Thompson Sampling with Less Exploration is Fast and Optimal

Tianyuan Jin [1] **Xianglin Yang** [1] **Xiaokui Xiao** [1] **Pan Xu** [2]

Abstract

We propose ϵ-Exploring Thompson Sampling (ϵ-TS), a modified version of the Thompson Sampling (TS) algorithm (Agrawal & Goyal, 2017) for multi-armed bandits. In ϵ-TS, arms are selected greedily based on empirical mean rewards with probability $1-\epsilon$, and based on posterior samples obtained from TS with probability ϵ. Here, $\epsilon \in (0,1)$ is a user-defined constant. By reducing exploration, ϵ-TS improves computational efficiency compared to TS while achieving better regret bounds. We establish that ϵ-TS is both minimax optimal and asymptotically optimal for various popular reward distributions, including Gaussian, Bernoulli, Poisson, and Gamma. A key technical advancement in our analysis is the relaxation of the requirement for a stringent anti-concentration bound of the posterior distribution, which was necessary in recent analyses that achieved similar bounds (Jin et al., 2021b; 2022). As a result, ϵ-TS maintains the posterior update structure of TS while minimizing alterations, such as clipping the sampling distribution or solving the inverse of the Kullback-Leibler (KL) divergence between reward distributions, as done in previous work. Furthermore, our algorithm is as easy to implement as TS, but operates significantly faster due to reduced exploration. Empirical evaluations confirm the efficiency and optimality of ϵ-TS.

1. Introduction

Multi-Armed Bandit (MAB) is an elementary model to trade off exploration and exploitation in sequential decision-making problems. In such problems, an agent has a set $[K] = \{1, 2, \cdots, K\}$ of K arms to play with, where each arm $i \in [K]$ is associated with a reward distribution with an unknown mean value μ_i. The agent will pull arms sequentially for T time steps. At each time $t \in [T]$, the agent first chooses an arm $A_t \in [K]$ to play based on its past observations and then receives a reward r_t, which is independently sampled from the reward distribution of arm A_t. The goal is to maximize the cumulative reward over T times, which is equivalent to minimizing the regret, $R_\mu(T)$, defined as follows.

$$R_\mu(T) := T \cdot \max_{i\in[K]} \mu_i - \sum_{t=1}^{T} \mu_{A_t}. \tag{1.1}$$

We assume without loss of generality in this paper that arm 1 is the optimal arm with the largest mean reward, i.e., $\mu_1 = \max_{i\in[K]} \mu_i$. For a fixed bandit instance (namely, $\mu_1, \ldots, \mu_K$ are fixed but unknown), when T goes to infinity, Lai & Robbins (1985) show that the regret of any algorithm is at least $C(\mu)\log(T)(1-o(1))$, where

$$C(\mu) = \sum_{i>1} \frac{\Delta_i}{\mathrm{kl}(\mu_i, \mu_1)}, \tag{1.2}$$

$\Delta_i = \mu_1 - \mu_i$, and $\mathrm{kl}(\mu_i, \mu_1)$ is the KL-divergence between two reward distributions with mean μ_i and μ_1 respectively. A bandit algorithm is said to be *asymptotically optimal* if its regret can be upper bounded by $C(\mu)\log(T)(1-o(1))$ for some constant $C(\mu)$. Well-known algorithms such as Thompson Sampling (Agrawal & Goyal, 2017; Kaufmann et al., 2012; Korda et al., 2013), KL-UCB (Garivier & Cappé, 2011; Ménard & Garivier, 2017), and Bayes-UCB (Kaufmann, 2016) are all shown to be asymptotically optimal. When the time horizon T is fixed, Auer et al. (2002) show that no algorithm can achieve a worst-case regret lower than $C\sqrt{KT}$ for some universal constant C. Here the worst-case regret is defined as the maximum regret of the algorithm on any possible bandit instance. A bandit algorithm that achieves the worst-case regret $O(\sqrt{KT})$ is said to be *minimax optimal*.

In this paper, we focus on the setting where reward distributions are from the one-parameter family distributions (Korda et al., 2013; Garivier et al., 2016). In this setting, a bandit instance can be parameterized by $\{\theta_1, \cdots, \theta_K\}$, and arm $i \in [K]$ has reward distribution $p(\cdot|\theta_i)$ and mean reward $\mu_i = \mu(\theta_i)$. This family contains most common distributions such as Bernoulli, Gaussian, Poisson, and Gamma,

[1]National University of Singapore [2]Duke University. Correspondence to: Tianyuan Jin <tianyuan@u.nus.edu>, Pan Xu <pan.xu@duke.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

as listed in Table 1, where the function map from the parameter θ to the mean reward $\mu(\cdot)$ is also instantiated. To minimize the regret defined in (1.1), we need to estimate the parameter θ_i and choose the arm that could give us the highest mean reward to play. We study the Thompson Sampling (TS for short) strategy (Thompson, 1933), which is one of the most widely used bandit algorithms due to its simplicity in implementation and superior performance in practice (Chapelle & Li, 2011). Thompson sampling starts with a prior distribution and a posterior distribution of θ_i for each arm $i \in [K]$. At time step $t = 1, 2, \ldots$ of the bandit problem, TS samples $\theta_i(t)$ for all $i \in [K]$ from the corresponding posterior distributions and chooses the arm $A_t = \arg\max_i \mu(\theta_i(t))$ to play. After observing the reward for arm A_t, TS will update the posterior distribution based on all the observations collected for each arm.

Theoretical Analyses of Thompson Sampling. The regret analysis of Thompson Sampling has been the subject of extensive research. Agrawal & Goyal (2012) provided the first finite-time instance-dependent regret bound for Thompson Sampling with $[0, 1]$ bounded rewards, albeit with a constant slightly larger than the one in (1.2). Kaufmann et al. (2012) improved upon this result and established the asymptotic optimality of Thompson Sampling. In particular, they derived the following regret bound for Thompson Sampling with Beta posterior (TS-Beta): for any constant $\delta > 0$,

$$R_\mu(T) \le (1+\delta)\sum_{i>1}\frac{\Delta_i(\log T + \log\log T)}{\mathrm{kl}(\mu_i, \mu_1)} + C(\delta, \mathcal{P}), \tag{1.3}$$

where $C(\delta, \mathcal{P})$ is a quantity dependent on δ and the bandit instance $\mathcal{P} = \{\mu_1, \cdots, \mu_n\}$. While this regret bound is asymptotically optimal due to matching the coefficient in (1.2) for the first term, the second term $C(\delta, \mathcal{P})$ can be large for worst-case instances, leading to suboptimal worst-case regret bounds. For one-parameter exponential family reward distributions, Korda et al. (2013) demonstrated the asymptotic optimality of Thompson Sampling with Jeffery's prior. However, their regret analysis, similar to (1.3), does not provide a closed-form solution for the term related to $C(\delta, \mathcal{P})$, thus failing to offer a worst-case guarantee on the regret.

Agrawal & Goyal (2017) established a worst-case regret bound of $O(\sqrt{KT\log T})$ for TS-Beta with Bernoulli rewards, making it the first near-minimax optimal regret bound for Thompson Sampling. They also provided a lower bound showing that Thompson Sampling with Gaussian posterior (TS-Gaussian) has a worst-case regret of at least $\Omega(\sqrt{KT\log K})$ for Bernoulli rewards. More recently, Jin et al. (2022) extended these results and proved worst-case regret bounds of $O(\sqrt{KT\log K})$ for TS-Beta with Bernoulli rewards and TS-Gaussian with Gaussian rewards. They also demonstrated the asymptotic optimality of TS-Beta and TS-Gaussian while achieving the aforementioned worst-case regret bounds simultaneously. However, none of these analyses establishes the exact minimax and asymptotic optimality of Thompson Sampling.

As observed by Agrawal & Goyal (2012); Kaufmann et al. (2012), the main hardness in the regret analysis of Thompson sampling is caused by the underestimation of the optimal arm. In particular, the underestimation error of the optimal arm can be measured by the following term

$$\sum_{s\ge 1}\mathbb{E}\bigg[\frac{1}{G_{1s}(\delta)} - 1\bigg], \tag{1.4}$$

where $G_{1s}(\delta) = \mathbb{P}(\mu(\theta_1(t)) \ge \mu_1 - \delta)$ is the probability that the estimation of μ_1, i.e., $\mu(\theta_1(t))$, is larger than $\mu_1 - \delta$. This term represents the expected number of pulls of suboptimal arms between two consecutive pulls of the optimal arm 1, effectively indicating the time distance between these pulls. Apparently, if the underestimation error in (1.4) is smaller, then the algorithm pulls the optimal arm more frequently, which leads to a smaller regret. Thus, finding a tight upper bound for (1.4) is crucial for deriving a more precise worst-case regret bound for Thompson Sampling.

In previous works (Agrawal & Goyal, 2017; Jin et al., 2021b; 2022), the focus has been on finding a lower bound for $G_{1s}(\delta)$, typically derived from the anti-concentration bound of the posterior distribution. For instance, Jin et al. (2021b) introduced MOTS-J, which uses a Rayleigh posterior distribution and clips the samples using the MOSS index (Audibert & Bubeck, 2009). MOTS-J achieves minimax and asymptotic optimality for Gaussian rewards. However, this clipping method is difficult to extend to more general reward distributions, as it relies on a tight anti-concentration bound that may not be available in the general case. More recently, Jin et al. (2022) proposed ExpTS$^+$, which employs an artificially designed sampling distribution instead of the posterior distribution used in TS. ExpTS$^+$ achieves minimax and asymptotic optimality for a broad class of one-parameter exponential family reward distributions. However, ExpTS$^+$ loses the posterior update structure of TS and requires solving the inverse of the KL divergence between two reward distributions, which can be infeasible. In practice, solving this inverse KL divergence problem, even approximately, can lead to a significant computational burden[1], as demonstrated in our experiments.

Our Approach. In this paper, we propose ϵ-Exploring

[1]Assume we use Newton's method to solve the KL equation. The number of iterations needed to find a solution with a precision of τ will be $O(\sqrt{\log(1/\tau)})$. Note that at iteration k, the Newton method updates with $x_{k+1} = x_k - f(x_k)/f'(x_k)$. Therefore, the total time complexity of finding the solution will be $c\sqrt{\log(1/\tau)}$, where c is the cost of calculating $f(x)/f'(x)$ with precision τ.

Thompson Sampling (ϵ-TS), a modified version of the Thompson Sampling algorithm. With ϵ-TS , exploration using samples from TS is performed only with probability ϵ, while the arm to play is greedily selected based on empirical mean rewards with probability $1-\epsilon$. Here, $\epsilon \in (0,1)$ is a user-defined parameter.

Our contributions can be summarized as follows:

- We introduce new proof techniques to address the underestimation of the optimal arm, which has been identified as a key challenge in regret analysis of TS (Agrawal & Goyal, 2012; Kaufmann et al., 2012). Our proof framework delves into the relationship between the reward distribution and the posterior distribution, without requiring the anti-concentration bound of the posterior distribution. We demonstrate the applicability of our approach to various popular reward distributions, including Gaussian, Bernoulli, Poisson, and Gamma.

- When $\epsilon = 1$, ϵ-TS recovers the original Thompson Sampling algorithm. We prove that TS with corresponding priors (TS-Beta, TS-Gaussian, TS-Poisson, and TS-Gamma) achieve minimax optimality up to a factor of $\sqrt{\log K}$ for Bernoulli, Gaussian, Poisson, and Gamma rewards. Our analysis also establishes their asymptotic optimality. Notably, this represents the first worst-case regret bound of posterior-based TS for Poisson and Gamma reward distributions.

- When $\epsilon = 1/K$, we demonstrate that ϵ-TS achieves both minimax and asymptotic optimality for Bernoulli, Gaussian, Poisson, and Gamma rewards. This is in contrast to existing work (Agrawal & Goyal, 2017; Jin et al., 2021b), which only achieves near-optimal regret bounds under different parameter settings.

- Importantly, ϵ-TS also outperforms state-of-the-art methods, including KL-UCB (Garivier & Cappé, 2011), KL-UCB^{++} (Ménard & Garivier, 2017), TS (Thompson, 1933), MOTS (Jin et al., 2021b), ExpTS, and ExpTS^{+} (Jin et al., 2022), in terms of empirical performance. By maintaining the posterior update structure of TS and offering a simple implementation similar to ϵ-Greedy, ϵ-TS achieves superior performance while running significantly faster than TS and other baseline methods due to reduced exploration.

Notations. We denote $[n] = \{1, 2, \cdots, n\}$. For any arm $i \in [K]$, $T_i(t) := \sum_{s=1}^{t} \mathbb{1}\{A_s = i\}$ is its number of pulls at time t, $\hat{\mu}_i(t) := \sum_{s=1}^{t} \mathbb{1}\{A_t = i\} \cdot r_t / T_i(t)$ is the sample mean reward at time t, and $\hat{\mu}_{is}$ is its sample mean reward after its s-th pull. $\mathrm{kl}(\mu, \mu')$ denotes the KL divergence between two distributions with mean μ and μ' respectively.

1.1. Additional Related Work

In addition to the algorithms discussed above, we would like to mention some other relevant work related to our paper.

Existing approaches for regret minimization primarily focus on achieving minimax and asymptotic optimal regret. For minimax optimality, MOSS (Audibert & Bubeck, 2009) was the first algorithm proven to be minimax optimal. Degenne & Perchet (2016) introduced the anytime version of MOSS. In terms of asymptotic optimality, KL-UCB (Garivier & Cappé, 2011; Maillard et al., 2011) was shown to be asymptotically optimal. Subsequently, Thompson Sampling (Kaufmann et al., 2012; Korda et al., 2013), Bayes-UCB (Kaufmann, 2016), Double Explore-then-Commit (Jin et al., 2021c), and Maillard sampling (Bian & Jun, 2022) were also proven to be asymptotically optimal. KL-UCB^{++} was the first algorithm demonstrated to be simultaneously minimax and asymptotically optimal. Additionally, Lattimore (2018) introduced Ada-UCB, which considers a strong notion of regret called "instance optimality" and is shown to be near optimal for any bandit instance. For a comprehensive overview of the literature and techniques related to bandits, we recommend referring to Lattimore & Szepesvári (2020); Bubeck et al. (2012).

In the context of Thompson Sampling, apart from the frequentist regret bounds discussed in previous paragraphs, Russo & Van Roy (2014) analyzed the Bayesian regret of Thompson Sampling and proved that its Bayesian regret is no greater than the worst-case regret of any UCB algorithm. Furthermore, Bubeck & Liu (2013) improved the Bayesian regret to $O(\sqrt{KT})$ by incorporating the MOSS idea.

2. ϵ-Exploring Thompson Sampling

2.1. The Proposed Algorithm

Algorithm 1 shows the pseudo code for the proposed algorithm, ϵ-Exploring Thompson Sampling (denoted as ϵ-TS). It shares the same updating rule as the conventional TS algorithm (Agrawal & Goyal, 2017), where an estimated reward $a_i(t)$ for arm i is constructed based on the posterior distribution at each time step t, and the arm with the highest estimated reward is selected. However, ϵ-TS introduces a modification to the estimation process.

In TS, the estimated reward $a_i(t)$ is directly set as the sample drawn from the posterior distribution of arm i. In contrast, ϵ-TS sets $a_i(t)$ as $\hat{\mu}_i(t)$ with probability $1-\epsilon$ and as the sample from the posterior distribution with probability ϵ. Here, ϵ is a parameter that controls the exploration rate. By performing exploration with a probability of ϵ, ϵ-TS explores less compared to TS, making it more computationally efficient, especially when sampling from the posterior distribution is computationally expensive.

Table 1. Reward distributions and the corresponding choices of prior and posterior distributions. The posterior distribution after observing rewards $x_1, \cdots, x_s$ is a function of the number of observations s and the sample mean $\hat{\mu}_s = \sum_{k=1}^{s} x_k/s$, denoted as $f_{\text{post}}(s, \hat{\mu}_s)$.

REWARD	DISTRIBUTION	PARAMETER	$\mu(\theta)$	PRIOR	POSTERIOR ($f_{\text{POST}}(s, \hat{\mu}_s)$)
BERNOULLI	$\mu^x(1-\mu)^{1-x}\delta_{0,1}$	μ	θ	BETA(1,1)	BETA$(1+s\hat{\mu}_s, 1+s-s\hat{\mu}_s)$
GAUSSIAN	$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}$	μ	θ	$\propto 1$	$\mathcal{N}(\hat{\mu}_s, \sigma^2/s)$
POISSON	$\frac{\mu^k e^{-\mu}}{k!}(k \in \mathbb{N})$	μ	θ	$\propto 1$	GAMMA$\big(1+s\hat{\mu}_s, s\big)$
GAMMA	$\frac{x^{\alpha-1}e^{-\beta x}\beta^\alpha}{\Gamma(\alpha)} \mathbb{1}\{x>0\}$	β	α/θ	$\propto 1/\beta^2$	GAMMA$(\alpha s - 1, s\hat{\mu}_s)$, $s > 1/\alpha$.

This modification in ϵ-TS introduces a balance between exploration and exploitation. With a small exploration probability, ϵ-TS focuses on exploiting the arm with the highest empirical mean most of the time, leading to reduced computational overhead while still allowing for occasional exploration to mitigate the underestimation of the optimal arm. This design advantage of ϵ-TS enables it to achieve a balance between computational efficiency and regret performance.

In the following sections, we will demonstrate that ϵ-TS achieves both minimax and asymptotic optimality, providing a tighter worst-case regret bound compared to traditional Thompson Sampling.

Algorithm 1 ϵ-Exploring Thompson Sampling

1: Initialize the prior distributions based on Table 1.
2: For all $i \in [K]$, $\hat{\mu}_i(1) = 0$ and $T_i(1) = 0$.
3: **for** $t = 1, 2, \cdots, T$ **do**
4: For all $i \in [K]$, update the posterior according to Table 1: $P^i_{\text{Posterior}}(t) = f_{\text{post}}(T_i(t), \hat{\mu}_i(t))$, and obtain

$$a_i(t) = \begin{cases} \theta_i(t) \sim P^i_{\text{Posterior}}(t) & \text{with prob. } \epsilon \\ \hat{\mu}_i(t) & \text{with prob. } 1-\epsilon \end{cases}$$

5: Pull the arm $A_t = \arg\max_{i\in[K]} a_i(t)$, and observe the corresponding reward r_t;
6: For all $i \in [K]$, $T_i(t+1) = T_i(t) + \mathbb{1}\{i = A_t\}$, $\hat{\mu}_i(t+1) = \frac{T_i(t)\hat{\mu}_i(t) + r_t\,\mathbb{1}\{i=A_t\}}{T_i(t+1)}$.
7: **end for**

2.2. Minimax and Asymptotic Optimality

Recall the definition of regret in Equation (1.1). Algorithm 1 exhibits the following result.

Theorem 2.1. *For Gaussian, Bernoulli, Poisson, and Gamma reward distributions, and $\epsilon \in [1/K, 1]$, there exists a universal constant $C > 0$ such that the regret of ϵ-TS is bounded as follows:*

$$R_\mu(T) \le C\big(\sqrt{VKT\log(eK\epsilon)}\big) + 2\sum_{i>1}\Delta_i,$$

where $V = \sigma^2$ for Gaussian, $V = 1/4$ for Bernoulli, $V = \mu_1$ for Poisson, and $V = \mu_1^2$ for Gamma. Moreover,

$$\lim_{T\to\infty} \frac{R_\mu(T)}{\log T} = \sum_{i>1} \frac{\Delta_i}{kl(\mu_i, \mu_1)}.$$

Remark 2.2 (Regret of ϵ-TS with $\epsilon = 1$). When ϵ is set to 1, ϵ-TS reduces to the original Thompson Sampling (TS) algorithm (Thompson, 1933; Korda et al., 2013; Agrawal & Goyal, 2017). For bandit problems with Gaussian, Bernoulli, Poisson, and Gamma reward distributions, Theorem 2.1 implies that TS with the corresponding prior is minimax optimal up to a factor of $\sqrt{\log K}$ and is asymptotically optimal under the same algorithm setting. Notably, this is the first worst-case regret bound for the original Thompson Sampling algorithm for Poisson and Gamma rewards.

Remark 2.3 (Regret of ϵ-TS with $\epsilon = 1/K$). When ϵ is set to $1/K$ in Algorithm 1, the results from Theorem 2.1 imply that $1/K$-TS is simultaneously minimax and asymptotically optimal for a large class of reward distributions, including Gaussian, Bernoulli, Poisson, and Gamma.

It is worth noting that Jin et al. (2022) recently proposed the ExpTS$^+$ algorithm, which samples from an artificially designed distribution $\mathcal{P}(\hat{\mu}_i(t), T_i(t))$ (see Equation (4.1)) for each arm at time step t. This sampling is achieved by solving the inverse of $\text{kl}(\hat{\mu}_i(t), x)$ for the variable x, where $\text{kl}(\hat{\mu}_i(t), x)$ represents the KL divergence between two distributions with means $\hat{\mu}_i(t)$ and x, respectively. ExpTS$^+$ makes decisions based on samples from $\mathcal{P}(\hat{\mu}_i(t), T_i(t))$ with probability $1/K$ and based on the sample average reward $\hat{\mu}_i(t)$ with probability $1 - 1/K$. This reduced exploration also leads to the minimax and asymptotic optimality of ExpTS$^+$. However, sampling from $\mathcal{P}$ becomes infeasible when the KL divergence is not invertible, such as for Bernoulli, Poisson, and Gamma distributions. As we will demonstrate in our experiments, approximately solving this inversion can be computationally expensive

For Bernoulli and Gaussian rewards, one may replace the sampling distribution $\mathcal{P}$ in ExpTS$^+$ (Jin et al., 2022) with the posterior presented in Table 1 and thus obtain the ϵ-

TS algorithm proposed in this paper. However, as pointed out by Jin et al. (2022), their regret analysis highly relies on the anti-concentration inequality of $\mathcal{P}$ and changing $\mathcal{P}$ with the posterior distributions to Table 1 breaks such property and thus one cannot derive the same minimax and asymptotic regret bound for ϵ-TS as they did for ExpTS$^+$.

3. Proof Sketch for the Regret Bound of ϵ-TS

In this section, we provide the roadmap for proving the regret bound presented in Theorem 2.1.

Let $S_j = \{i \in [K] \mid 2^{-(j+1)} \leq \Delta_i < 2^{-j}\}$ be the set of arms whose gaps from the optimal arm lie in the interval $[2^{-(j+1)}, 2^{-j})$. In what follows, we aim to prove that the regret of any arm $i \in S_j$ is smaller than $C|S_j|\log(T\epsilon\Delta_i^2)/\Delta_i$, where C is some constant. Let $\delta_0 = \max\{\log(K\epsilon), 1\}$. The idea of dividing the arms into groups based on exponential gaps is inspired by Jin et al. (2022). Define $\gamma = 1/2\log_2(T/(VK\delta_0)) - 3$. For any arm i with $\Delta_i > 4\sqrt{VK\delta_0/T} = 2^{-(\gamma+1)}$, there exists an index $j \leq \gamma$ such that $i \in S_j$. We decomposed the regret as:

$$
\begin{aligned}
R_\mu(T) &= \sum_{i:\Delta_i>0} \Delta_i \cdot \mathbb{E}[T_i(T)] \\
&\leq \sum_{i:\Delta_i>4\sqrt{VK\delta_0/T}} \Delta_i \cdot \mathbb{E}[T_i(T)] + \max_{i:\Delta_i<4\sqrt{VK\delta_0/T}} \Delta_i \cdot T \qquad (3.1) \\
&< \sum_{j\leq\gamma}\sum_{i\in S_j} 2^{-j} \cdot \mathbb{E}[T_i(T)] + 4\sqrt{VK\delta_0 T}. \qquad (3.2)
\end{aligned}
$$

Let $E_{i,\delta}(t) = \{a_i(t) \leq \mu_1 - \delta\}$ be the event that the estimation of the mean reward of arm i at t-th step is lower than $\mu_1 - \delta$. Based on event $E_{i,\delta}(t)$, the expected number of pulls of arms in S_j can be decomposed as follows.

$$
\begin{aligned}
\sum_{i\in S_j} \mathbb{E}[T_i(T)] = \underbrace{\sum_{i\in S_j} \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{A_t = i, E_{i,\delta_j}(t)\}\right]}_{I_1} \\
+ \underbrace{\sum_{i\in S_j} \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{A_t = i, E^c_{i,\delta_j}(t)\}\right]}_{I_2}, \qquad (3.3)
\end{aligned}
$$

where $\delta_j \geq \sqrt{VK\delta_0/T}$ and E^c is the complement of an event E. In what follows, we bound I_1 and I_2 respectively.

3.1. Bounding term I_1

Let η_s be defined in the following way.

$$
\begin{aligned}
\eta_s = \operatorname*{argmax}_{x>0}\{x : \text{such that} \\
\mathrm{kl}(\mu_1 - \delta_j - x, \mu_1 - \delta_j) \leq 2\log_+(T\epsilon/s)/s\}, \qquad (3.4)
\end{aligned}
$$

where $\log_+(x) = \max\{0, \log x\}$. Let $a_{iT_i(t)} = a_i(t)$ and $G_{is}(\delta) = \mathbb{P}(a_{is} \geq \mu_1 - \delta)$. We use the following lemma to bound term I_1, which is proved by Jin et al. (2022).

Lemma 3.1 (Lemma E.1 in Jin et al. (2022))**.** *For any $\delta_j \geq \sqrt{VK\delta_0/T}$, it holds that*

$$
\begin{aligned}
&\sum_{i\in S_j} \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{A_t = i, E_{i,\delta_j}(t)\}\right] \\
&\leq \underbrace{\sum_{s=1}^{T} \mathbb{E}\left[\left(\frac{1}{G_{1s}(\delta_j)} - 1\right) \cdot \mathbb{1}\{\hat{\mu}_{1s} \in L_s\}\right]}_{J} \\
&\quad + T \cdot \mathbb{P}(\exists s \in [T] : \hat{\mu}_{1s} \notin L_s), \qquad (3.5)
\end{aligned}
$$

where $L_s = \{x \mid x \geq \mu_1 - \delta_j - \eta_s\}$.

We present two useful lemmas for further bounding the results in (3.5). The first lemma characterizes the concentration properties of the optimal arm, which upper bounds the second term on the right hand side of (3.5).

Lemma 3.2. *Let L_s and δ_j be the same as in Lemma 3.1. Then we have $T \cdot \mathbb{P}(\exists s \in [T] : \hat{\mu}_{1s} \notin L_s) \leq 30V/(\epsilon\delta_j^2)$.*

If a tight anti-concentration bounds of the posterior distribution is available, following the idea in Agrawal & Goyal (2017); Jin et al. (2022), we can obtain a tight upper bound of term J, the first term on the right hand side (R.H.S.) of (3.5). However, for many posterior distributions other than Gaussian and Beta, deriving a tight anti-concentration bound might be infeasible. In this paper, we bound J by exploring in depth the relationship between the reward distribution and the posterior distribution and do not require the anti-concentration bounds. Specifically, the term J is bounded as in the following lemma, whose proof will be provided in the last part of this section.

Lemma 3.3. *Let δ_j and L_s be the same as in Lemma 3.1. For Gaussian, Bernoulli, Poisson, and Gamma reward distributions, there exists a universal constant C_1, such that*

$$
J \leq C_1\left(\frac{V}{\epsilon\delta_j^2} + \min\left\{\frac{V}{\epsilon\delta_j^2}\log\left(\frac{T\epsilon\delta_j^2}{V}\right), \frac{\sqrt{V^3\log T}}{\epsilon\delta_j^3}\right\}\right).
$$

Substituting the results in Lemmas 3.2 and 3.3 back into (3.5) yields

$$
\begin{aligned}
&\sum_{i\in S_j} \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{A_t = i, E_{i,\delta_j}(t)\}\right] \\
&\leq C_1\left(\frac{V}{\epsilon\delta_j^2} + \min\left\{\frac{V}{\epsilon\delta_j^2}\log\left(\frac{T\epsilon\delta_j^2}{V}\right), \frac{\sqrt{V^3\log T}}{\epsilon\delta_j^3}\right\}\right) + \frac{30V}{\epsilon\delta_j^2}. \qquad (3.6)
\end{aligned}
$$

3.2. Bounding Term I_2.

Let us define $s_0 = \log(T\epsilon\delta_j^2/V)/\mathrm{kl}(\mu_i + \delta_j, \mu_1 - \delta_j)$. The following lemma shows the concentration properties of the suboptimal arms.

Lemma 3.4. *Let $\delta_j > 0$. Assume $s \geq s_0 + 2 + V/\delta_j^2$ and $\hat{\mu}_{is} \leq \mu_i + \delta_j$, then*

$$\mathbb{P}(a_{is} \geq \mu_1 - \delta_j) \leq \frac{V}{T\delta_j^2}.$$

Furthermore,

$$\sum_{s=1}^{T} \mathbb{1}\{G_{is}(\epsilon) > V/(T\delta_j^2)\} \leq 2 + s_0 + \frac{3V}{\delta_j^2}. \tag{3.7}$$

Define $\mathcal{T} = \{t \in [T] : G_{iT_i(t)}(\delta_j) > V/(T\delta_j^2)\}$. Then, based on whether $t \in \mathcal{T}$, we can derive

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{A_t = i, E^c_{i,\delta_j}(t)\}\right] \\
&\leq \mathbb{E}\left[\sum_{t\in\mathcal{T}} \mathbb{1}\{A_t = i\}\right] + \mathbb{E}\left[\sum_{t\notin\mathcal{T}} \mathbb{1}\{E^c_{i,\delta_j}(t)\}\right] \\
&\leq \sum_{s=1}^{T} \mathbb{1}\{G_{is}(\delta_j) > V/(T\delta_j^2)\} + \mathbb{E}\left[\sum_{t\notin\mathcal{T}} \frac{V}{T\delta_j^2}\right] \\
&\leq \sum_{s=1}^{T} \mathbb{1}\{G_{is}(\delta_j) > V/(T\delta_j^2)\} + \frac{V}{\delta_j^2}. \qquad (3.8)
\end{aligned}$$

Applying (3.7) to (3.8), we obtain

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{A_t = i, E^c_{i,\delta_j}(t)\}\right] \\
&\leq \frac{\log(T\epsilon\delta_j^2/V)}{\mathrm{kl}(\mu_i + \delta_j, \mu_1 - \delta_j)} + \frac{4V}{\delta_j^2} + 2 \qquad (3.9) \\
&\leq \frac{2V\log(T\epsilon\delta_j^2/V)}{(\Delta_i - 2\delta_j)^2} + \frac{4V}{\delta_j^2} + 2, \qquad (3.10)
\end{aligned}$$

where the last inequality is from (B.4).

Proof of the Worst-Case Regret Bound. Choose $\delta_j = 2^{-(j+1)}/4 = 2^{-(j-\gamma)}\sqrt{VK\delta_0/T}$. Substituting (3.10) and (3.6) into (3.3) and combining the result with (3.2), we obtain

$$\begin{aligned}
R_\mu(T) &< \sum_{j\leq\gamma}\sum_{i\in S_j} 2^{-j}\cdot\mathbb{E}[T_i(T)] + 4\sqrt{VK\delta_0 T} \\
&\leq \sum_{j\leq\gamma}(2C_1+10)\left(\frac{VK\log(T\epsilon\delta_j^2/V)}{\delta_j} + \frac{KV}{\delta_j}\right) \\
&\quad + 4\sqrt{VK\delta_0 T} + 2\sum_{i>1}\Delta_i \\
&\leq (2C_1+20)\left(\sqrt{VKT}\frac{\log(K\epsilon\delta_0)}{\sqrt{\delta_0}}\sum_{j:j\geq\gamma}\frac{1}{2^{j-\gamma}}\right. \\
&\quad \left. + \sqrt{VKT}\sum_{j:j\geq\gamma}\frac{1}{2^{j-\gamma}}\right) + 4\sqrt{VK\delta_0 T} + 2\sum_{i>1}\Delta_i \\
&\leq C\sqrt{VKT\log(eK\epsilon)} + 2\sum_{i>1}\Delta_i,
\end{aligned}$$

where C is a universal constant and the third inequality is because $x\log(ax^2)$ is monotonically decreasing for $x \geq e/\sqrt{a}$ and $\delta_j \geq \sqrt{VK\delta_0/T}$.

Proof of Asymptotic Optimality Substituting (3.9) and (3.6) into (3.3) and then combining it with (3.1), we have that there exists a universal constant C_0 such that

$$\begin{aligned}
R_\mu(T) &\leq \sum_{i>1}\frac{\Delta_i\log(T\epsilon\delta_j^2/V)}{\mathrm{kl}(\mu_i+\delta_j,\mu_1-\delta_j)} + \max_{i:\Delta_i<4\sqrt{\frac{VK\delta_0}{T}}}\Delta_i\cdot T \\
&\quad + C_0\Delta_i\left(\frac{\sqrt{V^3\log T}}{\epsilon\delta_j^3} + \frac{V}{\epsilon\delta_j^2} + 1\right).
\end{aligned}$$

Let $\delta_j = 1/\log\log T$. We obtain

$$\lim_{T\to\infty}\frac{R_\mu(T)}{\log T} \leq \sum_{i>1}\frac{\Delta_i}{\mathrm{kl}(\mu_i,\mu_1)}.$$

3.3. Proof of Lemma 3.3 for Poisson Rewards

As we discussed in the previous subsection, Lemma 3.3 is the key novelty in our analysis such that we do not require a tight anti-concentration bound of the posterior distribution as Agrawal & Goyal (2017); Jin et al. (2022) do. We provide its proof in this subsection. Due to the space limit, we only provide the proof of Lemma 3.3 for Poisson Rewards here. The proofs for other reward distributions are similar and thus deferred to the appendix.

Let k be the sum of s independent random variables sampled from the Poisson distribution with parameter μ, and $f_{s,\mu}(\cdot)$ be the probability mass function of k. Then, we have

$$f_{s,\mu}(k) = (s\mu)^k\cdot e^{-s\mu}/k!.$$

The posterior distribution is Gamma$(1+k, s)$ with PDF

$$p(\mu) = \mu^k s^{k+1}e^{-s\mu}/k! = s f_{s,\mu}(k).$$

Let $F^G_{\alpha,\beta}(\cdot)$ be the CDF of Gamma distribution with parameter α and β. Let θ_{is} be the sample from the posterior distribution after arm i is pulled s times and $G'_{is}(\delta_j) = \mathbb{P}(\mu(\theta_{is}) \geq \mu_1 - \delta_j)$. We have

$$\begin{aligned}
G_{1s}(\delta_j) &= \mathbb{P}(a_{1s} \geq \mu_1 - \delta_j) \\
&\geq \epsilon G'_{1s}(\delta_j) + (1-\epsilon)\,\mathbb{1}\{\hat{\mu}_{1s} \geq \mu_1 - \delta_j\}. \qquad (3.11)
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\mathbb{E}\left[\left(\frac{1}{G_{1s}(\delta_j)} - 1\right)\cdot\mathbb{1}\{\hat{\mu}_{1s}\in L_s\}\right] \\
&\leq \underbrace{\sum_{k=\lceil s(\mu_1-\delta_j-\eta_s)\rceil}^{\lfloor s(\mu_1-\delta_j/2)\rfloor}\frac{f_{s,\mu_1}(k)}{\epsilon(1-F^G_{k+1,s}(\mu_1-\delta_j))}}_{J_1} \\
&\quad + \underbrace{\sum_{k=\lceil s(\mu_1-\delta_j/2)\rceil}^{\infty}\frac{f_{s,\mu_1}(k)}{1-F^G_{k+1,s}(\mu_1-\delta_j)} - f_{s,\mu_1}(k)}_{J_2}.
\end{aligned} \tag{3.12}$$

For term J_1,

$$\frac{f_{s,\mu_1}(k)}{1-F^G_{k+1,s}(\mu_1-\delta_j)} \le \frac{f_{s,\mu_1}(k)}{\int_{\mu_1-\delta_j}^{\infty} s f_{s,\mu}(k)\mathrm{d}\mu} \le \frac{f_{s,\mu_1}(k)}{\int_{\mu_1-\delta_j/2}^{\mu_1-\delta_j/4} s f_{s,\mu_1}(k)\cdot \frac{f_{s,\mu}(k)}{f_{s,\mu_1}(k)}\mathrm{d}\mu}. \tag{3.13}$$

Noting that for Poisson, $\mathrm{kl}(\mu,\mu') = \mu\log\left(\frac{\mu}{\mu'}\right)+\mu'-\mu$, we further obtain

$$e^{s\mathrm{kl}(k/s,\mu_1)-s\mathrm{kl}(k/s,\mu)} = e^{k\log(\mu/\mu_1)+s\mu_1-s\mu} = \frac{f_{s,\mu}(k)}{f_{s,\mu_1}(k)}.$$

Based on (3.13), we obtain

$$\begin{aligned}\text{r.h.s. of (3.13)} &= \frac{f_{s,\mu_1}(k)}{\int_{\mu_1-\delta_j/2}^{\mu_1-\delta_j/4} s f_{s,\mu_1}(k)\cdot e^{s\mathrm{kl}(k/s,\mu_1)-s\mathrm{kl}(k/s,\mu)}\mathrm{d}\mu} \\ &\le \frac{4e^{-s\mathrm{kl}(\mu_1-\delta_j/4,\mu_1)}}{s\delta_j} \\ &\le \frac{4e^{-s\delta_j^2/(32V)}}{s\delta_j},\end{aligned}$$

where the first inequality holds because from (B.5), $\mathrm{kl}(k/s,\mu_1)-\mathrm{kl}(k/s,\mu)\ge \mathrm{kl}(\mu,\mu_1)\ge \mathrm{kl}(\mu_1-\delta_j/4,\mu_1)$ for $k/s\le \mu_1-\delta_j/2$ and $\mu\in(\mu_1-\delta_j/2,\mu_1-\delta_j/4)$, and the last inequality is due to (B.4). For term J_1, we have

$$\begin{aligned}&\sum_{k=\lceil s(\mu_1-\delta_j-\eta_s)\rceil}^{\lfloor s(\mu_1-\delta_j/2)\rfloor} \frac{f_{s,\mu_1}(k)}{1-F_{k+1,s}(\mu_1-\delta_j)} \\ &\le \sum_{k=\lceil s(\mu_1-\delta_j-\eta_s)\rceil}^{\lfloor s(\mu_1-\delta_j/2)\rfloor} \frac{4e^{-s\delta_j^2/(32V)}}{s\delta_j} \\ &\le e^{-s\delta_j^2/(32V)}(2+4\eta_s/\delta_j).\end{aligned} \tag{3.14}$$

For term J_2, note that the median denoted as m of Gamma$(1+k,s)$ satisfies $m\ge (1+k)/s-1/(3s)\ge \mu_1-\delta_j$ for $k\ge \mu_1-\delta_j/2$ (Chen & Rubin, 1986). Hence,

$$\sum_{k:k\ge\lceil s(\mu_1-\delta_j/2)\rceil} \frac{f_{s,\mu_1}(k)}{1-F^G_{1+k,s}(\mu_1-\delta_j)} - f_{s,\mu_1}(k) \le 1. \tag{3.15}$$

Besides, let $F^{\mathrm{Poi}}_{\mu}(\cdot)$ be the CDF of the Poisson with parameter μ. Note that $F^G_{1+k,s}(x) = 1-F^{\mathrm{Poi}}_{sx}(k)$. From Lemma B.1, for $k\ge s(\mu_1-\delta_j/2)$,

$$\begin{aligned}1-F^G_{k+1,s}(\mu_1-\delta_j) &= F^{\mathrm{Poi}}_{s(\mu_1-\epsilon)}(k) \\ &\ge 1-e^{-s\mathrm{kl}(k/s,\mu_1-\delta_j)} \\ &\ge 1-e^{-s\delta_j^2/(8V)},\end{aligned}$$

where the second inequality is due to Lemma B.1 and the fact that $F^{\mathrm{Poi}}_{s(\mu_1-\delta_j)}(k)$ is probability that the sum of s independent random variables from Poisson with parameter $\mu_1-\delta_j$ is lower than k. Note that for any $c>0$,we have

$$\sum_{s=1} e^{-s\delta_j^2/(cV)} \le \frac{1}{e^{\delta_j^2/(cV)}-1} \le \frac{cV}{\delta_j^2}. \tag{3.16}$$

Here the last inequality is due to $e^x\ge x+1$ for any $x>0$. Therefore, for $s\ge 8V/\delta_j^2$,

$$\begin{aligned}&\sum_{k:k\ge\lceil s(\mu_1-\delta_j/2)\rceil} \frac{f_{s,\mu_1}(k)}{1-F^G_{1+k,s}(\mu_1-\delta_j)} - f_{s,\mu_1}(k) \\ &\le \frac{1}{1-e^{-s\delta_j^2/(8V)}}-1 \\ &\le 2e^{-s\delta_j^2/(8V)},\end{aligned} \tag{3.17}$$

where the last inequality is from (3.16). By substituting (3.14), (3.15), and (3.17) to (3.12), we obtain that there exists a universal constant C_1 such that

$$\begin{aligned}&\sum_{s=1}^{T}\mathbb{E}\left[\left(\frac{1}{G'_{1s}(\delta_j)}-1\right)\cdot \mathbb{1}\{\hat{\mu}_{1s}\in L_s\}\right] \\ &\le (1/\epsilon)\sum_{s=1}^{T} e^{-s\delta_j^2/(32V)}\left(2+\frac{4\eta_s}{\delta_j}\right) + \sum_{s\le 8V/\delta_j^2} 1 \\ &\quad + 2\sum_{s=\lceil 8V/\delta_j^2\rceil}^{T} e^{-s\delta_j^2/(8V)} \\ &\le C_1\left(\frac{V}{\epsilon\delta_j^2}+\min\left(\frac{V}{\epsilon\delta_j^2}\log\left(\frac{T\epsilon\delta_j^2}{V}\right), \frac{\sqrt{V^3\log T}}{\epsilon\delta_j^3}\right)\right),\end{aligned} \tag{3.18}$$

where (3.18) is due to (3.16) and Lemma 3.5 presented in the following.

Lemma 3.5. *Let η_s be the same as defined in (3.4). Then, for any $c\ge 1$,*

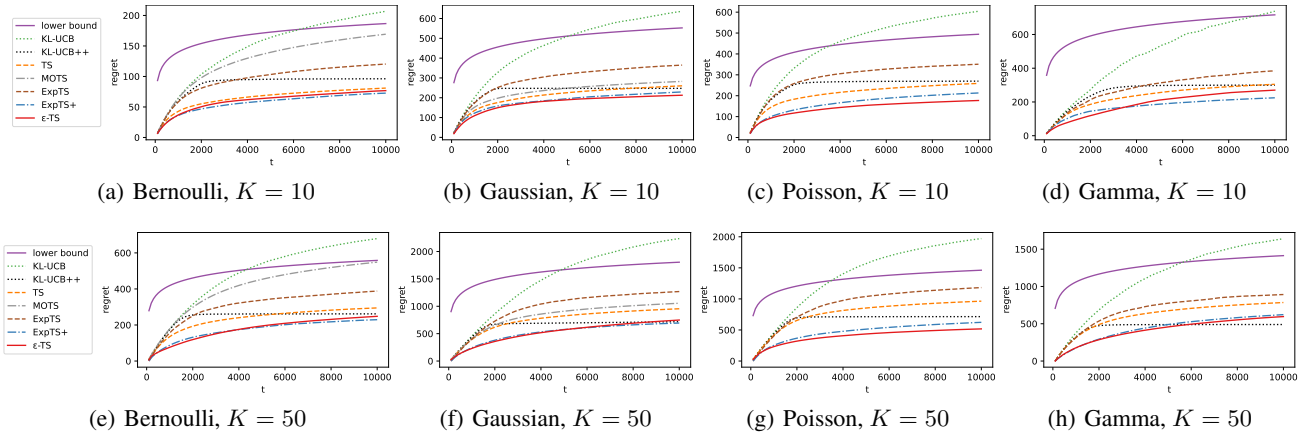
$$\begin{aligned}&\sum_{s=1}^{T} e^{-s\delta_j^2/(cV)}\frac{\eta_s}{\delta_j} \\ &\le \min\left\{\frac{\sqrt{4c}V}{\delta_j^2}\left(\log\left(\frac{T\epsilon\delta_j^2}{V}\right)+5\right), \frac{\sqrt{4cV^3}}{\delta_j^3}\sqrt{\log T}\right\}.\end{aligned}$$

4. Experiments

In this section, we conduct experiments to show that the proposed algorithm ϵ-TS achieves comparable or better performance than state-of-the-art MAB algorithms. All experiments were conducted on a Linux machine equipped with 72 threads, powered by two 18-core Intel Xeon(R) Gold 6240 CPUs @ 2.60GHz and 376GB RAM. We implemented all methods in Python.

4.1. Performance of ϵ-TS on Various Reward Distributions

Baselines In particular, we compare our algorithm with KL-UCB (Garivier & Cappé, 2011), KL-UCB^{++} (Ménard & Garivier, 2017), TS (Agrawal & Goyal, 2017), MOTS (Jin et al., 2021b), ExpTS and ExpTS^{+} (Jin et al., 2022), which all achieve optimal or nearly optimal worst-case regret in multi-armed bandits. All the above strategies select arms at



(a) Bernoulli, $K = 10$ (b) Gaussian, $K = 10$ (c) Poisson, $K = 10$ (d) Gamma, $K = 10$

(e) Bernoulli, $K = 50$ (f) Gaussian, $K = 50$ (g) Poisson, $K = 50$ (h) Gamma, $K = 50$

Figure 1. Comparison of different algorithms with $K = 10$ and $K = 50$ arms under different reward distributions.

round t according to $A_t = \arg\max_{i\in[K]} a_i(t)$, where the reward estimation $a_i(t)$ is defined as follows for different methods respectively.

- KL-UCB (Garivier & Cappé, 2011):

$$a_i(t) = \sup\{\mu : T_i(t)\mathrm{kl}(\hat{\mu}_i(t), \mu) \le \log t + 3\log\log t\}.$$

- KL-UCB^{++} (Ménard & Garivier, 2017):

$$a_i(t) = \sup\Bigg\{\mu : T_i(t)\mathrm{kl}(\hat{\mu}_i(t), \mu) \le \log_+\left(\frac{T}{KT_i(t)}\left(\log_+^2\left(\frac{T}{KT_i(t)}\right) + 1\right)\right)\Bigg\}.$$

- TS (Agrawal & Goyal, 2017): $a_i(t) = \mu(\theta_i(t))$ and $\theta_i(t) \sim P_i(t)$, where $\mu(\cdot)$ is a function that maps the posterior sample to the mean estimation and $P_i(t)$ is the posterior distribution. Both could be found in Table 1. In particular, $\mu(\theta_i(t)) = \theta_i(t)$ for Gaussian rewards.
- MOTS (Jin et al., 2021b):

$$a_i(t) = \min\left\{\mu, \hat{\mu}_i(t) + \sqrt{\frac{4\sigma^2}{T_i(t)}\log_+\left(\frac{T}{KT_i(t)}\right)}\right\},$$

  where $\mu \sim \mathcal{N}(\hat{\mu}_i(t), \sigma^2/(\rho T_i(t)))$.
- ExpTS (Jin et al., 2022): $a_i(t) \sim P(\hat{\mu}_i(t), T_i(t))$, where $P(\mu, n)$ is a distribution with its CDF defined as

$$F(x) = \begin{cases} 1 - 1/2e^{-(n-1)\cdot\mathrm{kl}(\mu,x)} & x \ge \mu, \\ 1/2e^{-(n-1)\cdot\mathrm{kl}(\mu,x)} & x \le \mu. \end{cases} \tag{4.1}$$

- ExpTS$^+$ (Jin et al., 2022): $a_i(t) \sim P(\hat{\mu}_i(t), T_i(t))$ with probability $1/K$ and $a_i(t) = \hat{\mu}_i(t)$ with probability $1 - 1/K$, where $P(\mu, n)$ is the same as in ExpTS.

Implementation To evaluate all the methods, we generate datasets under 4 reward distributions presented in Table 1 and 2 choices of K ($K = 10$ and 50 respectively). The mean rewards are generated as follows.

- Bernoulli: if $K = 10$, we set $\mu_1 = 0.9$ and $\mu_i = 0.8$ for $i \in \{2, 3, \cdots, 10\}$; if $K = 50$, we generate the first 10 arms the same way as just described, and we sample $\mu_i \sim \mathrm{Unif}[0.5, 0.7]$ for $i \in [K] \setminus [10]$.
- Gaussian: similarly, we set $\mu_1 = 1$, $\mu_i = 0.7$ for $i \in \{2, 3, \cdots, 10\}$, and $\mu_i \sim \mathrm{Unif}[0.3, 0.5]$ for $i \in [K] \setminus [10]$.
- Poisson: we set $\mu_1 = 1$, $\mu_i = 0.7$ for $i \in \{2, 3, \cdots, 10\}$, and $\mu_i \sim \mathrm{Unif}[0.3, 0.5]$ for $i \in [K] \setminus [10]$.
- Gamma: we set $\mu_1 = 1$, $\mu_i = 0.8$ for $i \in \{2, 3, \cdots, 10\}$, and $\mu_i \sim \mathrm{Unif}[0.3, 0.5]$ for $i \in [K] \setminus [10]$.

For Gaussian rewards, the variance is set to be 1, and for Gamma rewards, the shape parameter is chosen as $\alpha = 1$.

Since MOTS is only designed for subGaussian, we test it for Gaussian and Bernoulli rewards. Note that KL-UCB needs to solve the inverse of KL divergence for all arms at each time step t. To speed up KL-UCB, we only solve the inequalities involving the inverse of KL divergence at time steps $t = 2, 2^2, 2^3, \cdots$. The KL equations were solved using the scipy.optimize.newton function. We set $\epsilon = 1/K$ for ϵ-TS throughout our experiments. For all algorithms, the experimental results are averaged over 1000 repetitions.

Results Figure 1 presents the cumulative regrets of different algorithms for both the $K = 10$ and $K = 50$ settings. The purple line represents the asymptotic lower bound, $\sum_{i>1} \Delta_i \cdot \log t/\mathrm{kl}(\mu_i, \mu_1)$, up to some constants.

For the $K = 10$ setting, the results are shown in Figures 1(a), 1(b), 1(c), and 1(d) for Bernoulli, Gaussian, Poisson, and Gamma rewards, respectively. It is evident that ϵ-TS and ExpTS$^+$ consistently outperform the other baseline algorithms. The regret of ϵ-TS is comparable to that of ExpTS$^+$ for Bernoulli and Gamma rewards, and performs better for Gaussian and Poisson rewards.

For the $K = 50$ setting, the results are displayed in Figures

Table 2. Average running time of different algorithms on various reward distributions (in seconds).

REWARD DISTRIBUTION	# OF ARMS	TS	ϵ-TS	MOTS	EXPTS	EXPTS$^+$	KL-UCB	KL-UCB^{++}
BERNOULLI	$K = 10$	0.25	**0.11**	0.49	441	43	134	134
	$K = 50$	0.98	**0.27**	1.20	1654	38	127	126
GAUSSIAN	$K = 10$	0.25	**0.11**	0.33	0.27	0.15	0.32	0.36
	$K = 50$	1.05	**0.28**	1.20	1.13	0.35	1.25	1.18
POISSON	$K = 10$	0.23	**0.13**	–	349	36	108	102
	$K = 50$	0.92	**0.30**	–	1266	33	107	92
GAMMA	$K = 10$	0.22	**0.10**	–	394	43	117	134
	$K = 50$	0.96	**0.27**	–	1238	38	112	126

1(e), 1(f), 1(g), and 1(h) for Bernoulli, Gaussian, Poisson, and Gamma rewards, respectively. The regret of ϵ-TS is comparable to that of ExpTS$^+$ for Gaussian and Gamma rewards, and is smaller for Bernoulli and Poisson rewards. Notably, ϵ-TS consistently outperforms TS, MOTS, ExpTS, and KL-UCB across all settings.

4.2. Computational Efficiency of ϵ-TS

We conducted a comparison of the computational efficiency of all the methods by recording the CPU time required to run each algorithm for $T = 10000$ steps. The results are summarized in Table 2. In our experiments, we utilized parallelization for multiple repetitions of the algorithms, taking advantage of the 72 available threads on our server. The reported time of 1654 seconds for ExpTS represents the total running time divided by the number of repeated experiments and then multiplied by 72.

From Table 2, we observe that ϵ-TS is significantly more efficient than all the other baselines. This efficiency stems from the simple structure of our algorithm, where we only need to update the posterior distribution, similar to TS. In contrast, ExpTS, ExpTS$^+$, KL-UCB, and KL-UCB^{++} all require solving the inverse of the KL divergence between two reward distributions, which becomes computationally expensive when a closed-form solution is not available. For example, compared to ExpTS$^+$, our algorithm is more than $390\times$ faster for $K = 10$ and $140\times$ faster for $K = 50$ when the reward distribution is Bernoulli. One exception is for Gaussian rewards, where the KL divergence between two reward distributions has a closed-form inverse.

It is worth noting that ϵ-TS is also significantly faster than TS and MOTS, which do not require solving the inverse KL divergence. This is because in ϵ-TS , we only sample from the posterior distribution with a small probability, and most of the time, the algorithm only needs to calculate the empirical mean reward.

Due to the space limit, we defer more comprehensive experimental results to Appendix C.

5. Conclusion

In this paper, we investigated the performance of Thompson Sampling (TS) for popular reward distributions, including Gaussian, Bernoulli, Poisson, and Gamma. We proposed the ϵ-TS algorithm, which only performs TS exploration with probability ϵ and performs greedily otherwise. When $\epsilon = 1$, ϵ-TS is equivalent to the original TS algorithm. We provided both worst-case regret bounds and asymptotic regret bounds for ϵ-TS. When $\epsilon = 1$, our results demonstrate that TS is minimax optimal up to a factor of $\sqrt{\log K}$ and is asymptotically optimal. Notably, this study provides the first worst-case regret bound for TS with Poisson and Gamma reward distributions. When $\epsilon = 1/K$, our results show that ϵ-TS achieves simultaneous minimax and asymptotic optimality. Our experiments confirm the superior performance of ϵ-TS compared to state-of-the-art methods. Importantly, our proposed algorithm, ϵ-TS , exhibits significantly improved computational efficiency compared to existing multi-armed bandit algorithms, such as ExpTS$^+$ and KL-UCB^{++}, while achieving the same minimax and asymptotic optimality properties.

Acknowledgements

We thank the anonymous reviewers for their helpful comments. This research is supported by the National Research Foundation, Singapore under its AI Singapore Program (AISG Award No: AISG-PhD/2021-01-004[T]), by Singapore Ministry of Education Academic Research Fund Tier 3 under MOE's official grant number MOE2017-T3-1-007, and by the Whitehead Scholars Program at Duke University School of Medicine. In particular, T. Jin is supported by the National Research Foundation, Singapore under its AI Singapore Program (AISG Award No: AISG-PhD/2021-01-004[T]). X. Xiao is supported by Singapore Ministry of Education Academic Research Fund Tier 3 under MOE's official grant number MOE2017-T3-1-007. P. Xu is supported by the Whitehead Scholars Program. The views and conclusions in this paper are those of the authors and should not be interpreted as representing any funding agencies.

References

Agrawal, S. and Goyal, N. Analysis of Thompson sampling for the multi-armed bandit problem. In *Conference on Learning Theory*, pp. 39–1, 2012. (pp. 2 and 3.)

Agrawal, S. and Goyal, N. Near-optimal regret bounds for Thompson sampling. *Journal of the ACM (JACM)*, 64(5): 30, 2017. (pp. 1, 2, 3, 4, 5, 6, 7, and 8.)

Audibert, J.-Y. and Bubeck, S. Minimax policies for adversarial and stochastic bandits. In *COLT*, pp. 217–226, 2009. (pp. 2 and 3.)

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002. (p. 1.)

Bian, J. and Jun, K.-S. Maillard sampling: Boltzmann exploration done optimally. In *International Conference on Artificial Intelligence and Statistics*, pp. 54–72. PMLR, 2022. (p. 3.)

Bubeck, S. and Liu, C.-Y. Prior-free and prior-dependent regret bounds for thompson sampling. In *Advances in Neural Information Processing Systems*, pp. 638–646, 2013. (p. 3.)

Bubeck, S., Cesa-Bianchi, N., et al. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5 (1):1–122, 2012. (p. 3.)

Chapelle, O. and Li, L. An empirical evaluation of Thompson sampling. In *Advances in Neural Information Processing Systems*, pp. 2249–2257, 2011. (p. 2.)

Chen, J. and Rubin, H. Bounds for the difference between median and mean of gamma and poisson distributions. *Statistics & probability letters*, 4(6):281–283, 1986. (pp. 7 and 15.)

Degenne, R. and Perchet, V. Anytime optimal algorithms in stochastic multi-armed bandits. In *International Conference on Machine Learning*, pp. 1587–1595. PMLR, 2016. (p. 3.)

Garivier, A. and Cappé, O. The KL-UCB algorithm for bounded stochastic bandits and beyond. In *Proceedings of the 24th Annual Conference on Learning Theory*, pp. 359–376, 2011. (pp. 1, 3, 7, and 8.)

Garivier, A., Lattimore, T., and Kaufmann, E. On explore-then-commit strategies. In *Advances in Neural Information Processing Systems*, pp. 784–792, 2016. (p. 1.)

Jin, T., Tang, J., Xu, P., Huang, K., Xiao, X., and Gu, Q. Almost optimal anytime algorithm for batched multi-armed bandits. In *International Conference on Machine Learning*, pp. 5065–5073. PMLR, 2021a. (p. 19.)

Jin, T., Xu, P., Shi, J., Xiao, X., and Gu, Q. Mots: Minimax optimal Thompson sampling. In *International Conference on Machine Learning*, pp. 5074–5083. PMLR, 2021b. (pp. 1, 2, 3, 7, and 8.)

Jin, T., Xu, P., Xiao, X., and Gu, Q. Double explore-then-commit: Asymptotic optimality and beyond. In *Conference on Learning Theory*, pp. 2584–2633. PMLR, 2021c. (p. 3.)

Jin, T., Xu, P., Xiao, X., and Anandkumar, A. Finite-time regret of Thompson sampling algorithms for exponential family multi-armed bandits. In *Advances in Neural Information Processing Systems*, 2022. (pp. 1, 2, 3, 4, 5, 6, 7, and 8.)

Kaufmann, E. On bayesian index policies for sequential resource allocation. *arXiv preprint arXiv:1601.01190*, 2016. (pp. 1 and 3.)

Kaufmann, E., Korda, N., and Munos, R. Thompson sampling: An asymptotically optimal finite-time analysis. In *International Conference on Algorithmic Learning Theory*, pp. 199–213. Springer, 2012. (pp. 1, 2, and 3.)

Korda, N., Kaufmann, E., and Munos, R. Thompson sampling for 1-Dimensional exponential family bandits. *Advances in Neural Information Processing Systems*, 26, 2013. (pp. 1, 2, 3, and 4.)

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

Lattimore, T. Refining the confidence level for optimistic bandit strategies. *The Journal of Machine Learning Research*, 19(1):765–796, 2018. (p. 3.)

Lattimore, T. and Szepesvári, C. *Bandit Algorithms*. Cambridge University Press, 2020. (pp. 3, 11, and 20.)

Maillard, O.-A., Munos, R., and Stoltz, G. A finite-time analysis of multi-armed bandits problems with kullback-leibler divergences. In *Proceedings of the 24th Annual Conference on Learning Theory*, pp. 497–514, 2011. (p. 3.)

Ménard, P. and Garivier, A. A minimax and asymptotically optimal algorithm for stochastic bandits. In *International Conference on Algorithmic Learning Theory*, pp. 223–237, 2017. (pp. 1, 3, 7, 8, and 19.)

Russo, D. and Van Roy, B. Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39 (4):1221–1243, 2014. (p. 3.)

Thompson, W. R. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933. (pp. 2, 3, and 4.)

A. Proof of Supporting Lemmas

A.1. Proof of Lemma 3.2

Let $\mathrm{kl}_+(x,y) = \mathrm{kl}(x,y)\,\mathbb{1}(x \le y)$. From (B.5) and (B.4),

$$\mathrm{kl}_+(\hat{\mu}_{1s}, \mu_1 - \delta_j) \le \mathrm{kl}_+(\hat{\mu}_{1s}, \mu_1) - \mathrm{kl}(\mu_1 - \delta_j, \mu_1) \le \mathrm{kl}_+(\hat{\mu}_{1s}, \mu_1) - \frac{\delta_j^2}{2V}. \tag{A.1}$$

Similar to the proof of Lemma 9.3 in Lattimore & Szepesvári (2020), we have

$$\begin{aligned}
&\mathbb{P}\bigg(\exists s \ge 1 : \mathrm{kl}_+(\hat{\mu}_{1s}, \mu_1 - \delta_j) \ge 2\log_+(T\epsilon/s)/s\bigg) \\
&\le \mathbb{P}\bigg(\exists s > 1 : \mathrm{kl}_+(\hat{\mu}_{1s}, \mu_1) - \frac{\delta_j^2}{2V} \ge 2\log_+(T\epsilon/s)/s\bigg) \\
&\le \sum_{n=0}^{\infty} \mathbb{P}\bigg(\exists s \in [2^n, 2^{n+1}] : \mathrm{kl}_+(\hat{\mu}_{1s}, \mu_1) - \frac{\delta_j^2}{2V} \ge 2\log_+(T\epsilon/s)/s\bigg) \\
&\le \sum_{n=0}^{\infty} \mathbb{P}\bigg(\exists s \in [2^n, 2^{n+1}] : \mathrm{kl}_+(\hat{\mu}_{1s}, \mu_1) - \frac{\delta_j^2}{2V} \ge \frac{2\log_+(T\epsilon/2^{n+1})}{2^{n+1}}\bigg) \\
&\le \sum_{n=0}^{\infty} \exp\bigg(-2^n \cdot \frac{2\log_+(T\epsilon/2^{n+1})}{2^{n+1}} - 2^n \cdot \frac{\delta_j^2}{2V}\bigg),
\end{aligned} \tag{A.2}$$

where the first inequality is due to (A.1) and the last inequality is due to Lemma B.1. The rest part is purely algebraic:

$$\begin{aligned}
\text{r.h.s of (A.2)} &\le 1/(T\epsilon) \sum_{n=0}^{\infty} 2^{n+1} \cdot \exp\bigg(-\frac{\delta_j^2}{2V} \cdot 2^{j-2}\bigg) \\
&\le \frac{16V}{eT\epsilon\delta_j^2} + 1/(T\epsilon) \int_0^{\infty} 2^{s+1} \exp\big(-\delta_j^2/(2V) \cdot 2^{s-2}\big) \mathrm{d}s \\
&\le \frac{30V}{T\epsilon\delta_j^2},
\end{aligned}$$

The third inequality is due to the fact that the integrand is unimodal and has a maximum value $\frac{16V}{eT\epsilon\delta_j^2}$. For such function f, we have $\sum_{s=a}^{b} f(s) \le \int_a^b f(s)\mathrm{d}s + \max_{s\in[a,b]} f(s)$. This completes the proof.

A.2. Proof of Lemma 3.3 for Bernoulli, Gaussian, and Gamma Reward Distributions

Bernoulli Rewards: Let $f_{n,\mu}(\cdot)$ ($F_{n,\mu}(\cdot)$) be the PMF (CDF) of binomial distribution with parameter n and μ and $f^{beta}_{\alpha,\beta}(\cdot)$ ($F^{beta}_{\alpha,\beta}(\cdot)$) be the PDF (CDF) of Beta distribution with parameter α and β. For Bernoulli reward distribution with parameter μ, after s number of pulls, we let α be the number of successes (the reward is 1) and β be the number of failures (the reward is 0). The posterior distribution is

$$p(\mu) = f^{beta}_{\alpha+1,\beta+1}(\mu) = (s+1) f_{s,\mu}(\alpha).$$

From (3.11),

$$\begin{aligned}
&\mathbb{E}\bigg[\bigg(\frac{1}{G_{1s}(\delta_j)} - 1\bigg) \cdot \mathbb{1}\{\hat{\mu}_{1s} \in L_s\}\bigg] \\
&\le \underbrace{\sum_{\alpha=\lceil s(\mu_1-\delta_j-\eta_s)\rceil}^{\lfloor s(\mu_1-\delta_j/2)\rfloor} \frac{f_{s,\mu_1}(\alpha)}{\epsilon(1 - F^{beta}_{\alpha+1,s-\alpha+1}(\mu_1 - \delta_j))}}_{J_1} + \underbrace{\sum_{\alpha=\lceil s(\mu_1-\delta_j/2)\rceil}^{s} \frac{f_{s,\mu_1}(\alpha)}{1 - F^{beta}_{\alpha+1,s-\alpha+1}(\mu_1 - \delta_j)} - f_{s,\mu_1}(\alpha)}_{J_2}.
\end{aligned} \tag{A.3}$$

For term J_1, we bound $1 - F^{beta}_{\alpha+1,s-\alpha+1}(\mu_1 - \delta_j)$ as follows.

$$\begin{aligned}
1 - F^{beta}_{\alpha+1,s-\alpha+1}(\mu_1 - \delta_j) &= \int_{\mu_1-\delta_j}^{1} f^{beta}_{\alpha+1,s-\alpha+1}(x)\mathrm{d}x \\
&\geq \int_{\mu_1-\delta_j/2}^{\mu_1-\delta_j/4} f^{beta}_{\alpha+1,s-\alpha+1}(x)\mathrm{d}x \\
&= \int_{\mu_1-\delta_j/2}^{\mu_1-\delta_j/4} \frac{x^\alpha(1-x)^{s-\alpha}}{\mu_1^\alpha(1-\mu_1)^{s-\alpha}} \cdot (s+1) f_{s,\mu_1}(\alpha)\mathrm{d}x. \qquad \text{(A.4)}
\end{aligned}$$

Note that for Bernoulli, $\mathrm{kl}(\mu,\mu') = \mu\log\left(\frac{\mu}{\mu'}\right) + (1-\mu)\log\left(\frac{1-\mu}{1-\mu'}\right)$. Therefore, $e^{s\mathrm{kl}(\alpha/s,\mu_1)} = \left(\frac{\alpha/s}{\mu_1}\right)^\alpha \left(\frac{1-\alpha/s}{1-\mu_1}\right)^{s-\alpha}$ and $e^{s\mathrm{kl}(\alpha/s,x)} = \left(\frac{\alpha/s}{x}\right)^\alpha \left(\frac{1-\alpha/s}{1-x}\right)^{s-\alpha}$. We further obtain that for $\alpha/s < x < \mu_1 - \delta_j/4$,

$$\begin{aligned}
\frac{x^\alpha(1-x)^{s-\alpha}}{\mu_1^\alpha(1-\mu_1)^{s-\alpha}} &= e^{s\mathrm{kl}(\alpha/s,\mu_1) - s\mathrm{kl}(\alpha/s,x)} \\
&\geq e^{s\mathrm{kl}(x,\mu_1)} \\
&\geq e^{s\mathrm{kl}(\mu_1-\delta_j/4,\mu_1)} \\
&\geq e^{s\delta_j^2/(32V)},
\end{aligned}$$

where the first and second inequalities are due to (B.5) and the third inequality is due to (B.4). Based on (A.4), we obtain

$$\begin{aligned}
\text{r.h.s. of (A.4)} &\geq \int_{\mu_1-\delta_j/2}^{\mu_1-\delta_j/4} (s+1) e^{s\delta_j^2/(32V)} f_{s,\mu_1}(\alpha)\mathrm{d}x \\
&= e^{s\delta_j^2/(32V)} \delta_j (s+1) f_{s,\mu_1}(\alpha)/4.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\sum_{\alpha=\lceil s(\mu_1-\delta_j-\eta_s)\rceil}^{\lfloor s(\mu_1-\delta_j/2)\rfloor} \frac{f_{s,\mu_1}(\alpha)}{1 - F^{beta}_{\alpha+1,s-\alpha+1}(\mu_1-\delta_j)} &\leq e^{-s\delta_j^2/(32V)} \sum_{\alpha=\lceil s(\mu_1-\delta_j-\eta_s)\rceil}^{\lfloor s(\mu_1-\delta_j/2)\rfloor} \frac{4 f_{s,\mu_1}(\alpha)}{\delta_j(s+1) f_{s,\mu_1}(\alpha)} \\
&\leq e^{-s\delta_j^2/(32V)} \left(2 + \frac{4\eta_s}{\delta_j}\right). \qquad \text{(A.5)}
\end{aligned}$$

Now, we bound term J_2. A nice property of Beta distribution is $1 - F^{beta}_{\alpha+1,s-\alpha+1}(\mu_1 - \delta_j) = F_{s+1,\mu_1-\delta_j}(\alpha)$ for any $\alpha \in \mathbb{N}^+$. For $s < 2/\delta_j$ and $\alpha \geq \lceil s(\mu_1 - \delta_j/2)\rceil$, we obtain

$$\begin{aligned}
1 - F^{beta}_{j+1,s-j+1}(\mu_1 - \delta_j) &= F_{s+1,\mu_1-\delta_j}(\alpha) \\
&\geq (1 - (\mu_1 - \delta_j)) \cdot F_{s,\mu_1-\delta_j}(\alpha) \\
&\geq (1 - (\mu_1 - \delta_j))/2 \\
&\geq \delta_j/2 \qquad \text{(A.6)}
\end{aligned}$$

where the second inequality is because $\alpha \geq \lceil s(\mu_1 - \delta_j)\rceil$ and the median of the binomial distribution with parameters s and $\mu_1 - \delta_j$ is lower than $\lceil s(\mu_1 - \delta_j)\rceil$.
For $2/\delta_j < s \leq \lfloor 8V/\delta_j^2 \rfloor$ and $\alpha \geq \lceil s(\mu_1 - \delta_j/2)\rceil$, the median of binomial distribution with parameter $s+1$ and $\mu_1 - \delta_j$ is lower than $\lceil s(\mu_1 - \delta_j)\rceil$. We obtain

$$F_{s+1,\mu_1-\delta_j}(\alpha) \geq \frac{1}{2},$$

and thus

$$\sum_{\alpha=\lceil s(\mu_1-\delta_j/2)\rceil}^{s} \frac{f_{s,\mu_1}(\alpha)}{1 - F^{beta}_{\alpha+1,s-\alpha+1}(\mu_1-\delta_j)} - f_{s,\mu_1}(\alpha) \leq 1 \qquad \text{(A.7)}$$

Furthermore, for $s > \lceil 8V/\delta_j^2 \rceil$ and $\alpha \geq \lceil s(\mu_1 - \delta_j/2) \rceil$, we have

$$\begin{aligned} 1 - F^{beta}_{\alpha+1,s-\alpha+1}(\mu_1 - \delta_j) &= F_{s+1,\mu_1-\delta_j}(\alpha) \\ &\geq 1 - e^{-(s+1)\mathrm{kl}(\mu_1-\delta_j/2,\mu_1-\delta_j)} \\ &\geq 1 - e^{-s\delta_j^2/(8V)}, \end{aligned} \tag{A.8}$$

where the first inequality is due to Lemma B.1 and the fact that $F_{s+1,\mu_1-\delta_j}(\alpha)$ is the probability that the sum of $s+1$ Bernoulli random variables with parameter $\mu_1 - \delta_j$ is lower than α, the second inequality is due to (B.5). Therefore, for $s > \lceil 8V/\delta_j^2 \rceil$,

$$\begin{aligned} \sum_{\alpha=\lceil s(\mu_1-\delta_j/2)\rceil}^{s} \frac{f_{s,\mu_1}(\alpha)}{1 - F^{beta}_{\alpha+1,s-\alpha+1}(\mu_1 - \delta_j)} - f_{s,\mu_1}(\alpha) &\leq \frac{1}{1 - e^{-s\delta_j^2/(8V)}} - 1 \\ &\leq 2e^{-s\delta_j^2/(8V)}. \end{aligned} \tag{A.9}$$

Putting everything together, we obtain that there exists a universal constant $C_1 > 0$ such that

$$\begin{aligned} &\sum_{s=1}^{T} \mathbb{E}\left[\left(\frac{1}{G_{1s}(\delta_j)} - 1\right) \cdot \mathbb{1}\{\hat{\mu}_{1s} \in L_s\}\right] \\ \leq &(1/\epsilon) \cdot \sum_{s=1}^{T} e^{-s\delta_j^2/(32V)}\left(2 + \frac{4\eta_s}{\delta_j}\right) + \sum_{s<2/\delta_j} \frac{2}{\delta_j} + \sum_{s\leq 8V/\delta_j^2} 1 + 2 \sum_{s=\lceil 8V/\delta_j^2 \rceil}^{T} e^{-s\delta_j^2/(8V)} \\ \leq &C_1\left(\frac{V}{\epsilon\delta_j^2} + \frac{V}{\epsilon\delta_j^2}\log\left(\frac{T\epsilon\delta_j^2}{V}\right)\right), \end{aligned} \tag{A.10}$$

where the first inequality follows by substituting (A.5), (A.6), (A.7), and (A.9) to (A.3), the second inequality is due to Lemma 3.5 and (3.16)

Gaussian Rewards: The posterior distribution of arm 1 after n-th pull is

$$p(\mu) = \sqrt{\frac{n}{2\sigma^2\pi}} \exp\left(-\frac{n(\mu - \hat{\mu}_{1n})^2}{2\sigma^2}\right).$$

Also, note that the PDF of the random variable $\hat{\mu}_n$ is

$$f_{n,\mu_1}(\mu) = \sqrt{\frac{n}{2\sigma^2\pi}} \exp\left(-\frac{n(\mu - \mu_1)^2}{2\sigma^2}\right).$$

Therefore, $p(\mu) = f_{n,\hat{\mu}_{1n}}(\mu)$. Let $F_{n,\mu_1}(\cdot)$ be the CDF of $\hat{\mu}_n$. From (3.11), we have

$$\begin{aligned} &\mathbb{E}\left[\left(\frac{1}{G_{1s}(\delta_j)} - 1\right) \cdot \mathbb{1}\{\hat{\mu}_{1s} \in L_s\}\right] \\ &\leq \underbrace{\int_{\mu_1-\delta_j-\eta_s}^{\mu_1-\delta_j/2} \frac{f_{s,\mu_1}(x)}{\epsilon(1 - F_{s,x}(\mu_1 - \epsilon))}\mathrm{d}x}_{J_1} + \underbrace{\int_{\mu_1-\delta_j/2}^{\infty} \frac{f_{s,\mu_1}(x)}{1 - F_{s,x}(\mu_1 - \epsilon)} - f_{s,\mu_1}(x)\mathrm{d}x}_{J_2}. \end{aligned} \tag{A.11}$$

For term J_1, we have

$$\begin{aligned} \int_{\mu_1-\delta_j-\eta_s}^{\mu_1-\delta_j/2} \frac{f_{s,\mu_1}(x)}{1 - F_{s,x}(\mu_1 - \epsilon)}\mathrm{d}x &\leq \int_{\mu_1-\delta_j-\eta_s}^{\mu_1-\delta_j/2} \frac{f_{s,\mu_1}(x)}{\int_{\mu_1-\delta_j/2}^{\mu_1-\delta_j/4} f_{s,x}(t)\mathrm{d}t}\mathrm{d}x \\ &= \int_{\mu_1-\delta_j-\eta_s}^{\mu_1-\delta_j/2} \frac{f_{s,\mu_1}(x)}{\int_{\mu_1-\delta_j/2}^{\mu_1-\delta_j/4} f_{s,\mu_1}(x) e^{-\frac{s(t-x)^2}{2\sigma^2} + \frac{s(x-\mu_1)^2}{2V}}\mathrm{d}t}\mathrm{d}x \end{aligned}$$

$$\leq \int_{\mu_1-\delta_j-\eta_s}^{\mu_1-\delta_j/2} \frac{f_{s,\mu_1}(x)}{\int_{\mu_1-\delta_j/2}^{\mu_1-\delta_j/4} f_{s,\mu_1}(x) e^{\frac{s(\mu_1-t)^2}{2V}} \mathrm{d}t} \mathrm{d}x$$
$$= e^{-s\delta_j^2/(32V)} \left(2 + \frac{4\eta_s}{\delta_j}\right). \tag{A.12}$$

For term J_2, we have

$$\int_{\mu_1-\delta_j}^{\infty} \frac{f_{n,\mu_1}(x)}{1 - F_{s,x}(\mu_1 - \epsilon)} - f_{s,\mu_1}(x)\mathrm{d}x \leq \int_{\mu_1-\delta_j}^{\infty} f_{s,\mu_1}(x)\mathrm{d}x$$
$$\leq 1, \tag{A.13}$$

where the first inequality is due to $1 - F_{s,x}(\mu_1 - \delta_j) \geq \frac{1}{2}$ for $x \geq \mu_1 - \delta_j$. Moreover, for $s \geq \lceil 8V/\epsilon^2 \rceil$ and $x \geq \mu_1 - \delta_j/2$,

$$1 - F_{s,x}(\mu_1 - \delta_j) \geq 1 - e^{-s(x-(\mu_1-\delta_j))^2/(2V)}$$
$$\geq 1 - e^{-s\delta_j^2/(8V)},$$

where the first inequality is due to Lemma B.1 and the fact that $1 - F_{s,x}(\mu_1 - \delta_j)$ is the probability that the mean of s independent random variables from $\mathcal{N}(x, \sigma^2)$ is larger than $\mu_1 - \delta_j$. Therefore, for $s > \lceil 8V/\delta_j^2 \rceil$,

$$\int_{\mu_1-\delta_j}^{\infty} \frac{f_{n,\mu_1}(x)}{1 - F_{s,x}(\mu_1 - \epsilon)} - f_{s,\mu_1}(x)\mathrm{d}x \leq \frac{1}{1 - e^{-s\delta_j^2/(8V)}} - 1$$
$$\leq 2e^{-s\delta_j^2/(8V)}, \tag{A.14}$$

where the last inequality is from (3.16). Similar to (A.10), by substituting (A.12), (A.13), and (A.14) to (A.11), we obtain that exists a universal constant C_1 such that

$$\sum_{s=1}^{T} \mathbb{E}\left[\left(\frac{1}{G_{1s}(\delta_j)} - 1\right) \cdot \mathbb{1}\{\hat{\mu}_{1s} \in L_s\}\right]$$
$$\leq (1/\epsilon) \sum_{s=1}^{T} e^{-s\delta_j^2/(32V)} \left(2 + \frac{4\eta_s}{\delta_j}\right) + \sum_{s \leq 8V/\delta_j^2} 1 + 2 \sum_{s=\lceil 8V/\delta_j^2 \rceil}^{T} e^{-s\delta_j^2/(8V)}$$
$$\leq C_1 \left(\frac{V}{\epsilon\delta_j^2} + \frac{V}{\epsilon\delta_j^2} \log\left(\frac{T\epsilon\delta_j^2}{V}\right)\right), \tag{A.15}$$

where the last inequality is due to Lemma 3.5 and (3.16).

Gamma Rewards: Let $f_{n,\alpha,\beta}(\cdot)$ be the PDF of the sum of n gamma distribution with parameter β, α and $f^G_{\alpha,\beta}(\cdot)$ ($F^G_{\alpha,\beta}(\cdot)$) be the PDF (CDF) of Gamma distribution with parameter α and β respectively. Let z_1 be the random variable of the sum of the reward after s-th pull of the arm. Then, $z_1 \sim \text{Gamma}(s\alpha, \beta_1)$

$$f_{s,\alpha,\beta_1}(z_1) = \frac{z_1^{s\alpha-1} e^{-\beta_1 z_1} \beta_1^{s\alpha}}{\Gamma(s\alpha)},$$

where $\beta_1 = \alpha/\mu_1$. For $s \leq 1/\alpha$, we have

$$p(\beta) \propto e^{-\beta z_1} \beta^{s\alpha-2}.$$

Noting that

$$\int_0^{\beta_1} e^{-\beta z_1} \beta^{s\alpha-2} \mathrm{d}\beta \geq \int_0^{\beta_1} e^{-\beta z_1} \beta^{-1} \mathrm{d}\beta \geq e^{-\beta_1 z_1} \cdot \int_0^{\beta_1} \beta^{-1} \mathrm{d}\beta = \infty.$$

and

$$\int_{\beta_1}^{\infty} e^{-\beta z_1} \beta^{s\alpha-2} \mathrm{d}\beta \leq \int_{\beta_1}^{\infty} e^{-\beta z_1} \beta^{-1} \mathrm{d}\beta < \infty.$$

Therefore, $\int_0^{\beta_1} p(\beta)\mathrm{d}\beta = 1$. From (3.11), we obtain that for $s \le 1/\alpha$,

$$\mathbb{E}\left[\left(\frac{1}{G_{1s}(\delta_j)} - 1\right) \cdot \mathbb{1}\{\hat{\mu}_{1s} \in L_s\}\right] = 1/\epsilon. \tag{A.16}$$

For $s > 1/\alpha$, the posterior distribution is the Gamma distribution with parameters $s\alpha - 1$ and z_1, i.e.,

$$p(\beta; s\alpha - 1, z_1) = \frac{\beta^{s\alpha-2} z_1^{s\alpha-1} e^{-\beta z_1}}{\Gamma(s\alpha-1)} = \frac{f_{s,\alpha,\beta}(z_1)\cdot\Gamma(s\alpha)}{\beta^2\cdot\Gamma(s\alpha-1)} = \frac{f_{s,\alpha,\beta}(z_1)\cdot(s\alpha-1)}{\beta^2}.$$

From (3.11), we have

$$\begin{aligned}
&\mathbb{E}\left[\left(\frac{1}{G'_{1s}(\delta_j)} - 1\right) \cdot \mathbb{1}\{\hat{\mu}_{1s} \in L_s\}\right] \\
&\le \underbrace{\int_{s(\mu_1-\delta_j-\eta_s)}^{s(\mu_1-\delta_j)} \frac{f_{s,\alpha,\beta_1}(z_1)}{\epsilon\cdot F^G_{\alpha s-1,z_1}(\alpha/(\mu_1-\delta_j))}\mathrm{d}z_1}_{J_1} + \underbrace{\int_{s(\mu_1-\delta_j)}^{\infty} \frac{f_{s,\alpha,\beta_1}(z_1)}{F^G_{\alpha s-1,z_1}(\alpha/(\mu_1-\delta_j))}\mathrm{d}z_1}_{J_2}.
\end{aligned} \tag{A.17}$$

For term J_1, we have

$$\begin{aligned}
\int_{s(\mu_1-\delta_j-\eta_s)}^{s(\mu_1-\delta_j/2)} \frac{f_{s,\alpha,\beta_1}(z_1)}{F^G_{s\alpha-1,z_1}(\alpha/(\mu_1-\delta_j))}\mathrm{d}z_1 &= \int_{s(\mu_1-\delta_j-\eta_s)}^{s(\mu_1-\delta_j/2)} \frac{f_{s,\alpha,\beta_1}(z_1)}{\int_0^{\frac{\alpha}{\mu_1-\delta_j}} f_{s,\alpha,\beta}(z_1)\cdot(s\alpha-1)/\beta^2\,\mathrm{d}\beta}\mathrm{d}z_1 \\
&\le \int_{s(\mu_1-\delta_j-\eta_s)}^{s(\mu_1-\delta_j/2)} \frac{f_{s,\alpha,\beta_1}(z_1)}{\int_{\frac{\alpha}{\mu_1-\delta_j/2}}^{\frac{\alpha}{\mu_1-\delta_j/4}} f_{s,\alpha,\beta_1}(z_1)\cdot\frac{f_{s,\alpha,\beta}(z_1)}{f_{s,\alpha,\beta_1}(z_1)}\cdot(s\alpha-1)/\beta^2\,\mathrm{d}\beta}\mathrm{d}z_1.
\end{aligned} \tag{A.18}$$

Note that $\mathrm{kl}(\mu,\mu') = \alpha\ln(\mu'/\mu) + \alpha\mu/\mu' - \alpha$, we obtain

$$\begin{aligned}
e^{s\mathrm{kl}(z_1/s,\alpha/\beta_1) - s\mathrm{kl}(z_1/s,\alpha/\beta)} &= e^{\alpha s\ln(\beta/\beta_1) + z_1(\beta_1-\beta)} \\
&= \left(\frac{\beta}{\beta_1}\right)^{\alpha s}\frac{e^{-z_1\beta}}{e^{-z_1\beta_1}} \\
&= \frac{f_{s,\alpha,\beta}(z_1)}{f_{s,\alpha,\beta_1}(z_1)}.
\end{aligned}$$

Recall that $\alpha/\beta_1 = \mu_1$. Besides, from (B.5), we have $s\mathrm{kl}(z_1/s,\alpha/\beta_1) - s\mathrm{kl}(z_1/s,\alpha/\beta) \ge s\mathrm{kl}(\alpha/\beta,\alpha/\beta_1) \ge s\mathrm{kl}(\mu_1 - \delta_j/4,\mu_1)$ for $\alpha/\beta \in (\mu_1-\delta_j/2, \mu_1-\delta_j/4)$ and $z_1/s \le \mu_1 - \delta_j/2$. Based on (A.18),

$$\begin{aligned}
\text{r.h.s. of (A.18)} &\le \frac{\alpha^2}{(s\alpha-1)(\mu_1-\delta_j/2)^2}\int_{s(\mu_1-\delta_j-\eta_s)}^{s(\mu_1-\delta_j/2)} \frac{f_{s,\alpha,\beta_1}(z_1)}{\int_{\frac{\alpha}{\mu_1-\delta_j/2}}^{\frac{\alpha}{\mu_1-\delta_j/4}} f_{s,\alpha,\beta_1}(z_1)\cdot\frac{f_{s,\alpha,\beta}(z_1)}{f_{s,\alpha,\beta_1}(z_1)}\,\mathrm{d}\beta}\mathrm{d}z_1 \\
&\le \frac{\alpha^2\cdot e^{-s\mathrm{kl}(\mu_1-\delta_j/4,\mu_1)}}{(s\alpha-1)(\mu_1-\delta_j/2)^2}\int_{s(\mu_1-\delta_j-\eta_s)}^{s(\mu_1-\delta_j/2)} \frac{(\mu_1-\delta_j/4)(\mu_1-\delta_j/2)}{\alpha\delta_j/4}\mathrm{d}z_1 \\
&\le e^{-s\mathrm{kl}(\mu_1-\delta_j/4,\mu_1)}\cdot\big(4\delta_j + 8\eta_s\big) \\
&\le e^{-s\delta_j^2/(32V)}\cdot\big(4\delta_j + 8\eta_s\big),
\end{aligned} \tag{A.19}$$

where the third inequality is due to that for $\delta_j \le \mu_1$, $\mu_1 - \delta_j/4/(\mu_1-\delta_j/2) \le 2$ and the last inequality is due to (B.4). For term J_2, note that the median denoted as m of Gamma$(\alpha s - 1, z_1)$ satisfies $m \le (\alpha s - 1)/z_1$. Therefore, for $z_1 \ge s(\mu_1 - \delta_j/2)$, $m \le (\alpha s-1)/z_1 \le \frac{\alpha}{\mu_1-\delta_j/2}$ (Chen & Rubin, 1986). Hence,

$$\int_{s(\mu_1-\delta_j/2)}^{\infty} \frac{f_{s,\alpha,\beta_1}(z_1)}{F^G_{s\alpha-1,z_1}(\alpha/(\mu_1-\delta_j))} - f_{s,\alpha,\beta_1}(z_1)\mathrm{d}z_1 \le \int_{s(\mu_1-\delta_j/2)}^{\infty} f_{s,\alpha,\beta_1}(z_1)\mathrm{d}z_1 \le 1. \tag{A.20}$$

Besides, for $x \geq \alpha/(\mu_1 - \delta_j)$, $z_1 \geq s(\mu_1 - \delta_j/2)$, and $s\alpha \geq \alpha_1 \geq \alpha_2$,

$$\frac{f^G_{\alpha_1,z_1}(x)}{f^G_{\alpha_2,z_1}(x)} = z_1^{\alpha_1-\alpha_2} x^{\alpha_1-\alpha_2} \cdot \frac{\Gamma(\alpha_2)}{\Gamma(\alpha_1)} \geq (s\alpha)^{\alpha_1-\alpha_2} \frac{\Gamma(\alpha_2)}{\Gamma(\alpha_1)} \geq 1. \tag{A.21}$$

Therefore, for $z_1 \geq s(\mu_1 - \delta_j/2)$,

$$\begin{aligned} F^G_{s\alpha-1,z_1}\left(\frac{\alpha}{\mu_1-\delta_j}\right) &= 1 - \int_{\alpha/(\mu_1-\delta_j)}^{\infty} f^G_{s\alpha-1,z_1}(x)\mathrm{d}x \\ &\geq 1 - \int_{\alpha/(\mu_1-\delta_j)}^{\infty} f^G_{\alpha s,z_1}(x)\mathrm{d}x \\ &= F^G_{\alpha s,z_1}\left(\frac{\alpha}{\mu_1-\delta_j}\right). \end{aligned} \tag{A.22}$$

Gamma$(\alpha s, z_1)$ is the empirical mean of s random variables i.i.d. according to Gamma$\big(\alpha, z_1/s\big)$. From Lemma B.1, for $z_1 \geq s(\mu_1 - \delta_j/2)$,

$$\begin{aligned} F^G_{\alpha s,z_1}\left(\frac{\alpha}{\mu_1-\delta_j}\right) &\geq 1 - e^{-s\mathrm{kl}\left(\frac{\alpha}{\mu_1-\delta_j}, \frac{\alpha s}{z_1}\right)} \\ &\geq 1 - e^{-s\mathrm{kl}\left(\frac{\alpha}{\mu_1-\delta_j}, \frac{\alpha}{\mu_1-\delta_j/2}\right)} \\ &= 1 - e^{-s\mathrm{kl}\left(\mu_1-\delta_j/2, \mu_1-\delta_j\right)} \\ &\geq 1 - e^{-s\delta_j^2/(8V)}, \end{aligned} \tag{A.23}$$

where the equality is due to $\mathrm{kl}(\mu, \mu') = \alpha \ln(\mu'/\mu) + \alpha\mu/\mu' - \alpha$ and the last inequality is due to (B.5). Therefore, for $s \geq 8V/\delta_j^2$

$$\int_{s(\mu_1-\delta_j/2)}^{\infty} \frac{f_{s,\alpha,\beta_1}(z_1)}{F^G_{s\alpha-1,z_1}(\alpha/(\mu_1-\delta_j))} - f_{s,\alpha,\beta_1}(z_1)\mathrm{d}z_1 \leq \frac{1}{1-e^{-s\delta_j^2/(8V)}} - 1 \leq 2e^{-(s-1)\delta_j^2/(8V)}. \tag{A.24}$$

Substituting (A.19),(A.20), and (A.24) into (A.17) and then combine it with (A.16), we obtain that exists a universal constant C_1 such that

$$\begin{aligned} &\sum_{s=1}^{T} \mathbb{E}\left[\left(\frac{1}{G'_{1s}(\delta_j)} - 1\right) \cdot \mathbb{1}\{\hat{\mu}_{1s} \in L_s\}\right] \\ &\leq \epsilon \sum_{s=1}^{T} e^{-s\delta_j^2/(32V)}\left(4 + \frac{8\eta_s}{\delta_j}\right) + \sum_{s \leq 8V/\delta_j^2 + 1/\alpha} 1 + 2 \sum_{s=\lceil 8V/\delta_j^2 \rceil + \lceil 1/\alpha \rceil}^{T} e^{-(s-1)\delta_j^2/(8V)} \\ &\leq (C_1 - 1)\left(\frac{V}{\epsilon\delta_j^2} + \frac{V}{\epsilon\delta_j^2}\log\left(\frac{T\epsilon\delta_j^2}{V}\right) + 1/\alpha\right) \\ &\leq C_1\left(\frac{V}{\epsilon\delta_j^2} + \frac{V}{\epsilon\delta_j^2}\log\left(\frac{T\epsilon\delta_j^2}{V}\right)\right), \end{aligned} \tag{A.25}$$

where the second inequality is due to Lemma 3.5 and (3.16) and the last inequality is due to the fact

$$\frac{V}{\delta_j^2} \geq \frac{V}{\mu_1^2} = \frac{\alpha}{\beta_1^2} \cdot \frac{\beta_1^1}{\alpha^2} = \frac{1}{\alpha}. \tag{A.26}$$

A.3. Proof of Lemma 3.5

Proof. From the definition of η_s, we have

$$\frac{2\log_+(T\epsilon/s)}{s} \geq \mathrm{kl}(\mu_1 - \epsilon - \eta_s, \mu_1 - \delta_j) \geq \frac{\eta_s^2}{2V},$$

where the last inequality is due to (B.4). Therefore,

$$\eta_s \leq \sqrt{\frac{4V \log_+(T\epsilon/s)}{s}} \leq \log_+(T\epsilon/s) \cdot \sqrt{\frac{4V}{s}}. \tag{A.27}$$

Let $d = \min\{\lfloor cV/\delta_j^2 \rfloor, \lfloor T\epsilon \rfloor\}$. We have

$$\begin{aligned}
\sum_{s=1}^{d} \frac{\log_+\left(T\epsilon/s\right)}{\sqrt{s}} &= \sum_{s=1}^{d} \frac{\log\left(T\epsilon\right)}{\sqrt{s}} - \sum_{s=1}^{d} \frac{\log s}{\sqrt{s}} \\
&\leq \int_1^d \frac{\log(T\epsilon)}{\sqrt{x}} \mathrm{d}x + \log(T\epsilon) - \int_1^d \frac{\log x}{\sqrt{x}} \mathrm{d}x + \frac{1}{\sqrt{e}} \\
&\leq \sqrt{d} \log\left(T\epsilon\right) - \int_0^{\log d} te^{t/2}\, \mathrm{d}t + 1 \\
&= \sqrt{d} \log\left(T\epsilon\right) - \left(2te^{t/2} - 4e^{t/2}\right)\Big|_0^{\log d} + 1 \\
&\leq \sqrt{d} \log\left(T\epsilon\right) - 2 \log d\sqrt{d} + 4\sqrt{d} \\
&\leq \sqrt{d} \cdot \left(\log_+\left(\frac{T\epsilon\delta_j^2}{cV}\right) + 4 \right),
\end{aligned} \tag{A.28}$$

where the first equality is due to the fact $d \leq T\epsilon$ and thus $\log_+(T\epsilon/s) \geq 0$,the first inequality is because for monotone function f, $\sum_{i=a}^{b} f(i) \leq \int_a^b f(x)\mathrm{d}x + \max_{i\in[a,b]} f(x)$, and the second inequality follows by noting that the integrand is unimodal and has a maximum value of $\frac{1}{\sqrt{e}}$ and for such function $\sum_{i=a}^{b} f(i) \leq \int_a^b f(x)\mathrm{d}x + \max_{i\in[a,b]} f(x)$.
For $s \geq d$, we have

$$\begin{aligned}
\sum_{s=d+1}^{T} e^{-s\delta_j^2/(cV)} \frac{\log_+\left(T\epsilon/s\right)}{\sqrt{s}} &\leq \frac{\delta_j}{\sqrt{cV}} \sum_{s=d+1}^{T} e^{-s\delta_j^2/(cV)} \\
&\leq \frac{\sqrt{cV}}{\delta_j},
\end{aligned} \tag{A.29}$$

where the last inequality follows by the reason as shown in (3.16). Therefore,

$$\begin{aligned}
e^{-s\delta_j^2/(cV)} \sum_{s=1}^{T} \frac{\eta_s}{\delta_j} &\leq \frac{\sqrt{4V}}{\delta_j} \sum_{s=1}^{d} \frac{\log_+\left(T\epsilon/s\right)}{\sqrt{s}} + \frac{\sqrt{4V}}{\delta_j} \sum_{s=d+1}^{T} e^{-s\delta_j^2/(cV)} \frac{\log_+\left(T\epsilon/s\right)}{\sqrt{s}} \\
&\leq \frac{\sqrt{4c}V}{\delta_j^2} \left(\log_+\left(\frac{T\epsilon\delta_j^2}{cV}\right) + 5 \right),
\end{aligned}$$

where the first inequality is due to (A.27) and the last inequality is due to (A.28), (A.29), and the fact $d \leq cV/\delta_j^2$.
We also have

$$e^{-s\delta_j^2/(cV)} \sum_{s=1}^{\infty} \frac{\eta_s}{\delta_j} \leq \frac{\sqrt{4V}}{\delta_j} \sqrt{\log T} \cdot \sum_{s=1}^{\infty} e^{-s\delta_j^2/(cV)} \leq \frac{\sqrt{4cV^3}}{\delta_j^3} \cdot \sqrt{\log T},$$

where the first inequality is due to (A.27) and the last inequality is due to (A.29). □

A.4. Proof of Lemma 3.4

Proof. Recall that θ_{is} is the sample from the posterior distribution after arm i is pulled s times and $G'_{is}(\delta_j) = \mathbb{P}(\mu(\theta_{is}) \geq \mu_1 - \delta_j)$. Since $\hat{\mu}_{is} \leq \mu_1 - \delta_j$, $G_{is}(\delta_j) = \epsilon G'_{is}(\delta_j)$
Bernoulli rewards. Since $\hat{\mu}_{is} \leq \mu_i + \delta_i$. $\alpha \leq s(\mu_i + \delta_i)$.

$$G_{is}(\delta_j) = \epsilon G'_{is}(\delta_j)$$

$$
\begin{aligned}
&= \epsilon(1 - F^{beta}_{\alpha+1,s-\alpha+1}(\mu_1 - \delta_j)) \\
&= \epsilon F_{s+1,\mu_1-\delta_j}(\alpha) \\
&\leq \epsilon e^{-(s+1)\mathrm{kl}(\alpha/s,\mu_1-\delta_j)} \\
&\leq \frac{V}{T\delta_j^2},
\end{aligned} \tag{A.30}
$$

where the inequality is due to Lemma B.1 and the fact that $F_{s+1,\mu_1-\delta_j}(\alpha)$ is the probability that the sum of $s+1$ Bernoulli random variables with parameter $\mu_1 - \delta_j$ is lower than α, and last inequality is due to $s \geq s_0 = \frac{\log(T\epsilon\delta_j^2/V)}{\mathrm{kl}(\mu_i+\delta_j,\mu_1-\delta_j)}$.

Gaussian rewards.

$$
\begin{aligned}
G_{is}(\delta_j) &= \epsilon G'_{is}(\delta_j) \\
&= \epsilon(1 - F_{s,\hat{\mu}_{is}}(\mu_1 - \delta_j)) \\
&\leq \epsilon e^{-s\mathrm{kl}(\mu_1-\epsilon_j,\hat{\mu}_{is})} \\
&= \epsilon e^{-s\mathrm{kl}(\hat{\mu}_{is},\mu_1-\epsilon_j)} \\
&\leq \frac{V}{T\delta_j^2},
\end{aligned}
$$

where the inequality is due to Lemma B.1 and the fact that $1 - F_{s,x}(\mu_1 - \delta_j)$ is the probability that the mean of s independent random variables from $\mathcal{N}(x, \sigma^2)$ is larger than $\mu_1 - \delta_j$.

Poisson rewards. We have $k = s\hat{\mu}_{is}$.

$$
\begin{aligned}
G_{is}(\delta_j) &= \epsilon G'_{is}(\delta_j) \\
&= \epsilon\big(1 - F^{G}_{k+1,s}\big(\mu_1 - \delta_j\big)\big) \\
&= \epsilon F^{\mathrm{Poi}}_{s(\mu_1-\delta_j)}(k) \\
&\leq \epsilon e^{-s\mathrm{kl}(\hat{\mu}_{is},\mu_1-\delta_j)} \\
&\leq \frac{V}{T\delta_j^2},
\end{aligned}
$$

where the inequality is due to Lemma B.1 and the fact that $F^{\mathrm{Poi}}_{s(\mu_1-\delta_j)}(k)$ is the probability that the sum of s independent random variables from Poisson with parameter $\mu_1 - \delta_j$ is lower than k.

Gamma rewards.

$$
G_{is}(\delta_j) = \epsilon G'_{is}(\delta_j) = \epsilon F^{G}_{\alpha s-1,z_i}\left(\frac{\alpha}{\mu_1 - \delta_j}\right),
$$

where z_i is the sum of rewards of arm i after its s-th pull. Let $s' = s - \lceil 1/\alpha \rceil$. For $x \leq \alpha/(\mu_1 - \delta_j)$, $z_i \leq s(\mu_i + \delta_j)$, and $s\alpha \geq \alpha_1 \geq \alpha_2$,

$$
\frac{f^{G}_{\alpha_1,z_i}(x)}{f^{G}_{\alpha_2,z_i}(x)} = z_i^{\alpha_1-\alpha_2} x^{\alpha_1-\alpha_2} \cdot \frac{\Gamma(\alpha_2)}{\Gamma(\alpha_1)} \leq (s\alpha)^{\alpha_1-\alpha_2} \frac{\Gamma(\alpha_2)}{\Gamma(\alpha_1)} \leq 1. \tag{A.31}
$$

Therefore, for $z_i \leq s(\mu_1 + \delta_j)$,

$$
F^{G}_{s\alpha-1,z_i}\left(\frac{\alpha}{\mu_1 - \delta_j}\right) \leq F^{G}_{\alpha s',z_i}\left(\frac{\alpha}{\mu_1 - \delta_j}\right). \tag{A.32}
$$

Gamma$(\alpha s', z_i)$ is the empirical mean of s' random variables i.i.d. according to Gamma$\big(\alpha, z_i/s'\big)$. From Lemma B.1, for $z_i \leq s(\mu_1 + \delta_j)$,

$$
F^{G}_{\alpha s',z_i}\left(\frac{\alpha}{\mu_1 - \delta_j}\right) \leq e^{-s'\mathrm{kl}\left(\mu_i+\delta_j,\mu_1-\delta_j\right)} \leq \frac{V}{T\epsilon\delta_j^2}, \tag{A.33}
$$

where the last inequality is because that from (A.26), $s \geq \frac{\log(T\epsilon\delta_j^2/V)}{\mathrm{kl}(\mu_i+\delta_j,\mu_1-\delta_j)}+1+V/\delta_j^2 \geq \frac{\log(T\epsilon\delta_j^2/V)}{\mathrm{kl}(\mu_i+\delta_j,\mu_1-\delta_j)}+\lceil 1/\alpha \rceil$. Therefore, $G_{is}(\delta_j) \leq V/(T\delta_j^2)$. Now, we prove the second statement. From Lemma 3.4,

$$\begin{aligned}
\sum_{s=1}^{T} \mathbb{1}\{G_{is}(\epsilon) > V/(T\delta_j^2)\} &\leq \sum_{s:s\geq s_0} \mathbb{1}\{G_{is}(\epsilon) > V/(T\delta_j^2), \hat{\mu}_{is} \leq \mu_i + \delta_j\} \\
&\quad + \sum_{s:s\geq s_0} \{\hat{\mu}_{is} > \mu_i + \delta_j\} + 2 + s_0 + \frac{V}{\delta_j^2} \\
&\leq 2 + s_0 + \frac{V}{\delta_j^2} + \sum_{s:s\geq 1} \{\hat{\mu}_{is} > \mu_i + \delta_j\} \\
&\leq 2 + s_0 + \frac{V}{\delta_j^2} + \sum_{s:s\geq 1} \exp(-s\delta_j^2/(2V)) \\
&\leq 2 + s_0 + \frac{V}{\delta_j^2} + \frac{1}{e^{\delta_j^2/(2V)}-1} \\
&\leq 2 + s_0 + \frac{3V}{\delta_j^2}, \qquad \text{(A.34)}
\end{aligned}$$

where the second inequality is from Lemma 3.4, the third inequality is from Lemma B.1, and the last inequality is due to the fact $e^x \geq x+1$ for any x. □

B. Some Useful Inequalities

Lemma B.1 (Maximal Inequality (Ménard & Garivier, 2017)). *Let N and M be two real numbers in $\mathbb{R}^+ \times \overline{\mathbb{R}^+}$, let $\gamma > 0$, and $\hat{\mu}_n$ be the empirical mean of n random variables i.i.d. according to some distribution in exponential family with mean μ. Let V be the maximum variance of the distribution with mean $\mu \in [x, \mu]$. Then, for $x \leq \mu$,*

$$\begin{aligned}
\mathbb{P}(\exists N \leq n \leq M, \hat{\mu}_n \leq x) &\leq e^{-N\cdot kl(x,\mu)}, \\
\mathbb{P}(\exists N \leq n \leq M, \hat{\mu}_n \leq x) &\leq e^{-N(x-\mu)^2/(2V)}.
\end{aligned} \qquad \text{(B.1)}$$

Meanwhile, for every $x \geq \mu$,

$$\mathbb{P}(\exists N \leq n \leq M, \hat{\mu}_n \geq x) \leq e^{-N\cdot kl(x,\mu)}, \qquad \text{(B.2)}$$

$$\mathbb{P}(\exists N \leq n \leq M, \hat{\mu}_n \geq x) \leq e^{-N(x-\mu)^2/(2V)}. \qquad \text{(B.3)}$$

Lemma B.2 (Jin et al. (2021a)). *For any $\mu_1 \leq \mu_2$ (or $\mu_1 \geq \mu_2$),*

$$kl(\mu_1, \mu_2) \geq (\mu_1 - \mu_2)^2/(2V), \qquad \text{(B.4)}$$

where V is the maximum variance for the reward distribution with mean $\mu \in [\mu_1, \mu_2]$ (or $\mu \in [\mu_2, \mu_1]$). In addition, for $\epsilon > 0$ and $\mu_1 \leq \mu_2 - \epsilon$, we have

$$\begin{aligned}
& kl(\mu_1, \mu_2) - kl(\mu_1, \mu_2 - \epsilon) \geq kl(\mu_2 - \epsilon, \mu_2), \\
& kl(\mu_1, \mu_2 + \epsilon) \geq kl(\mu_1, \mu_2) \geq kl(\mu_1, \mu_2 - \epsilon), \\
\text{and} \quad & kl(\mu_1 + \epsilon, \mu_2) \leq kl(\mu_1, \mu_2) \leq kl(\mu_1 - \epsilon, \mu_2).
\end{aligned} \qquad \text{(B.5)}$$

C. Additional Experimental Results

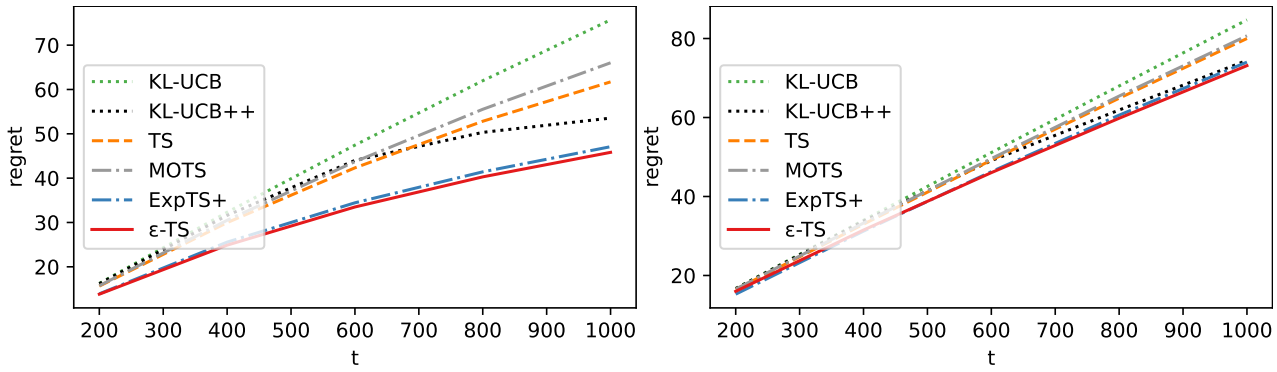
In this section, we provide more comprehensive experimental results on the proposed ϵ-TS algorithm.

C.1. Hard Bandit Instances

We conducted experiments on two challenging bandit instances: (1) Gaussian rewards with unit variance, and (2) Bernoulli rewards. The mean rewards were set as $(\mu_1, \mu_2, \cdots, \mu_K) = (0, 3, 0.2, \cdots, 0.2)$ for $K = 10$ and $T = 1000$. In these

instances, we have $\Delta_i = 0.1 = \sqrt{K/T}$, which is similar to the instance used in the proof of the minimax lower bound for multi-armed bandits (Lattimore & Szepesvári, 2020). This particular instance is designed to test the worst-case performance of bandit algorithms and represents a challenging scenario in practice. The experimental results are averaged over 1000 repetitions for all algorithms.

In Figure 2(a) and 2(b), we provide the average regret of each algorithm. In Table 3 and 4, we report the 95% confidence intervals for these algorithms. It can be seen that, the regret of ϵ-TS with Gaussian rewards at time 200 is 16.0 with a confidence interval $(6.4, 19.0)$, which means that the average regret at time 200 is 16.0, with at most 5% of experiments exhibiting a regret lower than 6.4 and at most 5% of experiments showing a regret greater than 19. These additional experiments reveal that the confidence interval of ϵ-TS is slightly larger than that of other algorithms. The regret of ϵ-TS consistently outperforms those of other algorithms.



(a) Bernoulli, $K = 10$ (b) Gaussian, $K = 10$

Figure 2. Comparison of different algorithms for hard bandit instances under Gaussian and Bernoulli reward distributions.

Table 3. Confidence intervals of different algorithms for hard bandit instances with Gaussian rewards.

ALGORITHMS / TIMES	200	400	600	800	1000
KL-UCB	16.7 (14.9, 18.1)	34.0 (30.5, 37.0)	51.1 (45.9, 55.8)	68.0 (61.2, 74.2)	84.7 (76.1, 92.5)
KL-UCB^{++}	16.7 (14.6, 18.3)	33.8 (29.6, 37.4)	49.1 (32.4, 56.7)	61.9 (32.4, 76.6)	74.5 (32.4, 96.6)
TS	16.4 (12.9, 18.7)	33.0 (25.7, 38.3)	49.2 (37.5, 57.8)	64.9 (48.9, 77.4)	80.0 (58.5, 96.9)
MOTS	16.6 (13.5, 18.7)	33.3 (26.2, 38.3)	49.6 (38.8, 57.8)	65.4 (50.1, 77.1)	80.7 (60.8, 96.3)
EXPTS	15.8 (13.5, 17.4)	32.9 (27.5, 36.7)	49.4 (41.0, 55.9)	65.7 (54.4, 74.9)	81.7 (67.4, 93.9)
EXPTS^{+}	15.3 (7.8, 17.8)	31.2 (15.4, 37.6)	46.3 (21.6, 57.3)	60.6 (27.6, 77.0)	74.0 (33.4, 96.7)
ϵ-TS	16.0 (6.4, 19.0)	31.5 (10.9, 38.9)	46.0 (15.6, 58.8)	59.8 (19.6, 78.8)	73.1 (22.9, 98.7)

Table 4. Confidence intervals of different algorithms for hard bandit instances with Bernoulli rewards.

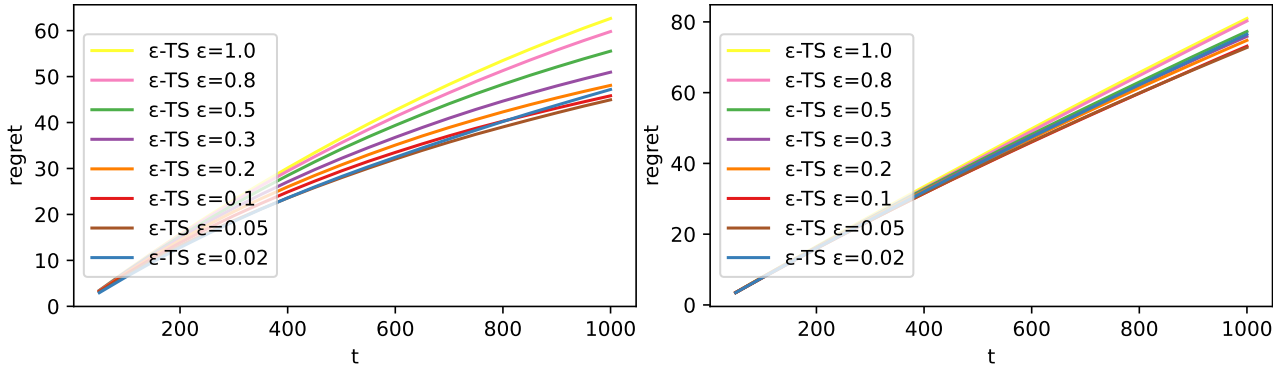
ALGORITHMS / TIMES	200	400	600	800	1000
KL-UCB	16.3 (14.6, 17.8)	32.3 (28.3, 35.9)	47.4 (41.2, 53.3)	62.0 (53.0, 70.1)	75.7 (64.0, 86.5)
KL-UCB^{++}	16.2 (14.2, 17.8)	31.6 (26.3, 36.4)	44.0 (30.6, 53.9)	50.3 (32.2, 71.6)	53.5 (32.8, 89.6)
TS	15.6 (12.0, 18.3)	30.0 (21.0, 37.1)	42.3 (28.4, 55.8)	52.8 (33.7, 73.6)	61.7 (39.1, 90.0)
MOTS	15.7 (12.3, 18.3)	30.5 (22.9, 37.0)	43.7 (31.4, 55.2)	55.5 (38.7, 73.2)	66.0 (44.7, 90.5)
EXPTS	15.2 (12.8, 17.0)	30.5 (24.4, 35.7)	44.3 (34.4, 53.5)	56.7 (43.2, 70.8)	67.9 (50.4, 86.9)
EXPTS^{+}	13.9 (7.6, 17.4)	25.6 (12.2, 36.9)	34.4 (15.8, 56.4)	41.4 (18.9, 75.8)	47.1 (21.6, 95.2)
ϵ-TS	13.8 (5.3, 18.7)	24.9 (8.9, 38.4)	33.5 (11.4, 58.1)	40.3 (14.0, 77.8)	45.8 (15.8, 97.6)

C.2. Ablation Study on the Choice of ϵ

We also conducted a series of ablation studies on the ϵ-TS algorithm, evaluating the influence of different ϵ values on its performance. The bandit instances used in these experiments are the same as those presented in Section C.1.

Figures 3(a) and 3(b) present the regret associated with different algorithms at each time step. Tables 6 and 5 display the corresponding confidence intervals for these algorithms. The tables specifically illustrate the average regret for each algorithm at time steps $T = 800, 1200, 1600, 2000$. We also present the confidence intervals. For example, the regret for the ϵ-TS algorithm with $\epsilon = 0.1$ at time step 1000 is 40.3, with a lower bound of 14.0 and an upper bound of 77.8, representing the range in which 90% of the experimental results fall.

The results reveal that as ϵ decreases from 1.0 to 0.1 (i.e., $1/K$), the associated regret diminishes. However, when ϵ values fall below 0.1, the regret might tend to increase. For instance, when the reward distribution is Bernoulli, the regret associated with ϵ-TS for $\epsilon = 0.02$ at $T = 800$ is higher than the regret of ϵ-TS for $\epsilon = 0.1$. This could be attributed to insufficient exploration when ϵ is too small.



(a) Bernoulli, $K = 10$ (b) Gaussian, $K = 10$

Figure 3. Ablation study of ϵ-TS under Gaussian and Bernoulli reward distributions.

Table 5. Confidence intervals of ϵ-TS with different ϵ in Bernoulli reward environments.

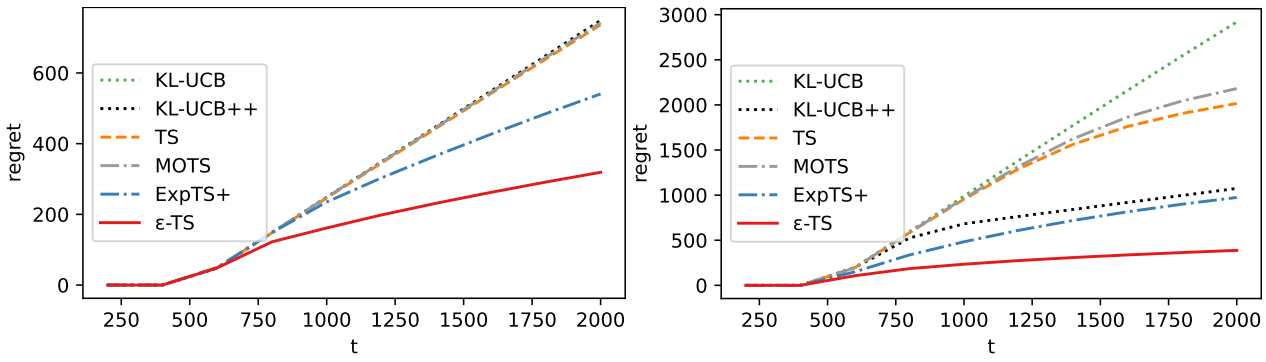
ϵ-TS / TIMES	200	400	800	1000
$\epsilon = 1.0$	15.7 (12.2, 18.2)	30.1 (21.5, 37.2)	42.7 (28.9, 55.8)	53.5 (35.4, 73.1)
$\epsilon = 0.8$	15.5 (11.5, 18.2)	29.4 (20.4, 37.2)	41.3 (27.0, 55.7)	51.3 (31.6, 73.8)
$\epsilon = 0.5$	15.2 (10.3, 18.3)	28.5 (18.1, 37.6)	39.4 (23.9, 56.4)	48.3 (28.4, 74.4)
$\epsilon = 0.3$	14.9 (8.5, 18.4)	27.1 (14.6, 37.9)	36.8 (19.4, 57.3)	44.7 (23.1, 76.3)
$\epsilon = 0.2$	14.4 (7.4, 18.5)	26.0 (12.5, 38.0)	35.1 (16.7, 57.2)	42.3 (20.2, 76.5)
$\epsilon = 0.1$	13.8 (5.3, 18.7)	24.9 (8.9, 38.4)	33.5 (11.4, 58.1)	40.3 (14.0, 77.8)
$\epsilon = 0.05$	13.2 (3.3, 18.8)	23.6 (5.8, 38.6)	32.0 (8.1, 58.4)	39.0 (10.0, 78.2)
$\epsilon = 0.02$	12.8 (1.7, 19.0)	23.6 (3.2, 38.9)	32.4 (4.3, 58.7)	40.2 (5.4, 78.6)

Table 6. Confidence intervals of ϵ-TS with different ϵ in Gaussian reward environments.

ε-TS / TIMES	200	400	800	1000
$\epsilon = 1.0$	16.5 (12.9, 18.8)	33.4 (25.3, 38.3)	49.8 (36.9, 58.1)	65.7 (47.9, 77.7)
$\epsilon = 0.8$	16.4 (12.0, 18.8)	33.0 (23.7, 38.4)	49.1 (34.9, 58.2)	64.8 (44.2, 77.8)
$\epsilon = 0.5$	16.4 (11.3, 18.8)	32.7 (21.9, 38.6)	48.2 (30.5, 58.3)	63.0 (38.7, 77.9)
$\epsilon = 0.3$	16.2 (9.9, 18.9)	32.3 (18.8, 38.6)	47.6 (26.5, 58.3)	62.1 (33.4, 78.1)
$\epsilon = 0.2$	16.2 (8.8, 18.9)	32.2 (15.7, 38.8)	47.1 (21.4, 58.7)	61.3 (27.5, 78.5)
$\epsilon = 0.1$	16.0 (6.4, 19.0)	31.5 (10.9, 38.9)	46.0 (15.6, 58.8)	59.8 (19.6, 78.8)
$\epsilon = 0.05$	15.9 (4.6, 19.0)	31.6 (7.6, 39.0)	46.2 (11.5, 58.9)	59.9 (14.0, 78.9)
$\epsilon = 0.02$	16.0 (3.2, 19.0)	32.0 (4.8, 39.0)	47.5 (7.4, 59.0)	62.2 (9.1, 79.0)

C.3. Large Number of Arms

We further test a 500-armed bandit under two settings: (1) Gaussian rewards with a unit variance where the mean rewards are $(2, 0, 0, \cdots, 0)$, and (2) Bernoulli rewards with mean rewards $(0, 75, 0.25, \cdots, 0.25)$. We set $T = 2000$.



(a) Bernoulli, $K = 500$ (b) Gaussian, $K = 500$

Figure 4. Comparison of different algorithms for bandit instances with a large number of arms under Gaussian and Bernoulli reward distributions.

We present the regret of different algorithms at each time step in Figures 4(a) and 4(b) and the confidence interval of the different algorithms in Tables 7 and 8. In the tables, we present the average regret of each algorithm at time steps $T = 800, 1200, 1600, 2000$ respectively. Additionally, we report the confidence intervals for these algorithms. For instance, the result of ϵ-TS with Gaussian rewards at time $T = 2000$ in Table 7 is displayed as 364 (146, 1456), which implies that the average regret at time $T = 2000$ is 364, with at most 5% of experiments exhibiting a regret lower than 146 and at most 5% of experiments showing a regret greater than 1456. Due to the fact that ExpTS fails to produce all results for Bernoulli/Gaussian rewards within a 72-hour timeframe, it is not included in the tables and figures.

The results indicate that for a larger number of arms, specifically $K = 500$, the regret of ϵ-TS is significantly smaller compared to the baselines. In particular, for Gaussian rewards, at time $T = 2000$, the regret of ϵ-TS is approximately $2.5\times$ smaller than the regret of ExpTS$^+$, about $5.2\times$ smaller than the regret of TS, and roughly $2.7\times$ smaller than the regret of KL-UCB^{++}. For Bernoulli rewards, at time $T = 2000$, the regret of ϵ-TS is approximately $1.7\times$ smaller than the regret of ExpTS$^+$ and about $2.2\times$ smaller than the other baselines.

Table 7. Confidence intervals for different algorithms dealing with large arm sets in Gaussian reward environments.

ALGORITHMS / TIMES	800	1200	1600	2000
KL-UCB	594 (590, 598)	1383 (1372, 1394)	2161 (2136, 2184)	2544 (2508, 2574)
KL-UCB^{++}	525 (254, 598)	764 (254, 1396)	919 (254, 2196)	997 (254, 2596)
TS	583 (558, 598)	1290 (1178, 1387)	1761 (1570, 2012)	1905.5 (1706, 2160)
MOTS	586 (568, 598)	1318 (1226, 1390)	1865 (1688, 2096)	2044 (1846, 2299)
ExpTS^{+}	337 (222, 600)	610 (458, 1359)	816 (652, 1498)	900 (734, 1558)
ϵ-TS	185 (50, 600)	275 (92, 1398)	337 (130, 1454)	364 (146, 1456)

Table 8. Confidence intervals for different algorithms dealing with large arm sets in Bernoulli reward environments.

ALGORITHMS / TIMES	800	1200	1600	2000
KL-UCB	148.3 (145.0, 149.5)	348.3 (345.0, 349.5)	548.3 (545.0, 549.5)	648.3 (645.0, 649.5)
KL-UCB^{++}	148.3 (145.0, 149.5)	348.3 (345.0, 349.5)	548.3 (545.0, 549.5)	648.3 (645.0, 649.5)
TS	147.7 (142.5, 150.0)	345.2 (336.5, 350.0)	541.9 (529.0, 549.5)	639.4 (623.4, 649.0)
MOTS	148.8 (147.5, 150.0)	347.7 (345.0, 350.0)	545.4 (540.5, 549.0)	643.9 (637.9, 649.0)
ExpTS^{+}	148.4 (145.0, 150.0)	303.2 (260.4, 350.0)	426.8 (361.5, 549.0)	484.9 (410.0, 649.0)
ϵ-TS	122.2 (83.0, 150.0)	198.4 (97.5, 350.0)	262.3 (113.0, 550.0)	291.4 (120.5, 650.0)