Finite-Time Regret of Thompson Sampling Algorithms for Exponential Family Multi-Armed Bandits

Tianyuan Jin
National University of Singapore
tianyuan@u.nus.edu

Pan Xu
Duke University
pan.xu@duke.edu

Xiaokui Xiao
National University of Singapore
xkxiao@nus.edu.sg

Anima Anandkumar
California Institute of Technology
anima@caltech.edu

Abstract

We study the regret of Thompson sampling (TS) algorithms for exponential family bandits, where the reward distribution is from a one-dimensional exponential family, which covers many common reward distributions including Bernoulli, Gaussian, Gamma, Exponential, etc. We propose a Thompson sampling algorithm, termed ExpTS, which uses a novel sampling distribution to avoid the under-estimation of the optimal arm. We provide a tight regret analysis for ExpTS, which simultaneously yields both the finite-time regret bound as well as the asymptotic regret bound. In particular, for a K-armed bandit with exponential family rewards, ExpTS over a horizon T is sub-UCB (a strong criterion for the finite-time regret that is problem-dependent), minimax optimal up to a factor $\sqrt{\log K}$, and asymptotically optimal, for exponential family rewards. Moreover, we propose ExpTS$^+$, by adding a greedy exploitation step in addition to the sampling distribution used in ExpTS, to avoid the over-estimation of sub-optimal arms. ExpTS$^+$ is an anytime bandit algorithm and achieves the minimax optimality and asymptotic optimality simultaneously for exponential family reward distributions. Our proof techniques are general and conceptually simple and can be easily applied to analyze standard Thompson sampling with specific reward distributions.

1 Introduction

The Multi-Armed Bandit (MAB) problem is centered around a fundamental model for balancing the exploration versus exploitation trade-off in many online decision problems. In this problem, the agent is given an environment with a set of K arms $[K] = \{1, 2, \cdots, K\}$. At each time step t, the agent pulls an arm $A_t \in [K]$ based on observations of previous $t-1$ time steps, and then a reward r_t is revealed at the end of the step. In real-world applications, reward distributions often have different forms such as Bernoulli, Gaussian, etc. As suggested by Auer et al. [8, 9], Agrawal and Goyal [5], Lattimore [29], Garivier et al. [16], a good bandit strategy should be general enough to cover a sufficiently rich family of reward distributions. In this paper, we assume the reward r_t is independently generated from some canonical one-parameter exponential family of distributions with a mean value μ_{A_t}. It is a rich family that covers many common distributions including Bernoulli, Gaussian, Gamma, Exponential, and others.

The goal of a bandit strategy is usually to maximize the cumulative reward over T time steps, which is equivalent to minimizing the regret, defined as the expected cumulative difference between playing

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

the best arm and playing the arm according to the strategy: $R_\mu(T) = T \cdot \max_{i\in[K]} \mu_i - \mathbb{E}[\sum_{t=1}^T r_t]$. We assume, without loss of generality, $\mu_1 = \max_{i\in[K]} \mu_i$ is the best arm throughout this paper. For a fixed bandit instance (i.e., mean rewards $\mu_1, \cdots, \mu_K$ are fixed), Lai and Robbins [26] shows that for distributions that are continuously parameterized by their means,

$$\lim_{T\to\infty} \frac{R_\mu(T)}{\log T} \geq \sum_{i>1} \frac{\mu_1 - \mu_i}{\mathrm{kl}(\mu_i, \mu_1)}, \tag{1.1}$$

where $\mathrm{kl}(\mu_i, \mu_1)$ is the Kullback-Leibler divergence between two distributions with mean μ_i and μ_1. A bandit strategy satisfying $\lim_{T\to\infty} R_\mu(T)/\log T = \sum_{i>1} \frac{\mu_1-\mu_i}{\mathrm{kl}(\mu_i,\mu_1)}$ is said to be *asymptotically optimal* or achieve the asymptotic optimality in regret. The asymptotic optimality is one of the most important statistical properties in regret minimization, which shows that an algorithm is consistently good when it is played for infinite steps and thus should be a basic theoretical requirement of any good bandit strategy [8].

In practice, we can only run the bandit algorithm for a finite number T steps, which is the time horizon of interest in real-world applications. Therefore, the finite-time regret is the ultimate property of a practical bandit strategy in regret minimization problems. A strong notion of finite-time regret bounds is called *the sub-UCB criteria* [29]. An algorithm is sub-UCB if there exist universal constants $C_1, C_2 > 0$ such that for any problem instances,

$$R_\mu(T) = C_1 \sum_{i\in[K]:\Delta_i>0} \Delta_i + C_2 \sum_{i\in[K]:\Delta_i>0} \frac{\log T}{\Delta_i}, \tag{1.2}$$

where $\Delta_i = \mu_1 - \mu_i$ is the sub-optimal gap between arm 1 and arm i. Note that the regret bound in (1.2) is a problem-dependent bound since it depends on the bandit instance and the sub-optimal gaps. Sub-UCB is an important metric for finite-time regret bound and has been adopted by recent work of Lattimore [29], Bian and Jun [10]. Another special type of finite-time bounds is called the worst-case regret, which is defined as the finite-time regret of an algorithm on any possible bandit instance within a bandit class. Specifically, for a finite time horizon T, Auer et al. [8] proves that any strategy has at least worst-case regret $\Omega(\sqrt{KT})$ for a K-armed bandit. We say the strategy that achieves a worst-case regret $O(\sqrt{KT})$ is *minimax optimal* or achieves the minimax optimality. Different from the asymptotic optimality, the minimax optimality characterizes the worst-case performance of the bandit strategy in finite steps.

A vast body of literature in multi-armed bandits [6, 5, 22, 32, 16, 29] have been pursing the aforementioned theoretical properties of bandit algorithms: generality, asymptotic optimality, problem-dependent finite-time regret, and minimax optimality. However, most of them focus on one or two properties and sacrifice the others. Moreover, many of existing theoretical analyses of bandit strategies are for optimism-based algorithm. The theoretical analysis of Thompson sampling (TS) is much less understood until recently, which has been shown to exhibit superior practical performances compared to the state-of-the-art methods [13, 34]. Specifically, its finite-time regret, asymptotic optimality, and near minimax optimality have been studied by Agrawal and Goyal [3, 4, 5] for Bernoulli rewards. Jin et al. [20] proved the minimax optimality of TS for sub-Gaussian rewards. For exponential family reward distributions, the asymptotic optimality is shown by Korda et al. [25], but no finite-time regret of TS is provided. See Table 1 for a comprehensive comparison of these results.

In this paper, we study the regret of Thompson sampling for exponential family reward distributions and address all the theoretical properties of TS. We propose a variant of TS algorithm with a general sampling distribution and a tight analysis for frequentist regret bounds. Our analysis simultaneously yields both the finite-time regret bound and the asymptotic regret bound.
Specifically, the **main contributions** of this paper are summarized as follows:

- We propose ExpTS, a general variant of Thompson sampling, that uses a novel sampling distribution with a tight anti-concentration bound to avoid the under-estimation of the optimal arm and a tight concentration bound to avoid the over-estimation of sub-optimal arms. For exponential family of reward distributions, we prove that ExpTS is the first Thompson sampling algorithm achieving the sub-UCB criteria, which is a strong notion of problem-dependent finite-time regret bounds. We further show that ExpTS is also simultaneously minimax optimal up to a factor of $\sqrt{\log K}$, as well as asymptotically optimal, where K is the number of arms.
- We also propose ExpTS$^+$, which explores between the sample generated in ExpTS and the empirical mean reward for each arm, to get rid of the extra $\sqrt{\log K}$ factor in the worst-case regret.

Table 1: Comparisons of different Thompson sampling algorithms on K-armed bandits over a horizon T. For any algorithm, *Asym. Opt* is the indicator whether it is asymptotically optimal, *minimax ratio* is the scaling of its worst-case regret w.r.t. the minimax optimal regret $O(\sqrt{VKT})$, where V is the variance of reward distributions, and sub-UCB is the indicator if it satisfies the sub-UCB criteria.

Algorithm	Reward Type	Asym. Opt	Finite-Time Regret		Anytime	Reference
			Minimax Ratio	Sub-UCB		
TS	Bernoulli	yes	$\sqrt{\log T}$	–*	yes	[4]
TS	Bernoulli	–	$\sqrt{\log K}$	–*	yes	[5]
TS	Exponential Family	yes	–	–	yes	[25]
MOTS	sub-Gaussian	no	1	no	no	[20]
MOTS-$\mathcal{J}$	Gaussian	yes	1	no	no	[20]
ExpTS	Exponential Family	yes	$\sqrt{\log K}$	yes	yes	This paper
ExpTS$^+$	Exponential Family	yes	1	no	yes	This paper

* [4, 5] did not explicitly show that their regret bounds are sub-UCB. However, the intermediate results in their proofs might imply sub-UCB regret bounds.

Thus ExpTS$^+$ is the first Thompson sampling algorithm that is simultaneously minimax and asymptotically optimal for exponential family of reward distributions.

- Our regret analysis of ExpTS can be easily extended to analyze standard Thompson sampling with common reward distributions. We prove that standard Thompson sampling without inflating the posterior distribution[1] is minimax optimal up to a factor of $\sqrt{\log K}$, which matches the regret lower bound for standard Thompson sampling in Agrawal and Goyal [5]. Similar to the idea of ExpTS$^+$, we can add a greedy exploration step to the posterior distributions used in these variants of TS, and then the algorithms are simultaneously minimax and asymptotically optimal.

Our techniques are novel and conceptually simple. First, we introduce a lower confidence bound in the regret decomposition to avoid the under-estimation of the optimal arm, which is important in obtaining the finite-time regret bound. Specifically, Jin et al. [20] (Lemma 5 in their paper) shows that for Gaussian reward distributions, Gaussian-TS has a regret bound at least in the order of $\Omega(\sqrt{KT\log T})$ if the standard regret decomposition in existing analysis of Thompson sampling [5, 30, 20] is adopted. With our new regret decomposition that is conditioned on the lower confidence bound introduced in this paper, we improve the worst-case regret of Gaussian-TS for Gaussian reward distributions to $O(\sqrt{KT\log K})$.

Second, we do not require the closed form of the reward distribution, but only make use of the corresponding concentration bounds. This means our results can be readily extended to other reward distributions. For example, we can extend ExpTS$^+$ to sub-Gaussian reward distributions and the algorithm is simultaneously minimax and asymptotically optimal[2], which improve the results of MOTS proposed by Jin et al. [20] (see Table 1).

Third, the idea of ExpTS$^+$ is simple and can be used to remove the extra $\sqrt{\log K}$ factor in the worst-case regret. We note that MOTS [20] can also achieve the minimax optimal via the clipped Gaussian. However, it is not clear how to generalize the clipping idea to the exponential family of reward distribution. Moreover, it uses the MOSS [6] index for clipping, which needs to know the horizon T in advance and thus cannot be extended to the anytime setting, while ExpTS$^+$ is an anytime bandit algorithm which does not need to know the horizon length in advance.

Notations. We let T be the total number of time steps, K be the total number of arms, and $[K] = \{1, 2, \cdots, K\}$. For simplicity, we assume arm 1 is the optimal throughout this paper, i.e., $\mu_1 = \max_{i\in[K]} \mu_i$. We denote $\log^+(x) = \max\{0, \log x\}$ and $\Delta_i := \mu_1 - \mu_i, i \in [K] \setminus \{1\}$ for the

[1]This is a common trick in the literature. In particular, for Bernoulli rewards, instead of using Beta posterior, Agrawal and Goyal [5] consider Thompson sampling with Gaussian posterior, whose variance is larger. Moreover, Jin et al. [20] inflate the variance of the sampling distribution by a factor $1/\rho$, where $\rho < 1$. However, both of these methods lose the asymptotic optimality.

[2]Note that sub-Gaussian is a non-parametric family and thus the lower bound (1.1) by Lai and Robbins [26] does not directly apply to a general sub-Gaussian distribution. Following similar work in the literature [20], in this paper, when we say an algorithm achieves the asymptotic optimality for sub-Gaussian rewards, we mean its regret matches the asymptotic lower bound for Gaussian rewards, which is a stronger notion.

gap between arm 1 and arm i. We let $T_i(t) := \sum_{j=1}^{t} \mathbb{1}\{A_t = i\}$ be the number of pulls of arm i at the time step t, $\widehat{\mu}_i(t) := 1/T_i(t) \sum_{j=1}^{t} \big[r_j \cdot \mathbb{1}\{A_t = i\}\big]$ be the average reward of arm i at the time step t, and $\widehat{\mu}_{is}$ be the average reward of arm i after its s-th pull. We reserve notations $C_1, C_2, \cdots$ to be positive universal constants that are independent of problem parameters.

2 Related Work

There are series of works pursuing the asymptotic regret bound and worst-case regret bound for MAB. For asymptotic optimality, UCB algorithms [15, 31, 5, 29], Thompson sampling [23, 25, 5, 20], Bayes-UCB [22], and other methods [21, 10] are all shown to be asymptotically optimal. Among them, only a few [15, 12, 25] can be extended to exponential families of distributions. One notable result in Cappé et al. [12] shows that for $[0, 1]$ bounded distribution, there exists an algorithm that has regret $\sum_{i>1} \Delta_i \log T/\mathrm{kl}(\mu_i, \mu_1) + O(\sum_{i>1} (\log T)^{4/5} \log\log T \cdot \Delta_i)^3$, which is better than (1.2). It is an interesting problem whether we can achieve such regret for unbounded reward distributions. For the worst-case regret, MOSS [6] is the first algorithm proved to be minimax optimal. Later, KL-UCB^{++} [5], AdaUCB [29], MOTS [20] also join the family. The anytime version of MOSS is studied by Degenne and Perchet [14]. There are also some works that focus on the near optimal problem-dependent regret bound [27, 28]. As far as we know, no algorithm has been proved to achieve the sub-UCB criteria, asymptotic optimality, and minimax optimality simultaneously for exponential family reward distributions.

For Thompson sampling, Russo and Van Roy [33] studied the Bayesian regret. They show that the Bayesian regret of Thompson sampling is never worse than the regret of UCB. Bubeck and Liu [11] further showed the Bayesian regret of Thompson sampling is optimal using the regret analysis of MOSS. There are also a line of works focused on the frequentist regret of TS. Agrawal and Goyal [3] proposed the first finite time regret analysis for TS. Kaufmann et al. [23], Agrawal and Goyal [4] proved that TS with Beta posteriors is asymptotically optimal for Bernoulli reward distributions. Korda et al. [25] extended the asymptotic optimality to the exponential family of reward distributions. Subsequently, for Bernoulli rewards, Agrawal and Goyal [5] proved that TS with Beta prior is asymptotically optimal and has worst-case regret $O(\sqrt{KT \log T})$. Besides, they showed that TS with Gaussian posteriors can achieve a better worst-case regret bound $O(\sqrt{KT \log K})$. They also proved that for Bernoulli rewards, TS with Gaussian posteriors has a worst-case regret at least $\Omega(\sqrt{KT \log K})$. Very recently, Jin et al. [20] proposed the MOTS algorithm that can achieve the minimax optimal regret $O(\sqrt{KT})$ for multi-armed bandits with sub-Gaussian rewards but at the cost of losing the asymptotic optimality by a multiplicative factor of $1/\rho$, where $0 < \rho < 1$ is an arbitrarily fixed constant. For bandits with Gaussian rewards, Jin et al. [20] proved that MOTS combined with a Rayleigh distribution can achieve the minimax optimality and the asymptotic optimality simultaneously. We refer readers to Tables 1 and 2 for more details.

3 Preliminary on Exponential Family Distributions

A one-dimensional canonical exponential family [15, 17, 32] is a parametric set of probability distributions with respect to some reference measure, with the density function given by

$$p_\theta(x) = \exp(x\theta - b(\theta) + c(x)),$$

where θ is the model parameter, and c is a real function. Denote the measure of $p_\theta(x)$ as ν_θ. Then, the above definition can be rewritten as

$$\frac{\mathrm{d}\nu_\theta}{\mathrm{d}\rho}(x) = \exp(x\theta - b(\theta)),$$

for some measure ρ and $b(\theta) = \log(\int e^{x\theta} \mathrm{d}\rho(x))$. We make the classic assumption used by Garivier and Cappé [15], Ménard and Garivier [32] that $b(\theta)$ is twice differentiable with a continuous second derivative. Then, we can verify that exponential families have the following properties:

$$b'(\theta) = \mathbb{E}[\nu_\theta] \qquad \text{and} \qquad b''(\theta) = \mathrm{Var}[\nu_\theta] > 0.$$

[3]Cappé et al. [12] used a more general notation $\mathcal{K}_{\inf}(\cdot, \cdot)$ for any distribution supported in $[0, 1]$, which is equivalent to the $\mathrm{kl}(\cdot, \cdot)$ notation for the one-exponential family distribution studied in our paper.

Let $\mu = \mathbb{E}[\nu_\theta]$. The above equality means that the mapping between the mean value μ of $\nu(\theta)$ and the parameter θ is one-to-one. Hence, exponential family of distributions can also be parameterized by the mean value $\mu = b'(\theta)$. Note that $b''(\theta) > 0$ for all θ, which implies $b'(\cdot)$ is invertible and its inverse function b'^{-1} satisfies $\theta = b'^{-1}(\mu)$. In this paper, we will use the notion of Kullback-Leibler (KL) divergence. The KL divergence between two exponential family distributions with parameter θ and θ' respectively is defined as follows:

$$\mathrm{KL}(\nu_\theta, \nu_{\theta'}) = b(\theta') - b(\theta) - b'(\theta)(\theta' - \theta). \tag{3.1}$$

Recall that the mapping $\theta \mapsto \mu$ is one-to-one. We can define an equivalent notion of the KL divergence between random variables ν_θ and $\nu_{\theta'}$ as a function of the mean values μ and μ' respectively:

$$\mathrm{kl}(\mu, \mu') = \mathrm{KL}(\nu_\theta, \nu_{\theta'}),$$

where $\mathbb{E}[\nu_\theta] = \mu$ and $\mathbb{E}[\nu_{\theta'}] = \mu'$. Similarly, we define $V(\mu) = \mathrm{Var}(\nu_{b'^{-1}(\mu)})$ as the variance of an exponential family random variable ν_θ with mean μ. We assume the variances of exponential family distributions used in this paper are bounded by a constant $V > 0$: $0 < V(\mu) \leq V < +\infty$. We have the following property of the KL divergence between exponential family distributions.

Proposition 3.1 (Harremoës [17]). *Let μ and μ' be the mean values of two exponential family distributions. The Kullback-Leibler divergence between them can be calculated as follows:*

$$kl(\mu, \mu') = \int_{\mu}^{\mu'} \frac{x - \mu}{V(x)} dx. \tag{3.2}$$

Based on Proposition 3.1, we can also verify the following properties.

Proposition 3.2 (Jin et al. [19]). *For all μ and μ', we have*

$$kl(\mu, \mu') \geq (\mu - \mu')^2/(2V). \tag{3.3}$$

In addition, for $\epsilon > 0$ and $\mu \leq \mu' - \epsilon$, we can obtain that

$$kl(\mu, \mu') \geq kl(\mu, \mu' - \epsilon) \quad \text{and} \quad kl(\mu, \mu') \leq kl(\mu - \epsilon, \mu'). \tag{3.4}$$

Exponential families cover many of the most common distributions used in practice such as Bernoulli, exponential, Gamma, and Gaussian distributions. In particular, for two Gaussian distributions with the same known variance σ^2 but different means μ and μ', we can choose $V(\cdot) = \sigma^2$, and it holds that $\mathrm{kl}(\mu, \mu') = (\mu - \mu')^2/(2\sigma^2)$. For two Bernoulli distributions with means μ and μ' respectively, the variance upper bound is set as $V = 1/4$. Thus we can recover the result in Proposition 3.1 as $\mathrm{kl}(\mu, \mu') = \mu \log(\mu/\mu') + (1 - \mu) \log((1 - \mu)/(1 - \mu'))$. For exponential and Gamma distributions, it suffices to ensure the variance id bounded as long as we assume the mean value is bounded.

The definition of one-dimensional exponential family in our paper is $p_\theta(x) = \exp(x\theta - b(\theta) + c(x))$, which is the same as that used by Garivier and Cappé [15], Harremoës [17], Jin et al. [19], Ménard and Garivier [32] as well as Cappé et al. [12]. The one-dimensional exponential family considered in Korda et al. [25] ($p_\theta(x) = \exp(T(x)\theta - b(\theta) + c(x))$) is more general than that in the aforementioned papers (see page 4 in Korda et al. [25]). Lai and Robbins [26] considers parametric distributions that satisfies some mild conditions, which is also more general than ours. Moreover, Cappé et al. [12] also considered the general reward distributions supported in $[0, 1]$, which is not compatible to ours.

4 Thompson Sampling for Exponential Family Reward Distributions

We present a general variant of Thompson sampling for exponential family rewards in Algorithm 1, named as ExpTS. At round t, ExpTS maintains an estimate of a sampling distribution for each arm, denoted as $\mathcal{P}$. The algorithm generates a sample parameter $\theta_i(t)$ for each arm i independently from their sampling distribution and chooses the arm that attains the largest sample parameter. For each arm $i \in [K]$, the sampling distribution $\mathcal{P}$ is usually defined as a function of the total number of pulls $T_i(t)$ and the empirical average reward $\widehat{\mu}_i(t)$. After pulling the chosen arm, the algorithm updates $T_i(t)$ and $\widehat{\mu}_i(t)$ for each arm based on the reward r_t it receives and proceeds to the next round.

Since we study the frequentist regret bound of Algorithm 1, ExpTS is not restricted as a Bayesian method. It has been shown [1, 20, 24, 36] that the sampling distribution does not have to be a posterior distribution derived from a pre-defined prior distribution. Therefore, we call $\mathcal{P}$ the sampling distribution instead of the posterior distribution as in Bayesian regret analysis of Thompson sampling [33, 11]. To obtain the finite-time regret bound of ExpTS for exponential family rewards, we will discuss the choice of a general sampling distribution and a new proof technique.

Algorithm 1 Exponential Family Thompson Sampling (ExpTS)

Input: Arm set $[K]$
Initialization: Play each arm once and set $T_i(K) = 1$; let $\widehat{\mu}_i(K)$ be the observed reward of playing arm i
for $t = K+1, K+2, \cdots$ **do**
  For all $i \in [K]$, sample $\theta_i(t)$ independently from $\mathcal{P}(\widehat{\mu}_i(t), T_i(t))$
  Play arm $A_t = \arg\max_{i \in [K]} \theta_i(t)$ and observe the reward r_t
  For all $i \in [K]$, update the mean reward estimator and the number of pulls:

$$\hat{\mu}_i(t) = \frac{T_i(t-1) \cdot \widehat{\mu}_i(t-1) + r_t \mathbb{1}\{i = A_t\}}{T_i(t-1) + \mathbb{1}\{i = A_t\}}, \quad T_i(t) = T_i(t-1) + \mathbb{1}\{i = A_t\}$$

end for

4.1 Challenges in Regret Analysis for Exponential Family Bandits

Before we choose a specific sampling distribution $\mathcal{P}$ for ExpTS, we first discuss the main challenges in the finite-time regret analysis of Thompson sampling, which is the main motivation for our design of $\mathcal{P}$ in the next subsection.

Under-Estimation of the Optimal Arm. Denote $\widehat{\mu}_{is}$ as the average reward of arm i after its s-th pull, $T_i(t)$ as the number of pulls of arm i at time t, and $\mathcal{P}(\widehat{\mu}_{is}, s)$ as the sampling distribution of arm i after its s-th pull. The regret of the algorithm contributed by pulling arm i is $\Delta_i \mathbb{E}[T_i(T)]$, where $T_i(T)$ is the total number of pulls of arm i. All existing analyses of finite-time regret bounds for TS [3–5, 20] decompose this regret term as $\Delta_i \mathbb{E}[T_i(T)] \le D_i + h_i(\Delta_i, T, \theta_i(1), \ldots, \theta_i(T))$, where $h_i()$ is a quantity characterizing the over-estimation of arm i which can be easily dealt with by some concentration properties of the sampling distribution (see Lemma A.3 for more details). The term D_i characterizes the under-estimation of the optimal arm 1, which is usually bounded as follows.

$$D_i = \Delta_i \sum_{s=1}^{T} \mathbb{E}_{\widehat{\mu}_{1s}} \left[\frac{1}{G_{1s}(\epsilon)} - 1 \right], \tag{4.1}$$

where $G_{1s}(\epsilon) = 1 - F_{1s}(\mu_1 - \epsilon)$, F_{1s} is the CDF of the sampling distribution $\mathcal{P}(\widehat{\mu}_{1s}, s)$, and $\epsilon = \Theta(\Delta_i)$. In other words, $G_{1s}(\epsilon) = \mathbb{P}(\theta_1(t) > \mu_1 - \epsilon)$ is the probability that the best arm will *not be under-estimated* from the mean reward by a margin ϵ. Furthermore, we can interpret the quantity in (4.1) as the result of a union bound indicating how many samples TS requires to ensure that at least one sample of the best arm $\{\theta_1(t)\}_{t=1}^T$ is larger than $\mu_1 - \epsilon$. If $G_{1s}(\epsilon)$ is too small, arm 1 could be significantly under-estimated, and thus D_i will be unbounded. In fact, as shown in Lemma 5 by Jin et al. [20], for MAB with Gaussian rewards, TS using Gaussian posteriors will unavoidably suffer from the lower bound of $K \cdot D_i = \Omega(\sqrt{KT \log T})$.

To address the above issue, we introduce a lower confidence bound for measuring the under-estimation problem. We use a new decomposition of the regret that bounds D_i with the following term

$$\Delta_i \sum_{s=1}^{T} \mathbb{E}_{\widehat{\mu}_{1s}} \left[\left(\frac{1}{G_{1s}(\epsilon)} - 1 \right) \cdot \mathbb{1}\{\widehat{\mu}_{1s} \ge Low_s\} \right], \tag{4.2}$$

where Low_s is a lower confidence bound of $\widehat{\mu}_{1s}$. Intuitively, due to the concentration of arm 1's rewards, the probability of $\widehat{\mu}_{1s} \le Low_s$ is very small. Thus, even when $G_{1s}(\epsilon)$ is small, the overall regret can be well controlled.

In the regret analysis of TS, we can bound (4.2) from two facets: (1) the lower confidence bound can be proved using the concentration property of the reward distribution; and (2) the term $G_{1s}(\epsilon) = \mathbb{P}(\theta_1(t) > \mu_1 - \epsilon)$ can be upper bounded by the anti-concentration property for the sampling distribution $\mathcal{P}$. To achieve an optimal regret, one needs to carefully balance the interplay between these two bounds. For a specific reward distribution (e.g., Gaussian, Bernoulli) as is studied by Agrawal and Goyal [5], Jin et al. [20], there are already tight anti-concentration inequalities for the reward distribution, and thus the lower confidence bound is tight. Therefore, by choosing Gaussian or Bernoulli as the prior (which leads to a Gaussian or Beta sampling distribution $\mathcal{P}$), we can use existing anti-concentration bounds for Gaussian [2, Formula 7.1.13] or Beta [18, Prop. A.4] distributions to obtain a tight bound of $G_{1s}(\epsilon)$.

In this paper, we study the general exponential family of reward distributions, which has no closed form. Thus we cannot obtain a tight concentration bound for $\hat{\mu}_{1s}$ as in special cases such as Gaussian or Bernoulli rewards. This increases the hardness of tightly bounding term (4.2) and it is imperative for us to design a sampling distribution $\mathcal{P}$ with a tight anti-concentration bound that can carefully control $G_{1s}(\epsilon)$ without any knowledge of the closed form distribution of the average reward $\hat{\mu}_{1s}$. Due to the generality of exponential family distributions, it is challenging and nontrivial to find such a sampling distribution to obtain a tight finite-time regret bound.

4.2 Sampling Distribution Design in Exponential Family Bandits

In this subsection, we show how to choose a sampling distribution $\mathcal{P}$ that has a tight anti-concentration bound to overcome the under-estimation of the optimal arm and concentration bound to overcome the over-estimation of the suboptimal arms.

For the simplicity of notation, we denote $\mathcal{P}(\mu, n)$ as the sampling distribution, where μ and n are some input parameters. In particular, for ExpTS, we will choose $\mu = \widehat{\mu}_i(t)$ and $n = T_i(t)$ for arm $i \in [K]$ at round t. We define $\mathcal{P}(\mu, n)$ as a distribution with PDF

$$f(x; \mu, n) = 1/2|(nb_n \cdot \mathrm{kl}(\mu, x))'|e^{-nb_n \cdot \mathrm{kl}(\mu,x)} = \frac{nb_n \cdot |x-\mu|}{2V(x)} e^{-nb_n \cdot \mathrm{kl}(\mu,x)}, \tag{4.3}$$

where $(\mathrm{kl}(\mu, x))'$ denotes the derivative of $\mathrm{kl}(\mu, x)$ with respect to x, and b_n is a function of n and will be chosen later.

We assume the reward is supported in $[R_{\min}, R_{\max}]$. Note that $R_{\min} = 0$, and $R_{\max} = 1$ for Bernoulli rewards, and $R_{\min} = -\infty$, and $R_{\max} = \infty$ for Gaussian rewards. Let $p(x)$ and $q(x)$ be the density functions of two exponential family distributions with mean values μ_p and μ_q respectively. By the definition in Section 3, we have $\mathrm{kl}(\mu_p, \mu_q) = \mathrm{KL}(p(x), q(x)) = \int_{R_{\min}}^{R_{\max}} p(x) \log \frac{p(x)}{q(x)} \mathrm{d}x$.

Proposition 4.1. *If the mean reward of $q(x)$ is equal to the maximum value in its support, i.e., $\mu_q = R_{\max}$, we will have $\mathrm{kl}(\mu, R_{\max}) = \infty$ for any $\mu < R_{\max}$.*

Proof. First consider the case that $R_{\max} < \infty$. Since the mean value concentrates on the maximum value, we must have $q(x) = 0$ for all $x < R_{\max}$, which immediately implies $\mathrm{kl}(\mu, R_{\max}) = \infty$ for any $\mu < R_{\max}$. For the case that $R_{\max} = \infty$, from (3.3) and the assumption that $V < \infty$, we also have $\mathrm{kl}(\mu, \infty) = (\infty - \mu)^2/V = \infty$. □

Similarly, we can also prove that $\mathrm{kl}(\mu, R_{\min}) = \infty$ for $\mu > R_{\min}$. Based on these properties, we can easily verify that a sample from the proposed sampling distribution $\theta \sim \mathcal{P}$ has the following tail bounds: for $z \in [\mu, R_{\max})$, it holds that

$$\mathbb{P}(\theta \geq z) = \int_z^{R_{\max}} f(x; \mu, n) \mathrm{d}x = -1/2 e^{-nb_n \cdot \mathrm{kl}(\mu,x)} \Big|_z^{R_{\max}} = 1/2 e^{-nb_n \cdot \mathrm{kl}(\mu,z)}, \tag{4.4}$$

and for $z \in (R_{\min}, \mu]$, it holds that

$$\mathbb{P}(\theta \leq z) = \int_{R_{\min}}^z f(x; \mu, n) \mathrm{d}x = 1/2 e^{-nb_n \cdot \mathrm{kl}(\mu,x)} \Big|_{R_{\min}}^z = 1/2 e^{-nb_n \cdot \mathrm{kl}(\mu,z)}. \tag{4.5}$$

Note that $\int_{R_{\min}}^{R_{\max}} f(x; \mu, n) \mathrm{d}x = \int_{R_{\min}}^{\mu} f(x; \mu, n) \mathrm{d}x + \int_{\mu}^{R_{\max}} f(x; \mu, n) \mathrm{d}x = 1$, which indicates the PDF of $\mathcal{P}$ is well-defined.

Intuition for the Design of the Sampling Distribution. The tail bounds in (4.4) and (4.5) provide proper anti-concentration and concentration bounds for the sampling distribution $\mathcal{P}$ as long as we have corresponding lower and upper bounds of $e^{-nb_n \cdot \mathrm{kl}(\mu,z)}$. When n is large, we will choose b_n to be close to 1, and thus (4.4) and (4.5) ensure that the sample of the corresponding arm concentrates in the interval $(\mu - \epsilon, \mu + \epsilon)$ with an exponentially small probability $e^{-n\mathrm{kl}(\mu-\epsilon,\mu+\epsilon)}$, which is crucial for achieving a tight finite-time regret.

How to Sample from $\mathcal{P}$. We show that sampling from $\mathcal{P}$ is tractable when the CDF of $\mathcal{P}$ is invertible. In particular, according to (4.4) and (4.5), the CDF of $\mathcal{P}(\mu, n)$ is

$$F(x) = \begin{cases} 1 - 1/2 e^{-nb_n \cdot \mathrm{kl}(\mu,x)} & x \geq \mu, \\ 1/2 e^{-nb_n \cdot \mathrm{kl}(\mu,x)} & x \leq \mu. \end{cases}$$

Table 2: Comparisons of different algorithms on K-armed bandits over a horizon T. For any algorithm, *Asym. Opt* is the indicator whether it is asymptotically optimal, *minimax ratio* is the scaling of its worst-case regret w.r.t. the minimax optimal regret $O(\sqrt{VKT})$, *sub-UCB* indicates whether it satisfies the sub-UCB criteria, and *Anytime* indicates whether it needs the knowledge of the horizon length T in advance.

Algorithm	Reward Type	Asym. Opt	Finite-Time Regret		Anytime	References
			Minimax Ratio	Sub-UCB		
MOSS	$[0, 1]$	no	1	no	no	[6]
Anytime MOSS	$[0, 1]$	no	1	no	yes	[14]
KL-UCB^{++}	Exponential Family	yes	1	no	no	[32]
OCUCB	sub-Gaussian	no	$\sqrt{\log\log T}$	yes	yes	[28]
AdaUCB	Gaussian	yes	1	yes	no	[29]
MS	sub-Gaussian	yes	$\sqrt{\log K}$	yes	yes	[10]
ExpTS	Exponential Family	yes	$\sqrt{\log K}$	yes	yes	This paper
ExpTS^{+}	Exponential Family	yes	1	no	yes	This paper

To sample from $\mathcal{P}(\mu, n)$, we can first pick y uniformly random from $[0, 1]$. Then, for $y \geq 1/2$, we solve the equation $y = 1 - 1/2e^{-nb_n\cdot \mathrm{kl}(\mu,x)}$ for x ($x \geq \mu$), which is equivalent to solving $\log(1/(2(1-y)))/(nb_n) = \mathrm{kl}(\mu, x)$. For $y \leq 1/2$, we solve the equation $y = 1/2e^{-nb_n\cdot \mathrm{kl}(\mu,x)}$ for x ($x \leq \mu$), which is equivalent to solving $\log(1/(2y))/(nb_n) = \mathrm{kl}(\mu, x)$. If $b(\theta)$ is reversible and the mapping $\theta \mapsto \mu$ is given[4], then according to (3.1), $\mathrm{kl}(\mu, x)$ is also reversible for x. We can obtain an exact sample from distribution $\mathcal{P}$ by solving $\mathrm{kl}(\mu, x) = \log(1/(2y))/(nb_n)$. Alternatively, we can also use approximate sampling methods such as Monte Carlo Markov Chain and Hastings-Metropolis [25] or gradient based Langevin Monte Carlo [35] to obtain samples from the target distribution.

4.3 Regret Analysis of ExpTS for Exponential Family Rewards

Now we present the regret bound of ExpTS for general exponential family bandits. The sampling distribution used in Algorithm 1 is defined in (4.3).

Theorem 4.2. *Let $b_n = (n-1)/n$. Let $\mathcal{P}$ be the sampling distribution defined in Section 4.2. There exist universal constants $C_0, C_1 > 0$ such that the regret of Algorithm 1 satisfies*

$$R_\mu(T) \leq C_0\bigg(\sum_{i\in[K]:\Delta_i>\lambda} \Delta_i + \frac{V\log(T\Delta_i^2/V)}{\Delta_i}\bigg) + \max_{i\in[K],\Delta_i\leq\lambda} \Delta_i \cdot T, \tag{4.6}$$

$$R_\mu(T) \leq C_1\bigg(\sum_{i=2}^{K} \Delta_i + \sqrt{VKT\log K}\bigg), \tag{4.7}$$

where $\lambda \geq 16\sqrt{V/T}$, and also satisfies the following asymptotic bound simultaneously:

$$\lim_{T\to\infty} \frac{R_\mu(T)}{\log T} = \sum_{i=2}^{K} \frac{\Delta_i}{kl(\mu_i, \mu_1)}. \tag{4.8}$$

Remark 4.3. *Similar to the argument by Auer and Ortner [7], we can see that the logarithm term in* (4.6) *is the main term for suitable λ. For instance, if we choose $\lambda = 16\sqrt{V/T}$, we will have $\max_{i\in[K],\Delta_i\leq\lambda} \Delta_i T \leq \sqrt{VT}$, which is in the order of $O(V/\Delta_i)$ due to $\Delta_i \leq \lambda$. Thus it is obvious to see that the regret in* (4.6) *satisfies the sub-UCB criteria.*

It is worth highlighting that ExpTS is an anytime algorithm and simultaneously satisfies the sub-UCB criteria in (1.2), the minimax optimal regret up to a factor $\sqrt{\log K}$, and the asymptotically optimal regret. ExpTS is also the first Thompson sampling algorithm that provides finite-time regret bounds for exponential family of rewards. Compared with state-of-the-art MAB algorithms listed in Table 2, ExpTS is comparable to the best known UCB algorithms that work for exponential family of reward distributions and no algorithms can dominate ExpTS. In particular, compared with MS [10] and

[4]This is true for distributions such as Gaussian with known variance, exponential distribution, and Bernoulli.

OCUCB [28], ExpTS is asymptotically optimal for exponential family of rewards, while MS is only asymptotically optimal for sub-Gaussian rewards and OCUCB is not asymptotically optimal. We note that Exponential Family does not cover the sub-Gaussian rewards. However, since we only use the tail bound to approximate the reward distribution, ExpTS can also be extended to solve sub-Gaussian reward bandits, which we leave as a future open direction.

4.4 Simple Variants for Gaussian and Bernoulli Reward Distributions

The choice of $\mathcal{P}$ in (4.3) seems complicated for a general exponential family reward distribution, even though we only need the sampling distribution to satisfy a nice tail bound derived from this reward distribution. When the reward distribution has a closed form such as Gaussian and Bernoulli distributions, we can replace $\mathcal{P}$ with the posterior in standard Thompson sampling and obtain the asymptotic and finite-time regrets in the previous section.

Theorem 4.4. *If the reward follows a Gaussian distribution with a known variance V, we can set the sampling distribution in Algorithm 1 as $\mathcal{N}(\widehat{\mu}_i(t), V/T_i(t))$. The resulting algorithm (denoted as Gaussian-TS) enjoys the same regret bounds presented in Theorem 4.2.*

Remark 4.5. *Lemma 5 in Jin et al. [20] shows for Gaussian rewards, Gaussian-TS has a regret bound at least $\Omega(\sqrt{VKT\log T})$ if the standard regret decomposition discussed in Section 4.1 is adopted in the proof [5, 30, 20]. With our new regret decomposition and the lower confidence bound introduced in* (4.2)*, we improve it to $O(\sqrt{VKT\log K})$.*

Jin et al. [20] also shows that their algorithms MOTS/MOTS-$\mathcal{J}$ can overcome the under-estimation issue of (4.1)*. However, they are either at the cost of sacrificing the asymptotic optimality or not generalizable to exponential family bandits. In specific, (1) For Gaussian rewards, MOTS [20] enlarges the variance of Gaussian posterior by a factor of $1/\rho$, where $\rho \in (0,1)$, which loses the asymptotic optimality by a factor of $1/\rho$ resultantly. (2) For Gaussian rewards, MOTS-$\mathcal{J}$ [20] introduces the Rayleigh posterior to overcome the under-estimation while maintaining the asymptotic optimality. However, it is not clear whether the idea can be generalized to exponential family rewards. Interestingly, their experimental results show that compared with Rayleigh posterior, Gaussian posterior actually has a smaller regret empirically. Therefore, to use a Gaussian sampling distribution, the new regret decomposition and the novel lower confidence bound in our paper is a better way to overcome the under-estimation issue of Gaussian-TS.*

Theorem 4.6. *If the reward distribution is Bernoulli, we can set the sampling distribution $\mathcal{P}$ in Algorithm 1 as Beta posterior $\mathcal{B}(S_i(t)+1, T_i(t)-S_i(t)+1)$, where $S_i(t)$ is the number of successes among the $T_i(t)$ plays of arm i. We denote the resulting algorithm as Bernoulli-TS, which enjoys the same regret bounds as in Theorem 4.2.*

Agrawal and Goyal [5] proved that for Bernoulli rewards, Thompson sampling with Beta posterior is asymptotically optimal and has a worst-case regret in the order of $O(\sqrt{KT\log T})$. Our regret analysis improves the worst-case regret to $O(\sqrt{KT\log K})$. They also proved that Gaussian-TS applied to the Bernoulli reward setting has a regret $O(\sqrt{KT\log K})$. However, no asymptotic regret was guaranteed in this setting.

5 Minimax Optimal Thompson Sampling for Exponential Family Rewards

In this section, in order to remove the extra logarithm term in the worst-case regret of ExpTS, we introduce a new sampling distribution that adds a greedy exploration step to the sampling distribution used in ExpTS. Specifically, the new algorithm ExpTS$^+$ is the same as ExpTS but uses a new sampling distribution $\mathcal{P}^+(\mu, n)$. A sample θ is generated from $\mathcal{P}^+(\mu, n)$ in the following way: $\theta = \mu$ with probability $1 - 1/K$ and $\theta \sim \mathcal{P}(\mu, n)$ with probability $1/K$.

Over-Estimation of Sub-Optimal Arms. We first elaborate the over-estimation issue of sub-optimal arms, which results in the extra $\sqrt{\log K}$ term in the worst-case regret of Thompson sampling. To explain, suppose that the sample of each arm i has a probability $p = \mathbb{P}(\theta_i(t) \geq \theta_1(t))$ to become larger than the sample of arm 1. Note that when this event happens, the algorithm chooses the wrong arm and thus incurs a regret. Intuitively, the probability of making a mistake will be $K-1$ times larger due to the union bound over $K-1$ sub-optimal arms, which leads to an additional $\sqrt{\log K}$ factor in the worst-case regret. To reduce the probability $\mathbb{P}(\theta_i(t) \geq \theta_1(t))$, ExpTS$^+$ adds a greedy step that chooses the ExpTS sample with probability $1/K$ and chooses the arm with the largest empirical average reward with probability $1 - 1/K$. Then we can prove that for sufficiently large s, with high probability we have $\hat{\mu}_{is} < \theta_1(t)$ and in this case it holds that $\mathbb{P}(\theta_i(t) \geq \theta_1(t)) = p/K$. Thus the extra factor $\sqrt{\log K}$ in regret is removed.

In specific, we have the following theorem showing that ExpTS$^+$ is asymptotically optimal and minimax optimal simultaneously.

Theorem 5.1. *Let $b_n = (n-1)/n$. There exists a constant $C_1 > 0$ such that ExpTS$^+$ satisfies*

$$R_\mu(T) \leq C_1\bigg(\sum_{i=2}^{K} \Delta_i + \sqrt{VKT}\bigg), \quad \textit{and} \quad \lim_{T\to\infty} \frac{R_\mu(T)}{\log T} = \sum_{i=2}^{K} \frac{\Delta_i}{kl(\mu_i, \mu_1)}.$$

This is the first time that the Thompson sampling algorithm achieves the minimax and asymptotically optimal regret for exponential family of reward distributions. Moreover, ExpTS$^+$ is also an anytime algorithm since it does not need to know the horizon T in advance.

Remark 5.2 (Sub-Gaussian Rewards). *In the proof of Theorem 5.1, we do not need the strict form of the PDF of the empirical mean reward $\widehat{\mu}_{is}$, but only need the maximal inequality (Lemma H.1). This means that the proof can be straightforwardly extended to sub-Gaussian reward distributions, where similar maximal inequality holds [21].*

It is worth noting that MOTS proposed by [20] (Thompson sampling with a clipped Gaussian posterior) also achieves the minimax optimal regret for sub-Gaussian rewards, but it can not keep the asymptotic optimality simultaneously with the same algorithm parameters. In particular, to achieve the minimax optimality, MOTS will have an additional $1/\rho$ factor in the asymptotic regret with $0 < \rho < 1$. Moreover, different from ExpTS$^+$, MOTS is only designed for fixed T setting and thus is not an anytime algorithm.

Remark 5.3 (Gaussian and Bernoulli Rewards). *Following the idea in Section 4.4, we can derive new algorithms Gaussian-TS$^+$ and Bernoulli-TS$^+$ for Gaussian and Bernoulli rewards by replacing the sampling distribution in ExpTS$^+$. However, the posterior distribution does not fully satisfy the properties shown in Section 4.2. In particular, the factor $b_n < 1$ in Theorem 5.1 is an essential requirement for the asymptotic analyses whereas the posterior distribution does not have this factor. Due to these extra challenges, the proof techniques used for Theorem 5.1 can not be directly applied to these two new algorithms, and it is interesting to further investigate whether they are simultaneously minimax and asymptotically optimal.*

6 Conclusions

We studied Thompson sampling for exponential family of reward distributions. We proposed the ExpTS algorithm and proved it satisfies the sub-UCB criteria for problem-dependent finite-time regret, as well as achieves the asymptotic optimality and the minimax optimality up to a factor of $\sqrt{\log K}$ for exponential family rewards. Furthermore, we proposed a variant of ExpTS, dubbed ExpTS$^+$, that adds a greedy exploration step to balance between the sample generated in ExpTS and the empirical mean reward for each arm. We proved that ExpTS$^+$ is simultaneously minimax and asymptotically optimal. We also extended our proof techniques to standard Thompson sampling with common posterior distributions and improved existing results. This work is mainly focused on the theoretical optimality of Thompson sampling type algorithms. It would be an interesting future direction to investigate the empirical performance of ExpTS and ExpTS$^+$.

Acknowledgement

We thank the anonymous reviewers and the area chair for their helpful comments. This research is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG-PhD/2021-01-004[T]), by ASTAR, Singapore under Grant A19E3b0099, and by Bren Named Chair Professorship at Caltech. In particular, T. Jin is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG-PhD/2021-01-004[T]). X. Xiao is supported by ASTAR, Singapore under Grant A19E3b0099. A. Anandkumar is partially supported by Bren Named Chair Professorship at Caltech. The views and conclusions contained in this paper are those of the authors and should not be interpreted as representing any funding agencies.

References

[1] Marc Abeille and Alessandro Lazaric. Linear thompson sampling revisited. In *Artificial Intelligence and Statistics*, pages 176–184. PMLR, 2017. (p. 5.)

[2] Milton Abramowitz and Irene A Stegun. *Handbook of mathematical functions with formulas, graphs, and mathematical tables*, volume 55. US Government printing office, 1964. (pp. 6 and 41.)

[3] Shipra Agrawal and Navin Goyal. Analysis of thompson sampling for the multi-armed bandit problem. In *Conference on learning theory*, pages 39–1, 2012. (pp. 2, 4, and 6.)

[4] Shipra Agrawal and Navin Goyal. Further optimal regret bounds for thompson sampling. In *Artificial intelligence and statistics*, pages 99–107, 2013. (pp. 2, 3, and 4.)

[5] Shipra Agrawal and Navin Goyal. Near-optimal regret bounds for thompson sampling. *Journal of the ACM (JACM)*, 64(5):30, 2017. (pp. 1, 2, 3, 4, 6, 9, 27, and 29.)

[6] Jean-Yves Audibert and Sébastien Bubeck. Minimax policies for adversarial and stochastic bandits. In *COLT*, pages 217–226, 2009. (pp. 2, 3, 4, and 8.)

[7] Peter Auer and Ronald Ortner. Ucb revisited: Improved regret bounds for the stochastic multi-armed bandit problem. *Periodica Mathematica Hungarica*, 61(1-2):55–65, 2010. (p. 8.)

[8] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 2002. (pp. 1 and 2.)

[9] Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002. (p. 1.)

[10] Jie Bian and Kwang-Sung Jun. Maillard sampling: Boltzmann exploration done optimally. *arXiv preprint arXiv:2111.03290*, 2021. (pp. 2, 4, and 8.)

[11] Sébastien Bubeck and Che-Yu Liu. Prior-free and prior-dependent regret bounds for thompson sampling. In *Advances in Neural Information Processing Systems*, pages 638–646, 2013. (pp. 4 and 5.)

[12] Olivier Cappé, Aurélien Garivier, Odalric-Ambrym Maillard, Rémi Munos, and Gilles Stoltz. Kullback-leibler upper confidence bounds for optimal sequential allocation. *The Annals of Statistics*, pages 1516–1541, 2013. (pp. 4 and 5.)

[13] Olivier Chapelle and Lihong Li. An empirical evaluation of thompson sampling. In *Advances in neural information processing systems*, pages 2249–2257, 2011. (p. 2.)

[14] Rémy Degenne and Vianney Perchet. Anytime optimal algorithms in stochastic multi-armed bandits. In *International Conference on Machine Learning*, pages 1587–1595. PMLR, 2016. (pp. 4 and 8.)

[15] Aurélien Garivier and Olivier Cappé. The kl-ucb algorithm for bounded stochastic bandits and beyond. In *Proceedings of the 24th annual conference on learning theory*, pages 359–376, 2011. (pp. 4 and 5.)

[16] Aurélien Garivier, Hédi Hadiji, Pierre Menard, and Gilles Stoltz. Kl-ucb-switch: optimal regret bounds for stochastic bandits from both a distribution-dependent and a distribution-free viewpoints. *arXiv preprint arXiv:1805.05071*, 2018. (pp. 1 and 2.)

[17] Peter Harremoës. Bounds on tail probabilities in exponential families. *arXiv preprint arXiv:1601.05179*, 2016. (pp. 4 and 5.)

[18] Emil Jeřábek. Dual weak pigeonhole principle, boolean complexity, and derandomization. *Annals of Pure and Applied Logic*, 129(1-3):1–37, 2004. (pp. 6 and 27.)

[19] Tianyuan Jin, Jing Tang, Pan Xu, Keke Huang, Xiaokui Xiao, and Quanquan Gu. Almost optimal anytime algorithm for batched multi-armed bandits. In *International Conference on Machine Learning*, pages 5065–5073. PMLR, 2021. (p. 5.)

[20] Tianyuan Jin, Pan Xu, Jieming Shi, Xiaokui Xiao, and Quanquan Gu. MOTS: Minimax Optimal Thompson Sampling. In *International Conference on Machine Learning*, pages 5074–5083. PMLR, 2021. (pp. 2, 3, 4, 5, 6, 9, and 10.)

[21] Tianyuan Jin, Pan Xu, Xiaokui Xiao, and Quanquan Gu. Double explore-then-commit: Asymptotic optimality and beyond. In *Conference on Learning Theory*, pages 2584–2633. PMLR, 2021. (pp. 4 and 10.)

[22] Emilie Kaufmann. On bayesian index policies for sequential resource allocation. *arXiv preprint arXiv:1601.01190*, 2016. (pp. 2 and 4.)

[23] Emilie Kaufmann, Nathaniel Korda, and Rémi Munos. Thompson sampling: An asymptotically optimal finite-time analysis. In *International conference on algorithmic learning theory*, pages 199–213. Springer, 2012. (p. 4.)

[24] Wonyoung Kim, Gi-soo Kim, and Myunghee Cho Paik. Doubly robust thompson sampling with linear payoffs. *Advances in Neural Information Processing Systems*, 34, 2021. (p. 5.)

[25] Nathaniel Korda, Emilie Kaufmann, and Remi Munos. Thompson sampling for 1-dimensional exponential family bandits. In *Advances in neural information processing systems*, pages 1448–1456, 2013. (pp. 2, 3, 4, 5, and 8.)

[26] Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985. (pp. 2, 3, and 5.)

[27] Tor Lattimore. Optimally confident ucb: Improved regret for finite-armed bandits. *arXiv preprint arXiv:1507.07880*, 2015. (p. 4.)

[28] Tor Lattimore. Regret analysis of the finite-horizon gittins index strategy for multi-armed bandits. In *Conference on Learning Theory*, pages 1214–1245, 2016. (pp. 4, 8, and 9.)

[29] Tor Lattimore. Refining the confidence level for optimistic bandit strategies. *The Journal of Machine Learning Research*, 19(1):765–796, 2018. (pp. 1, 2, 4, and 8.)

[30] Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020. (pp. 3, 9, and 16.)

[31] Odalric-Ambrym Maillard, Rémi Munos, and Gilles Stoltz. A finite-time analysis of multi-armed bandits problems with kullback-leibler divergences. In *Proceedings of the 24th annual Conference On Learning Theory*, pages 497–514, 2011. (p. 4.)

[32] Pierre Ménard and Aurélien Garivier. A minimax and asymptotically optimal algorithm for stochastic bandits. In *International Conference on Algorithmic Learning Theory*, pages 223–237, 2017. (pp. 2, 4, 5, 8, and 41.)

[33] Daniel Russo and Benjamin Van Roy. Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39(4):1221–1243, 2014. (pp. 4 and 5.)

[34] Siwei Wang and Wei Chen. Thompson sampling for combinatorial semi-bandits. In *International Conference on Machine Learning*, pages 5114–5122, 2018. (p. 2.)

[35] Pan Xu, Hongkai Zheng, Eric V Mazumdar, Kamyar Azizzadenesheli, and Animashree Anandkumar. Langevin monte carlo for contextual bandits. In *International Conference on Machine Learning*, pages 24830–24850. PMLR, 2022. (p. 8.)

[36] Tong Zhang. Feel-good thompson sampling for contextual bandits and reinforcement learning. *arXiv preprint arXiv:2110.00871*, 2021. (p. 5.)

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes] All the proofs are provided in the appendix.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Proof of the Finite-Time Regret Bound of ExpTS

In this section, we prove the finite-time regret bound of ExpTS presented in Theorem 4.2. Specifically, we prove the sub-UCB property of ExpTS in (4.6) and the nearly minimax optimal regret of ExpTS in (4.7).

A.1 Proof of the Main Results

We first focus on bounding the number of pulls of arm i for the case that $\Delta_i > 16\sqrt{V/T}$. We start with the decomposition. Note that due to the warm start of Algorithm 1, each arm has been pulled once in the first K steps. For any $\epsilon > 8\sqrt{V/T}$, define event $E_{i,\epsilon}(t) = \{\theta_i(t) \le \mu_1 - \epsilon\}$, for all $i \in [K]$, which indicates that the estimate of arm i at time step t is smaller than the lower bound of the true mean reward of arm 1 ($\mu_1 - \epsilon \le \mu_1$). The expected number of times that Algorithm 1 plays arms i is bounded as follows.

$$\begin{aligned}\mathbb{E}[T_i(T)] &= 1 + \mathbb{E}\left[\sum_{t=K+1}^{T} \mathbb{1}\{A_t = i, E_{i,\epsilon}(t)\} + \sum_{t=K+1}^{T} \mathbb{1}\{A_t = i, E_{i,\epsilon}^c(t)\}\right] \\ &= 1 + \underbrace{\mathbb{E}\left[\sum_{t=K+1}^{T} \mathbb{1}\{A_t = i, E_{i,\epsilon}(t)\}\right]}_{A} + \underbrace{\mathbb{E}\left[\sum_{t=K+1}^{T} \mathbb{1}\{A_t = i, E_{i,\epsilon}^c(t)\}\right]}_{B}, \qquad \text{(A.1)}\end{aligned}$$

where E^c is the complement of an event E, $\epsilon > 8\sqrt{V/T}$ is an arbitrary constant, and we used the fact $T_i(T) = \sum_{t=1}^{T} \mathbb{1}\{A_t = i\}$. In what follows, we bound these terms individually.

Bounding Term A: Let us define

$$\alpha_s = \sup_{x \in [0, \mu_1 - \epsilon - R_{\min})} \mathrm{kl}(\mu_1 - \epsilon - x, \mu_1) \le 4\log(T/s)/s. \qquad \text{(A.2)}$$

We decompose the term $\mathbb{E}\left[\sum_{t=K+1}^{T} \mathbb{1}\{A_t = i, E_{i,\epsilon}(t)\}\right]$ by the following lemma.

Lemma A.1. *Let $M = \lceil 16V\log(T\epsilon^2/V)/\epsilon^2 \rceil$ and α_s be the same as defined in* (A.2)*. Then, there exists a universal constant $C_2 > 0$,*

$$\mathbb{E}\left[\sum_{t=K+1}^{T} \mathbb{1}\{A_t = i, E_{i,\epsilon}(t)\}\right] \le \sum_{s=1}^{M} \mathbb{E}\left[\left(\frac{1}{G_{1s}(\epsilon)} - 1\right) \cdot \mathbb{1}\{\widehat{\mu}_{1s} \in L_s\}\right] + \frac{C_2 V}{\epsilon^2},$$

where $G_{is}(\epsilon) = 1 - F_{is}(\mu_1 - \epsilon)$, F_{is} is the CDF of $\mathcal{P}(\widehat{\mu}_{is}, s)$, and $L_s = \big(\mu_1 - \epsilon - \alpha_s, R_{\max}\big]$.

The first term on the right hand side could be further bounded as follows.

Lemma A.2. *Let M, $G_{1s}(\epsilon)$, and L_s be the same as defined in Lemma A.1. Then there a universal constant $C_3 > 0$, such that*

$$\sum_{s=1}^{M} \mathbb{E}_{\widehat{\mu}_{1s}}\left[\left(\frac{1}{G_{1s}(\epsilon)}\right) \cdot \mathbb{1}\{\widehat{\mu}_{1s} \in L_s\}\right] \le \frac{C_3 \cdot V\log(T\epsilon^2/V)}{\epsilon^2}.$$

Combining Lemma A.1 and Lemma A.2 together, we have the upper bound of term A in (A.1).

$$A = \frac{(C_3 + C_2)\log(T\epsilon^2/V)}{\epsilon^2}.$$

Bounding Term B: To bound the second term in (A.1), we first prove the following lemma that bounds the number of time steps when the empirical average reward of arm i deviates from its mean value.

Lemma A.3. *Let $N = \min\{1/(1 - kl(\mu_i + \rho_i, \mu_1 - \epsilon)/\log(T\epsilon^2/V)), 2\}$. For any $\rho_i, \epsilon > 0$ that satisfies $\epsilon + \rho_i < \Delta_i$, then*

$$\mathbb{E}\left[\sum_{t=K+1}^{T} \mathbb{1}\{A_t = i, E_{i,\epsilon}^c(t)\}\right] \le 1 + \frac{2V}{\rho_i^2} + \frac{V}{\epsilon^2} + \frac{N\log(T\epsilon^2/V)}{kl(\mu_i + \rho_i, \mu_1 - \epsilon)}.$$

Applying Lemma A.3, we have the following bound for term B in (A.1).

$$\mathbb{E}\left[\sum_{t=K+1}^{T}\mathbb{1}\{A_t=i,E_{i,\epsilon}^c(t)\}\right]\le 1+\frac{2V}{\rho_i^2}+\frac{V}{\epsilon^2}+\frac{N\log(T\epsilon^2/V)}{\mathrm{kl}(\mu_i+\rho_i,\mu_1-\epsilon)}$$
$$\le 1+\frac{2V}{\rho_i^2}+\frac{V}{\epsilon^2}+\frac{4V\log(T\epsilon^2/V)}{(\Delta_i-\epsilon-\rho_i)^2},$$

where the last inequality is due to (3.3) and $N\le 2$.

Putting It Together: Substituting the bounds of terms A and B back into (A.1), we have

$$\mathbb{E}[T_i(T)]=1+\frac{4V\log(T\epsilon^2/V)}{(\Delta_i-\epsilon-\rho_i)^2}+\frac{2V}{\rho_i^2}+\frac{V}{\epsilon^2}+\frac{(C_3+C_2)V\log(T\epsilon^2/V)}{\epsilon^2}.$$

Let $\epsilon=\rho_i=\Delta_i/4$, we have

$$\mathbb{E}[T_i(T)]=1+\frac{(C_3+C_2+64)V\log(T\Delta_i^2/V)}{\Delta_i^2}.$$

Note that we have assumed $\Delta_i>16\sqrt{V/T}$ at the beginning of the proof. Therefore, there exists a universal constant $C_0>0$ such that

$$R_\mu(T)\le C_0\cdot\sum_{i\in[K]:\Delta_i>\lambda}O\bigg(\Delta_i+\frac{V\log(T\Delta_i^2/V)}{\Delta_i}\bigg)+\max_{i\in[K],\Delta_i\le\lambda}\Delta_i\cdot T,$$

for any $\lambda\ge 16\sqrt{V/T}$. By choosing $\lambda=16\sqrt{VK\log K/T}$, we obtain the following worst-case regret: $R_\mu(T)\le C_1\cdot\sqrt{VKT\log K}$ for some universal constant C_1. This completes the proof of the finite-time regret bounds of ExpTS.

A.2 Proof of Supporting Lemmas

In this subsection, we prove the lemmas used in the proof of our main results in this section.

A.2.1 Proof of Lemma A.1

Define $\mathcal{E}$ to be the event such that $\widehat{\mu}_{1s}\in L_s$ holds for all $s\in[T]$. The proof of Lemma A.1 needs the following lemma, which is used for bounding $\mathbb{P}(\mathcal{E}^c)$.

Lemma A.4. *Let $\epsilon>0$, $b\in[K]$ and $f(\epsilon)=\lceil 16V\log(T\epsilon^2/(bV))/\epsilon^2\rceil$. Assume $T\ge bf(\epsilon)$. Then, there exists a universal constant C_2 such that*

$$\mathbb{P}\big(\exists 1\le s\le f(\epsilon):\hat{\mu}_{1s}\le\mu_1-\epsilon,kl(\hat{\mu}_{1s},\mu_1)\ge 4\log(T/(bs))/s\big)\le\frac{C_2bV}{T\epsilon^2}.$$

The proof of Lemma A.4 could be found in Section G. Now, we are ready to prove Lemma A.1.

Proof of Lemma A.1. The indicator function can be decomposed based on $\mathcal{E}$, that is

$$\begin{aligned}&\mathbb{E}\left[\sum_{t=K+1}^{T}\mathbb{1}\{A_t=i,E_{i,\epsilon}(t)\}\right]\\&\le T\cdot\mathbb{P}(\mathcal{E}^c)+\mathbb{E}\left[\sum_{t=K+1}^{T}\big[\mathbb{1}\{A_t=i,E_{i,\epsilon}(t)\}\cdot\mathbb{1}\{\widehat{\mu}_{1T_i(t-1)}\in L_{T_i(t-1)}\}\big]\right]\\&\le\frac{C_2V}{T\epsilon^2}+\mathbb{E}\left[\sum_{t=K+1}^{T}\big[\mathbb{1}\{A_t=i,E_{i,\epsilon}(t)\}\cdot\mathbb{1}\{\widehat{\mu}_{1T_i(t-1)}\in L_{T_i(t-1)}\}\big]\right],\end{aligned}\tag{A.3}$$

where the second inequality is due to Lemma A.4 with $b=1$ and from the fact $\epsilon>8\sqrt{V/T}$, $T\ge f(\epsilon)$. Let $\mathcal{F}_t=\sigma(A_1,r_1,\cdots,A_t,r_t)$ be the filtration. Note that $\theta_i(t)$ is sampled from

$\mathcal{P}(\hat{\mu}_i(t-1), T_i(t-1))$. Recall the definition, we know that $\hat{\mu}_i(t-1) = \hat{\mu}_{is}$ as long as $s = T_i(t-1)$. By the definition of $G_{is}(x)$, it holds that

$$G_{1T_1(t-1)}(\epsilon) = \mathbb{P}(\theta_1(t) \geq \mu_1 - \epsilon \mid \mathcal{F}_{t-1}). \tag{A.4}$$

Consider two cases. **Case 1:** $t : T_1(t-1) \leq M$**.** The proof of this case is similar to that of [30, Theorem 36.2]. Let $A'_t = \arg\max_{i \neq 1} \theta_i(t)$. Then

$$\begin{aligned}
\mathbb{P}(A_t = 1 \mid \mathcal{F}_{t-1}) &\geq \mathbb{P}(\{\theta_1(t) \geq \mu_1 - \epsilon\} \cap \left\{A'_t = i, E_{i,\epsilon}(t)\right\} \mid \mathcal{F}_{t-1}) \\
&= \mathbb{P}(\theta_1(t) \geq \mu_1 - \epsilon \mid \mathcal{F}_{t-1}) \cdot \mathbb{P}(A'_t = i, E_{i,\epsilon}(t) \mid \mathcal{F}_{t-1}) \\
&\geq \frac{G_{1T_1(t-1)}(\epsilon)}{1 - G_{1T_1(t-1)}(\epsilon)} \cdot \mathbb{P}(A_t = i, E_{i,\epsilon}(t) \mid \mathcal{F}_{t-1}),
\end{aligned} \tag{A.5}$$

The first inequality is due to the fact that when both event $\{\theta_1(t) \geq \mu_1 - \epsilon\}$ and event $\{A'_t = i, E_{i,\epsilon}(t)\}$ hold, we must have $\{A_t = 1\}$. The first equality is due to $\theta_1(t)$ is conditionally independent of A'_t and $E_{i,\epsilon}(t)$ given $\mathcal{F}_{t-1}$. For the last inequality, let $C = \{A_t = i, E_{i,\epsilon}(t) \text{ occurs}\}$, $A = \{A'_t = i, E_{i,\epsilon}(t) \text{ occurs}\}$ and $B = \{\theta_1(t) \leq \mu_1 - \epsilon\}$. Then A and B are conditionally independent given $\mathcal{F}_{t-1}$. Besides, if C happens, then $A_t = i$ and $\theta_i(t) \leq \mu_1 - \epsilon$. This implies the $\theta_1(t) \leq \mu_1 - \epsilon$ (otherwise, we will have $A_t \neq i$). Therefore, if C happens, we must have $A'_t = i, E_{i,\epsilon}(t)$ occurs and $\theta_1(t) \leq \mu_1 - \epsilon$. Therefore, $C \subseteq A \cap B$ and

$$\begin{aligned}
&\mathbb{P}(A_t = i, E_{i,\epsilon}(t) \text{ occurs} \mid \mathcal{F}_{t-1}) \\
&\quad = \mathbb{P}(C \mid \mathcal{F}_{t-1}) \\
&\quad \leq \mathbb{P}(A \cap B \mid \mathcal{F}_{t-1}) \\
&\quad \leq \mathbb{P}(A \mid \mathcal{F}_{t-1}) \cdot \mathbb{P}(B \mid \mathcal{F}_{t-1}) \\
&\quad = \mathbb{P}(A'_t = i, E_{i,\epsilon}(t) \text{ occurs} \mid \mathcal{F}_{t-1}) \cdot \mathbb{P}(\theta_1(t) \leq \mu_1 - \epsilon \mid \mathcal{F}_{t-1}) \\
&\quad = \mathbb{P}(A'_t = i, E_{i,\epsilon}(t) \text{ occurs} \mid \mathcal{F}_{t-1}) \cdot \big(1 - \mathbb{P}(\theta_1(t) > \mu_1 - \epsilon \mid \mathcal{F}_{t-1})\big).
\end{aligned}$$

Note that from (A.4), $G_{1T_1(t-1)}(\epsilon) = \mathbb{P}(\theta_1(t) \geq \mu_1 - \epsilon \mid \mathcal{F}_{t-1})$. Therefore, the above equation implies the last inequality of (A.5). Therefore, we have

$$\begin{aligned}
\mathbb{E}\left[\sum_{t:T_1(t-1) \leq M} \mathbb{1}\{A_t = i, E_{i,\epsilon}(t)\}\right] &\leq \mathbb{E}\left[\sum_{t:T_1(t-1) \leq M} \left(\frac{1}{G_{1T_1(t-1)}(\epsilon)} - 1\right) \mathbb{P}(A_t = 1 \mid \mathcal{F}_{t-1})\right] \\
&= \mathbb{E}\left[\sum_{t:T_1(t-1) \leq M} \left(\frac{1}{G_{1T_1(t-1)}(\epsilon)} - 1\right) \mathbb{1}\{A_t = 1\}\right] \\
&\leq \mathbb{E}\left[\sum_{s=1}^{M} \left(\frac{1}{G_{1s}(\epsilon)} - 1\right)\right].
\end{aligned} \tag{A.6}$$

The first inequality is from (A.5). The first equality is due to $\mathbb{E}[\mathbb{1}\{A_t = 1\}] = \mathbb{P}(A_t = 1 \mid \mathcal{F}_{t-1})$. For the last inequality, note that due to the indicator function, the summation in first inequality is not zero only when $\mathbb{1}\{A_t = 1\} = 1$. And $\mathbb{1}\{A_t = 1\} = 1$ further means that we have pulled the best arm (arm 1) at time t. Therefore, the summation over all $T_1(t-1)$ conditional on $\mathbb{1}\{A_t = 1\} = 1$ is equivalent to the summation over s, which is the number of pulls of arm 1.

Case 2: $t : T - 1 \geq T_1(t-1) > M$**.** For this case, we have

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t:T_1(t-1) > M}^{T-1} \mathbb{1}\{A_t = i, E_{i,\epsilon}(t)\}\right] \\
&\leq \mathbb{E}\left[\sum_{t:T_1(t-1) > M}^{T} \mathbb{1}\{\theta_1(t) < \mu_1 - \epsilon\}\right] \\
&\leq T \cdot \mathbb{P}\big(\exists s > M : \widehat{\mu}_{1s} < \mu_1 - \epsilon/2\big) \\
&\qquad + \mathbb{E}\left[\left(\sum_{t:T_1(t-1) > M}^{T} \mathbb{1}\{\theta_1(t) < \mu_1 - \epsilon)\}\right) \mathbb{1}\left\{\forall t \in \{t \mid T_1(t-1) > M\} : \hat{\mu}_{1T_1(t-1)} \geq \mu_1 - \epsilon/2\right\}\right]
\end{aligned}$$

$$
\begin{aligned}
&\le T \cdot e^{-M(\mu_1-(\mu_1-\epsilon/2))^2/(2V)} \\
&\qquad + \mathbb{E}\Big[\sum_{t:T_1(t-1)>M} \mathbb{P}\big[\theta_1(t) < \mu_1 - \epsilon \mid \widehat{\mu}_{1T_1(t-1)} \ge \mu_1 - \epsilon/2\big]\Big] \\
&\le \frac{V}{\epsilon^2} + T \cdot e^{-M\epsilon^2/(16V)} \\
&\le \frac{2V}{\epsilon^2}.
\end{aligned} \tag{A.7}
$$

In the first inequality, we use the fact that $\{A_t = i, E_{i,\epsilon}(t)\} \subseteq \{\theta_1(t) < \mu_1 - \epsilon\}$. In the second inequality, we decompose the term into two events. Event one: there exists a t with $T_1(t-1) > M$ and $\hat{\mu}_{1T_1(t-1)} < \mu_1 - \epsilon/2$. Event two: for all $T_1(t-1) > M$, $\hat{\mu}_{1T_1(t-1)} \ge \mu_1 - \epsilon/2$. In the third inequality, we use Lemma H.1 and following facts

$$
\begin{aligned}
&\mathbb{E}\Big[\Big(\sum_{t:T_1(t-1)>M}^{T} \mathbb{1}\{\theta_1(t) < \mu_1 - \epsilon\}\Big) \mathbb{1}\left\{\forall t \in \{t \mid T_1(t-1) > M\} : \hat{\mu}_{1T_1(t-1)} \ge \mu_1 - \epsilon/2\right\}\Big] \\
=&\mathbb{E}\Big[\sum_{t:T_1(t-1)>M}^{T} \Big(\mathbb{1}\{\theta_1(t) < \mu_1 - \epsilon\} \cdot \mathbb{1}\left\{\forall t \in \{t \mid T_1(t-1) > M\} : \hat{\mu}_{1T_1(t-1)} \ge \mu_1 - \epsilon/2\right\}\Big)\Big] \\
\le&\mathbb{E}\Big[\sum_{t:T_1(t-1)>M}^{T} \Big(\mathbb{1}\{\theta_1(t) < \mu_1 - \epsilon\} \cdot \mathbb{1}\{\hat{\mu}_{1T_1(t-1)} \ge \mu_1 - \epsilon/2\}\Big)\Big] \\
=&\mathbb{E}\Big[\sum_{t:T_1(t-1)>M}^{T} \Big(\mathbb{1}\left\{\theta_1(t) < \mu_1 - \epsilon, \hat{\mu}_{1T_1(t-1)} \ge \mu_1 - \epsilon/2\right\}\Big)\Big] \\
=&\mathbb{E}\Big[\sum_{t:T_1(t-1)>M}^{T} \mathbb{E}\Big[\mathbb{1}\left\{\theta_1(t) < \mu_1 - \epsilon, \hat{\mu}_{1T_1(t-1)} \ge \mu_1 - \epsilon/2\right\} \Big| T_1(t-1)\Big]\Big] \\
=&\mathbb{E}\Big[\sum_{t:T_1(t-1)>M}^{T} \mathbb{P}(\theta_1(t) < \mu_1 - \epsilon, \hat{\mu}_{1T_1(t-1)} \ge \mu_1 - \epsilon/2)\Big] \\
\le&\mathbb{E}\Big[\sum_{t:T_1(t-1)>M}^{T} \mathbb{P}\big(\theta_1(t) < \mu_1 - \epsilon \mid \hat{\mu}_{1T_1(t-1)} \ge \mu_1 - \epsilon/2\big)\Big],
\end{aligned}
$$

where the first inequality is due to the fact $\mathbb{1}\left(\forall t \in T_1(t-1) > M : \hat{\mu}_{1T_1(t-1)} \ge \mu_1 - \epsilon/2\right) \le \mathbb{1}\left(\hat{\mu}_{1T_1(t-1)} \ge \mu_1 - \epsilon/2\right)$ for any $t \in \{t \mid T_1(t-1) > M\}$, and the second inequality is due to $\mathbb{P}(A, B) = \mathbb{P}(A) \cdot \mathbb{P}(A \mid B) \le \mathbb{P}(A \mid B)$.

In the fourth inequality of (A.7), we apply the following results

$$
\begin{aligned}
&\mathbb{E}\Big[\sum_{t:T_1(t-1)>M} \mathbb{P}(\theta_1(t) \le \mu_1 - \epsilon \mid \hat{\mu}_{1T_1(t-1)} \ge \mu_1 - \epsilon/2)\Big] \\
&= \mathbb{E}\left[\sum_{t:T_1(t-1)>M} 1/2 e^{-T_1(t-1) b_{T_1(t-1)} \mathrm{kl}(\hat{\mu}_{1T_1(t-1)}, \mu_1 - \epsilon)}\right] \\
&\le \mathbb{E}\left[\sum_{t:T_1(t-1)>M} e^{-\frac{M}{2} \mathrm{kl}(\hat{\mu}_{1T_1(t-1)}, \mu_1 - \epsilon)}\right] \\
&\le \mathbb{E}\left[\sum_{t:T_1(t-1)>M} e^{-\frac{M}{2} \mathrm{kl}(\mu_1 - \epsilon/2, \mu_1 - \epsilon)}\right] \\
&\le T \cdot e^{-M\epsilon^2/(16V)},
\end{aligned}
$$

where the first equality is due to $\theta_1(t) \sim \mathcal{P}(\hat{\mu}_{1T_1(t-1)}, T_1(t-1))$ and (4.5), the first inequality is due to the fact that $b_s \geq 1/2$ for any $s > 1$ and $T_1(t-1) > M$, the second inequality is due to $\hat{\mu}_{1T_1(t-1)} \geq \mu_1 - \epsilon/2$ and the fact that from Proposition 3.2, $\mathrm{kl}(x, \mu_1 - \epsilon)$ is increasing for $x > \mu_1 - \epsilon/2$, and the last inequality is due to (3.3). Combining (A.3), (A.6), and (A.7) together, we finish the proof of Lemma A.1. □

Note that in order to bound term A, we need the following lemma that states the upper bound of the first term in Lemma A.1.

A.2.2 Proof of Lemma A.2

Let $p(x)$ be the PDF of $\widehat{\mu}_{1s}$ and θ_{1s} be a sample from $\mathcal{P}(\widehat{\mu}_{1s}, s)$. We have

$$\begin{aligned}
&\sum_{s=1}^{M} \mathbb{E}_{\widehat{\mu}_{1s}}\bigg[\bigg(\frac{1}{G_{1s}(\epsilon)} - 1\bigg) \cdot \mathbb{1}\{\widehat{\mu}_{1s} \in L_s\}\bigg] \\
&\leq \underbrace{\sum_{s=1}^{M} \bigg(\int_{\mu_1-\epsilon/2}^{R_{\max}} p(x)/\mathbb{P}(\theta_{1s} \geq \mu_1 - \epsilon \mid \widehat{\mu}_{1s} = x)\mathrm{d}x - 1\bigg)}_{A_1} \\
&\quad + \underbrace{\sum_{s=1}^{M} \int_{\mu_1-\epsilon}^{\mu_1-\epsilon/2} p(x)/\mathbb{P}(\theta_{1s} \geq \mu_1 - \epsilon \mid \widehat{\mu}_{1s} = x)\mathrm{d}x}_{A_2} \\
&\quad + \underbrace{\sum_{s=1}^{M} \int_{\mu_1-\epsilon-\alpha_s}^{\mu_1-\epsilon} \Big[p(x)/\mathbb{P}(\theta_{1s} \geq \mu_1 - \epsilon \mid \widehat{\mu}_{1s} = x)\Big]\mathrm{d}x}_{A_3},
\end{aligned} \tag{A.8}$$

where the inequality is due to the definition of L_s[5].

Bounding term A_1. For term A_1, we divide $\sum_{s=1}^{M}$ into two term, i.e., $\sum_{s=1}^{\lfloor 32V/\epsilon^2 \rfloor}$ and $\sum_{s=\lceil 32V/\epsilon^2 \rceil}^{M}$. Intuitively, for $s \geq 32V/\epsilon^2$, $\mathbb{P}(\theta_{1s} \geq \mu_1 - \epsilon \mid \hat{\mu}_{1s} \geq \mu_1 - \epsilon/2)$ will be large. We have

$$\begin{aligned}
A_1 &= \sum_{s=1}^{M} \bigg(\int_{\mu_1-\epsilon/2}^{R_{\max}} \frac{p(x)}{\mathbb{P}(\theta_{1s} \geq \mu_1 - \epsilon \mid \widehat{\mu}_{1s} = x)}\mathrm{d}x - 1\bigg) \\
&\leq \frac{32V}{\epsilon^2} + \sum_{s=\lceil 32V/\epsilon^2 \rceil}^{M} \bigg(\int_{\mu_1-\epsilon/2}^{R_{\max}} \frac{p(x)}{\mathbb{P}(\theta_{1s} \geq \mu_1 - \epsilon \mid \widehat{\mu}_{1s} = x)}\mathrm{d}x - 1\bigg) \\
&\leq \frac{32V}{\epsilon^2} + \sum_{s=\lceil 32V/\epsilon^2 \rceil}^{M} \bigg(\frac{1}{1 - e^{-s/2 \cdot \mathrm{kl}(\mu_1-\epsilon/2, \mu_1-\epsilon)}} - 1\bigg) \\
&\leq \frac{32V}{\epsilon^2} + \sum_{s=\lceil 32V/\epsilon^2 \rceil}^{M} \bigg(\frac{1}{1 - e^{-s\epsilon^2/(16V)}} - 1\bigg) \\
&= \frac{16V}{\epsilon^2} + \sum_{s=\lceil 32V/\epsilon^2 \rceil}^{M} \frac{1}{e^{s\epsilon^2/(16V)} - 1} \\
&\leq \frac{16V}{\epsilon^2} + \frac{16V}{\epsilon^2} \sum_{s=1}^{\infty} \frac{1}{e^{1+s} - 1} \\
&\leq \frac{32V}{\epsilon^2}.
\end{aligned} \tag{A.9}$$

[5] For the discrete reward distribution, we can use the Dirac delta function for the integral.

For the first inequality, we used the fact $\mathbb{P}(\theta_{1s} \geq \mu_1 - \epsilon \mid \hat{\mu}_{1s} \geq \mu_1 - \epsilon) \geq 1/2$, which is due to (4.4). The second inequality is due to (4.5) and the fact $b_s \geq 1/2$. The third inequality is due to (3.3).

Bounding term A_2. We have

$$A_2 = \sum_{s=1}^{M} \int_{\mu_1-\epsilon}^{\mu_1-\epsilon/2} \frac{p(x)}{\mathbb{P}(\theta_{1s} \geq \mu_1 - \epsilon \mid \widehat{\mu}_{1s} = x)} \mathrm{d}x \leq 2 \sum_{s=1}^{\infty} e^{-s\epsilon^2/(8V)} \leq \frac{2}{e^{\epsilon^2/(8V)} - 1} \leq \frac{16V}{\epsilon^2}, \tag{A.10}$$

where the first inequality is due to $\mathbb{P}(\theta_{1s} \geq \mu_1 - \epsilon \mid \hat{\mu}_{1s} \geq \mu_1 - \epsilon) \geq 1/2$ and from Lemma H.1, $\mathbb{P}(\hat{\mu}_{1s} \leq \mu_1 - \epsilon/2) \leq e^{-s\epsilon^2/(8V)}$, and the last inequality is due to $e^x - 1 \geq x$ for all $x > 0$.

Bounding term A_3. Note that the closed form of the probability density function of $\widehat{\mu}_{1s}$ is hard to compute. Nevertheless, we only need to find an upper bound of the integration in A_3. In the following lemma, we show that it is possible to find such an upper bound with an explicit form.

Lemma A.5. *Let $q(x) = |(s \cdot kl(x, \mu_1))'|e^{-s \cdot kl(x,\mu_1)} = s \int_x^{\mu_1} 1/V(t)dt \cdot e^{-s \cdot kl(x,\mu_1)}$, $g(x) = e^{sb_s \cdot kl(x,\mu_1-\epsilon)}$ and $p(x)$ be the PDF of distribution of $\widehat{\mu}_{1s}$, then*

$$\int_{\mu_1-\epsilon-\alpha_s}^{\mu_1-\epsilon} q(x)g(x)dx + e^{-s \cdot kl(\mu_1-\epsilon-\alpha_s,\mu_1)} \cdot g(\mu_1 - \epsilon - \alpha_s) \geq \int_{\mu_1-\epsilon-\alpha_s}^{\mu_1-\epsilon} p(x)g(x)dx.$$

The proof of Lemma A.5 could be found in Section G. Besides, we need the following inequality on kl-divergence, which resembles the three-point identity property. In particular, for $\mu_1 - \epsilon > x$, we have

$$\begin{aligned}
-\mathrm{kl}(x, \mu_1) + \mathrm{kl}(x, \mu_1 - \epsilon) &= -\int_x^{\mu_1} \frac{t - x}{V(t)} \mathrm{d}t + \int_x^{\mu_1-\epsilon} \frac{t - x}{V(t)} \mathrm{d}t \\
&= -\int_{\mu_1-\epsilon}^{\mu_1} \frac{t - x}{V(t)} \mathrm{d}t \\
&\leq -\int_{\mu_1-\epsilon}^{\mu_1} \frac{t - (\mu_1 - \epsilon)}{V(t)} \mathrm{d}t \\
&= -\mathrm{kl}(\mu_1 - \epsilon, \mu_1),
\end{aligned} \tag{A.11}$$

where the first and the last equality is due to (3.2). For term A_3, we have

$$\begin{aligned}
A_3 &\leq \sum_{s=1}^{M} \int_{\mu_1-\epsilon-\alpha_s}^{\mu_1-\epsilon} p(x)e^{sb_s \cdot \mathrm{kl}(x,\mu_1-\epsilon)} \mathrm{d}x \qquad \text{(A.12)}\\
&\leq \sum_{s=1}^{M} \int_{\mu_1-\epsilon-\alpha_s}^{\mu_1-\epsilon} \Big[q(x) \cdot e^{s \cdot \mathrm{kl}(x,\mu_1-\epsilon)}\Big] \mathrm{d}x + \sum_{s=1}^{M} e^{-s \cdot \mathrm{kl}(\mu_1-\epsilon-\alpha_s,\mu_1)} \cdot e^{s \cdot \mathrm{kl}(\mu_1-\epsilon-\alpha_s,\mu_1-\epsilon)} \\
&\leq \sum_{s=1}^{M} \int_{\mu_1-\epsilon-\alpha_s}^{\mu_1-\epsilon} \Big[|s \cdot \mathrm{kl}(x, \mu_1)'| \cdot e^{-s \cdot \mathrm{kl}(x,\mu_1)} \cdot e^{s \cdot \mathrm{kl}(x,\mu_1-\epsilon)}\Big] \mathrm{d}x + e^{-s\epsilon^2/(2V)} \cdot M \\
&\leq \sum_{s=1}^{M} e^{-s \cdot \mathrm{kl}(\mu_1-\epsilon,\mu_1)} \cdot \int_{\mu_1-\epsilon-\alpha_s}^{\mu_1} \big[|s \cdot \mathrm{kl}(x, \mu_1)|'\big] \mathrm{d}x + e^{-s\epsilon^2/(2V)} \cdot M \\
&\leq \sum_{s=1}^{M} e^{-s\epsilon^2/(2V)}(1 + s \cdot \mathrm{kl}(\mu_1 - \epsilon - \alpha_s, \mu_1)) \\
&\leq \sum_{s=1}^{M} e^{-s\epsilon^2/(2V)}(1 + 4\log(T/s)), \qquad \text{(A.13)}
\end{aligned}$$

where the first inequality is due to (4.4), the second inequality is due to Lemma A.5 and $b_s \leq 1$, the third inequality is due to (A.11), the fourth inequality is due to (A.11), and the last inequality is due to Lemma A.4 and the definition of α_s. Let $d = \lceil V/\epsilon^2 \rceil$. For term $\sum_{s=1}^{d} \log(T/s)$, we have

$$\sum_{s=1}^{d} \log(T/s) = d \log T - \sum_{s=1}^{d} \log s$$

$$
\begin{aligned}
&\le d\log T - \left((s\log s - s)\Big|_1^d - \log d\right)\\
&\le d\log(T/d) + d + \log d\\
&\le \frac{2V\log(T\epsilon^2/V)}{\epsilon^2},
\end{aligned}
\tag{A.14}
$$

where the first inequality is due to $\sum_{x=a}^{b} f(x) \ge \int_a^b f(x)\mathrm{d}x - \max_{x\in[a,b]} f(x)$ for monotone function f. For term $\sum_{s=d}^{M} e^{-s\epsilon^2/(2V)}\log(T/s)$, we have

$$
\begin{aligned}
\sum_{s=d}^{M} e^{-s\epsilon^2/(2V)}\log(T/s) &\le \log(T/d)\sum_{s=d}^{M} e^{-s\epsilon^2/2V}\\
&\le \log(T/d)\sum_{s=1}^{\infty} e^{-s\epsilon^2/2V}\\
&\le \frac{\log(T/d)}{e^{\epsilon^2/(2V)}-1}\\
&\le \frac{2V\log(T/d)}{\epsilon^2}\\
&\le \frac{2V\log(T\epsilon^2/V)}{\epsilon^2},
\end{aligned}
\tag{A.15}
$$

where the fourth inequality is due to $e^x \ge 1+x$ for $x>0$. Substituting (A.15) and (A.14) to (A.13), we have $A_3 = 16V\log(T\epsilon^2/V)/\epsilon^2$. Substituting the bounds of A_1, A_2, and A_3 to (A.8), we have that there exists a constant C_3,

$$
\sum_{s=1}^{M}\mathbb{E}_{\widehat{\mu}_{1s}}\left[\left(\frac{1}{G_{1s}(\epsilon)}-1\right)\cdot \mathbb{1}\{\widehat{\mu}_{1s}\in L_s\}\right] \le C_3\cdot\left(\frac{V\log(T\epsilon^2/V)}{\epsilon^2}\right),
$$

which completes the proof.

A.2.3 Proof of Lemma A.3

Let $\mathcal{T} = \{t\in[K+1,T] : 1 - F_{iT_i(t-1)}(\mu_1-\epsilon) > V/(T\epsilon^2)\}$. Then,

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{t=K+1}^{T}\mathbb{1}\{A_t=i, E_{i,\epsilon}^c(t)\}\right]\\
&\le \mathbb{E}\left[\sum_{t\in\mathcal{T}}\mathbb{1}\{A_t=i\}\right] + \mathbb{E}\left[\sum_{t\notin\mathcal{T}}\mathbb{1}\{E_{i,\epsilon}^c(t)\}\right]\\
&\le \mathbb{E}\left[\sum_{t\ge K+1}\left(\mathbb{1}\{A_t=i\}\cdot\mathbb{1}\{1-F_{iT_i(t-1)}(\mu_1-\epsilon) > V/(T\epsilon^2)\}\right)\right] + \mathbb{E}\left[\sum_{t\notin\mathcal{T}} V/(T\epsilon^2)\right]\\
&\le \mathbb{E}\left[\sum_{s\in[T]}\mathbb{1}\{1-F_{is}(\mu_1-\epsilon) > V/(T\epsilon^2)\}\right] + \frac{V}{\epsilon^2}\\
&\le \mathbb{E}\left[\sum_{s=1}^{T}\mathbb{1}\{G_{is}(\epsilon) > V/(T\epsilon^2)\}\right] + \frac{V}{\epsilon^2}.
\end{aligned}
\tag{A.16}
$$

Let $s \ge N\log(T\epsilon^2/V)/\mathrm{kl}(\mu_i+\rho_i, \mu_1-\epsilon)$. Note that

$$
\frac{1}{N} = \max\left\{1 - \mathrm{kl}(\mu_i+\rho_i,\mu_1-\epsilon)/\log(T\epsilon^2/V), \frac{1}{2}\right\}.
$$

For case $1/N = 1/2$, we have

$$
b_s \ge 1/2 = 1/N.
$$

For case $1/N = 1 - \mathrm{kl}(\mu_i + \rho_i, \mu_1 - \epsilon)/\log(T\epsilon^2/V)$, we have

$$\begin{aligned} b_s &\geq 1 - 1/s \\ &\geq 1 - \mathrm{kl}(\mu_i + \rho_i, \mu_1 - \epsilon)/(N \log(T\epsilon^2/V)) \\ &\geq 1 - \mathrm{kl}(\mu_i + \rho_i, \mu_1 - \epsilon)/(\log(T\epsilon^2/V)) \\ &= 1/N, \end{aligned}$$

where the second inequality is due to $s \geq N \log(T\epsilon^2/V)/\mathrm{kl}(\mu_i + \rho_i, \mu_1 - \epsilon)$, the third inequality is due to $N > 1$. Let X_{is} be a sample from the distribution $\mathcal{P}(\widehat{\mu}_{is}, s)$, if $\widehat{\mu}_{is} \leq \mu_i + \rho_i$, we have

$$\mathbb{P}(X_{is} \geq \mu_1 - \epsilon) \leq \exp\big(-sb_s\mathrm{kl}(\widehat{\mu}_{is}, \mu_1 - \epsilon)\big) \leq \exp\big(-sb_s\mathrm{kl}(\mu_i + \rho_i, \mu_1 - \epsilon)\big) \leq \frac{V}{T\epsilon^2}, \tag{A.17}$$

where the first inequality is from (4.4), the second inequality is due to the assumption $\widehat{\mu}_{is} \leq \mu_i + \rho_i$, and the last inequality is due to $s \geq N \log(T\epsilon^2/V)/\mathrm{kl}(\mu_i + \rho_i, \mu_1 - \epsilon)$ and $b_s \geq 1/N$. Note that when $\mathbb{P}(X_{is} \geq \mu_1 - \epsilon) \leq V/(T\epsilon^2)$ holds, $\mathbb{1}\{G_{is}(\epsilon) > V/(T\epsilon^2)\} = 0$. Now, we check the assumption $\hat{\mu}_{is} \leq \mu_i + \rho_i$ that is needed for (A.17). From Lemma H.1, we have $\mathbb{P}(\widehat{\mu}_{is} > \mu_i + \rho_i) \leq \exp(-s\rho_i^2/(2V))$. Furthermore, it holds that

$$\sum_{s=1}^{\infty} e^{-\frac{s\rho_i^2}{2V}} \leq \frac{1}{e^{\rho_i^2/(2V)} - 1} \leq \frac{2V}{\rho_i^2}, \tag{A.18}$$

where the last inequality is due to the fact $1 + x \leq e^x$ for all x. Let Y_{is} be the event that $\widehat{\mu}_{is} \leq \mu_i + \rho_i$ and $m = N \log(T\epsilon^2/V)/\mathrm{kl}(\mu_i + \rho_i, \mu_1 - \epsilon)$. We further obtain

$$\begin{aligned} \mathbb{E}\left[\sum_{s=1}^{T} \mathbb{1}\{G_{is}(\epsilon) > V/(T\epsilon^2)\}\right] &\leq \mathbb{E}\left[\sum_{s=1}^{T} [\mathbb{1}\{G_{is}(\epsilon) > V/(T\epsilon^2)\} \mid Y_{is}]\right] + \sum_{s=1}^{T}(1 - \mathbb{P}[Y_{is}]) \\ &\leq \mathbb{E}\left[\sum_{s=\lceil m \rceil}^{T} [\mathbb{1}\{\mathbb{P}(X_{is} > \mu_1 - \epsilon) > V/(T\epsilon^2))\} \mid Y_{is}]\right] \\ &\quad + \lceil m \rceil + \sum_{s=1}^{T}(1 - \mathbb{P}[Y_{is}]) \\ &\leq \lceil m \rceil + \sum_{s=1}^{T}(1 - \mathbb{P}[Y_{is}]) \\ &\leq 1 + \frac{2V}{\rho_i^2} + \frac{N \log(T\epsilon^2/V)}{\mathrm{kl}(\mu_i + \rho_i, \mu_1 - \epsilon)}, \end{aligned} \tag{A.19}$$

where the first inequality is due to the fact that $\mathbb{P}(A) \leq \mathbb{P}(A \mid B) + 1 - \mathbb{P}(B)$, the third inequality is due to (A.17) and the last inequality is due to (A.18). Substituting (A.19) into (A.16), we complete the proof.

B Proof of the Asymptotic Optimality of ExpTS

Now we prove the asymptotic regret bound (4.8) of ExpTS presented in Theorem 4.2.

B.1 Proof of the Main Result

The proof in this section shares many components with the finite-time regret analysis presented in Section A. We reuse the decomposition (A.1) by specifying $\epsilon = 1/\log\log T$. In what follows, we bound terms A and B, respectively.

Bounding Term A: We reuse Lemma A.1. Then, it only remains term $\sum_{s=1}^{M} \mathbb{E}\big[\big(1/G_{1s}(\epsilon) - 1\big) \cdot \mathbb{1}\{\widehat{\mu}_{1s} \in L_s\}\big]$ to be bounded. We bound this term by the following lemma.

Lemma B.1. *Let $\epsilon = 1/\log\log T$. Let M, $G_{1s}(\epsilon)$, and L_s be the same as defined in Lemma A.1. Then,*

$$\sum_{s=1}^{M} \mathbb{E}_{\widehat{\mu}_{1s}} \left[\left(\frac{1}{G_{1s}(\epsilon)} - 1 \right) \cdot \mathbb{1}\{\widehat{\mu}_{1s} \in L_s\} \right] = O(V^2(\log\log T)^6 + V(\log\log T)^2 + 1).$$

Let $\epsilon = 1/\log\log T$. Combining Lemma B.1 and Lemma A.1 together, we have

$$A = O\big(V^2(\log\log T)^6 + V(\log\log T)^2 + 1\big).$$

Bounding Term B: Let $\rho_i = \epsilon = 1/\log\log T$. Applying Lemma A.3, we have

$$\begin{aligned} B &= \mathbb{E}\left[\sum_{t=K+1}^{T} \mathbb{1}\{A_t = i, E_{i,\epsilon}^c(t)\} \right] \\ &= O(1 + V(\log\log T)^2) + \frac{N \log(T/(V(\log\log T)^2))}{\mathrm{kl}(\mu_i + 1/\log\log T, \mu_1 - 1/\log\log T)}. \end{aligned} \tag{B.1}$$

Putting It Together: Substituting the bound of term A and B into (A.1), we have

$$\mathbb{E}[T_i(T)] = O(1 + V^2(\log\log T)^6 + V(\log\log T)^2) + \frac{N \log(T/(V(\log\log T)^2))}{\mathrm{kl}(\mu_i + 1/\log\log T, \mu_1 - 1/\log\log T)}.$$

Note that for $T \to +\infty$, $N \to 1$. Therefore,

$$\lim_{T \to +\infty} \frac{\mathbb{E}[T_i(T)]}{\log T} = \frac{1}{\mathrm{kl}(\mu_i, \mu_1)}.$$

This completes the proof of asymptotic regret.

B.2 Proof of Lemma B.1

The proof of this part shares many elements with the proof of Lemma A.2. The difference starts at bounding term A_3.

Bounding term A_3. We need to bound the term $\int_{\mu_1-\epsilon-\alpha_s}^{\mu_1-\epsilon} p(x) e^{s\mathrm{kl}(x,\mu_1-\epsilon)} \mathrm{d}x$. We divide the interval $[\mu_1 - \epsilon - \alpha_s, \mu_1 - \epsilon]$ into n sub-intervals $[x_0, x_1), [x_1, x_2), \cdots, [x_{n-1}, x_n]$, such that $x_0 \le x_1 \le \cdots \le x_n$. For $i \in [n-1]$, we let

$$x_i = \sup_{x: x \le \mu_1 - \epsilon} 4\log(T/e^{i+1})/s < \mathrm{kl}(x, \mu_1) \le 4\log(T/e^i)/s. \tag{B.2}$$

Let $n = \lceil \log T \rceil$ and $x_n = \mu_1$. Then, from definition of α_s, $\mathrm{kl}(x_0, \mu_1) \ge \mathrm{kl}(\mu_1 - \epsilon - \alpha_s, \mu_1)$. Thus, $x_0 \le \mu_1 - \epsilon - \alpha_s$. Now, continue on (A.12), we have

$$\begin{aligned} \int_{\mu_1-\epsilon-\alpha_s}^{\mu_1-\epsilon} p(x) e^{sb_s \cdot \mathrm{kl}(x,\mu_1-\epsilon)} \mathrm{d}x &\le \sum_{i=0}^{n} \int_{x_i}^{x_{i+1}} p(x) e^{sb_s \mathrm{kl}(x,\mu_1-\epsilon)} \mathrm{d}x \\ &\le \sum_{i=0}^{n} e^{sb_s \mathrm{kl}(x_i,\mu_1)} \int_{x_i}^{x_{i+1}} p(x) \mathrm{d}x \\ &\le \sum_{i=0}^{n} e^{sb_s \mathrm{kl}(x_i,\mu_1)} e^{-s \cdot \mathrm{kl}(x_{i+1},\mu_1)} \\ &\le \sum_{i=0}^{n} \left(\frac{T}{e^i}\right)^{b_s} \left(\frac{e^{i+1}}{T}\right) \\ &= O\bigg(\int_0^{\ln T} \left(\frac{T}{e^x}\right)^{b_s} \cdot \frac{e^{x+1}}{T} \mathrm{d}x + e \bigg) \\ &= O\bigg(\frac{1}{1 - b_s} \bigg) \end{aligned}$$

$$= O(s), \tag{B.3}$$

where the first inequality is due to $x_0 \le \mu_1 - \epsilon - \alpha_s$ and $x_n = \mu_1 \ge \mu_1 - \epsilon$, the fourth inequality is due to the definition of x_i, and the first equality is due to $\sum_{x=a}^{b} f(x) \le \int_a^b f(x)\mathrm{d}x + \max_{x\in[a,b]} f(x)$ for monotone function f. Now, we bound term A_3 as follows.

$$\begin{aligned}
A_3 &\le \sum_{s=1}^{M} \int_{\mu_1-\epsilon-\alpha_s}^{\mu_1-\epsilon} p(x) e^{sb_s\cdot \mathrm{kl}(x,\mu_1-\epsilon)} \mathrm{d}x \\
&\le \sum_{s=1}^{\lceil 4V(\log\log T)^3\rceil} \int_{\mu_1-\epsilon-\alpha_s}^{\mu_1-\epsilon} p(x) e^{sb_s\cdot \mathrm{kl}(x,\mu_1-\epsilon)} \mathrm{d}x \\
&\quad + \sum_{s=\lceil 4V(\log\log T)^3\rceil}^{M} \int_{\mu_1-\epsilon-\alpha_s}^{\mu_1-\epsilon} p(x) e^{sb_s\cdot \mathrm{kl}(x,\mu_1-\epsilon)} \mathrm{d}x \\
&\le \underbrace{O\left(\sum_{s=1}^{\lceil 4V(\log\log T)^3\rceil} s\right)}_{I_1} + \underbrace{\sum_{s=\lceil 4V(\log\log T)^3\rceil}^{M} e^{-s\epsilon^2/(2V)}(1+4\log T)}_{I_2},
\end{aligned} \tag{B.4}$$

where the first inequality is from (A.12) and the last inequality is from (B.3) and (A.13). For term I_1, we have $I_1 = O(V^2(\log\log T)^6 + 1)$. Let $\epsilon = 1/\log\log T$, then $M \le O(V \log T \cdot (\log\log T)^2)$. For $s \ge 4V(\log\log T)^3$, we have $e^{-s\epsilon^2/(2V)} = 1/\log^2 T$. Thus, $I_2 = O(M/\log T) = O(V(\log\log T)^2)$. Therefore,

$$A_3 = O(V^2(\log\log T)^6 + V(\log\log T)^2 + 1). \tag{B.5}$$

From (A.9) and (A.10), we have

$$A_1 + A_2 = O(V(\log\log T)^2).$$

Substituting the bound of A_1, A_2 and A_3 to (A.8), we have

$$\sum_{s=1}^{M} \mathbb{E}_{\widehat{\mu}_{1s}}\left[\left(\frac{1}{G_{1s}(\epsilon)} - 1\right) \cdot \mathbb{1}\{\widehat{\mu}_{1s} \in L_s\}\right] \le A_1 + A_2 + A_3 = O(V^2(\log\log T)^6 + V(\log\log T)^2 + 1).$$

This completes the proof.

C Proof of Theorem 4.4 (Gaussian-TS)

The proof of Theorem 4.4 is similar to that of Theorem 4.2. Thus we reuse the notation in the proofs of Theorem 4.2 presented in Sections A and F. However, the sampling distribution $\mathcal{P}$ in Theorem 4.4 is chosen as a Gaussian distribution, and therefore, the concentration and anti-concentration inequalities for Gaussian-TS are slightly different from those used in previous sections. This further affects the results of the supporting lemmas whose proofs depend on the concentration bound of $\mathcal{P}$. In this section, we will prove the regret bounds of Gaussian-TS by showing the existence of counterparts of these lemmas for Gaussian-TS.

C.1 Proof of the Finite-Time Regret Bound

From Lemma H.1, the Gaussian posterior $\mathcal{N}(\mu, V/n)$ satisfies $\mathbb{P}(\theta \le \mu - x) \le e^{-nx^2/(2V)}$. Hence, A.1 also holds for Gaussian-TS. The proof of Lemma A.2 needs to call (4.4) and (4.5). However, the tail bound for Gaussian distribution has a different form. We need to replace Lemma A.2 with the following variant.

Lemma C.1. *Let M, $G_{1s}(\epsilon)$, and L_s be the same as defined in Lemma A.1. Then, there exists a universal constant $C_3 > 0$ such that*

$$\sum_{s=1}^{M} \mathbb{E}_{\widehat{\mu}_{1s}}\left[\left(\frac{1}{G_{1s}(\epsilon)} - 1\right) \cdot \mathbb{1}\{\widehat{\mu}_{1s} \in L_s\}\right] \le C_3 \cdot \left(\frac{V \log(T\epsilon^2/V)}{\epsilon^2}\right).$$

In Section A, the proof of Lemma A.3 only uses the following property of the sampling distribution: let X_{is} be a sample from $\mathcal{P}(\widehat{\mu}_{is}, s)$ and if $\widehat{\mu}_{is} \le \mu_1 - \epsilon$, then

$$\mathbb{P}(X_{is} \ge \mu_1 - \epsilon) \le \exp(-sb_s \cdot \mathrm{kl}(\widehat{\mu}_{is}, \mu_1 - \epsilon)),$$

where the $\mathrm{kl}(\cdot)$ function is defined for Gaussian distribution with variance V. For Gaussian distribution, let X_{is} be a sample from $\mathcal{N}(\widehat{\mu}_{is}, V/s)$. Then from Lemma H.1

$$\mathbb{P}(X_{is} \ge \mu_1 - \epsilon) \le \exp(-s \cdot \mathrm{kl}(\widehat{\mu}_{is}, \mu_1 - \epsilon)) \le \exp(-sb_s \cdot \mathrm{kl}(\widehat{\mu}_{is}, \mu_1 - \epsilon)),$$

where the last inequality is due to $b_s \le 1$ The other parts of the proof of the finite-time bound are the same as that of Theorem 4.2 and thus are omitted.

C.2 Proof of the Asymptotic Regret Bound

The proof of Lemma B.1 needs to call (4.4) and (4.5). However, the tail bound for Gaussian distribution has a different form. We need to replace Lemma A.2 with the following variant.

Lemma C.2. *Let M, $G_{1s}(\epsilon)$, and L_s be the same as defined in Lemma A.1 and let $\epsilon = 1/\log\log T$. Then,*

$$\lim_{T\to\infty} \sum_{s=1}^{M} \mathbb{E}_{\widehat{\mu}_{1s}}\left[\left(\frac{1}{G_{1s}(\epsilon)} - 1\right) \cdot \mathbb{1}\{\widehat{\mu}_{1s} \in L_s\}\right] / \log T = 0.$$

The other parts of asymptotic regret bound are the same as that in Theorem 4.2 and are omitted.

C.3 Proof of Supporting Lemmas

C.3.1 Proof of Lemma C.1

Let Z be a sample from $\mathcal{N}(\widehat{\mu}_{1s}, V/s)$ and $\widehat{\mu}_{1s} = \mu_1 + x$. For $x \le -\epsilon$, applying Lemma H.2 with $z = -\sqrt{s/V}(\epsilon + x) > 0$ yields: for $0 < z \le 1$,

$$G_{1s}(\epsilon) = \mathbb{P}(Z > \mu_1 - \epsilon) \ge \frac{1}{2\sqrt{2\pi}} \exp\left(-\frac{s(\epsilon + x)^2}{2V}\right). \tag{C.1}$$

Besides, for $z > 1$,

$$G_{1s}(\epsilon) \ge \frac{1}{\sqrt{2\pi}} \frac{z}{z^2+1} e^{-\frac{z^2}{2}} \ge \frac{1}{2\sqrt{2\pi} \cdot z} e^{-\frac{z^2}{2}} = \frac{\sqrt{V}}{-2\sqrt{2\pi}\sqrt{s}(\epsilon + x)} \exp\left(-\frac{s(\epsilon+x)^2}{2V}\right). \tag{C.2}$$

Since $\widehat{\mu}_{1s} \sim \mathcal{N}(\mu_1, V/s)$, $x \sim \mathcal{N}(0, V/s)$. Let $p(x)$ be the PDF of $\mathcal{N}(0, V/s)$. Note that $G_{1s}(\epsilon)$ is a random variable with respect to $\widehat{\mu}_{1s}$ and $\widehat{\mu}_{1s} = \mu_1 + x$. We have

$$\begin{aligned}
\mathbb{E}_{\widehat{\mu}_{1s}}\left[\left(\frac{1}{G_{1s}(\epsilon)} - 1\right) \cdot \mathbb{1}\{\widehat{\mu}_{1s} \in L_s\}\right] &\le \int_{-\epsilon}^{+\infty} \frac{p(x)}{G_{1s}(\epsilon)} \mathrm{d}x - 1 + \int_{-\epsilon-\alpha_s}^{-\epsilon} \frac{p(x)}{G_{1s}(\epsilon)} \mathrm{d}x \\
&\le 1 + \int_{-\epsilon-\alpha_s}^{-\epsilon} \frac{p(x)}{G_{1s}(\epsilon)} \mathrm{d}x \\
&\le 1 + \underbrace{\int_{-\epsilon-\alpha_s}^{-\epsilon} p(x)\left(2\sqrt{2\pi} \cdot \exp\left(\frac{s(\epsilon+x)^2}{2V}\right)\right) \mathrm{d}x}_{I_1} \\
&\quad + \underbrace{\int_{-\epsilon-\alpha_s}^{-\epsilon} p(x)\left(2\sqrt{2\pi}\sqrt{s/V}(-\epsilon - x) \cdot \exp\left(\frac{s(\epsilon+x)^2}{2V}\right)\right) \mathrm{d}x}_{I_2}.
\end{aligned} \tag{C.3}$$

The second inequality is due to the fact that for $\widehat{\mu}_{1s} \ge \mu_1 - \epsilon$, $G_{1s}(\epsilon) = \mathbb{P}(Z \ge \mu_1 - \epsilon) \ge 1/2$. The last inequality is due to (C.1) and (C.2). For term I_1, we have

$$I_1 = \int_{-\alpha_s-\epsilon}^{-\epsilon} \left(2\sqrt{\frac{s}{V}} \exp\left(\frac{-sx^2}{2V}\right) \exp\left(\frac{s(\epsilon+x)^2}{2V}\right)\right) \mathrm{d}x$$

$$\leq 2\sqrt{\frac{s}{V}}\exp\left(\frac{s\epsilon^2}{2V}\right)\int_{-\infty}^{-\epsilon}\exp(s\epsilon x/V)\mathrm{d}x$$
$$= \frac{2\sqrt{V}e^{-s\epsilon^2/(2V)}}{\sqrt{s}\epsilon}. \tag{C.4}$$

For term I_2, we have

$$I_2 \leq \int_{-\alpha_s-\epsilon}^{-\epsilon}\left(2s/V(-\epsilon-x)\exp\left(\frac{-sx^2}{2V}\right)\exp\left(\frac{s(\epsilon+x)^2}{2V}\right)\right)\mathrm{d}x$$
$$\leq 2s/V\exp\left(\frac{s\epsilon^2}{2V}\right)\int_{-\alpha_s-\epsilon}^{-\epsilon}(-\epsilon-x)\exp(s\epsilon x/V)\mathrm{d}x$$
$$\leq 2s/V\exp\left(\frac{-s\epsilon^2}{2V}\right)\int_{-\alpha_s-2\epsilon}^{-2\epsilon}-x\exp(s\epsilon x/V)\mathrm{d}x$$
$$\leq 2e\cdot\exp\left(\frac{-s\epsilon^2}{2V}\right)\alpha_s/\epsilon, \tag{C.5}$$

where the last inequality is due to $h(x) = -x\exp(s\epsilon x/V)$ on $x < 0$ achieve is maximum at $x = -V/(s\epsilon)$. We further obtain that

$$\sum_{s=1}^{M}\mathbb{E}\bigg[\bigg(\frac{1}{G_{1s}(\epsilon)}-1\bigg)\cdot\mathbb{1}\{\widehat{\mu}_{1s}\in L_s\}\bigg] = 2e\bigg(\sum_{s=1}^{M}\alpha_s/\epsilon+\sum_{s=1}^{M}\frac{\sqrt{V}}{\sqrt{s}\epsilon}+M\bigg)$$
$$= 2e\bigg(\sum_{s=1}^{M}\alpha_s/\epsilon+\int_{s=1}^{M}\frac{\sqrt{V}}{\sqrt{s}\epsilon}\mathrm{d}s+M\bigg)$$
$$= 2e\bigg(\sum_{s=1}^{M}\alpha_s/\epsilon+\frac{\sqrt{VM}}{\epsilon}+M\bigg)$$
$$= 2e\bigg(\sum_{s=1}^{M}\alpha_s/\epsilon+\frac{V\log(T\epsilon^2/V)}{\epsilon^2}\bigg). \tag{C.6}$$

Note that

$$\mathrm{kl}(\mu_1-\epsilon-4\sqrt{V\log(T/s)/s},\mu_1) \geq \mathrm{kl}(\mu_1-\epsilon-4\sqrt{V\log(T/s)/s},\mu_1-\epsilon) = 8V\log(T/s)/s, \tag{C.7}$$

where the equality is due to (3.3). Thus, from the definition of α_s in (A.2), we have $\alpha_s \leq 4\sqrt{V\log(T/s)/s}$. For term $\sum_{s=1}^{M}\alpha_s$, we have

$$\sum_{s=1}^{M}\alpha_s/(4\sqrt{V}) \leq \sum_{s=1}^{M}\frac{\sqrt{\log(T/s)}}{\sqrt{s}}$$
$$\leq \sum_{j=0}^{\lceil\log M-1\rceil-1}\sum_{s=\lceil e^j\rceil}^{\lceil e^{j+1}\rceil}\frac{\sqrt{\log(T/e^j)}}{\sqrt{s}}$$
$$\leq \sum_{j=0}^{\lceil\log M-1\rceil}\sqrt{\log(T/e^j)}\int_{e^j}^{e^{j+1}}\frac{1}{\sqrt{s}}\mathrm{d}s+\sum_{j=0}^{\lceil\log M-1\rceil}\frac{\sqrt{\log(T/e^j)}}{e^{j/2}}$$
$$\leq 2\sum_{j=0}^{\lceil\log M-1\rceil}e^{(j+1)/2}\cdot\log(T/e^j)$$
$$\leq 2\sqrt{e}\int_0^{\log M}(\log T-x)e^{x/2}\mathrm{d}x+2\sqrt{eM}\log(T/M)$$
$$= 2\sqrt{e}\bigg(2\log(e^2T)e^{x/2}-2xe^{x/2}\bigg|_0^{\log M}\bigg)+2\sqrt{eM}\log(T/M)$$

$$= 20\left(\frac{\sqrt{V}\log(T\epsilon^2/V)}{\epsilon}\right), \tag{C.8}$$

where the third and sixth inequality is due to $\sum_{x=a}^{b} f(x) \le \int_a^b f(x)\mathrm{d}x + \max_{x\in[a,b]} f(x)$ for monotone function f. Substituting (C.8) to (C.6), we have that there exists a constant $C_3 > 0$ such that

$$\sum_{s=1}^{M} \mathbb{E}\bigg[\bigg(\frac{1}{G_{1s}(\epsilon)} - 1\bigg)\cdot \mathbb{1}\{\widehat{\mu}_{1s} \in L_s\}\bigg] \le C_3 \cdot \bigg(\frac{V\log(T\epsilon^2/V)}{\epsilon^2}\bigg).$$

This completes the proof.

C.3.2 Proof of Lemma C.2

The proof of this part is similar to the proof of Lemma C.1. We reuse the notation defined in the Lemma C.1. Recall Z is a sample from $\mathcal{N}(\widehat{\mu}_{1s}, V/s)$. For $\widehat{\mu}_{1s} = \mu_1 - \epsilon/2$, from (H.1)

$$\mathbb{P}(Z \le \mu_1 - \epsilon) \le \exp\big(-s\epsilon^2/(8V)\big). \tag{C.9}$$

We have

$$\begin{aligned}
&\mathbb{E}_{\widehat{\mu}_{1s}}\bigg[\bigg(\frac{1}{G_{1s}(\epsilon)} - 1\bigg)\cdot \mathbb{1}\{\widehat{\mu}_{1s} \in L_s\}\bigg] \\
&\le \int_{-\epsilon/2}^{+\infty} \frac{p(x)}{G_{1s}(\epsilon)}\mathrm{d}x + \int_{-\epsilon}^{-\epsilon/2} \frac{p(x)}{G_{1s}(\epsilon)}\mathrm{d}x - 1 + \int_{-\epsilon-\alpha_s}^{-\epsilon} \frac{p(x)}{G_{1s}(\epsilon)}\mathrm{d}x \\
&\le e^{-s\epsilon^2/(8V)} \int_{-\epsilon/2}^{+\infty} p(x)\mathrm{d}x + 2\int_{-\epsilon}^{-\epsilon/2} p(x)\mathrm{d}x + \int_{-\epsilon-\alpha_s}^{-\epsilon} \frac{p(x)}{G_{1s}(\epsilon)}\mathrm{d}x \\
&\le e^{-s\epsilon^2/(8V)} + e^{-s\epsilon^2/(8V)} + \underbrace{\int_{-\epsilon-\alpha_s}^{-\epsilon} p(x)\bigg(2\sqrt{2\pi}\cdot\exp\bigg(\frac{s(\epsilon+x)^2}{2V}\bigg)\bigg)\mathrm{d}x}_{I_1} \\
&\quad + \underbrace{\int_{-\epsilon-\alpha_s}^{-\epsilon} p(x)\bigg(2\sqrt{2\pi}\sqrt{s}(-\epsilon-x)\cdot\exp\bigg(\frac{s(\epsilon+x)^2}{2V}\bigg)\bigg)\mathrm{d}x}_{I_2},
\end{aligned} \tag{C.10}$$

where the second inequality is due to (C.9) and the fact that for $x \ge -\epsilon$, $G_{1s}(\epsilon) \ge 1/2$, the third inequality is due to $x \sim \mathcal{N}(0, V/s)$ and from (H.1), $\mathbb{P}(x \le -\epsilon/2) \le \exp(-s\epsilon^2/(8V))$. Further, we have

$$\sum_{s=1}^{\infty} \exp(-s\epsilon^2/(8V)) \le \frac{1}{e^{\epsilon^2/(8V)} - 1} \le \frac{8V}{\epsilon^2}.$$

By applying (C.4) and (C.5) to bound term I_1 and I_2, we obtain

$$\begin{aligned}
\sum_{s=1}^{M} \mathbb{E}_{\widehat{\mu}_{1s}}\bigg[\bigg(\frac{1}{G_{1s}(\epsilon)} - 1\bigg)\cdot \mathbb{1}\{\widehat{\mu}_{1s} \in L_s\}\bigg] &\le \frac{8V}{\epsilon^2} + O\bigg(\sum_{s=1}^{M} \frac{e^{-s\epsilon^2/(2V)}}{\sqrt{s}\epsilon} + \sum_{s=1}^{M} \frac{e^{-s\epsilon^2/(2V)}\alpha_s}{\epsilon}\bigg) \\
&= O\bigg(\frac{V}{\epsilon^2} + 2\sqrt{\log T}/\epsilon \sum_{s=1}^{\infty} e^{-s\epsilon^2/(2V)}\bigg) \\
&= O\bigg(\frac{V}{\epsilon^2} + \frac{V\sqrt{\log T}}{\epsilon^3}\bigg),
\end{aligned}$$

where the first equality is due to (C.7). Let $\epsilon = 1/\log\log T$, we have

$$\lim_{T\to\infty} \sum_{s=1}^{M} \mathbb{E}_{\widehat{\mu}_{1s}}\bigg[\bigg(\frac{1}{G_{1s}(\epsilon)} - 1\bigg)\cdot \mathbb{1}\{\widehat{\mu}_{1s} \in L_s\}\bigg]/\log T = 0,$$

which competes the proof.

D Proof of Theorem 4.6 (Bernoulli-TS)

Similar to the proof strategy used in Section C, we will prove the regret bounds of Bernoulli-TS via providing a counterpart of the supporting lemma used in the proof of Theorem 4.6 that depends on the concentration bound of the sampling distribution $\mathcal{P}$.

D.1 Proof of the Finite-Time Regret Bound

Due to the same reason shown in Section C.1, we only need to replace Lemma A.2 with the following variant. The rest of the proof remains the same as that of Theorem 4.2.

Lemma D.1. *Let M, $G_{1s}(\epsilon)$, and L_s be the same as defined in Lemma A.1. Let $\epsilon = \Delta/4$. There exists a universal constant C_3 such that*

$$\sum_{s=1}^{M} \mathbb{E}_{\widehat{\mu}_{1s}}\left[\left(\frac{1}{G_{1s}(\epsilon)} - 1\right) \cdot \mathbb{1}\{\widehat{\mu}_{1s} \in L_s\}\right] \leq C_3 \cdot \left(\frac{\log(T\epsilon^2)}{\epsilon^2}\right).$$

D.2 Proof of the Asymptotic Regret Bound

We note that Agrawal and Goyal [5] has proved the asymptotic optimality for Beta posteriors under Bernoulli rewards. One can find the details therein and we omit the proofs here.

D.3 Proof of Lemma D.1

We first define some notations. Let $F^B_{n,p}(\cdot)$ denote the CDF, and $f^B_{n,p}(\cdot)$ denote the probability mass function of binomial distribution with parameters n, p respectively. We also let $F^{beta}_{\alpha,\beta}(\cdot)$ denote the CDF of the beta distribution with parameters α, β. The following equality gives the relationship between $F^{beta}_{\alpha,\beta}(\cdot)$ and $F^B_{n,p}(\cdot)$.

$$F^{beta}_{\alpha,\beta}(y) = 1 - F^B_{\alpha+\beta-1,y}(\alpha - 1). \tag{D.1}$$

Let $y = \mu_1 - \epsilon$. Let $j = S_i(t)$ and $s = T_i(t)$. From (D.1), we have $G_{1s}(\epsilon) = \mathbb{P}(\theta_1(t) > y) = F^B_{s+1,y}(j)$. Note that for Bernoulli distribution, we can set $V = 1/4$. Besides,

$$\mathrm{kl}(\mu_1 - \epsilon - 4\sqrt{V\log(T/s)/s}, \mu_1) \geq \mathrm{kl}(\mu_1 - \epsilon - 4\sqrt{V\log(T/s)/s}, \mu_1 - \epsilon) \geq 8V\log(T/s)/s,$$

where the inequality is due to (3.3). Thus, from the definition of α_s in (A.2), we have $\alpha_s \leq 2\sqrt{\log(T/s)/s}$. For $j/s \in L_s$, we have $j/s \geq \mu_1 - \epsilon - \sqrt{\frac{2\log(T/s)}{s}}$. Hence,

$$j \geq ys - 2\sqrt{s\log(T/s)}.$$

Let $\gamma_s = \lceil ys - 2\sqrt{s\log(T/s)}\rceil$. Therefore,

$$\mathbb{E}_{\widehat{\mu}_{1s}}\left[\left(\frac{1}{G_{1s}(\epsilon)}\right) \cdot \mathbb{1}\{\widehat{\mu}_{1j} \in L_s\}\right] \leq \sum_{j=\gamma_s}^{s} \frac{f_{s,\mu_1}(j)}{F^B_{s+1,y}(j)}.$$

In the derivation below, we abbreviate $F^B_{s+1,y}(j)$ as $F_{s+1,y}(j)$.
Case $s < 8/\epsilon$. From Lemma 2.9 of Agrawal and Goyal [5], we have

$$\sum_{j=1}^{s} \frac{f_{s,\mu_1}(j)}{F^B_{s+1,y}(j)} \leq \frac{3}{\epsilon}. \tag{D.2}$$

Case $s \geq 8/\epsilon$. We divide the $Sum(\gamma_s, s) = \sum_{j=\gamma_s}^{s} \frac{f_{s,\mu_1}(j)}{F^B_{s+1,y}(j)}$ into four partial sums: $Sum(\gamma_s, \lfloor ys\rfloor)$, $Sum(\lfloor ys\rfloor, \lfloor ys\rfloor)$, $Sum(\lceil ys\rceil, \lfloor \mu_1 s - \frac{\epsilon}{2}s\rfloor)$, and $Sum(\lceil \mu_1 s - \frac{\epsilon}{2}s\rceil, \lfloor s\rfloor)$ and bound them respectively. We need the following bounds on the CDF of Binomial distribution [18] [Prop. A.4].
There exists a universal constant $C_4 > 0$ such that for $j \leq y(s+1) - \sqrt{(s+1)y(1-y)}$,

$$F_{s+1,y}(j) \leq C_4 \cdot \left(\frac{y(s+1-j)}{y(s+1)-j}\binom{s+1}{j} y^j (1-y)^{s+1-j}\right); \tag{D.3}$$

and for $j \geq y(s+1) - \sqrt{(s+1)y(1-y)}$,

$$F_{s+1,y}(j) \leq C_4.$$

Bounding $Sum(\gamma_s, \lfloor ys \rfloor)$**.** Let $R = \frac{\mu_1(1-y)}{y(1-\mu_1)}$. Then we have $R > 1$. Using the bounds above, we have for any j, there exists constant C_4 such that

$$\begin{aligned}\frac{f_{s,\mu_1}(j)}{F_{s+1,y}(j)} &\leq C_4 \cdot \left(\frac{f_{s,\mu_1}(j)}{\frac{y(s+1-j)}{y(s+1)-j}\binom{s+1}{j}y^j(1-y)^{s+1-j}}\right) + C_4 \cdot f_{s,\mu_1}(j) \\ &= C_4 \cdot \left(\left(1 - \frac{j}{y(s+1)}\right)R^j \frac{(1-\mu_1)^s}{(1-y)^{s+1}}\right)\right) + C_4 \cdot f_{s,\mu_1}(j).\end{aligned}$$

This applies that for $s \leq \lfloor ys \rfloor$,

$$\begin{aligned}\left(1 - \frac{j}{y(s+1)}\right)R^j \frac{(1-\mu_1)^s}{(1-y)^{s+1}} &= \frac{y(s+1)-j}{y(1-y)(s+1)} R^{j-ys} R^{ys} \frac{(1-\mu_1)^s}{(1-y)^s} \\ &= \frac{e^{-s\cdot \mathrm{kl}(y,\mu_1)}}{y(1-y)(s+1)}(y(s+1)-j)R^{j-ys}, \end{aligned} \tag{D.4}$$

where the last equality is due to the fact for Bernoulli distribution, $\mathrm{kl}(y,\mu_1) = y\log(y/\mu_1) + (1-y)\log((1-y)/(1-\mu_1))$. Next, we prove $(y(s+1)-j)R^{j-ys} \leq \frac{2R}{R-1} + e/\ln 2$. Consider the following two cases.

Case 1: $1/\ln R \leq y$. We have

$$(y(s+1)-j)R^{j-ys} = (y(s+1)-j)R^{j-y(s+1)}R^y \leq yR^{-y}R^y \leq 1,$$

where the inequality is due to xR^{-x} is monotone increasing on $x \in (0, 1/\ln R)$ and $y(s+1) - j \geq y(s+1) - ys = y \geq 1/\ln R$.

Case 2: $1/\ln R \geq y$**.** We will divide it into the following three intervals of R:

For $R \geq e^2$, we have

$$\begin{aligned}(y(s+1)-j)R^{j-ys} &= (y(s+1)-j)R^{j-y(s+1)}R^y \\ &\leq \frac{1}{\ln R}R^{-1/\ln R}R^y \\ &\leq \frac{1}{\ln R}R^{-y}R^y \\ &\leq \frac{1}{\ln R} \\ &\leq 1,\end{aligned}$$

where the first inequality is due to xa^{-x} achieve its maximum at $1/\ln a$.

For $2 < R < e^2$, we have

$$R^{-1/\ln R}/\ln R \leq 1/(e\ln 2) \Leftrightarrow -1 \leq \ln(\ln R/(e\ln 2)) \Leftrightarrow R \geq 2.$$

Therefore,

$$\begin{aligned}(y(s+1)-j)R^{j-ys} &= (y(s+1)-j)R^{j-y(s+1)}R^y \\ &\leq \frac{1}{\ln R}R^{-1/\ln R}R \\ &\leq R/(e\ln 2) \\ &\leq e/\ln 2.\end{aligned}$$

For $1 < R < 2$, we have $\ln R \geq (R-1) - (R-1)^2/2$. Further,

$$\frac{R^{-1/\ln R}}{\ln R} \leq \frac{1}{\ln R} \leq \frac{1}{(R-1)-(R-1)^2/2} \leq \frac{1}{(R-1)(1-(R-1)/2)} \leq \frac{2}{R-1}.$$

We have

$$(y(s+1)-j)R^{j-ys} \leq (y(s+1)-j)R^{j-y(s+1)}R$$

$$\leq \frac{R}{\ln R} R^{-1/\ln R}$$
$$\leq \frac{2R}{R-1}.$$

Combining **Case 1** and **Case 2** together, we have $(y(s+1)-j)R^{j-ys} \leq \frac{2R}{R-1} + e/\ln 2$. Substituting this into (D.4), we have

$$\begin{aligned}\left(1-\frac{j}{y(s+1)}\right)R^j\frac{(1-\mu_1)^s}{(1-y)^{s+1}} &\leq \frac{e^{-s\cdot\mathrm{kl}(y,\mu_1)}}{y(1-y)(s+1)}\left(\frac{2R}{R-1}+e/\ln 2\right)\\ &\leq \frac{2\mu_1 e^{-s\cdot\mathrm{kl}(y,\mu_1)}}{y(\mu_1-y)(s+1)} + \frac{8e^{-s\cdot\mathrm{kl}(y,\mu_1)}}{y(1-y)(s+1)}\\ &\leq \frac{20e^{-s\cdot\mathrm{kl}(y,\mu_1)}}{\epsilon(s+1)}.\end{aligned} \tag{D.5}$$

The second inequality is due to $\frac{R}{R-1} = \frac{\mu_1(1-y)}{\mu_1-y}$. The last inequality is due to

$$\frac{\mu_1}{y} = \frac{\mu_1}{\mu_1-\epsilon} = \frac{\mu_1}{\mu_1-\Delta_i/4} \leq 4/3 < 2,$$

and

$$y(1-y) \geq \Delta_i/4(1-\Delta_i/4) = \epsilon(1-\epsilon) \geq \epsilon/2,$$

where the first inequality is because $y(1-y)$ is decreasing for $y \geq 1/2$ and increasing for $y \leq 1/2$ and $y = \mu_1 - \Delta_i/4 \in [3/(4\Delta_i), 1-\Delta_i/4]$, since $\mu_1 \in [0,1]$ and $\mu_1 \geq \Delta_i$ by definition, the last inequality is due to the fact $\epsilon = \Delta_i/4 \leq 1/4$. Therefore, we have

$$\begin{aligned}Sum(\gamma_s, \lfloor ys\rfloor) &= \sum_{j=\gamma_s}^{\lfloor ys\rfloor}\frac{f_{s,\mu_1}(j)}{F^B_{s+1,y}(j)}\\ &\leq C_4\cdot\left(\sum_{j=\gamma_s}^{\lfloor ys\rfloor}\left(1-\frac{j}{y(s+1)}\right)R^j\frac{(1-\mu_1)^s}{(1-y)^{s+1}}\right)\Bigg) + C_4\cdot\sum_{j=1}^{s} f_{s,\mu_1}(j)\\ &\leq 20C_4\cdot\left(\frac{e^{-s\cdot\mathrm{kl}(y,\mu_1)}(ys-\gamma_s)}{\epsilon(s+1)}\right)\Bigg) + C_4\\ &= \frac{40C_4\sqrt{\log(T/s)/s}}{\epsilon} + C_4,\end{aligned} \tag{D.6}$$

where the second equality is due to (D.5).
Bounding $Sum(\lfloor ys\rfloor, \lfloor ys\rfloor)$ **and** $Sum(\lceil ys\rceil, \lfloor \mu_1 s - \frac{\epsilon}{2}s\rfloor)$**.** From Lemma 2.9 of Agrawal and Goyal [5], we have

$$Sum(\lfloor ys\rfloor, \lfloor ys\rfloor) \leq 3e^{-s\mathrm{kl}(y,\mu_1)} \leq 3e^{-2s\epsilon^2}, \tag{D.7}$$

and there exist a universal constant $C_5 > 0$ such that

$$Sum\left(\lceil ys\rceil, \left\lfloor \mu_1 s - \frac{\epsilon}{2}s\right\rfloor\right) \leq C_5\cdot e^{-s\epsilon^2/2}. \tag{D.8}$$

Bounding $Sum(\lceil \mu_1 s - \frac{\epsilon}{2}s\rceil, s)$**.** For $j \in [\lceil \mu_1 s - \frac{\epsilon}{2}s\rceil, s]$, $F_{s+1,y}(j) \leq C_4$. Hence,

$$Sum(\lceil \mu_1 s - \frac{\epsilon}{2}s\rceil, s) \leq C_4. \tag{D.9}$$

Combining (D.6), (D.7), (D.8) and (D.9) together, we have that for $s \geq 8/\epsilon$, there exists a universal constant $C_6 > 0$ such that

$$\mathbb{E}_{\widehat{\mu}_{1s}}\left[\left(\frac{1}{G_{1s}(\epsilon)}-1\right)\cdot \mathbb{1}\{\widehat{\mu}_{1j}\in L_s\}\right] \leq C_6\cdot\left(1+e^{-s\epsilon^2/2}+e^{-2s\epsilon^2}+\frac{\sqrt{\log(T/s)/s}}{\epsilon}\right). \tag{D.10}$$

Combining (D.2) and (D.10) together, we have that there exists a universal constant $C_3 > 0$ such that

$$
\begin{aligned}
&\sum_{s=1}^{M} \mathbb{E}_{\widehat{\mu}_{1s}}\left[\left(\frac{1}{G_{1s}(\epsilon)} - 1\right) \cdot \mathbb{1}\{\widehat{\mu}_{1s} \in L_s\}\right] \\
&= \sum_{s:1\le s<8/\epsilon} \mathbb{E}_{\widehat{\mu}_{1s}}\left[\left(\frac{1}{G_{1s}(\epsilon)} - 1\right) \cdot \mathbb{1}\{\widehat{\mu}_{1s} \in L_s\}\right] \\
&\quad + \sum_{s:s\ge 8/\epsilon} \mathbb{E}_{\widehat{\mu}_{1s}}\left[\left(\frac{1}{G_{1s}(\epsilon)} - 1\right) \cdot \mathbb{1}\{\widehat{\mu}_{1s} \in L_s\}\right] \\
&\le \frac{24}{\epsilon^2} + C_6 \cdot \left(M + \sum_{s=1}^{\infty} e^{-2s\epsilon^2} + \sum_{s=1}^{\infty} e^{-s\epsilon^2/2} + \sum_{s=1}^{M} \frac{\sqrt{\log(T/s)/s}}{\epsilon}\right) \\
&\le C_3 \cdot \left(\frac{\log(T\epsilon^2)}{\epsilon^2}\right),
\end{aligned}
$$

where the last inequality is due to the fact $\sum_{s=1}^{\infty} e^{-2s\epsilon^2} \le 1/(e^{2\epsilon^2} - 1) \le \frac{1}{2\epsilon^2}$ and $\sum_{s=1}^{M} \sqrt{\frac{\log(T/s)}{s}} \le 80(\epsilon^{-1} \log(T\epsilon^2))$ from (C.8).

E Proof of the Minimax Optimality of ExpTS$^+$

In this section, we prove the worst case regret bound of ExpTS$^+$ presented in Theorem 5.1.

E.1 Proof of the Main Result

Regret Decomposition: For simplicity, we reuse the notations in Section A. Let $S_j = \{i \in [K] \mid 2^{-(j+1)} \le \Delta_i < 2^{-j}\}$ be the set of arms whose gaps from the optimal arm are bounded in the interval $[2^{-(j+1)}, 2^{-j})$. Define $\gamma = 1/2 \log_2(T/(VK)) - 3$. Then we know that for any arm $i \in [K]$ that $\Delta_i > 4\sqrt{VK/T} = 2^{-(\gamma+1)}$, there must exist some $j \le \gamma$ such that $i \in S_j$. Therefore, the regret of ExpTS$^+$ can be decomposed as follows.

$$
\begin{aligned}
R_\mu(T) &= \sum_{i:\Delta_i>0} \Delta_i \cdot \mathbb{E}[T_i(T)] \\
&\le \sum_{i:\Delta_i>4\sqrt{VK/T}} \Delta_i \cdot \mathbb{E}[T_i(T)] + \max_{i:\Delta_i<4\sqrt{VK/T}} \Delta_i \cdot T && \text{(E.1)} \\
&< \sum_{j<\gamma} \sum_{i\in S_j} 2^{-j} \cdot \mathbb{E}[T_i(T)] + 4\sqrt{VKT}, && \text{(E.2)}
\end{aligned}
$$

where in the first inequality we used the fact that $\sum_i \mathbb{E}[T_i(T)] = T$, and in the last inequality we used the fact that $\Delta_i < 2^{-j}$ for $\Delta_i \in S_j$. The expected number of times that Algorithm 1 plays arms in set S_j with $j < \gamma$ is bounded as follows.

$$
\begin{aligned}
\sum_{i\in S_j} \mathbb{E}[T_i(T)] &= |S_j| + \sum_{i\in S_j} \mathbb{E}\left[\sum_{t=K+1}^{T} \mathbb{1}\{A_t = i, E_{i,\epsilon_j}(t)\} + \sum_{t=K+1}^{T} \mathbb{1}\{A_t = i, E_{i,\epsilon_j}^c(t)\}\right] \\
&= |S_j| + \underbrace{\sum_{i\in S_j} \mathbb{E}\left[\sum_{t=K+1}^{T} \mathbb{1}\{A_t = i, E_{i,\epsilon_j}(t)\}\right]}_{A} + \underbrace{\sum_{i\in S_j} \mathbb{E}\left[\sum_{t=K+1}^{T} \mathbb{1}\{A_t = i, E_{i,\epsilon_j}^c(t)\}\right]}_{B},
\end{aligned}
\tag{E.3}
$$

where $\epsilon_j > \sqrt{8VK/T}$ is an arbitrary constant.

Bounding Term A: Define

$$\alpha_s = \sup_{x\in[0,\mu_1-\epsilon-R_{\min})} \mathrm{kl}(\mu_1-\epsilon-x,\mu_1) \le 4\log^+(T/(Ks))/s, \tag{E.4}$$

where $\log^+(x) = \max\{0,\log x\}$. We decompose the term $\sum_{i\in S_j}\mathbb{E}\left[\sum_{t=K+1}^T \mathbb{1}\{A_t=i,E_{i,\epsilon}(t)\}\right]$ by the following lemma.

Lemma E.1. *Let $\epsilon_j = 2^{-j-2}$. Let $M_j = \lceil 16V\log(T\epsilon_j^2/(KV))/\epsilon_j^2\rceil$. Then, there exists a universal constant $C_2>0$,*

$$\sum_{i\in S_j}\mathbb{E}\left[\sum_{t=K+1}^T \mathbb{1}\{A_t=i,E_{i,\epsilon_j}(t)\}\right] \le \sum_{s=1}^{M_j}\mathbb{E}\left[\left(\frac{1}{G_{1s}(\epsilon_j)}-1\right)\cdot\mathbb{1}\{\widehat{\mu}_{1s}\in L_s\}\right] + \frac{C_2VK}{\epsilon_j^2},$$

where $G_{is}(\epsilon) = 1-F_{is}(\mu_1-\epsilon)$, F_{is} is the CDF of $\mathcal{P}(\widehat{\mu}_{is},s)$, and $L_s = \Big(\mu_1-\epsilon-\alpha_s, R_{\max}\Big]$.

Now, we bound the remaining term in Lemma E.1.

Lemma E.2. *Let M_j, $G_{1s}(\epsilon_j)$, and L_s be the same as defined in Lemma E.1. Then, there exists a universal constant $C_3>0$ such that*

$$\sum_{s=1}^{M_j}\mathbb{E}_{\widehat{\mu}_{1s}}\left[\left(\frac{1}{G_{1s}(\epsilon_j)}-1\right)\cdot\mathbb{1}\{\widehat{\mu}_{1s}\in L_s\}\right] \le C_3\cdot\left(\frac{VK\log(T\epsilon_j^2/(KV))}{\epsilon_j^2}\right).$$

Combining Lemma E.1 and Lemma E.2 together, we have

$$A \le \left(\frac{(C_3+C_2)VK\log(T\epsilon_j^2/(KV))}{\epsilon_j^2}\right).$$

Bounding Term B: We have the following lemma that bounds the second term in (E.3).

Lemma E.3. *Let $N_i = \min\{1/(1-(kl(\mu_i+\rho_i,\mu_1-\epsilon_j))/\log(T\epsilon_j^2/V)),2\}$. For any $\rho_i,\epsilon_j>0$ that satisfies $\epsilon_j+\rho_i<\Delta_i$, then*

$$\mathbb{E}\left[\sum_{t=K+1}^T \mathbb{1}\{A_t=i,E^c_{i,\epsilon_j}(t)\}\right] \le 1+\frac{2V}{\rho_i^2}+\frac{V}{\epsilon_j^2}+\frac{N_i\log(T\epsilon_j^2/(VK))}{kl(\mu_i+\rho_i,\mu_1-\epsilon_j)}.$$

Putting it Together: Let $\rho_i=\epsilon_j$. Substituting Lemma E.1 and Lemma E.3 to the regret decomposition (E.2), we obtain

$$\begin{aligned}
R_\mu(T) &\le \sum_{j<\gamma}\sum_{i\in S_j}\epsilon_j\cdot\mathbb{E}[T_i(T)] + 4\sqrt{VKT}\\
&= \sum_{j<\gamma}\frac{(C_3+C_2+64)KV\log(T\epsilon_j^2/(VK))}{\epsilon_j} + \sqrt{VKT} + \sum_{i\ge 2}\Delta_i\\
&= 8(C_3+C_2+64)\sqrt{VKT}\cdot\sum_{n=0}^{\infty}\frac{\log 64+n\log 2}{2^n} + \sum_{i\ge 2}\Delta_i,
\end{aligned}$$

which completes the proof of the minimax optimality. Therefore, there exists a universal constant $C_1>0$ such that

$$R_\mu(T) \le C_1\cdot\left(\sum_{i>1}\Delta_i + \sqrt{VKT}\right).$$

E.2 Proof of Supporting Lemmas

E.2.1 Proof of Lemma E.1

The proof of this lemma shares many element with that of Lemma A.1. Let $\mathcal{F}_t = \sigma(A_1, r_1, \cdots, A_t, r_t)$ be the filtration. By the definition of $G_{is}(x)$, it holds that

$$G_{1T_1(t-1)}(\epsilon_j) = \mathbb{P}(\theta_1(t) \geq \mu_1 - \epsilon_j \mid \mathcal{F}_{t-1}). \tag{E.5}$$

Define $\mathcal{E}$ to be the event such that $\widehat{\mu}_{1s} \in L_s$ holds for all $s \in [T]$. The indicator function can be decomposed based on $\mathcal{E}$.

$$\begin{aligned}
&\sum_{i \in S_j} \mathbb{E}\left[\sum_{t=K+1}^{T} \mathbb{1}\{A_t = i, E_{i,\epsilon_j}(t)\}\right] \\
&\leq T \cdot \mathbb{P}(\mathcal{E}^c) + \sum_{i \in S_j} \mathbb{E}\left[\sum_{t=K+1}^{T} \left[\mathbb{1}\{A_t = i, E_{i,\epsilon_j}(t)\} \cdot \mathbb{1}\{\widehat{\mu}_{1T_1(t-1)} \in L_{T_i(t-1)}\}\right]\right] \\
&\leq \Theta\bigg(\frac{VK}{\epsilon_j^2}\bigg) + \sum_{i \in S_j} \mathbb{E}\left[\sum_{t=K+1}^{T} \left[\mathbb{1}\{A_t = i, E_{i,\epsilon_j}(t)\} \cdot \mathbb{1}\{\widehat{\mu}_{1T_1(t-1)} \in L_{T_i(t-1)}\}\right]\right],
\end{aligned} \tag{E.6}$$

where the second inequality is from Lemma A.4 with $b = K$. Let $A'_t = \arg\max_{i \neq 1} \theta_i(t)$. Then

$$\begin{aligned}
\mathbb{P}(A_t = 1 \mid \mathcal{F}_{t-1}) &\geq \mathbb{P}(\{\theta_1(t) \geq \mu_1 - \epsilon_j\} \cap \{\exists i \in S_j : A'_t = i, E_{i,\epsilon_j}(t)\} \mid \mathcal{F}_{t-1}) \\
&= \mathbb{P}(\theta_1(t) \geq \mu_1 - \epsilon_j \mid \mathcal{F}_{t-1}) \mathbb{P}\bigg(\bigcup_{i \in S_j} \{A'_t = i, E_{i,\epsilon_j}(t)\}\bigg) \\
&= \mathbb{P}(\theta_1(t) \geq \mu_1 - \epsilon_j \mid \mathcal{F}_{t-1}) \cdot \sum_{i \in S_j} \mathbb{P}(A'_t = i, E_{i,\epsilon_j}(t) \mid \mathcal{F}_{t-1}) \\
&\geq \frac{G_{1T_1(t-1)}}{1 - G_{1T_1(t-1)}} \cdot \sum_{i \in S_j} \mathbb{P}(A_t = i, E_{i,\epsilon_j}(t) \mid \mathcal{F}_{t-1}).
\end{aligned} \tag{E.7}$$

The first inequality is due to the fact when both event $\{\theta_1(t) \geq \mu_1 - \epsilon\}$ and event $\{\exists i \in S_j : A'_t = i, E_{i,\epsilon_j}(t)\}$ hold, we must have $\{A_t = 1\}$. The first equality is due to $\theta_1(t)$ is conditionally independent of A'_t and $E_{i,\epsilon_j}(t)$ given $\mathcal{F}_{t-1}$. The second equality is due to that these events are mutually exclusive. For the last inequality, let $C = \{\exists i \in S_j : A_t = i, E_{i,\epsilon}(t) \text{ occurs}\}$, $A = \{\exists i \in S_j : A'_t = i, E_{i,\epsilon}(t) \text{ occurs}\}$ and $B = \{\theta_1(t) \leq \mu_1 - \epsilon\}$. Then A and B are conditionally independent given $\mathcal{F}_{t-1}$. Besides, if C happens, then $A_t = i$ and $\theta_i(t) \leq \mu_1 - \epsilon$ for some $i \in S_j$. This implies $\theta_1(t) \leq \mu_1 - \epsilon$. Therefore, if C happens, we must have $A'_t = i, E_{i,\epsilon}(t)$ occurs for some $i \in S_j$ and $\theta_1(t) \leq \mu_1 - \epsilon$. Therefore, $C \subseteq A \cap B$ and

$$\begin{aligned}
&\sum_{i \in S_j} \mathbb{P}(A_t = i, E_{i,\epsilon_j}(t) \mid \mathcal{F}_{t-1}) \\
&= \mathbb{P}(\exists i \in S_j : A_t = i, E_{i,\epsilon}(t) \text{ occurs} \mid \mathcal{F}_{t-1}) \\
&= \mathbb{P}(C \mid \mathcal{F}_{t-1}) \\
&\leq \mathbb{P}(A \cap B \mid \mathcal{F}_{t-1}) \\
&\leq \mathbb{P}(A \mid \mathcal{F}_{t-1}) \cdot \mathbb{P}(B \mid \mathcal{F}_{t-1}) \\
&= \mathbb{P}(\exists i \in S_j : A'_t = i, E_{i,\epsilon}(t) \text{ occurs} \mid \mathcal{F}_{t-1}) \cdot \mathbb{P}(\theta_1(t) \leq \mu_1 - \epsilon \mid \mathcal{F}_{t-1}) \\
&= \sum_{i \in S_j} \mathbb{P}(A'_t = i, E_{i,\epsilon}(t) \text{ occurs} \mid \mathcal{F}_{t-1}) \cdot \big(1 - \mathbb{P}(\theta_1(t) > \mu_1 - \epsilon \mid \mathcal{F}_{t-1})\big).
\end{aligned} \tag{E.8}$$

Note that $G_{1T_1(t-1)}(\epsilon) = \mathbb{P}(\theta_1(t) \geq \mu_1 - \epsilon \mid \mathcal{F}_{t-1})$. (E.8) implies the last inequality of (E.7).

Consider two cases. **Case 1:** $t : T_1(t-1) \leq M_j$**.** We have

$$\mathbb{E}\left[\sum_{t:T_1(t-1) \leq M_j} \sum_{i \in S_j} \mathbb{P}(A_t = i, E_{i,\epsilon_j}(t))\right] \leq \mathbb{E}\left[\sum_{t:T_1(t-1) \leq M_j} \bigg(\frac{1}{G_{1T_1(t-1)}(\epsilon_j)} - 1\bigg) \mathbb{P}(A_t = 1 \mid \mathcal{F}_{t-1})\right]$$

$$
\begin{aligned}
&= \mathbb{E}\left[\sum_{t:T_1(t-1)\le M_j}\left(\frac{1}{G_{1T_1(t-1)}(\epsilon_j)}-1\right)\mathbb{1}\{A_t=1\}\right]\\
&\le \mathbb{E}\left[\sum_{s=1}^{M_j}\left(\frac{1}{G_{1s}(\epsilon_j)}-1\right)\right]. \qquad \text{(E.9)}
\end{aligned}
$$

The first inequality is from (E.7) The first equality is due to $\mathbb{E}[\mathbb{1}\{A_t=1\}]=\mathbb{P}(A_t=1\mid\mathcal{F}_{t-1})$. For the last inequality, note that due to the indicator function, the summation in first equality is not zero only when $\mathbb{1}\{A_t=1\}=1$. And $\mathbb{1}\{A_t=1\}=1$ further means that we have pulled the best arm (arm 1) at time t. Therefore, the summation over all $T_1(t-1)$ conditional on $\mathbb{1}\{A_t=1\}=1$ is equivalent to the summation over s, which is the number of pulls of arm 1.

Case 2: $t:T\ge T_1(t-1)>M_j$**.** For this case, we have

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{t:T_1(t-1)>M_j}^{T}\mathbb{1}\{A_t=i,E_{i,\epsilon_j}(t)\}\right]\\
&\le \mathbb{E}\left[\sum_{t:T_1(t-1)>M_j}^{T}\mathbb{1}\{\theta_1(t)<\mu_1-\epsilon_j\}\right]\\
&\le T\cdot\mathbb{P}\big(\exists s>M_j:\widehat{\mu}_{1s}<\mu_1-\epsilon_j/2\big)\\
&\quad+\mathbb{E}\left[\sum_{t:T_1(t-1)>M_j}\mathbb{P}\Big(\{\theta_1(t)<\mu_1-\epsilon_j\mid\widehat{\mu}_{1T_1(t-1)}\ge\mu_1-\epsilon_j/2\}\Big)\right]\\
&\le T\cdot e^{-M_j(\mu_1-(\mu_1-\epsilon_j/2))^2/(2V)}+T\cdot e^{-M_j\epsilon^2/(16V)}\\
&\le \frac{2VK}{\epsilon_j^2}, \qquad \text{(E.10)}
\end{aligned}
$$

In the first inequality, we use the fact that $\{A_t=i,E_{i,\epsilon}(t)\}\subseteq\{\theta_1(t)<\mu_1-\epsilon\}$. In the second inequality we use the same argument as the third inequality of (A.7). In the third inequality, we use Lemma H.1 and the following results

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{t:T_1(t-1)>M_j}\mathbb{P}(\theta_1(t)\le\mu_1-\epsilon_j\mid\hat{\mu}_{1T_1(t-1)}\ge\mu_1-\epsilon_j/2)\right]\\
&\le \mathbb{E}\left[\sum_{t:T_1(t-1)>M_j}1/2e^{-T_1(t-1)b_{T_1(t-1)}\mathrm{kl}(\hat{\mu}_{1T_1(t-1)},\mu_1)}\right]\\
&\le \mathbb{E}\left[\sum_{t:T_1(t-1)>M_j}e^{-\frac{M_j}{2}\mathrm{kl}(\hat{\mu}_{1T_1(t-1)},\mu_1)}\right]\\
&\le \mathbb{E}\left[\sum_{t:T_1(t-1)>M_j}e^{-\frac{M_j}{2}\mathrm{kl}(\mu_1-\epsilon_j/2,\mu_1)}\right]\\
&\le T\cdot e^{-M_j\epsilon_j^2/(16V)},
\end{aligned}
$$

where the first inequality is due to the facts that $\theta_1(t)\sim\mathcal{P}^+(\hat{\mu}_{1T_1(t-1)},T_1(t-1))$, $\hat{\mu}_{1T_1(t-1)}\ge\mu_1-\epsilon_j$, and (4.5), the second inequality is due to the facts that $b_s\ge1/2$ for any $s>1$ and $T_1(t-1)>M_j$, the third inequality is due to $\hat{\mu}_{1T_1(t-1)}\ge\mu_1-\epsilon_j/2$ and the fact that from Proposition 3.2, $\mathrm{kl}(x,\mu_1-\epsilon)$ is increasing for $x>\mu_1-\epsilon/2$, and the last inequality is due to (3.3). Combining (E.6), (E.9), and (E.10) together, we complete the proof of this lemma.

E.2.2 Proof of Lemma E.2

Let $p(x)$ be the PDF of $\widehat{\mu}_{1s}$ and θ_{1s} be a sample from $\mathcal{P}(\widehat{\mu}_{1s}, s)$. We have

$$\begin{aligned}
&\sum_{s=1}^{M_j} \mathbb{E}_{\widehat{\mu}_{1s}}\left[\left(\frac{1}{G_{1s}(\epsilon)} - 1\right) \cdot \mathbb{1}\{\widehat{\mu}_{1s} \in L_s\}\right] \\
&\leq \underbrace{\sum_{s=1}^{M_j} \left(\int_{\mu_1 - \epsilon_j/2}^{R_{\max}} p(x)/\mathbb{P}(\theta_{1s} \geq \mu_1 - \epsilon_j \mid \widehat{\mu}_{1s} = x) \mathrm{d}x - 1 \right)}_{A_1} \\
&\quad + \underbrace{\sum_{s=1}^{M_j} \left(\int_{\mu_1 - \epsilon_j}^{\mu_1 - \epsilon_j/2} p(x)/\mathbb{P}(\theta_{1s} \geq \mu_1 - \epsilon_j \mid \widehat{\mu}_{1s} = x) \mathrm{d}x \right)}_{A_2} \\
&\quad + \underbrace{\sum_{s=1}^{M_j} \int_{\mu_1 - \epsilon_j - \alpha_s}^{\mu_1 - \epsilon_j} \Big[p(x)/\mathbb{P}(\theta_{1s} \geq \mu_1 - \epsilon_j \mid \widehat{\mu}_{1s} = x) \Big] \mathrm{d}x}_{A_2}, \qquad \text{(E.11)}
\end{aligned}$$

where the inequality is due to the definition of L_s.

Bounding term A_1. Similar to the bounding term A_1 in Lemma A.2, we divide $\sum_{s=1}^{M_j}$ into two term, i.e., $\sum_{s=1}^{\lfloor 32V/\epsilon_j^2 \rfloor}$ and $\sum_{s=\lceil 32V/\epsilon_j^2 \rceil}^{M_j}$. We have

$$\begin{aligned}
A_1 &= \sum_{s=1}^{M_j} \left(\int_{\mu_1 - \epsilon_j/2}^{R_{\max}} \frac{p(x)}{\mathbb{P}(\theta_{1s} \geq \mu_1 - \epsilon_j \mid \widehat{\mu}_{1s} = x)} \mathrm{d}x - 1 \right) \\
&\leq \frac{32V}{\epsilon_j^2} + \sum_{s=\lceil 32V/\epsilon_j^2 \rceil}^{M_j} \left(\int_{\mu_1 - \epsilon_j/2}^{R_{\max}} \frac{p(x)}{\mathbb{P}(\theta_{1s} \geq \mu_1 - \epsilon_j \mid \widehat{\mu}_{1s} = x)} \mathrm{d}x - 1 \right) \\
&\leq \frac{32V}{\epsilon_j^2} + \sum_{s=\lceil 32V/\epsilon_j^2 \rceil}^{M_j} \left(\frac{1}{1 - e^{-s/2 \cdot \mathrm{kl}(\mu_1 - \epsilon_j/2, \mu_1 - \epsilon_j)}} - 1 \right) \\
&\leq \frac{32V}{\epsilon_j^2} + \sum_{s=\lceil 32V/\epsilon_j^2 \rceil}^{M_j} \left(\frac{1}{1 - e^{-s\epsilon_j^2/(16V)}} - 1 \right) \\
&= \frac{16V}{\epsilon_j^2} + \sum_{s=\lceil 32V/\epsilon_j^2 \rceil}^{M_j} \frac{1}{e^{s\epsilon_j^2/(16V)} - 1} \\
&\leq \frac{32V}{\epsilon_j^2}, \qquad \text{(E.12)}
\end{aligned}$$

For the first inequality, we use the fact that with probability at least $1 - 1/K \geq 1/2$, $\theta_{1s} = \hat{\mu}_{1s} \geq \mu_1 - \epsilon$. For second inequality we use the fact that for $\theta_{1s} = \hat{\mu}_{1s}$, $\theta_{1s} \geq \mu_1 - \epsilon$; for $\theta_{1s} \sim \mathcal{P}$, from (4.5), $\mathbb{P}(\theta_{1s} \geq \mu_1 - \epsilon \mid \widehat{\mu}_{1s} = x) \geq 1 - e^{-sb_s\epsilon^2/(16V)}$. The third inequality is due to (3.3).

Bounding term A_2. This part is the same as the bounding term A_2 in Lemma A.2, thus we omit the details.

Bounding term A_3. The proof of bounding term A_3 is similar to that of proofs in Lemma A.2. We also omit the details. The results are as follows.

$$A_3 \leq K \sum_{s=1}^{M_j} \int_{\mu_1 - \epsilon_j - \alpha_s}^{\mu_1 - \epsilon_j} p(x) e^{sb_s \cdot \mathrm{kl}(x, \mu_1 - \epsilon_j)} \mathrm{d}x \qquad \text{(E.13)}$$

$$\leq K \sum_{s=1}^{M_j} e^{-s\epsilon_j^2/(2V)} \Big(1 + 4 \log(T/(Ks)) \Big) \qquad \text{(E.14)}$$

$$= 16\left(\frac{VK\log(T\epsilon_j^2/(KV))}{\epsilon_j^2}\right), \tag{E.15}$$

where the first inequality is due to the fact that with probability $1/K$, we sample from $\mathcal{P}$. Substituting the bound of A_1, A_1 and A_3 to (E.11), we have that there exists a universal constant C_3 such that

$$\sum_{s=1}^{M_j}\mathbb{E}_{\widehat{\mu}_{1s}}\left[\left(\frac{1}{G_{1s}(\epsilon_j)}-1\right)\cdot\mathbb{1}\{\widehat{\mu}_{1s}\in L_s\}\right]\le C_3\cdot\left(\frac{VK\log(T\epsilon_j^2/(VK))}{\epsilon_j^2}\right), \tag{E.16}$$

which completes the proof.

E.2.3 Proof of Lemma E.3

Similar to (A.16), we can obtain

$$\mathbb{E}\left[\sum_{t=K+1}^{T}\mathbb{1}\{A_t=i,E_{i,\epsilon_j}^c(t)\}\right]\le\mathbb{E}\left[\sum_{s=1}^{T}\mathbb{1}\{G_{is}(\epsilon_j)>V/(T\epsilon_j^2)\}\right]+\frac{V}{\epsilon_j^2}. \tag{E.17}$$

Let $s\ge N_i\log(T\epsilon_j^2/(VK))/\mathrm{kl}(\mu_i+\rho_i,\mu_1-\epsilon_j)$. Let $s\ge N\log(T\epsilon^2/V)/\mathrm{kl}(\mu_i+\rho_i,\mu_1-\epsilon)$. Note that

$$\frac{1}{N_i}=\max\left\{1-\mathrm{kl}(\mu_i+\rho_i,\mu_1-\epsilon_j)/\log(T\epsilon_j^2/V),\frac{1}{2}\right\}.$$

For case $1/N_i=1/2$, we have

$$b_s\ge 1/2=1/N_i. \tag{E.18}$$

For case $1/N_i=1-\mathrm{kl}(\mu_i+\rho_i,\mu_1-\epsilon_j)/\log(T\epsilon_j^2/V)$, we have

$$\begin{aligned}b_s&\ge 1-1/s\\&\ge 1-\mathrm{kl}(\mu_i+\rho_i,\mu_1-\epsilon_j)/(N\log(T\epsilon_j^2/V))\\&\ge 1-\mathrm{kl}(\mu_i+\rho_i,\mu_1-\epsilon_j)/(\log(T\epsilon_j^2/V))\\&=1/N_i,\end{aligned} \tag{E.19}$$

where the second inequality is due to $s\ge N_i\log(T\epsilon_j^2/V)/\mathrm{kl}(\mu_i+\rho_i,\mu_1-\epsilon_j)$ and the third inequality is due to $N_i>1$. Let X_{is} be a sample from the distribution $\mathcal{P}^+(\widehat{\mu}_{is},s)$. Assume $\widehat{\mu}_{is}\le\mu_i+\rho_i$. Then from definition of $\mathcal{P}^+(\widehat{\mu}_{is},s)$, with probability $1-1/K$, $\widehat{\mu}_{is}\le\mu_i+\rho_i$; with probability $1/K$, X_{is} is a random sample from $\mathcal{P}(\widehat{\mu}_{is},s)$. Therefore if $\widehat{\mu}_{is}\le\mu_i+\rho_i$ and $s\ge N_i$, we have

$$\begin{aligned}\mathbb{P}(X_{is}\ge\mu_1-\epsilon_j)&\le\exp(-sb_s\mathrm{kl}(\widehat{\mu}_{is},\mu_1-\epsilon_j))/K\\&\le\exp(-sb_s\mathrm{kl}(\mu_i+\rho_i,\mu_1-\epsilon_j))/K\\&\le\frac{V}{T\epsilon_j^2},\end{aligned} \tag{E.20}$$

where the first inequality is from (4.4) and the definition of $\mathcal{P}^+(\mu,n)$, the second inequality is due to the assumption $\widehat{\mu}_{is}\le\mu_i+\rho_i$, and the last inequality is due to $s\ge N_i\log(T\epsilon_j^2/(VK))/\mathrm{kl}(\mu_i+\rho_i,\mu_1-\epsilon_j)$ and $b_s\ge 1/N_i$ from (E.18) and (E.19). The rest of proofs are similar to the proofs in Theorem 4.2. Note that when $\mathbb{P}(X_{is}\ge\mu_1-\epsilon_j)\le V/(T\epsilon_j^2)$ holds, term $\mathbb{1}\{G_{is}(\epsilon_j)>V/(T\epsilon_j^2)\}=0$. Now, we check the assumption $\hat{\mu}_{is}\le\mu_i+\rho_i$ that is needed for (E.20). From Lemma H.1, we have $\mathbb{P}(\widehat{\mu}_{is}>\mu_i+\rho_i)\le\exp(-s\rho_i^2/(2V))$. Furthermore, it holds that

$$\sum_{s=1}^{\infty}e^{-\frac{s\rho_i^2}{2V}}\le\frac{1}{e^{\rho_i^2/(2V)}-1}\le\frac{2V}{\rho_i^2}, \tag{E.21}$$

where the last inequality is due to the fact $1+x\le e^x$ for all x. Let Y_{is} be the event that $\widehat{\mu}_{is}\le\mu_i+\rho_i$ and $m=N_i\log(T\epsilon_j^2/(VK))/\mathrm{kl}(\mu_i+\rho_i,\mu_1-\epsilon_j)$. We further obtain

$$\mathbb{E}\left[\sum_{s=1}^{T}\mathbb{1}\{G_{is}(\epsilon)>V/(T\epsilon_j^2)\}\right]\le\mathbb{E}\left[\sum_{s=1}^{T}[\mathbb{1}\{G_{is}(\epsilon)>V/(T\epsilon_j^2)\}\mid Y_{is}]\right]+\sum_{s=1}^{T}(1-\mathbb{P}[Y_{is}])$$

$$\le \mathbb{E}\left[\sum_{s=\lceil m\rceil}^{T} [\mathbb{1}\{\mathbb{P}(X_{is} > \mu_1 - \epsilon_j) > V/(T\epsilon_j^2))\} \mid Y_{is}]\right]$$
$$+ \lceil m\rceil + \sum_{s=1}^{T}(1 - \mathbb{P}[Y_{is}])$$
$$\le \lceil m\rceil + \sum_{s=1}^{T}(1 - \mathbb{P}[Y_{is}])$$
$$\le 1 + \frac{2V}{\rho_i^2} + \frac{N_i \log(T\epsilon_j^2/V)}{\mathrm{kl}(\mu_i + \rho_i, \mu_1 - \epsilon_j)}, \tag{E.22}$$

where the first inequality is due to the fact that $\mathbb{P}(A) \le \mathbb{P}(A \mid B) + 1 - \mathbb{P}(B)$, the third inequality is due to (E.20) and the last inequality is due to (E.21). Substituting (E.22) into (E.17), we complete the proof.

F Proof of the Asymptotic Optimality of ExpTS$^+$

Now we prove the asymptotic regret bound of ExpTS$^+$ presented in Theorem 5.1.

F.1 Proof of the Main Result

The proof of the this part shares many elements with finite time regret analysis. In what follows, we bound terms A and B, respectively.

Bounding Term A: We reuse the Lemma E.1. Then, it only remains term $\sum_{s=1}^{M_j} \mathbb{E}\big[\big(1/G_{1s}(\epsilon) - 1\big) \cdot \mathbb{1}\{\widehat{\mu}_{1s} \in L_s\}\big]$ to be bounded. We bound this term by the following lemma.

Lemma F.1. *Let M_j, $G_{1s}(\epsilon_j)$, and L_s be the same as defined in Lemma E.1.*

$$\sum_{s=1}^{M_j} \mathbb{E}_{\widehat{\mu}_{1s}}\left[\left(\frac{1}{G_{1s}(\epsilon_j)} - 1\right) \cdot \mathbb{1}\{\widehat{\mu}_{1s} \in L_s\}\right] = O(V^2 K (\log\log T)^6 + V(\log\log T)^2 + K).$$

Combining Lemma E.1 and Lemma F.1 together and let $\epsilon = 1/\log\log T$, we have

$$A = O\big((V^2 K(\log\log T)^6 + V(\log\log T)^2 + K\big).$$

Bounding Term B: Let $\rho_i = \epsilon_j = 1/\log\log T$. Applying (E.3), we have

$$\mathbb{E}\left[\sum_{t=K+1}^{T} \mathbb{1}\{A_t = i, E_{i,\epsilon_j}^c(t)\}\right] = O(V\log^2\log T) + \frac{N_i \log(T\epsilon_j^2/V)}{\mathrm{kl}(\mu_i + 1/\log\log T, \mu_1 - 1/\log\log T)}. \tag{F.1}$$

Therefore, we have

$$B = \sum_{i\in S_j} \mathbb{E}\left[\sum_{t=K+1}^{T} \mathbb{1}\{A_t = i, E_{i,\epsilon_j}^c(t)\}\right]$$
$$\le O(VK\log^2\log T) + \sum_{i\in S_j} \frac{N_i \log(T/(V\log^2\log T))}{\mathrm{kl}(\mu_i + 1/\log\log T, \mu_1 - 1/\log\log T)}.$$

Putting It Together: Substituting the bound of term A and B into (E.2), we have

$$\sum_{i\in S_j} \mathbb{E}[T_i(T)] = O(V^2 K\log^6\log T + VK\log^2\log T + K) + \sum_{i\in S_j} \frac{N_i \log(T/(V\log^2\log T))}{\mathrm{kl}(\mu_i + 1/\log\log T, \mu_1 - 1/\log\log T)}.$$

Note that for $T \to \infty$, $N_i \to 1$. Therefore,

$$\lim_{T\to\infty} \sum_{i\in S_j} \frac{\mathbb{E}[T_i(T)]}{\log T} = \sum_{i\in S_j} \frac{1}{\mathrm{kl}(\mu_i, \mu_1)}.$$

This completes the proof of the asymptotic regret.

F.2 Proof of Lemma F.1

The proof of this Lemma shares many elements with the proof of Lemma E.2. We can use the bound of A_1 and A_2 in Lemma E.2. Let $\epsilon = 1/\log\log T$, we have

$$A_1 + A_2 = O(V(\log\log T)^2).$$

For term A_3, from (E.13), we have

$$A_3 \leq K \sum_{s=1}^{M_j} \int_{\mu_1-\epsilon_j-\alpha_s}^{\mu_1-\epsilon_j} p(x) e^{s b_s \cdot \mathrm{kl}(x, \mu_1-\epsilon_j)} \mathrm{d}x.$$

Then, similar to (B.3), we have

$$K \int_{\mu_1-\epsilon_j-\alpha_s}^{\mu_1-\epsilon_j} p(x) e^{s b_s \cdot \mathrm{kl}(x, \mu_1-\epsilon_j)} \mathrm{d}x = O(Ks). \tag{F.2}$$

Let $\epsilon = 1/\log\log T$. By dividing $\sum_{s=1}^{M_j}$ into two terms $\sum_{s=1}^{\lceil 4V(\log\log T)^3 \rceil}$ and $\sum_{s=\lceil 4V(\log\log T)^3 \rceil}^{M_j}$, from (B.4) and (B.5), we have

$$A_3 = O(V^2 K(\log\log T)^6 + VK(\log\log T)^2 + K).$$

Substituting the bound of A_1, A_1 and A_3 to (E.11), we have

$$A = O(V^2 K(\log\log T)^6 + VK(\log\log T)^2 + K).$$

This completes the proof.

G Proof of Technical Lemmas

In this section, we present the proofs of the remaining lemmas used in previous sections.

G.1 Proof of Lemma A.4

Let $\mathrm{kl}_+(x, y) = \mathrm{kl}(x, y)\, \mathbb{1}(x \leq y)$. We only need to prove

$$\mathbb{P}\bigg(\exists s \leq f(\epsilon) : \mathrm{kl}_+(\hat{\mu}_{1s}, \mu_1) \geq 4\log(T/(bs))/s\bigg) = O\bigg(\frac{bV}{T\epsilon^2}\bigg).$$

The proof of this step relies on the standard "peeling technique". We have

$$\begin{aligned}
&\mathbb{P}\bigg(\exists s \leq f(\epsilon) : \mathrm{kl}_+(\hat{\mu}_{1s}, \mu_1) \geq 4\log(T/(bs))/s\bigg) \\
&\leq \sum_{n=0}^{\infty} \mathbb{P}\bigg(\exists \frac{f(\epsilon)}{2^{n+1}} \leq s \leq \frac{f(\epsilon)}{2^n} : \mathrm{kl}_+(\hat{\mu}_{1s}, \mu_1) \geq 4\log(T/(bs))/s\bigg) \\
&\leq \sum_{n=0}^{\infty} \mathbb{P}\bigg(\exists \frac{f(\epsilon)}{2^{n+1}} \leq s \leq \frac{f(\epsilon)}{2^n} : \mathrm{kl}_+(\hat{\mu}_{1s}, \mu_1) \geq \frac{4\log(T/(b \cdot f(\epsilon)/2^n))}{M/2^n}\bigg) \\
&\leq \sum_{n=0}^{\infty} \exp\bigg(-\frac{f(\epsilon)}{2^{n+1}} \cdot \frac{4\log(T/(b \cdot f(\epsilon)/2^n))}{f(\epsilon)/2^n} \cdot\bigg) \\
&= \sum_{n=0}^{\infty} \exp\big(-2\log(T/(b \cdot f(\epsilon)/2^n))\big) \\
&\leq \sum_{n=0}^{\infty} \bigg(\frac{bf(\epsilon)}{T \cdot 2^n}\bigg)^2,
\end{aligned} \tag{G.1}$$

where the third inequality is due to Lemma H.1. Note that $f(\epsilon) \leq 32V\log(T\epsilon^2/(bV))/\epsilon^2$. We have

$$\frac{bf(\epsilon)}{T \cdot 2^n} \leq \frac{bf(\epsilon)}{T} \leq 32\log\bigg(\frac{T\epsilon^2}{bV}\bigg) \cdot \frac{bV}{T\epsilon^2}.$$

Continue on equation (G.1), we have

$$\begin{aligned}\sum_{n=0}^{\infty}\left(\frac{bf(\epsilon)}{T\cdot 2^n}\right)^2 &\le \sum_{n=0}^{\infty}\left(\frac{bf(\epsilon)}{T\cdot 2^n}\cdot 32\log\left(\frac{T\epsilon^2}{bV}\right)\cdot\frac{bV}{T\epsilon^2}\right)\\ &\le 32^2\sum_{n=0}^{\infty}\left(\frac{bV}{T\epsilon^2\cdot 2^n}\cdot\left(\log\left(\frac{T\epsilon^2}{bV}\right)\right)^2\cdot\frac{bV}{T\epsilon^2}\right)\\ &\le 32^2\cdot\frac{2bV}{T\epsilon^2},\end{aligned}$$

where the last inequality is due to $(\log(x))^2/x \le 1$ for $x \ge 1$.

G.2 Proof of Lemma A.5

We decompose the proof of Lemma A.5 into two cases: Case 1: $p(x)$ is in continues form, and Case 2: $p(x)$ is in discrete form. We first focus on the case that $p(x)$ is in continues form.

Divide the interval $[x_0, x_n]$ into n sub-intervals $[x_0, x_1)$,$[x_1, x_2), \cdots, [x_{n-1}, x_n]$, such that $\mu_1 - \epsilon - \alpha_s = x_0 \le x_1 \le \cdots \le x_{n-1} \le x_n = \mu_1 - \epsilon - b_s$, $\int_{x_0}^{x_1} q(x)\mathrm{d}x = \int_{x_{i-1}}^{x_i} q(x)\mathrm{d}x$ for all $i \in [n]$. We now define a new function $p_n(x)$. Assume $p_n(x)$ has been defined on $[x_i, x_n]$. We define $p_n(x)$ on $[x_{i-1}, x_i)$ in the following way. We consider two cases.
Case 1: $\int_{x_i}^{x_n} p_n(x)\mathrm{d}x + \int_{x_{i-1}}^{x_i} p(x)\mathrm{d}x \ge e^{-s\cdot\mathrm{kl}(x_n,\mu_1)} - e^{-s\cdot\mathrm{kl}(x_{i-1},\mu_1)}$. Then, we define the function $p_n(x) = p(x)$ for $x \in [x_{i-1}, x_i)$.
Case 2: $\int_{x_i}^{x_n} p_n(x)\mathrm{d}x + \int_{x_{i-1}}^{x_i} p(x)\mathrm{d}x < e^{-s\cdot\mathrm{kl}(x_n,\mu_1)} - e^{-s\cdot\mathrm{kl}(x_{i-1},\mu_1)}$. Let $\beta = e^{-s\cdot\mathrm{kl}(x_n,\mu_1)} - e^{-s\cdot\mathrm{kl}(x_{i-1},\mu_1)} - \int_{x_i}^{x_n} p_n(x)\mathrm{d}x - \int_{x_{i-1}}^{x_i} p(x)\mathrm{d}x$. Then, define $p_n(x) = p(x) + \beta/(x_i - x_{i-1})$. Hence, for case 2, it holds that

$$\int_{x_{i-1}}^{x_n} p_n(x)\mathrm{d}x = e^{-s\cdot\mathrm{kl}(x_n,\mu_1)} - e^{-s\cdot\mathrm{kl}(x_{i-1},\mu_1)}. \tag{G.2}$$

Let $y_n = x_n = \mu_1 - \epsilon$. For all $i \in [n]$, define $y_i = x_i$ if $\int_{x_i}^{y_n} p_n(x)\mathrm{d}x = e^{-s\cdot\mathrm{kl}(\mu_1-\epsilon-b_s,\mu_1)} - e^{-s\cdot\mathrm{kl}(x_i,\mu_1)}$. Otherwise, define y_i such that

$$\int_{y_i}^{y_n} p_n(x)\mathrm{d}x = e^{-s\cdot\mathrm{kl}(\mu_1-\epsilon-b_s,\mu_1)} - e^{-s\cdot\mathrm{kl}(x_i,\mu_1)}.$$

From the definition, we know

$$x_i \le y_i. \tag{G.3}$$

Since $p_n(x) \ge p(x)$ holds for any $x \in [x_0, x_n]$ and $g(x) \ge 0$, we have

$$\int_{\mu_1-\epsilon-\alpha_s}^{\mu_1-\epsilon-b_s} p(x)g(x)\mathrm{d}x \le \int_{\mu_1-\epsilon-\alpha_s}^{\mu_1-\epsilon-b_s} p_n(x)g(x)\mathrm{d}x. \tag{G.4}$$

Note that $g(x)$ is monotone decreasing

$$\begin{aligned}g'(x) = \left(e^{sb_s\cdot\mathrm{kl}(x,\mu_1-\epsilon)}\right)' &= sb_s(\mathrm{kl}(x,\mu_1-\epsilon))'e^{sb_s\cdot\mathrm{kl}(x,\mu_1-\epsilon)}\\ &= sb_s e^{sb_s\cdot\mathrm{kl}(x,\mu_1-\epsilon)}\left(\int_x^{\mu_1-\epsilon}\frac{t-x}{V(t)}\mathrm{d}t\right)'\\ &= sb_s e^{sb_s\cdot\mathrm{kl}(x,\mu_1-\epsilon)}\cdot\int_x^{\mu_1-\epsilon}\frac{-1}{V(t)}\mathrm{d}t\\ &\le 0.\end{aligned}$$

We have

$$\sum_{i=0}^{n-1}\int_{y_i}^{y_{i+1}} p_n(x)g(x)\mathrm{d}x \le \sum_{i=0}^{n-1} g(y_i)\int_{y_i}^{y_{i+1}} p_n(x)\mathrm{d}x$$

$$\leq \sum_{i=0}^{n-1} g(x_i) \int_{y_i}^{y_{i+1}} p_n(x)\mathrm{d}x = \sum_{i=0}^{n-1} g(x_i) \int_{x_i}^{x_{i+1}} q(x)\mathrm{d}x$$
$$\leq \sum_{i=0}^{n-1} \int_{x_i}^{x_{i+1}} q(x)g(x)\mathrm{d}x + \sum_{i=0}^{n-1}(g(x_i) - g(x_{i+1})) \int_{x_i}^{x_{i+1}} q(x)\mathrm{d}x$$
$$= \sum_{i=0}^{n-1} \int_{x_i}^{x_{i+1}} q(x)g(x)\mathrm{d}x + (g(x_0) - g(x_n]) \int_{x_0}^{x_1} q(x)\mathrm{d}x$$
$$\leq \sum_{i=0}^{n-1} \int_{x_i}^{x_{i+1}} q(x)g(x)\mathrm{d}x + (g(x_0) - g(x_n))/n. \tag{G.5}$$

In the first inequality, we use the fact that $g(x)$ is monotone decreasing. The second inequality is due to (G.3). The first equality is from the fact that

$$\int_{x_i}^{x_{i+1}} q(x)\mathrm{d}x = e^{-s\cdot\mathrm{kl}(x,\mu_1)}\Big|_{x_i}^{x_{i+1}} = e^{-s\cdot\mathrm{kl}(x_{i+1},\mu_1)} - e^{-s\cdot\mathrm{kl}(x_i,\mu_1)},$$

and the definition of y_i such that

$$\int_{y_i}^{y_{i+1}} p_n(x)\mathrm{d}x = \int_{y_i}^{y_n} p_n(x)\mathrm{d}x - \int_{y_{i+1}}^{y_n} p_n(x)\mathrm{d}x = e^{-s\cdot\mathrm{kl}(x_{i+1},\mu_1)} - e^{-s\cdot\mathrm{kl}(x_i,\mu_1)} = \int_{x_i}^{x_{i+1}} q(x)\mathrm{d}x.$$

The third inequality is due to $\sum_{i=0}^{n-1} \int_{x_{i+1}}^{x_i} q(x)g(x)\mathrm{d}x \geq \sum_{i=0}^{n-1} g(x_i) \int_{x_{i+1}}^{x_i} q(x)\mathrm{d}x$. Now, we focus on bounding term $\int_{x_0}^{y_0} p_n(x)g(x)\mathrm{d}x$. Note that

$$\int_{x_0}^{y_0} p_n(x)g(x)\mathrm{d}x \leq g(x_0) \int_{x_0}^{y_0} p_n(x)\mathrm{d}x.$$

Hence, we only need to bound $\int_{x_0}^{y_0} p_n(x)\mathrm{d}x$. Let

$$n' = \min\big\{j \in \{0, \cdots, n\} : p_n(x) = p(x) \text{ for all } \ x \in [x_0, x_j)\big\}.$$

From the definition, for $x < x_{n'}$, $p_n(x) = p(x)$. Besides, for $x \in [x_{n'}, x_{n'+1})$, it must belong to case 2 in the definition of $p_n(x)$. Hence,

$$\int_{x_{n'}}^{x_n} p_n(x)\mathrm{d}x = e^{-s\mathrm{kl}(x_n,\mu_1)} - e^{-s\mathrm{kl}(x_{n'},\mu_1)}.$$

Therefore, $y_{n'} = x_{n'}$. Further, from Lemma H.1, we have

$$\int_{x_0}^{y_{n'}} p_n(x)\mathrm{d}x = \int_{x_0}^{y_{n'}} p(x)\mathrm{d}x \leq \Pr(\widehat{\mu}_{1s} \leq y_{n'}) \leq e^{-s\cdot\mathrm{kl}(y_{n'},\mu_1)}. \tag{G.6}$$

Now, we have

$$\int_{x_0}^{y_0} p_n(x)\mathrm{d}x = \int_{x_0}^{y_{n'}} p_n(x)\mathrm{d}x - \int_{y_0}^{y_{n'}} p_n(x)\mathrm{d}x$$
$$\leq e^{-s\mathrm{kl}(y_{n'},\mu_1)} - \bigg(\int_{y_0}^{x_n} p_n(x)\mathrm{d}x - \int_{y_{n'}}^{x_n} p_n(x)\mathrm{d}x\bigg)$$
$$= e^{-s\mathrm{kl}(y_{n'},\mu_1)} - \big(e^{-s\mathrm{kl}(\mu_1-\epsilon,\mu_1)} - e^{-s\mathrm{kl}(x_n,\mu_1)} - e^{-s\mathrm{kl}(\mu_1-\epsilon,\mu_1)} + e^{-s\mathrm{kl}(x_{n'},\mu_1)}\big)$$
$$= e^{-s\mathrm{kl}(x_0,\mu_1)}, \tag{G.7}$$

where the first inequality is due to (G.6), and the last inequality we use the fact $y_{n'} = x_{n'}$. Finally, we have

$$\int_{\mu_1-\epsilon-\alpha_s}^{\mu_1-\epsilon} p(x)g(x)\mathrm{d}x$$
$$\leq \int_{\mu_1-\epsilon-\alpha_s}^{\mu_1-\epsilon} p_n(x)g(x)\mathrm{d}x$$

$$= \sum_{i=0}^{n-1} \int_{y_i}^{y_{i+1}} p_n(x)g(x)\mathrm{d}x + \int_{x_0}^{y_0} p_n(x)g(x)\mathrm{d}x$$
$$\leq \sum_{i=0}^{n-1} \int_{x_i}^{x_{i+1}} q(x)g(x)\mathrm{d}x + (g(x_0) - g(x_n])/n + g(x_0) \int_{x_0}^{y_0} p_n(x)\mathrm{d}x$$
$$\leq \sum_{i=0}^{n-1} \int_{x_i}^{x_{i+1}} q(x)g(x)\mathrm{d}x + (g(x_0) - g(x_n])/n + g(\mu_1 - \epsilon - \alpha_s)e^{-s\mathrm{kl}(\mu_1 - \epsilon - \alpha_s, \mu_1)}$$
$$= \int_{\mu_1 - \epsilon - \alpha_s}^{\mu_1 - \epsilon - b_s} q(x)g(x)\mathrm{d}x + g(\mu_1 - \epsilon - \alpha_s)e^{-s\mathrm{kl}(\mu_1 - \epsilon - \alpha_s, \mu_1)} + (g(x_0) - g(x_n))/n,$$

where the second inequality is due to (G.5), and the third inequality is due to (G.7). Note that

$$\lim_{n \to \infty} (g(x_0) - g(x_n))/n = 0.$$

Therefore, it holds that

$$\int_{\mu_1 - \epsilon - \alpha_s}^{\mu_1 - \epsilon} p(x)g(x)\mathrm{d}x \leq \int_{\mu_1 - \epsilon - \alpha_s}^{\mu_1 - \epsilon} q(x)g(x)\mathrm{d}x + g(\mu_1 - \epsilon - \alpha_s)e^{-s\mathrm{kl}(\mu_1 - \epsilon - \alpha_s, \mu_1)} + (g(x_0) - g(x_n))/n,$$

which completes the Lemma for continues form $p(x)$.
Next, we assume $p(x)$ is in discrete form. Let $z_0, \cdots, z_k$ be all the points such that $p(z_i) > 0$ and $z_i \in [\mu_1 - \epsilon - \alpha_s, \mu_1 - \epsilon]$. We need to prove

$$\int_{\mu_1 - \epsilon - \alpha_s}^{\mu_1 - \epsilon - b_s} q(x)g(x)\mathrm{d}x + e^{-s \cdot \mathrm{kl}(\mu_1 - \epsilon - \alpha_s, \mu_1)} \cdot g(\mu_1 - \epsilon - \alpha_s) \geq \sum_{i=0}^{k} p(z_i)g(z_i).$$

We assume $z_0 \leq z_1 \leq \cdots \leq z_k$. Define $h(z_i) = \int_{z_{i-1}}^{z_i} q(x)\mathrm{d}x$ for $i \in [K]$ and $h(z_0) = \int_{\mu_1 - \epsilon - \alpha_s}^{z_0} q(x)\mathrm{d}x$. We have

$$\begin{aligned}
\int_{\mu_1 - \epsilon - \alpha_s}^{\mu_1 - \epsilon} q(x)g(x)\mathrm{d}x - \sum_{i=0}^{k} p(z_i)g(z_i) &\geq \int_{\mu_1 - \epsilon - \alpha_s}^{z_k} q(x)g(x)\mathrm{d}x - \sum_{i=0}^{k} p(z_i)g(z_i) \\
&\geq \sum_{i=1}^{k} g(z_i) \int_{z_{i-1}}^{z_i} q(x)\mathrm{d}x + g(z_0) \int_{\mu_1 - \epsilon - \alpha_s}^{z_0} q(x)\mathrm{d}x \\
&\quad - \sum_{i=0}^{k} p(z_i)g(z_i) \\
&= \sum_{i=0}^{k} (h(z_i) - p(z_i))g(z_i). \qquad \text{(G.8)}
\end{aligned}$$

We define $k' = \min\{j \in \{0, \cdots, k\} : p(z_i) - h(z_i) \geq 0 \text{ for all } i \leq j\}$. If such k' does not exist, then $p(z_i) - h(z_i) < 0$ always holds. Hence, $\sum_{i=0}^{k} (h(z_i) - p(z_i))g(z_i) \geq 0$. Otherwise, k' exists. From lemma H.1, we have

$$\sum_{i=0}^{k'} p(z_i) \leq \Pr(\widehat{\mu}_{1s} \leq z_{k'}) \leq e^{-s \cdot \mathrm{kl}(z_{n'}, \mu_1)}. \qquad \text{(G.9)}$$

Continue on (G.8), we have

$$\begin{aligned}
\sum_{i=0}^{k} (h(z_i) - p(z_i))g(z_i) &\geq -\sum_{i=0}^{k'} (p(z_i) - h(z_i))g(z_i) \\
&\geq -g(\mu_1 - \epsilon - \alpha_s) \sum_{i=0}^{k'} p(z_i) - h(z_i)
\end{aligned}$$

$$
\begin{aligned}
&= -g(\mu_1 - \epsilon - \alpha_s)\bigg(\sum_{i=0}^{k'} p(z_i) - \sum_{i=0}^{k'} h(z_i)\bigg) \\
&\geq -g(\mu_1 - \epsilon - \alpha_s)\bigg(e^{-s\mathrm{kl}(z_{k'},\mu_1)} - \int_{\mu_1-\epsilon-\alpha_s}^{z_{k'}} q(x)\mathrm{d}x\bigg) \\
&= -g(\mu_1 - \epsilon - \alpha_s)\cdot e^{-s\mathrm{kl}(\mu_1-\epsilon-\alpha_s,\mu_1)}, \qquad \text{(G.10)}
\end{aligned}
$$

where the first inequality is from the definition of k', the second inequality is due to that $g(\cdot)$ is monotone decreasing, the first equality is due to $p(z_i) - h(z_i) \geq 0$ for all $i \in \{0, 1, \cdots, k'\}$, the third inequality is due to (G.9), and the last equality is due to

$$
\int_{\mu_1-\epsilon-\alpha_s}^{z_{k'}} q(x)\mathrm{d}x = e^{-s\mathrm{kl}(x,\mu_1)}\bigg|_{\mu_1-\epsilon-\alpha_s}^{z_{k'}} = e^{-s\mathrm{kl}(z_{k'},\mu_1)} - e^{-s\mathrm{kl}(\mu_1-\epsilon-\alpha_s,\mu_1)}.
$$

Combining (G.9) and (G.10), we complete the proof.

H Useful Inequalities

Lemma H.1 (Maximal Inequality [32])**.** *Let N and M be two real numbers in $\mathbb{R}^+ \times \overline{\mathbb{R}^+}$, let $\gamma > 0$, and $\widehat{\mu}_n$ be the empirical mean of n random variables i.i.d. according to the distribution $\nu_{b'^{-1}(\mu)}$. Then, for $x \leq \mu$,*

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

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

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

Lemma H.2 (Tail Bound for Gaussian Distribution)**.** *For a random variable $Z \sim \mathcal{N}(\mu, \sigma^2)$,*

$$
\frac{e^{-z^2/2}}{z\cdot\sqrt{2\pi}} \geq \mathbb{P}(Z > \mu + z\sigma) \geq \frac{1}{\sqrt{2\pi}}\frac{z}{z^2+1}e^{-\frac{z^2}{2}}. \qquad \text{(H.3)}
$$

Besides, for $0 \leq z \leq 1$,

$$
\mathbb{P}(Z > \mu + z\sigma) \geq \frac{1}{\sqrt{8\pi}}e^{-\frac{z^2}{2}}.
$$

Proof. (H.3) is from Abramowitz and Stegun [2]. For the second statement, we have that for $0 \leq z \leq 1$,

$$
\mathbb{P}(Z > \mu + z\sigma) \geq \mathbb{P}(Z > \mu + \sigma) \geq \frac{1}{\sqrt{8\pi}}e^{-\frac{z^2}{2}},
$$

where the last inequality is due to (H.3). □