third step upper bounds the final term, since $\frac{T-bN-1-L}{T-bN} < 1$.

Putting everything together. Finally, plugging in the bounds of Lemma 5, Lemma 6, and Equation (8) into Equation (5) gives the following bound on the expected number of times each arm i is pulled in a stream of length T. Recall that $L = \lceil \frac{72c}{\Delta_i^2} \log((T-bN)\Delta_i^2) - (b+1) \rceil$. Then,

$$\begin{aligned}
\mathbb{E}[n_{i,T}] \leq\ & b + \frac{9}{\Delta_i^2} e^{-2b\Delta_i^2/9} + \max\{0, \frac{18c\log((T-bN)\Delta_i^2)}{\Delta_i^2} - b\} + \frac{1}{\Delta_i^2} + 71L\mathbb{1}(L > 0) + \frac{4}{\Delta_i^2} \\
\leq\ & b + \frac{9}{\Delta_i^2} + \frac{18c\log((T-bN)\Delta_i^2)}{\Delta_i^2} + \frac{1}{\Delta_i^2} + 71 \cdot \frac{72c}{\Delta_i^2} \log((T-bN)\Delta_i^2) + \frac{4}{\Delta_i^2}. \\
=\ & b + \frac{14}{\Delta_i^2} + \frac{18c\log((T-bN)\Delta_i^2)}{\Delta_i^2} + \frac{5112c}{\Delta_i^2} \log((T-bN)\Delta_i^2) \\
=\ & b + \frac{14}{\Delta_i^2} + \frac{5130c}{\Delta_i^2} \log((T-bN)\Delta_i^2). \qquad (9)
\end{aligned}$$

The first inequality is plugging in the bounds of Lemma 5, Lemma 6, and Equation (8) into Equation (5). The second inequality holds because $\frac{18c\log((T-bN)\Delta_i^2)}{\Delta_i^2} > 0$ (because $T - bN > e^{1/4\pi}/\Delta_i^2$ by assumption) is an upper bound for $\max\{0, \frac{18c\log((T-bN)\Delta_i^2)}{\Delta_i^2} - b\}$, and $\frac{72c}{\Delta_i^2}\log((T-bN)\Delta_i^2) > 0$ is an upper bound for L. The third and fourth steps combine terms.

Note that in Equation (9), the term b comes from the pre-pulling stage which we will count separately from the Thompson Sampling stage. We focus on the Thompson Sampling stage for now, and upper bound the regret from the pre-pulling stage, which is at most bN later. Define $\tilde{\mathbb{E}}[n_{i,T}]$ to be the expected number of pulls of arm i excluding the b pre-pulls, from $t = 1$ to $T - bN$, in the Thompson Sampling stage (excluding the pre-pulls). Then,

$$\tilde{\mathbb{E}}[n_{i,T}] \leq \frac{14}{\Delta_i^2} + \frac{5130c}{\Delta_i^2}\log((T-bN)\Delta_i^2).$$

To obtain an upper bound on the expected regret due to arm i in the $T - bN$ timesteps of the Thompson Sampling stage, we multiply the above expression by Δ_i.

$$\Delta_i\tilde{\mathbb{E}}[n_{i,T}] \leq \frac{14}{\Delta_i} + \frac{5130c}{\Delta_i}\log((T-bN)\Delta_i^2). \tag{10}$$

From Equation (10), adding up the expected regret over the $N - 1$ suboptimal arms and adding back the bN maximum possible regret from the pre-pulling phase, we obtain the desired problem-dependent asymptotic bound $bN + \sum_{i=1}^N O(c\frac{\log((T-bN)\Delta_i^2)}{\Delta_i})$.

Moving to the problem-independent bound, we note that the first term in Equation (10) is decreasing in Δ_i and the second term will also be decreasing for large enough Δ_i. More precisely, define $f(\Delta_i) = \frac{\log\left((T-bN)\Delta_i^2\right)}{\Delta_i}$, so that $f'(\Delta_i) = \frac{2-\log((T-bN)\Delta_i^2)}{\Delta_i^2}$. We see that $f'(\Delta_i) < 0$ (i.e., the second term is decreasing in Δ_i) if $\Delta_i \geq \frac{e}{\sqrt{T-bN}}$. Therefore, if we consider those arms with $\Delta_i \geq \frac{e\sqrt{N\log N}}{\sqrt{T-bN}}$, the total regret these arms incur in the Thompson Sampling stage would be bounded by:

$$\begin{aligned}
\sum_{i=1}^N \left\{\frac{14}{\Delta_i} + 5130c \cdot \frac{\log\left((T-bN)\Delta_i^2\right)}{\Delta_i}\right\} &\leq \sum_{i=1}^N \left\{\left(\frac{14\sqrt{T-bN}}{e\sqrt{N\log N}} - \right) + 5130c \cdot \sqrt{T-bN} \cdot \frac{\log\left(e^2 N\log N\right)}{e\sqrt{N\log N}}\right\} \\
&= \frac{14\sqrt{N(T-bN)}}{e\sqrt{\log N}} + 5130c \cdot \sqrt{T-bN}\frac{\sqrt{N}\log\left(e^2 N\log N\right)}{e\sqrt{\log N}} \\
&= O\left(c\sqrt{N(T-bN)\log N}\right).
\end{aligned}$$

For every arm with $\Delta_i \leq \frac{e\sqrt{N\log N}}{\sqrt{T-bN}}$, the total regret due to all of these arms in $T - bN$ time steps is bounded by $(T-bN)\Delta_i \leq (T-bN)\frac{e\sqrt{N\log N}}{\sqrt{T-bN}} \leq e\sqrt{N(T-bN)\log N} = O(\sqrt{N(T-bN)\log N})$, because $bN \leq T$. Therefore, if we add up the regrets due to all arms (those with $\Delta_i \geq \frac{e\sqrt{N\log N}}{\sqrt{T-bN}}$ and those with $\Delta_i \leq \frac{e\sqrt{N\log N}}{\sqrt{T-bN}}$), in the Thompson Sampling stage, we get that the total regret is $O\left(c\sqrt{N(T-bN)\log N}\right) + O\left(\sqrt{N(T-bN)\log N}\right) = O\left(c\sqrt{N(T-bN)\log N}\right)$. Adding the regret from the pre-pulling stage – which is at most bN – and the regret from the Thompson Sampling stage, we conclude that the total regret in all T timesteps is bounded by $bN + O(c\sqrt{N(T-bN)\log N})$.

B.2 Proofs of Auxiliary Lemmas

Lemmas 5, Lemma 6 and 8 can be viewed as extended and refined versions of Lemmas 2.15, 2.16, and 2.13 in (Agrawal and Goyal, 2017), respectively.

Lemma 5. $\sum_{t=1}^{T-bN} \Pr\left[a_t = i, \overline{E_i^\mu(t)}\right] \leq \frac{9}{\Delta_i^2}e^{-2b\Delta_i^2/9}$.

Proof. Recall that $\overline{E_i^\mu(t)} = \{\hat{\mu}_{i,t-1} > x_i\}$. Let $\tau_{i,k}$ denote the time we pull arm i for the k-th time, excluding the pre-pulling stage. Note that $t \leq \tau_{i,t}$ for all $t \in \mathbb{N}$, and that for $k > n_{i,t}$ it holds that $\tau_{i,k} > t$.

$$\sum_{t=1}^{T-bN} \Pr[a_t = i, \overline{E_i^{\mu}(t)}] \leq \sum_{k=1}^{T-bN} \Pr[\overline{E_i^{\mu}(\tau_{i,k})}].$$

The above inequality holds true because if we are at a timestep t such that the pulled arm is not i, then the probability in the left-hand side sum is zero. So we should only count the probabilities at timesteps $\{t = \tau_{i,k}\}$ for $k = 1, 2, \ldots, T - bN$.

At time $\tau_{i,k}$, the empirical mean $\hat{\mu}_{i,\tau_{i,k}-1}$ used by the algorithm to make the decision is upper bounded by the average of the outcomes of $(b + k - 1)$ i.i.d. plays of arm i. We will use Hoeffding's inequality (Lemma 9) to obtain high probability bounds of the deviations between these empirical means and their true means.

Lemma 9 (Hoeffding's inequality)**.** *Let $X_1, \ldots, X_n \in [0, 1]$ be i.i.d. and $\mathbb{E}[X_i] = \mu, \forall i$. Let $S_n = X_1 + \ldots + X_n$. Then for all $a \geq 0$,*

$$\Pr[S_n \geq n\mu + a] \leq e^{-2a^2/n} \quad \textit{and} \quad \Pr[S_n \leq n\mu - a] \leq e^{-2a^2/n}.$$

Using the definition $\overline{E_i^{\mu}(\tau_{i,k})} = \{\hat{\mu}_{i,\tau_{i,k}-1} > \mu_i + \Delta_i/3\}$, followed by Hoeffding's Inequality (Lemma 9),

$$\begin{aligned}\sum_{k=1}^{T-bN} \Pr\left[\overline{E_i^{\mu}(\tau_{i,k})}\right] &= \sum_{k=1}^{T-bN} \Pr\left[\hat{\mu}_{i,\tau_{i,k}-1} - \mu_i > \frac{\Delta_i}{3}\right] \\ &\leq \sum_{k=1}^{T-bN} e^{-\frac{2(k+b-1)\Delta_i^2}{9}} \\ &= \sum_{k=0}^{T-bN-1} e^{-\frac{2(k+b)\Delta_i^2}{9}} \\ &= e^{-\frac{2b\Delta_i^2}{9}} \sum_{k=0}^{T-bN-1} (e^{-\frac{2\Delta_i^2}{9}})^k \\ &= \frac{e^{-2b\Delta_i^2/9}(1 - e^{-2(T-bN)\Delta_i^2/9})}{1 - e^{-2\Delta_i^2/9}}. \end{aligned} \tag{11}$$

The last equalities follow by simple manipulations and the computation of a geometric sum. Finally, noting that $\frac{1-e^{-xt}}{1-e^{-x}} \leq \frac{1}{1-e^{-x}} \leq \frac{2}{x}$ for $t \geq 0$ and $x \in (0, 2/9]$, we see that (11) implies

$$\sum_{k=1}^{T-bN} \Pr\left[\overline{E_i^{\mu}(\tau_{i,k})}\right] \leq \frac{9}{\Delta_i^2} e^{-2b\Delta_i^2/9}.$$

□

Lemma 6. *Let $T \geq bN + \frac{1}{\Delta_i^2} e^{\frac{1}{4\pi}}$. Then, $\sum_{t=1}^{T-bN} \Pr\left[a_t = i, E_i^{\mu}(t), \overline{E_i^{\theta}(t)}\right] \leq \max\{0, \frac{18c\log((T-bN)\Delta_i^2)}{\Delta_i^2} - b\} + \frac{1}{\Delta_i^2}$.*

Proof. First define $M = \frac{18c\log((T-bN)\Delta_i^2)}{\Delta_i^2}$. Then by the law of total probability,

$$\begin{aligned}\sum_{t=1}^{T-bN} \Pr\left[a_t = i, E_i^{\mu}(t), \overline{E_i^{\theta}(t)}\right] &= \sum_{t=1}^{T-bN} \Pr\left[a_t = i, n_{i,t-1} + b \leq M, E_i^{\mu}(t), \overline{E_i^{\theta}(t)}\right] \\ &\quad + \sum_{t=1}^{T-bN} \Pr\left[a_t = i, n_{i,t-1} + b > M, E_i^{\mu}(t), \overline{E_i^{\theta}(t)}\right]. \end{aligned} \tag{12}$$

We can bound the first term in Equation (12) by removing the conditioning, as follows,

$$\begin{aligned}\sum_{t=1}^{T-bN} \Pr\left[a_t = i, n_{i,t-1} + b \leq M, E_i^{\mu}(t), \overline{E_i^{\theta}(t)}\right] &\leq \mathbb{E}\left[\sum_{t=1}^{T-bN} \mathbb{1}\left(a_t = i, n_{i,t-1} + b \leq M\right)\right] \\ &\leq \max\{0, M - b\}. \end{aligned} \tag{13}$$

where the second step holds because for any arm i, the timesteps satisfying $a_t = i, n_{i,t-1} + b \leq M$ are those in which the arm was pulled at most $M - b$ times.

To bound the second term in Equation (12), we show that if $n_{i,t-1}$ is large and the event $E_i^{\mu}(t)$ is satisfied, then the probability that the event $E_i^{\theta}(t)$ is violated is small. Recall that $E_i^{\theta}(t)$ is the event that $\theta_{i,t} \leq y_i$.

$$\begin{aligned}
&\sum_{t=1}^{T-bN} \Pr\left[a_t = i, n_{i,t-1} + b > M, E_i^{\mu}(t), \overline{E_i^{\theta}(t)}\right] \\
\leq\ & \sum_{t=1}^{T-bN} \Pr\left(a_t = i, \overline{E_i^{\theta}(t)} \middle| n_{i,t-1} + b > M,\ E_i^{\mu}(t)\right) \\
=\ & \mathbb{E}\left[\sum_{t=1}^{T-bN} \Pr\left(a_t = i, \overline{E_i^{\theta}(t)} \middle| n_{i,t-1} + b > M,\ E_i^{\mu}(t)\right) \middle| \mathcal{F}_{t-1}\right] \\
=\ & \mathbb{E}\left[\sum_{t=1}^{T-bN} \Pr\left(a_t = i, \overline{E_i^{\theta}(t)} \middle| n_{i,t-1} + b > M,\ E_i^{\mu}(t),\ \mathcal{F}_{t-1}\right)\right] \\
=\ & \mathbb{E}\left[\sum_{t=1}^{T-bN} \Pr\left(\theta_{i,t} > y_i \middle| n_{i,t-1} + b > M,\ \hat{\mu}_{i,t-1} \leq x_i,\ \mathcal{F}_{t-1}\right)\right]. \qquad (14)
\end{aligned}$$

Next we will upper bound the probabilities inside the expectation in Equation (14) using that $Z \sim N(0,1)$ we have that

$$\begin{aligned}
\mathbb{E}\left[\Pr\left(\theta_{i,t} > y_i \middle| n_{i,t-1} + b > M,\ \hat{\mu}_{i,t-1} \leq x_i,\ \mathcal{F}_{t-1}\right)\right] &\leq \Pr\left[\hat{\mu}_{i,t-1} + Z\sqrt{\frac{c}{n_{i,t-1} + b + 1}} > y_i\right] \\
&\leq \Pr\left[x_i + Z\sqrt{\frac{c}{n_{i,t-1} + b + 1}} > y_i\right]. \qquad (15)
\end{aligned}$$

To bound the latter expression we will use Mill's inequality (Lemma 10) as it allows us to bound the deviations of centered Gaussian random variable $\theta_{i,t}$ around 0.

Lemma 10 (Mill's inequality)**.** *Let $X \sim N(\mu, \sigma^2)$. Then for any $t > 0$,*

$$\Pr[X - \mu > t] \leq \frac{\sigma}{\sqrt{2\pi}} \frac{e^{-\frac{t^2}{2\sigma^2}}}{t} \qquad \textit{and} \qquad \Pr[X - \mu < -t] \leq \frac{\sigma}{\sqrt{2\pi}} \frac{e^{-\frac{t^2}{2\sigma^2}}}{t}.$$

Let Z be a standard Gaussian random variable. Consecutively using $y_i - x_i = \frac{\Delta_i}{3}$, Mill's inequality (Lemma 10), the fact that $n_{i,t-1} + b + 1 > M = \frac{18c\log((T-bN)\Delta_i^2)}{\Delta_i^2}$ and $T \geq bN + \frac{1}{\Delta_i^2} e^{\frac{1}{4\pi}}$ we see that

$$\begin{aligned}
\Pr\left[x_i + Z\sqrt{\frac{c}{n_{i,t-1} + b + 1}} > y_i\right] &= \Pr\left[Z\sqrt{\frac{c}{n_{i,t-1} + b + 1}} > \frac{1}{3}\Delta_i\right] \\
&\leq \sqrt{\frac{c}{2\pi(n_{i,t-1} + b + 1)}} \frac{3}{\Delta_i} \exp\left(-\frac{\Delta_i^2}{18} \frac{(n_{i,t-1} + b + 1)}{c}\right) \\
&\leq \frac{1}{2\sqrt{\pi \log((T - bN)\Delta_i^2)}} \frac{1}{(T - bN)\Delta_i^2} \\
&\leq \frac{1}{(T - bN)\Delta_i^2}. \qquad (16)
\end{aligned}$$

We can now use (14), (15)(16) and sum all the probabilities over $t = 1, \ldots, T - bN$, yielding

$$\sum_{t=1}^{T-bN} \Pr\left[a_t = i, n_{i,t-1} + b > cM_i(T), E_i^{\mu}(t), \overline{E_i^{\theta}(t)}\right] \leq \frac{1}{\Delta_i^2}. \qquad (17)$$

Plugging in the bounds of Equations (13) and (17) into Equation (12) gives the desired bound. □

Lemma 8. *Let τ_k be the first time when arm 1 is played for the k-th time excluding the pre-pulls, i.e. $n_{1,\tau_k} = b+k$ and $n_{1,t} < b+k$ for $t < \tau_k$. Then for $T \geq bN + \frac{4}{\Delta_i^2}$,*

$$\mathbb{E}\left[\frac{1}{p_{i,\tau_k+1}}\right] - 1 \leq \begin{cases} 71 & \textit{for all } k, \\ \frac{4}{(T-bN)\Delta_i^2} & \textit{for } k \geq \max\{1, \frac{72c}{\Delta_i^2}\log((T-bN)\Delta_i^2) - (b+1)\}. \end{cases}$$

Proof. Recall that $p_{i,t} = \Pr[\theta_{1,t} > y_i | \mathcal{F}_{t-1}]$ and $\theta_{i,t} \sim \mathcal{N}(\hat{\mu}_{i,t-1}, \frac{c}{n_{i,t-1}+1})$. Let us introduce some useful notation. Let $\Theta_1, \ldots, \Theta_r \overset{iid}{\sim} \mathcal{N}(\hat{\mu}_{1,\tau_k}, \frac{c}{k+b+1})$ be a random sample identically distributed to θ_{i,τ_k} given $\mathcal{F}_{\tau_k}$, and let G_k be a geometric random variable that denotes the number of trials until $\Theta_k > y_i$. Note that $p_{i,\tau_k+1} = \Pr\left[\Theta_j > y_i | \mathcal{F}_{\tau_k}\right]$ and hence

$$\mathbb{E}\left[\frac{1}{p_{i,\tau_k+1}}\right] - 1 = \mathbb{E}\left[\mathbb{E}\left[G_k | \mathcal{F}_{\tau_k}\right]\right] = \mathbb{E}\left[G_k\right] = \sum_{r=0}^{\infty} \Pr(G_k \geq r).$$

We will therefore establish the desired upper bound by upper bounding the summands of the above expression or equivalently, lower bounding $\Pr(G_k < j)$. We will first establish a bound that holds for all k, and then we will establish a tighter bound that holds when k is sufficiently large.

Upper bound for all j: Define $\text{MAX}_r := \max_{1 \leq k \leq r}(\Theta_k)$ and $z = \sqrt{\log r}$. Then,

$$\begin{aligned}
\Pr[G_k < r] &\geq \Pr[\text{MAX}_r > y_i] \\
&\geq \Pr\left[\text{MAX}_r > \hat{\mu}_{1,\tau_k} + \sqrt{\frac{c}{k+b+1}} z \geq y_i\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\mathbb{1}\left(\text{MAX}_r > \hat{\mu}_{1,\tau_k} + \sqrt{\frac{c}{k+b+1}} z \geq y_i\right) \Big| \mathcal{F}_{\tau_k}\right]\right] \\
&= \mathbb{E}\left[\mathbb{1}\left(\hat{\mu}_{1,\tau_k} + \sqrt{\frac{c}{k+b+1}} z \geq y_i\right) \Pr\left(\text{MAX}_r > \hat{\mu}_{1,\tau_j} + \sqrt{\frac{c}{k+b+1}} z \Big| \mathcal{F}_{\tau_k}\right)\right]. \qquad (18)
\end{aligned}$$

To continue bounding this expression, we first lower bound $\Pr(\text{MAX}_r > \hat{\mu}_{1,\tau_k} + \sqrt{\frac{c}{k+b+1}} z | \mathcal{F}_{\tau_k})$ using Mill's inequality (Lemma 10). This lemma gives that for any instantiation F_{τ_k} of $\mathcal{F}_{\tau_k}$ and $r > 1$, then

$$\begin{aligned}
\Pr\left[\text{MAX}_r > \hat{\mu}_{1,\tau_k} + \sqrt{\frac{c}{k+b+1}} z | \mathcal{F}_{\tau_k} = F_{\tau_k}\right] &= 1 - \prod_{k=1}^{r} \Pr\left[\Theta_k \leq \hat{\mu}_{1,\tau_k} + \sqrt{\frac{c}{k+b+1}} z | \mathcal{F}_{\tau_k} = F_{\tau_k}\right] \\
&\geq 1 - \left(1 - \frac{1}{\sqrt{2\pi}} \frac{e^{-z^2/2}}{z}\right)^r \\
&= 1 - \left(1 - \frac{1}{\sqrt{2\pi r \log r}}\right)^r \\
&\geq 1 - e^{-\sqrt{\frac{r}{2\pi \log r}}} \qquad (19)
\end{aligned}$$

If $r \geq e^{11}$, we have $e^{-\sqrt{\frac{r}{2\pi \log r}}} \leq \frac{1}{r^2}$. Therefore, we can bound this term separately for $r < e^{11}$ and $r \geq e^{11}$,

$$\Pr\left[\text{MAX}_r > \hat{\mu}_{1,\tau_k} + \sqrt{\frac{c}{k+b+1}} z | \mathcal{F}_{\tau_k} = F_{\tau_k}\right] \geq \begin{cases} 1 - e^{-\sqrt{\frac{r}{2\pi \log r}}} & \text{for } 1 < r < e^{11} \\ 1 - \frac{1}{r^2} & \text{for } r \geq e^{11} \end{cases}.$$

Plugging this into Equation (18) gives,

$$\begin{aligned}
\Pr[G_k < r] &\geq \begin{cases} \left(1 - e^{-\sqrt{\frac{r}{2\pi \log r}}}\right) \mathbb{E}\left[\mathbb{1}\left(\hat{\mu}_{1,\tau_k} + \sqrt{\frac{c}{k+b+1}} z \geq y_i\right)\right] & \text{for } 1 < r < e^{11} \\ \left(1 - \frac{1}{r^2}\right) \mathbb{E}\left[\mathbb{1}\left(\hat{\mu}_{1,\tau_k} + \sqrt{\frac{c}{k+b+1}} z \geq y_i\right)\right] & \text{for } r \geq e^{11} \end{cases} \\
&= \begin{cases} \left(1 - e^{-\sqrt{\frac{r}{2\pi \log r}}}\right) \Pr\left(\hat{\mu}_{1,\tau_k} + \sqrt{\frac{c}{k+b+1}} z \geq y_i\right) & \text{for } 1 < r < e^{11} \\ \left(1 - \frac{1}{r^2}\right) \Pr\left(\hat{\mu}_{1,\tau_k} + \sqrt{\frac{c}{k+b+1}} z \geq y_i\right) & \text{for } r \geq e^{11} \end{cases}. \qquad (20)
\end{aligned}$$

Continuing lower bounding this expression, we have:

$$\begin{aligned}
\Pr\left[\hat{\mu}_{1,\tau_k} + \sqrt{\frac{c}{k+b+1}} z \geq y_i\right] &= \Pr\left[\hat{\mu}_{1,\tau_k} - \mu_1 \geq -\sqrt{\frac{c \log r}{b+k+1}} - \frac{1}{3}\Delta_i\right] \\
&\geq 1 - \exp\left(-2(b+k+1)\left(\frac{1}{3}\Delta_i + \sqrt{\frac{c\log r}{b+k+1}}\right)^2\right) \\
&= 1 - \frac{1}{r^{2c}} \exp\left(-\frac{2}{9}(b+k+1)\Delta_i^2 - \frac{4}{3}\Delta_i\sqrt{(b+k+1)c\log r}\right) \\
&\geq 1 - \frac{1}{r^{2c}} \\
&\geq 1 - \frac{1}{r^2} \qquad (21)
\end{aligned}$$

where the first step comes from the definitions of $y_i = \mu_1 - \frac{1}{3}\Delta_i$ and $z = \sqrt{\log r}$, the second inequality comes from an application the Hoeffding's inequality. . The third line expands and combines terms, the fourth line upper bounds the exponential term by 1, and the fifth line follows from $c \geq 1$.

We can return to bounding $\mathbb{E}[G_k] = \sum_{r=0}^{\infty} \Pr(G_k \geq r)$ using Equation (18) by plugging in Equations (20) and (21).

$$\begin{aligned}
\mathbb{E}[G_k] &= \sum_{r=0}^{\infty} \Pr[G_k \geq r] \\
&\leq 1 + 1 + \sum_{r=2}^{\infty}(1 - \Pr[G_k < r]) \\
&\leq 2 + \sum_{r=2}^{\lfloor e^{11}\rfloor}\left(1 - \left(1 - e^{-\sqrt{\frac{r}{2\pi\log r}}}\right)\left(1 - \frac{1}{r^2}\right)\right) + \sum_{r=\lceil e^{11}\rceil}^{\infty}\left(1 - \left(1 - \frac{1}{r^2}\right)^2\right) \\
&= 2 + \sum_{r=2}^{\lfloor e^{11}\rfloor}\left(e^{-\sqrt{\frac{r}{2\pi\log r}}}\left(1 - \frac{1}{r^2}\right) + \frac{1}{r^2}\right) + \sum_{r=\lceil e^{11}\rceil}^{\infty}\left(\frac{2}{r^2} - \frac{1}{r^4}\right) \\
&\leq 2 + 2\sum_{r=2}^{\infty}\frac{1}{r^2} + \sum_{r=2}^{\lfloor e^{11}\rfloor} e^{-\sqrt{\frac{r}{2\pi\log r}}} \\
&\leq 2 + \frac{\pi^2}{3} + 65.58 \\
&\leq 71.
\end{aligned}$$

Thus,

$$\mathbb{E}\left[\frac{1}{p_{i,\tau_k+1}}\right] - 1 = \mathbb{E}[G_k] \leq 71,$$

which is the first upper bound of the lemma.

Tighter upper bound for $k \geq \max\{1, \frac{72c}{\Delta_i^2}\log((T-bN)\Delta_i^2) - (b+1)\}$: Note that when k is large, there is an increased probability of the event $\theta_1 > y_i$, because $\hat{\mu}_{1,t}$ is closer to μ_1. Thus we give a tighter bound for this case when k is sufficiently large.

We first use the fact that $\Theta_k \sim \mathcal{N}(\hat{\mu}_{1,\tau_k}, \frac{c}{b+k+1})$ to apply Lemma 11, which bounds the tails of Θ_k.

Lemma 11. *Let $X \sim N(\mu, \sigma^2)$. Then for any $t > 0$, we have*

$$\Pr[X - \mu > t] \leq e^{-\frac{t^2}{2\sigma^2}} \qquad \textit{and} \qquad \Pr[X - \mu < -t] \leq e^{-\frac{t^2}{2\sigma^2}}.$$

We instantiate Lemma 11 on Θ_k with $t = \Delta_i/6$ to get the first inequality below, and use the assumed lower bound on k to get the second inequality below. Thus for any instantiation F_{τ_k} of $\mathcal{F}_{\tau_k}$,

$$\begin{aligned}\Pr\left[\Theta_k > \hat{\mu}_{1,\tau_k} - \frac{\Delta_i}{6} \big| \mathcal{F}_{\tau_k} = F_{\tau_k}\right] &\geq 1 - e^{-\frac{\Delta_i^2(b+k+1)}{72c}} \\ &\geq 1 - \frac{1}{(T-bN)\Delta_i^2}. \end{aligned} \tag{22}$$

Next define the event A_{t-1} to be the event that $\hat{\mu}_{1,t-1} - \frac{\Delta_i}{6} \geq y_i$. Note that A_{t-1} implicitly depends on the history $\mathcal{F}_{t-1}$. Now, consider an instantiation F_{τ_k} of $\mathcal{F}_{\tau_k}$ such that A_{τ_k} occurs. For such F_{τ_k}, from Equation (22) we have that,

$$\Pr[\Theta_k > y_k | \mathcal{F}_{\tau_k} = F_{\tau_k}] \geq 1 - \frac{1}{(T-bN)\Delta_i^2}. \tag{23}$$

Let $\mathcal{F}_{t-1}|_{A_{t-1}}$ denote the random variable $\mathcal{F}_{t-1}$ conditioned on the event A_{t-1} occurring. Then

$$\begin{aligned}\mathbb{E}\left[\frac{1}{p_{i,\tau_k+1}}\right] &= \mathbb{E}\left[\frac{1}{\Pr(\Theta_k > y_i | \mathcal{F}_{\tau_k})}\right] \\ &\leq \mathbb{E}\left[\frac{1}{\Pr(\Theta_k > y_i | \mathcal{F}_{\tau_k}|_{A_{\tau_k}}) \Pr(A_{\tau_k})}\right] \\ &\leq \mathbb{E}\left[\frac{1}{(1 - \frac{1}{(T-bN)\Delta_i^2}) \Pr(A_{\tau_k})}\right]. \end{aligned} \tag{24}$$

The second step is by law of total probability, and the third step applies the bound in Equation (23).

We can continue to bound $\Pr(A_{\tau_k})$ as follows. For any $t \geq \tau_k + 1$ and $j \geq \frac{72c}{\Delta_i^2} \log((T-bN)\Delta_i^2) - b - 1$,

$$\begin{aligned}\Pr(A_{t-1}) &= 1 - \Pr\left[\hat{\mu}_{1,t} < \mu_1 - \frac{\Delta_i}{6}\right] \\ &\geq 1 - \exp\left(-\frac{n_{1,t-1}\Delta_i^2}{18}\right) \\ &\geq 1 - \exp\left(-4c\log((T-bN)\Delta_i^2) + \frac{1}{18}\right) \\ &\geq 1 - e^{1/18}\frac{1}{(T-bN)^4\Delta_i^8}, \end{aligned} \tag{25}$$

where the first line follows from the definition of A_{t-1}, the second line is an application of Hoeffding's inequality (Lemma 9), the third line uses $n_{1,t-1} \geq b+k \geq \frac{72c}{\Delta_i^2}\log((T-bN)\Delta_i^2) - 1$ for any $t \geq \tau_k + 1$ and the last inequality used $c \geq 1$.

Hence for $T \geq bN + \frac{e^{1/54}}{\Delta_i^2}$, from (23) we obtain the lower bound

$$\Pr(A_{t-1}) \geq 1 - \frac{1}{(T-bN)\Delta_i^2}. \tag{26}$$

Finally, combining (24), (25) and (26), using that $(T-bN)\Delta_i^2 \geq 4$ by our assumption, we get that

$$\mathbb{E}\left[\frac{1}{p_{i,\tau_k}}\right] - 1 \leq \frac{1}{\left(1 - \frac{1}{(T-bN)\Delta_i^2}\right)^2} - 1 \leq \frac{4}{(T-bN)\Delta_i^2}.$$

□

C ALTERNATE PRIVACY ANALYSES

C.1 Alternative Method: Standard DP

In the privacy analysis presented in Section 3 and 4, we treat Thompson sampling at every time step as an instantiation of the Gaussian mechanism and analyze the privacy guarantee assuming that all samples $\{\theta_{i,t}\}_{i\in[N]}$ are output. However, in reality, only the index of the sample which with the max value needs to be published:

$$\arg\max_i \theta_{i,t} \quad \text{where} \;\; \theta_{i,t} \sim \mathcal{N}(\hat{\mu}_{i,t-1}, \frac{1}{n_{i,t-1}+1}) \;\; \text{for} \;\; i = 1, \ldots, N. \tag{27}$$

We are interested in whether this fact can be used to achieve a better bound. There is an existing algorithm, ReportNoisyMax (Dwork and Roth, 2014), presented in Algorithm 3 for differentially privately computing an argmax of several functions. This algorithm adds Laplace noise of the same parameters to each value, and then produces the argmax of the noisy values. To contrast, Equation (27) adds Gaussian noise with different variance to each empirical means. Thus a single round of Thompson Sampling can be viewed as a variant of the classic ReportNoisyMax algorithm, that adds heterogeneous Gaussian noise.

Algorithm 3 ReportNoisyMax

```
Input: Database $R$, queries $\{f_1, \ldots, f_n\}$ of sensitivity $s$, privacy parameter $\epsilon$
Sample $Z_1, \ldots, Z_n \sim \text{Lap}(s/\epsilon)$
Return $\arg\max_{i\in[n]}(f_i(R) + Z_i)$
```

Algorithm 3 achieves $(\epsilon, 0)$-differential privacy by tuning the Laplace noise parameter based on the sensitivity s of all functions and privacy parameter ϵ (Dwork and Roth, 2014). Algorithm 4 formally defines the Heterogeneous Gaussian ReportNoisyMax algorithm, which adds Gaussian noise of heterogeneous variances to a set of queries, where each query possesses different sensitivities.

Algorithm 4 Heterogeneous Gaussian ReportNoisyMax

```
Input: Database $R$, queries $\{f_1, \ldots, f_n\}$, where query $f_i$ has sensitivity $s_i$, noise variances $\{\sigma_1^2, \ldots, \sigma_n^2\}$
Sample $X_i \sim \mathcal{N}(0, \sigma_i^2)$ for $i \in [n]$
Return $\arg\max_{i\in[n]}(f_i(R) + X_i)$
```

In the context of Thompson sampling, the queries $f_i(R)$ are the empirical means $\hat{\mu}_{i,t}$ of the arms, given the history of observed rewards $\mathcal{F}_t$. Recall that neighboring histories $\mathcal{F}_t$ and $\mathcal{F}_t'$ contain databases of rewards R and R' that differ only in a single reward observation, so only one empirical mean will be different across these neighbors. Thus we prove differential privacy guarantees for Algorithm 4 under the assumption that for any pair of neighboring databases R, R', it holds that $f(R) = (f_1(R), \ldots, f_n(R))$ and $f(R') = (f_1(R'), \ldots, f_n(R'))$ differ at at most one function value, and the sensitivity of the j-th function is s_j.

Theorem 6 gives the privacy guarantee of Algorithm 4 under this assumption. The proof of this theorem follows closely to the structure of the proof of privacy of ReportNoisyMax in Dwork and Roth (2014), but is modified in key ways based on the different noise distribution, and the fact that Algorithm 4 satisfies (ϵ, δ)-DP for $\delta > 0$, while Algorithm 3 satisfies $(\epsilon, 0)$-DP.

Theorem 6. *Assume that for any pair of neighboring databases R and R', $f(R)$ and $f(R')$ differ at at most one entry. Then, Algorithm 4 is (ϵ, δ)-differentially private for $\epsilon \geq \frac{1}{2}\sqrt{\log\frac{n-1}{2\delta}}\max_{i\in[n]}\left(\frac{s_i}{\sigma_i}\right)$.*

Proof. Let R and R' be two neighboring databases, let $\mathcal{M}$ denote Algorithm 4, and let $c = f(D)$ and $c' = f(D')$ for the given collection of functions $f = \{f_1, \ldots, f_n\}$. Without loss of generality, let $c \geq c'$. Fix any $i \in [n]$. We will bound from above and below the ratio of the probabilities that i is selected with D and with D'. Fix X_{-i} to be a draw from the Gaussian distributions used for all queries except the i-th one.

We first argue that $\Pr[\mathcal{M}(R) = i|X_{-i}] \leq \Pr[\mathcal{M}(R') = i|X_{-i}] + \frac{1}{n-1}\delta$. Define X^* to be the minimum value of X_i such that $c_i + X_i \leq c_j' + X_j$ for all $j \neq i$. Then i is the output of Algorithm 4 if and only if $X_i > X^*$. Then for all $j \neq i$,

$$\begin{aligned} & c_i + X^* > c_j + X_j \\ \implies \quad & (s_i + c_i' + X^*) \geq c_i + X^* > c_j + X_j \geq c_j' + X_j \\ \implies \quad & c_i' + (s_i + X^*) \geq c_j' + X_j. \end{aligned}$$

Thus, if $X_i \geq X^* + s_i$, then the i-th noisy function value will be the maximum under database R' when the noise vector is (X_i, X_{-i}). Note that $X_i = \sigma_i Z$ in distribution, for $Z \sim \mathcal{N}(0,1)$.

We will next apply a lemma from Nissim et al. (2007) that bounds the closeness between a standard normal random variable Z and Z plus an additive shift.

Lemma 12 (Nissim et al. (2007))**.** *For a subset $\mathcal{S} \in \mathbb{R}^d$ and a vector $a \in \mathbb{R}^d$, we write $\mathcal{S} + a$ for the set $\{y + a : y \in \mathcal{S}\}$. The standard normal distribution $\mathcal{N}(0,1)$ satisfies that for all $||a||_1 \leq \frac{2\epsilon}{\sqrt{\log(1/2\delta)}}$ and subsets $\mathcal{S} \in \mathbb{R}^d$,*

$$\Pr_{Z \sim \mathcal{N}(0,1)}[Z \in \mathcal{S}] \leq e^{\epsilon} \Pr_{Z \sim \mathcal{N}(0,1)}[Z \in \mathcal{S} + a] + \delta.$$

Applying Lemma 12 with $d = 1$, $a = \frac{s_i}{\sigma_i}$, and for (ϵ, δ) such that $\frac{2\epsilon}{\sqrt{\log((n-1)/2\delta)}} \geq \max_{i \in [n]} \left(\frac{s_i}{\sigma_i}\right) \geq \frac{s_i}{\sigma_i}$,

$$\begin{aligned} & \Pr[Z \geq \frac{X^*}{\sigma_i}] \leq e^{\epsilon} \Pr[Z \geq \frac{X^*}{\sigma_i} + \frac{s_i}{\sigma_i}] + \frac{\delta}{n-1} \\ \implies & \Pr[x_i \geq x^*] \leq e^{\epsilon} \Pr[x_i \geq x^* + s_i] + \frac{\delta}{n-1}. \end{aligned}$$

Then,

$$\Pr[\mathcal{M}(D) = i | X_{-i}] = \Pr[X_i \geq X^*] \leq e^{\epsilon} \Pr[X_i \geq X^* + s_i] + \frac{\delta}{n-1} \leq e^{\epsilon} \Pr[\mathcal{M}(R') = i | X_{-i}] + \frac{\delta}{n-1}.$$

We now argue that $\Pr[\mathcal{M}(R') = i | X_{-i}] \leq \Pr[\mathcal{M}(R) = i | X_{-i}] + \frac{1}{n-1}\delta$. Similarly, define X^* to be the minimum value of X_i such that $c_i + X_i \leq c_j' + X_j$ for all $j \neq i$. This means that i is the output of $\mathcal{M}$ if and only if $X_i > X^*$. Then for all $j \neq i$, we have

$$\begin{aligned} & c_i' + X^* > c_j' + X_j \\ \implies \quad & s_i + c_i' + X^* > s_i + c_j' + X_j \\ \implies \quad & c_i' + (s_i + X^*) > (s_i + c_j') + X_j \\ \implies \quad & c_i + (s_i + X^*) \geq c_i' + (s_i + X^*) > (s_i + c_j') + X_j \geq c_j + X_j. \end{aligned}$$

Thus, if $X_i \geq X^* + s_i$ then the i-th noisy function value will be the maximum under database R when the noise vector is (X_i, X_{-i}). Again, we note that $X_i = \sigma_i Z$ in distribution, $Z \sim \mathcal{N}(0,1)$. Applying Lemma 12 with $d = 1$ and $a = \frac{s_i}{\sigma_i}$, then for (ϵ, δ) such that $\frac{2\epsilon}{\sqrt{\log((n-1)/2\delta)}} \geq \max_{i \in [n]} \left(\frac{s_i}{\sigma_i}\right) \geq \frac{s_i}{\sigma_i}$,

$$\begin{aligned} & \Pr[Z \geq \frac{X^*}{\sigma_i}] \leq e^{\epsilon} \Pr[Z \geq \frac{X^*}{\sigma_i} + \frac{s_i}{\sigma_i}] + \frac{\delta}{n-1} \\ \implies \quad & \Pr[X_i \geq X^*] \leq e^{\epsilon} \Pr[X_i \geq X^* + s_i] + \frac{\delta}{n-1}. \end{aligned}$$

Then,

$$\Pr[\mathcal{M}(R') = i | X_{-i}] = \Pr[X_i \geq X^*] \leq e^{\epsilon} \Pr[X_i \geq X^* + s_i] + \frac{\delta}{n-1} \leq e^{\epsilon} \Pr[\mathcal{M}(R) = i | X_{-i}] + \frac{\delta}{n-1}.$$

Using the law of total probability we can obtain $\Pr[\mathcal{M}(D) = i] \leq e^{\epsilon} \Pr[\mathcal{M}(D') = i] + \frac{\delta}{n-1}$ for all $i \in [n]$. Then, for any $\mathcal{S} = \{i_1, i_2, \ldots, i_k\}$, $\Pr[\mathcal{M}(D) \in \mathcal{S}] \leq e^{\epsilon} \Pr[\mathcal{M}(D') \in \mathcal{S}] + \frac{k}{n-1}\delta$. We have the worst case happening when $k = n - 1$ because when $k = n$, $\mathcal{S} = \mathit{Range}(\mathcal{M})$, both probabilities are 1. We conclude that Algorithm 4 is (ϵ, δ)-differentially private for $\epsilon \geq \frac{1}{2}\sqrt{\log \frac{n-1}{2\delta}} \max_{i \in [n]} \left(\frac{s_i}{\sigma_i}\right)$. □

Now, we can apply the above theorem in the context of Thompson Sampling.

Theorem 7. *At timestep t, the action of Algorithm 1 is (ϵ, δ)-differentially private for $\epsilon = \frac{1}{2\sqrt{2}}\sqrt{\log \frac{N-1}{2\delta}}$.*

Proof. At timestep t of Algorithm 1, we are instantiating Algorithm 4 with sensitivity $s_i = \frac{1}{n_{i,t-1}+1}$, and adding Gaussian noise of variance $\sigma_i^2 = \frac{1}{n_{i,t-1}+1}$. Plugging these into Theorem 6 gives that Algorithm 1 at timestep t is (ϵ, δ)-differentially private for $\epsilon \geq \frac{1}{2}\sqrt{\log \frac{N-1}{2\delta}} \max_i(\frac{1}{n_{i,t-1}+1}/\sqrt{\frac{1}{n_{i,t-1}+1}}) = \frac{1}{2}\sqrt{\log \frac{N-1}{2\delta}} \max_i\left(\frac{1}{\sqrt{n_{i,t-1}+1}}\right)$. Note that for privacy to be relevant, it must be that $n_{i,t-1} \geq 1$, otherwise there would no data to protect. Thus, $\max_i\left(\frac{1}{\sqrt{n_{i,t-1}+1}}\right) \leq \frac{1}{\sqrt{2}}$, and the algorithm is (ϵ, δ)-differentially private for any $\epsilon \geq \frac{1}{2\sqrt{2}}\sqrt{\log \frac{N-1}{2\delta}}$. □

Corollary 1 gives a bound on the complete Algorithm 1 across all timesteps, by utilizing Advanced Composition (Dwork et al., 2010) to compose the privacy guarantees of Theorem 7.

Corollary 1. *Given any ϵ, δ such that $\epsilon = \frac{1}{2\sqrt{2}}\sqrt{\log \frac{N-1}{2\delta}}$, Algorithm 1 is $(\epsilon_{TS}, \delta_{TS})$-differentially private for $\epsilon_{TS} = \epsilon\sqrt{2T\log(\frac{1}{\delta_{TS}-T\delta})} + T\epsilon(e^\epsilon - 1)$ for $\delta_{TS} > T\delta$.*

C.2 Alternative Method: RDP

Renyi Differential Privacy (RDP) (Mironov, 2017) also generalizes differential privacy, with the guarantees of closeness of outputs across neighboring databases based on *Renyi divergence.*

Definition 3. *(Renyi Differential Privacy (Mironov, 2017)). An algorithm $\mathcal{M}$ satisfies $(\alpha, \gamma(\alpha))$-RDP with $\alpha \geq 1$ if for any neighboring datasets D and D':*

$$D_\alpha(\mathcal{M}(D)||\mathcal{M}(D')) = \frac{1}{\alpha-1}\log \mathbb{E}_{x\sim\mathcal{M}(D)}\left[\left(\frac{\Pr[\mathcal{M}(D)=x]}{\Pr[\mathcal{M}(D')=x]}\right)^{\alpha-1}\right] \leq \gamma(\alpha),$$

where the Renyi divergence D_α between two distributions P and Q is

$$D_\alpha(P||Q) = \frac{1}{\alpha-1}\log \mathbb{E}_{x\sim Q}\left[(P(x)/Q(x))^\alpha\right] = \frac{1}{\alpha-1}\log \mathbb{E}_{x\sim P}[(P(x)/Q(x))^{\alpha-1}].$$

It is known that the Gaussian Mechanism that adds noise $\mathcal{N}(0, \sigma^2)$ to the value of a function with sensitivity s is $(\alpha, \gamma(\alpha))$-RDP where $\gamma(\alpha) = \frac{s^2}{2\sigma^2}\alpha$, for any $\alpha > 1$. In the context of a single-step of the Thompson Sampling algorithm, let arm j be the arm where the observed rewards different across two neighboring databases). Then,

$$\gamma(\alpha) = \frac{\alpha}{2\sigma_j^2}s_j^2 = \frac{1/(n_{j,t-1}+1)^2}{2/(n_{j,t-1}+1)}\alpha = \frac{1}{2(n_{j,t-1}+1)}\alpha \leq \frac{1}{4}\alpha.$$

The last inequality holds because $n_{j,t-1} \geq 1$, otherwise the dataset would be empty and there would be no data to protect.

The composition guarantees of RDP show that the adaptive composition of T mechanisms that each satisfy $(\alpha, \gamma(\alpha))$-RDP, will together satisfy $(\alpha, T\gamma(\alpha))$-RDP (Mironov, 2017). Thus T rounds of Thompson Sampling will together satisfy $(\alpha, \frac{1}{4}\alpha T)$-RDP. To convert the RDP guarantee back to DP, we use the fact from (Mironov, 2017) that if an algorithm $\mathcal{M}$ satisfies $(\alpha, \gamma(\alpha))$-RDP, then it also satisfies $(\gamma(\alpha)+\frac{\log(1/\delta)}{\alpha-1}, \delta)$-DP for any $\delta \in (0,1)$. This gives us the final privacy result of Theorem 8.

Theorem 8. *Algorithm 1 is (ϵ, δ)-differentially private for any $\delta \in (0,1)$, $\alpha > 1$, and $\epsilon = \frac{1}{4}\alpha T + \frac{\log(1/\delta)}{\alpha-1}$.*

C.3 Comparisons of the GDP, Standard DP and RDP Results

In Figure 4, we empirically evaluate the privacy guarantees provided by our three different privacy results: Theorem 3 (GDP), Corollary 1 (Standard DP), and Theorem 8 (RDP). We observe that the guarantees obtained by GDP and RDP are significantly tighter than the one obtained by the standard DP method. Unlike the standard DP guarantee, the GDP and RDP guarantees do not depend on N. RDP and GDP offer comparable guarantees, with GDP showing a slight advantage in specific regions, particularly when δ is small, and consistently performing no worse than RDP across all scenarios.

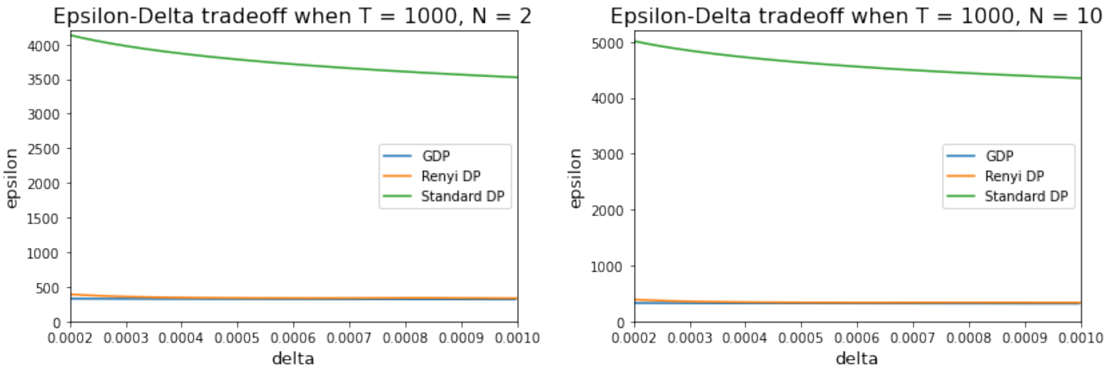

Figure 4: DP parameter ϵ as a function of δ when fixing $T = 1000$, obtained by three different analyses (GDP, RDP and Standard DP). The left plot has $N = 2$, and the right plot has $N = 10$.

D COMPARISON AGAINST OTHER PRIVATE BANDIT ALGORITHMS

We compare our modified TS algorithm against two recent non-TS-based DP algorithms for online bandit learning:, DP-SE (Sajed and Sheffet, 2019) and Anytime-Lazy-UCB (Hu et al., 2021), with empirical results in Figure 5 below. We use the same arm settings as in Section 5.1, and privacy parameters $\epsilon = 4.88, 35.57$, and 96.71, corresponding to 1-, 5- and 10-GDP if $\delta = 10^{-6}$ for all three algorithms. The (b, c) parameters used by TS are set via grid search and are omitted on the plots due to space limitation. We observe that when ϵ is small, corresponding to strongest privacy, TS is not the optimal algorithm; as ϵ increases, TS begins to substantially outperform the other methods.

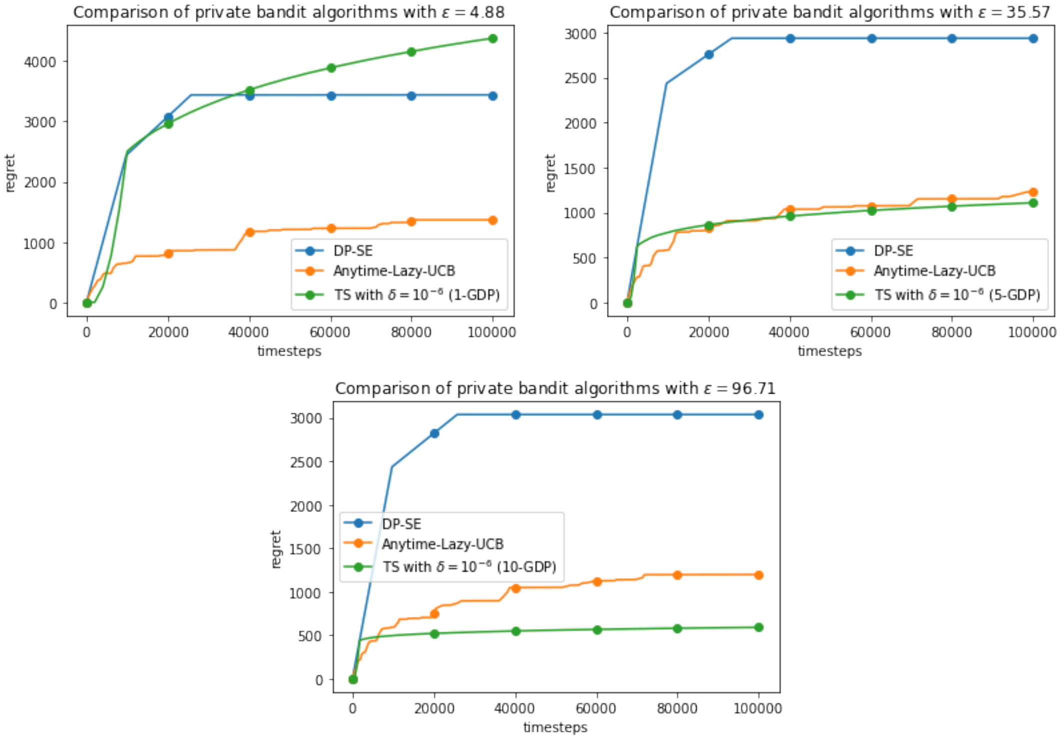

Figure 5: Comparisons of our modified TS algorithm against DP-SE (Sajed and Sheffet, 2019) and Anytime-Lazy-UCB (Hu et al., 2021), under fixed privacy levels $\epsilon = 4.88, 35.57$, and 96.71.