Near-optimal Regret Bounds for Thompson Sampling[1]

SHIPRA AGRAWAL, Microsoft Research [2]
NAVIN GOYAL, Microsoft Research

Thompson Sampling (TS) is one of the oldest heuristics for multi-armed bandit problems. It is a randomized algorithm based on Bayesian ideas, and has recently generated significant interest after several studies demonstrated that it has favorable empirical performance compared to the state of the art methods. In this paper, a novel and almost tight martingale-based regret analysis for Thompson Sampling is presented. Our technique simultaneously yield both problem-dependent and problem-independent bounds: (1) The first near-optimal problem-independent bound of $O(\sqrt{NT \ln T})$ on the expected regret. (2) The optimal problem-dependent bound of $(1+\epsilon)\sum_i \frac{\ln T}{d(\mu_i,\mu_1)} + O(\frac{N}{\epsilon^2})$ on the expected regret (this bound was first proven by Kaufmann et al. [Kaufmann et al. 2012b]).

Our technique is conceptually simple, and easily extends to distributions other than the Beta distribution used in the original TS algorithm. For the version of TS that uses Gaussian priors, we prove a problem-independent bound of $O(\sqrt{NT \ln N})$ on the expected regret, and show the optimality of this bound by providing a matching lower bound. This is the first lower bound on the performance of a natural version of Thompson Sampling that is away from the general lower bound of $\Omega(\sqrt{NT})$ for the multi-armed bandit problem.

1. INTRODUCTION

The Multi-Armed Bandit problem (MAB) models the exploration/exploitation trade-off inherent in sequential decision problems. Many versions and generalizations of MAB have been studied in the literature; in this paper we will consider a basic and well-studied version of this problem: the stochastic multi-armed bandit problem. Among many algorithms available for stochastic MAB, some popular ones include Upper Confidence Bound (UCB) family of algorithms, (e.g., [Lai and Robbins 1985; Auer et al. 2002], and more recently [Audibert and Bubeck 2009; Garivier and Cappé 2011; Maillard et al. 2011; Kaufmann et al. 2012a]), which have good theoretical guarantees, and the algorithm by [Gittins 1989], which gives optimal strategy under a Bayesian setting with known priors and geometric time-discounted rewards. In one of the earliest works on stochastic MAB, [Thompson 1933] proposed a natural randomized Bayesian algorithm to minimize regret. The basic idea is to assume a simple prior distribution on the parameters of the reward distribution of every arm, and at any time step, play an arm according to its posterior probability of being the best arm. This algorithm is known as *Thompson Sampling* (TS), and it is a member of the family of *randomized probability matching* algorithms. TS is a natural algorithm: the same idea has been rediscovered many times independently in the context of reinforcement learning, e.g., in [Wyatt 1997; Strens 2000; Ortega and Braun 2010].

Recently, TS has attracted considerable attention. Several studies (e.g., [Granmo 2010; Scott 2010; Graepel et al. 2010; Chapelle and Li 2011; May and Leslie 2011; Kaufmann et al. 2012b]) have empirically demonstrated the efficacy of TS. Despite being easy to implement, competitive to the state of the art methods, and being used in practice, TS lacked a strong theoretical analysis until very recently. [Granmo 2010; May et al. 2011] provide weak guarantees, namely, a bound of $o(T)$ on expected regret in time T. Significant progress was made in more recent work of [Agrawal and Goyal 2012] and [Kaufmann et al. 2012b]. In [Agrawal and Goyal 2012], the first logarithmic bound on expected regret of TS was proven. [Kaufmann et al. 2012b] pro-

[1] A preliminary version of this paper was presented at AISTATS 2013 [Agrawal and Goyal 2013a]
[2] Present address: Department of Industrial Engineering and Operations Research, Columbia University, New York, NY 10027.

vided a bound that matches the asymptotic lower bound of [Lai and Robbins 1985] for this problem. However, both of these bounds were problem-dependent, i.e. the regret bounds are logarithmic in the time horizon T when the problem parameters, namely the mean rewards for each arm, and their differences, are assumed to be constants. The problem-independent bounds implied by these works were far from optimal. Obtaining a problem-independent bound that is close to the lower bound of $\Omega(\sqrt{NT})$ was also posed as an open problem by [Li and Chapelle 2012].

In this paper, we give a regret analysis for TS that provides both optimal problem-dependent and near-optimal problem-independent regret bounds. Our novel martingale-based analysis technique is conceptually simple and arguably simpler than the previous work. Our technique easily extends to the distributions other than Beta distribution, and it also extends to the more general contextual bandits setting [Agrawal and Goyal 2013b].

Before stating our results, we describe the stochastic multi-armed bandit problem and TS formally.

1.1. The stochastic multi-armed bandit problem

We consider the stochastic multi-armed bandit problem: We are given a slot machine with N arms; at each time step $t = 1, 2, 3, \ldots$, one of the N arms must be chosen to be played. Each arm i, when played, yields a random real-valued reward according to some fixed unknown distribution associated with arm i with support in $[0, 1]$. The random rewards obtained from playing an arm repeatedly are i.i.d. and independent of the plays of the other arms. The reward is observed immediately after playing the arm.

An algorithm for stochastic MAB must decide which arm to play at each time step t, based on the outcomes of the previous $t-1$ plays. Let μ_i denote the (unknown) expected reward for arm i. A popular goal in designing algorithms for stochastic MAB is to maximize the expected total reward at time T, i.e., $\mathbb{E}[\sum_{t=1}^{T} \mu_{i(t)}]$, where $i(t)$ is the arm played in step t, and the expectation is over the random choices of $i(t)$ made by the algorithm. It is common to work with the equivalent measure of expected total *regret*: the amount we lose because of not playing optimal arm in each step. To formally define regret, let us introduce some notation. Let $\mu^* := \max_i \mu_i$, and $\Delta_i := \mu^* - \mu_i$. Let $k_i(t)$ denote the number of times arm i has been played up to step $t-1$; thus $k_i(t)$ is a random variable. Then the expected total regret in time T is given by

$$\mathbb{E}\left[\mathcal{R}(T)\right] = \mathbb{E}\left[\sum_{t=1}^{T}(\mu^* - \mu_{i(t)})\right] = \sum_i \Delta_i \cdot \mathbb{E}\left[k_i(T+1)\right].$$

To define different notions of expected regret used in this paper, let $\mathbb{E}\left[\mathcal{R}(T, \Theta)\right]$ denote the expected regret for a MAB instance Θ which is fully specified by the number of arms and the distributions for the arms. Fixing the problem instance fixes the reward distributions, and therefore the values of the means $\mu_i, i = 1, \ldots, N$. Then, *problem-dependent bounds* on regret are bounds on $\mathbb{E}\left[\mathcal{R}(T, \Theta)\right]$ for every problem instance Θ, in terms of $T, \mu_i, i = 1, \ldots, N$ and possibly other distribution parameters associated with Θ. *Problem-independent bounds* are bounds on the *worst-case* expected regret as a function of the number of arms N and time T, i.e.,

$$\max_{\Theta} \mathbb{E}\left[\mathcal{R}(T, \Theta)\right],$$

where the maximization is over MAB instances with N arms. These are the notions of regret considered in most classic works on the UCB algorithm for multi-armed bandit problem, e.g. [Auer et al. 2002; Lai and Robbins 1985], and therefore our bounds are directly comparable with those available for UCB. [Auer et al. 2002] provides

$O(\sum_{i:\mu_i<\mu^*} \frac{\log(T)}{\Delta_i})$ problem-dependent regret bound and $O(\sqrt{NT\log(T)})$ problem-independent regret bound for UCB.

A related notion of regret considered by many recent works on TS, e.g. [Bubeck and Liu 2014; Russo and Van Roy 2015, 2014; Russo et al. 2013], is *Bayesian Regret*. Bayesian regret is expected regret over a (known) prior f over the problem instances. Using the terminology above, Bayesian regret is defined as

$$\mathbb{E}_{\Theta\sim f}[\mathbb{E}\left[\mathcal{R}(T,\Theta)|\Theta\right]].$$

For priors on reward distributions with $[0,1]$ support, clearly, this is a weaker notion of regret than the worst-case problem-independent regret, in that any bound on the worst-case regret implies the same bound on Bayesian regret for any such prior f, but not vice-versa. However, we must note that these works allow more general priors and some of these even accommodate contexts and further complex information structures. Bayesian regret bounds in those more complex settings are incomparable to the worst-case regret bounds presented here.

1.2. Thompson Sampling

As mentioned before, TS is a natural algorithm for stochastic MAB. The basic idea behind TS is to assume a simple prior distribution on the underlying parameters of the reward distribution of every arm, and at every time step, play an arm according to its posterior probability of being the best arm. While TS is a specific algorithm due to Thompson, in this paper we will use TS more generally to refer to a class of algorithms that have a similar structure and include the original algorithm of Thompson as a special case. The general structure of TS involves the following elements (this description of TS follows closely that of [Chapelle and Li 2011]):

(1) a set ψ of parameters $\tilde{\mu}$;
(2) an assumed prior distribution $P(\tilde{\mu})$ on these parameters;
(3) past observations $\mathcal{D}$ consisting of (reward r) for the arms played in the past time steps;
(4) an assumed likelihood function $P(r|\tilde{\mu})$, which gives the probability of reward given a parameter $\tilde{\mu}\in\psi$;
(5) a posterior distribution $P(\tilde{\mu}|\mathcal{D})\propto P(\mathcal{D}|\tilde{\mu})P(\tilde{\mu})$, where $P(\mathcal{D}|\tilde{\mu})$ is the likelihood function.

The notation $P(\cdot)$ above denotes probability density function (or probability mass function for discrete random variables). TS maintains a posterior distribution for the underlying parameters μ_i, i.e. the expected reward of every arm i. In each round, TS plays an arm according to its posterior probability of being the best arm, that is, the posterior probability of having the highest value of μ_i. A simple way to achieve this is to produce a sample from the posterior distribution of every arm, and play the arm that produces the largest sample. Below we describe two versions of TS, using Beta priors and Bernoulli likelihood function, and using Gaussian priors and Gaussian likelihood.

We emphasize that the Beta priors and the Bernoulli likelihood model, or Gaussian priors and the Gaussian likelihood model for rewards are only used below to design the Thompson Sampling algorithm. Our analysis of these algorithms allows these models to be completely unrelated to the actual reward distribution. The assumptions on the actual reward distribution are only those mentioned in Section 1.1, namely the rewards are in the range $[0,1]$, and are generated i.i.d. upon playing an arm. In the description of TS using Beta priors and Bernoulli likelihood, for simplicity we do begin with the description of the algorithm for the Bernoulli bandit problem, i.e., when the

rewards are either 0 or 1, but as we explain later, the algorithm and its analysis extend to any distribution of rewards with $[0,1]$ support.

Thompson Sampling using Beta priors and Bernoulli likelihood. Consider the Bernoulli bandit problem, i.e., when the rewards are either 0 or 1, and the likelihood of reward 1 for arm i (the probability of success) is μ_i. Using Beta priors is convenient for Bernoulli rewards because if the prior is a Beta(α,β) distribution, then after observing a Bernoulli trial, the posterior distribution is simply Beta$(\alpha+1,\beta)$ or Beta$(\alpha,\beta+1)$, depending on whether the trial resulted in a success or failure, respectively.

TS initially assumes arm i to have prior Beta$(1,1)$ on μ_i, which is natural because Beta$(1,1)$ is the uniform distribution on $(0,1)$. Informally, this choice captures the fact that initially we have no knowledge about the μ_i. At time t, having observed $S_i(t)$ successes (reward = 1) and $F_i(t)$ failures (reward = 0) in $k_i(t)=S_i(t)+F_i(t)$ plays of arm i, the algorithm updates the distribution on μ_i to Beta$(S_i(t)+1,F_i(t)+1)$. The algorithm then generates independent samples from these posterior distributions of the μ_i's, and plays the arm with the largest sample value.

ALGORITHM 1: Thompson Sampling using Beta priors

For each arm $i=1,\ldots,N$ set $S_i=0,F_i=0$.
foreach $t=1,2,\ldots,$ **do**
  For each arm $i=1,\ldots,N$, sample $\theta_i(t)$ from the Beta(S_i+1,F_i+1) distribution.
  Play arm $i(t):=\arg\max_i \theta_i(t)$ and observe reward r_t.
  If $r_t=1$, then $S_{i(t)}:=S_{i(t)}+1$, else $F_{i(t)}:=F_{i(t)}+1$.
end

We have provided the details of TS with Beta priors for the Bernoulli bandit problem. A simple extension of this algorithm to general reward distributions with support $[0,1]$ is described in Agrawal and Goyal [2012]. In this extension, on observing a reward $r_t\in[0,1]$, we toss a coin with bias r_t, and use the $\{0,1\}$ outcome to update the beta distribution as above. It is easy to show that any expected regret bounds produced for Algorithm 1 will also hold for this extension [Agrawal and Goyal 2012].

Thompson Sampling using Gaussian priors and likelihood. As before, let $k_i(t)$ denote the number of plays of arm i until time $t-1$, and let $i(t)$ denote the arm played at time t. Let $r_i(t)$ denote the reward of arm i at time t; define $\hat{\mu}_i(t):=\frac{\sum_{\tau=1:i(\tau)=i}^{t-1} r_i(\tau)}{k_i(t)+1}$. Note that $\hat{\mu}_i(1)=0$. To derive TS with Gaussian priors, assume that the **likelihood** of reward $r_i(t)$ at time t, given parameter μ_i, is given by the pdf of Gaussian distribution $\mathcal{N}(\mu_i,1)$. Then, assuming that the **prior** for μ at time t is given by $\mathcal{N}(\hat{\mu}_i(t),\frac{1}{k_i(t)+1})$, and arm i is played at time t with reward r, it is easy to compute the **posterior** distribution $\Pr(\tilde{\mu}_i|r_i(t))\propto\Pr(r_i(t)|\tilde{\mu}_i)\Pr(\tilde{\mu}_i)$ as Gaussian distribution $\mathcal{N}(\hat{\mu}_i(t+1),\frac{1}{k_i(t+1)+1})$. In TS with Gaussian priors, for each arm i, we will generate an independent sample $\theta_i(t)$ from the distribution $\mathcal{N}(\hat{\mu}_i(t),\frac{1}{k_i(t)+1})$ at time t. The arm with maximum value of $\theta_i(t)$ will be played.

1.3. Our results

In this article, we bound the *finite time* expected regret of TS. Henceforth we will assume that the first arm is the unique optimal arm, i.e., $\mu^*=\mu_1>\arg\max_{i\neq 1}\mu_i$. Assuming that the first arm is an optimal arm is a matter of convenience for stating the results and for the analysis. The algorithms did not use this assumption. The assumption of *unique* optimal arm is also without loss of generality, since adding more

ALGORITHM 2: Thompson Sampling using Gaussian priors

For each arm $i = 1, \ldots, N$ set $k_i = 0, \hat{\mu}_i = 0$.
foreach $t = 1, 2, \ldots,$ **do**
  For each arm $i = 1, \ldots, N$, sample $\theta_i(t)$ independently from the $\mathcal{N}(\hat{\mu}_i, \frac{1}{k_i+1})$ distribution.
  Play arm $i(t) := \arg\max_i \theta_i(t)$ and observe reward r_t.
  Set $\hat{\mu}_{i(t)} := \frac{\hat{\mu}_{i(t)} k_{i(t)} + r_t}{k_{i(t)}+2}$, $k_{i(t)} := k_{i(t)} + 1$.
end

arms with $\mu_i = \mu^*$ can only decrease the expected regret; details of this argument were provided in [Agrawal and Goyal 2012].

THEOREM 1.1. *Fix $\epsilon \in (0, 1)$. For the N-armed stochastic bandit problem with the μ_i satisfying assumptions in the previous paragraph, Thompson Sampling using Beta priors has expected regret*

$$\mathbb{E}[\mathcal{R}(T)] \leq (1+\epsilon) \sum_{i=2}^{N} \frac{\ln T}{d(\mu_i, \mu_1)} \Delta_i + O(\frac{N}{\epsilon^2})$$

at time T, where $d(\mu_i, \mu_1) := \mu_i \log \frac{\mu_i}{\mu_1} + (1-\mu_i) \log \frac{(1-\mu_i)}{(1-\mu_1)}$. The big-Oh notation assumes $\mu_i, \Delta_i, i = 1, \ldots, N$ to be constants.

THEOREM 1.2. *For the N-armed stochastic bandit problem, Thompson Sampling using Beta priors has expected regret*

$$\mathbb{E}[\mathcal{R}(T)] \leq O(\sqrt{NT \ln T})$$

at time T. The big-Oh notation hides only the absolute constants.

THEOREM 1.3. *For the N-armed stochastic bandit problem, Thompson Sampling using Gaussian priors has expected regret*

$$\mathbb{E}[\mathcal{R}(T)] \leq O(\sqrt{NT \ln N})$$

at time $T \geq N$. The big-Oh notation hides only absolute constants.

THEOREM 1.4. *There exists an instance of the N-armed stochastic bandit problem, for which Thompson Sampling using Gaussian priors has expected regret*

$$\mathbb{E}[\mathcal{R}(T)] \geq \Omega(\sqrt{NT \ln N})$$

at time $T \geq N$. Here Ω hides only absolute constants.

1.4. Related work

Let us contrast our bounds with the previous work. Let us first consider the problem-dependent regret bounds, i.e., regret bounds that depend on the problem parameters N and $\mu_i, \Delta_i, i = 1, \ldots, N$. [Lai and Robbins 1985] essentially proved an asymptotic lower bound of $\left[\sum_{i=2}^{N} \frac{\Delta_i}{d(\mu_i, \mu_1)} + o(1)\right] \ln T$ for any algorithm for this problem. They also gave algorithms asymptotically achieving this guarantee. [Auer et al. 2002] gave the UCB1 algorithm, which achieves a finite time regret bound of $\left[8 \sum_{i=2}^{N} \frac{1}{\Delta_i}\right] \ln T + (1 + \pi^2/3) \left(\sum_{i=2}^{N} \Delta_i\right)$. More recently, Kaufmann et al. [2012a] gave Bayes-UCB algorithm, and [Garivier and Cappé 2011] and Maillard et al. [2011] gave UCB-like algorithms,

which achieve the lower bound of Lai and Robbins [1985]. Our regret bound in Theorem 1.1 achieves the lower bounds of Lai and Robbins [1985], and matches the upper bounds provided by [Kaufmann et al. 2012b] for TS.

Theorem 1.2 and 1.3 show that TS with Beta and Gaussian distributions achieve a problem independent regret bound of $O(\sqrt{NT \ln T})$ and $O(\sqrt{NT \ln N})$ respectively. This is the first analyis for TS that matches the $\Omega(\sqrt{NT})$ problem-independent lower bound (see Section 3.3 of [Bubeck and Cesa-Bianchi 2012]) for the stochastic MAB within logarithmic factors. The problem-dependent bounds can be used to derive problem-independent bounds. However, the previous work on TS implied only suboptimal problem-independent bounds: The results of Agrawal and Goyal [2012] implied a problem-independent bound of $\tilde{O}(N^{1/5}T^{4/5})$. In [Kaufmann et al. 2012b], the additive problem-dependent term was not explicitly calculated, which makes it difficult to derive the implied problem-independent bound, but a preliminary examination suggests that it would involve an even higher power of T.

To compare with other existing algorithms for this problem, note that the best known problem-independent bound for the expected regret of UCB1 is $O(\sqrt{NT \ln T})$ (see [Bubeck and Cesa-Bianchi 2012]). Our regret bound of $O(\sqrt{NT \ln N})$ for TS with Gaussian priors is an improvement over the bound for UCB1. More recently, [Audibert and Bubeck 2009] gave an algorithm MOSS, inspired by UCB1, with regret $O(\sqrt{NT})$ that matches the $\Omega(\sqrt{NT})$ problem-independent lower bound for the multi-armed bandit problem. However, their algorithm needs to know the time horizon T. It is unclear whether an $O(\sqrt{NT})$ regret can be achieved by an algorithm that does not know the time horizon. Interestingly, Theorem 1.4 shows that this is unachievable for TS with Gaussian priors, as there is a lower bound of $\Omega(\sqrt{NT \ln N})$ on its expected regret. This is the first lower bound for TS that differs from the general lower bound for the problem.

Much followup work has been conducted in understanding theoretical properties of Thompson Sampling since our work first appeared in public domain as [Agrawal and Goyal 2013a] (conference version). We mention some of this work: [Korda et al. 2013] study TS for the special case of exponential family of distributions, [Kocák et al. 2014] for spectral bandits, [Agrawal and Goyal 2013b; Russo and Van Roy 2014; Li 2013] for contextual bandits, and [Russo et al. 2013] for reinforcement learning.

2. PROOFS OF UPPER BOUNDS

In this section, we prove Theorems 1.1, 1.2 and 1.3. The proofs of the three theorems follow similar steps, and diverge only towards the end of the analysis.

Proof Outline:. Our proof uses a martingale based analysis. Essentially, we prove that conditioned on any history of execution in the preceding steps, the probability of playing any suboptimal arm i at the current step can be bounded by a linear function of the probability of playing the optimal arm at the current step. This is proven in Lemma 2.8, which forms the core of our analysis. Further, we show that the coefficient in this linear function decreases exponentially fast with the number of plays of the optimal arm (Lemma 2.9). This allows us to bound the number of plays of every suboptimal arm, which in turn bounds the regret. The differences between the analyses for obtaining the logarithmic problem-dependent bound of Theorem 1.1, and the problem-independent bound of Theorem 1.2 and Theorem 1.3 are technical, and occur only towards the end of the proof.

We recall some of the definitions introduced earlier and introduce some new ones.

Definition 2.1 ($F^B_{n,p}, f^B_{n,p}, F^{beta}_{\alpha,\beta}$). $F^B_{n,p}(\cdot)$ denotes the cdf, $f^B_{n,p}(\cdot)$ denotes the probability mass function of the binomial distribution with parameters n, p, and $F^{beta}_{\alpha,\beta}(\cdot)$ denotes the cdf of the beta distribution with parameters α, β.

Definition 2.2 (**Quantities** $k_i(t), i(t), S_i(t), \hat{\mu}_i(t)$). $i(t)$ denotes the arm played at time t, $k_i(t)$ denotes the number of plays of arm i until (and including) time $t-1$, $S_i(t)$ denotes the number of successes among the $k_i(t)$ plays of arm i until time $t-1$ for the Bernoulli bandit case (in other words, $S_i(t)$ is the number of times arm i gave reward 1). Finally, the empirical mean $\hat{\mu}_i(t)$ for arm i at time t is defined by $\hat{\mu}_i(t) := \frac{\sum_{\tau=1:i(\tau)=i}^{t-1} r_i(\tau)}{k_i(t)+1}$, where $r_i(\tau)$ is the reward for arm i at time τ (note that $\hat{\mu}_i(t) = 0$ when $k_i(t) = 0$). For Bernoulli bandits, $\hat{\mu}_i(t) = \frac{S_i(t)}{k_i(t)+1}$.

Definition 2.3 (**Quantities** $\theta_i(t)$). $\theta_i(t)$ denotes a sample generated independently for each arm i, from the posterior distribution at time t. For Algorithm 1, this is generated from posterior distribution $\text{Beta}(S_i(t)+1, k_i(t)-S_i(t)+1)$. For Algorithm 2, this is generated from posterior distribution $\mathcal{N}(\hat{\mu}_i(t), \frac{1}{k_i(t)+1})$.

Definition 2.4 (**Quantities** x_i, y_i). For each arm $i \neq 1$, we will choose two thresholds x_i and y_i such that $\mu_i < x_i < y_i < \mu_1$. The specific choice of these thresholds will depend on whether we are proving problem-dependent bound or problem-independent bound, and will be described at the appropriate points in the proof.

Definition 2.5 (**Events** $E^\mu_i(t)$ **and** $E^\theta_i(t)$). For $i \neq 1$, $E^\mu_i(t)$ is the event $\hat{\mu}_i(t) \leq x_i$, and $E^\theta_i(t)$ is the event $\theta_i(t) \leq y_i$.

Intuitively, $E^\mu_i(t)$, $E^\theta_i(t)$ are the events that the estimate $\hat{\mu}_i(t)$ and the sample value $\theta_i(t)$, respectively, are not too far above the mean μ_i. As we show later, these events will hold with high probability.

Definition 2.6 (**History** $\mathcal{F}_t$). For $t = 1, 2, \ldots$, define $\mathcal{F}_t$ as the history of plays until time t, i.e. the sequence

$$\mathcal{F}_t = \{i(\tau), r_{i(\tau)}(\tau), \tau = 1, \ldots, t\},$$

where $i(\tau)$ denotes the arm played at time τ, and $r_{i(\tau)}(\tau)$ denotes the reward observed for arm $i(\tau)$ at time τ. Define $\mathcal{F}_0 = \{\}$.

By definition, $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \cdots \subseteq \mathcal{F}_{T-1}$. Also by definition, for every arm i, the quantities $S_i(t)$ (this is only defined for the case of Bernoulli rewards), $k_i(t)$, $\hat{\mu}_i(t)$, the distribution of $\theta_i(t)$, and whether or not $E^\mu_i(t)$ is true, are determined by the history of plays until time $t-1$, i.e. by $\mathcal{F}_{t-1}$.

Definition 2.7. Define, $p_{i,t}$ as the probability

$$p_{i,t} := \Pr(\theta_1(t) > y_i | \mathcal{F}_{t-1}).$$

Note that $p_{i,t}$ is a random variable determined by $\mathcal{F}_{t-1}$; we do not explicitly indicate this dependence by using notation such as $p_{i,t}(\mathcal{F}_{t-1})$ for brevity.

We prove the following lemma for Thompson Sampling, independent of the type of priors (e.g., Beta or Gaussian) used.

LEMMA 2.8. *For all* $t, i \neq 1$ *and all instantiations* F_{t-1} *of* $\mathcal{F}_{t-1}$ *we have*

$$\Pr\left(i(t) = i, E^\mu_i(t), E^\theta_i(t) \,\middle|\, F_{t-1}\right) \leq \frac{(1-p_{i,t})}{p_{i,t}} \Pr\left(i(t) = 1, E^\mu_i(t), E^\theta_i(t) \,\middle|\, F_{t-1}\right).$$

PROOF. Recall that whether or not $E_i^\mu(t)$ is true is determined by the instantiation F_{t-1} of $\mathcal{F}_{t-1}$. Assume that history F_{t-1} is such that $E_i^\mu(t)$ is true (otherwise the probability on the left hand side is 0 and the inequality is trivially true). It then suffices to prove that for all such F_{t-1} we have

$$\Pr\left(i(t)=i \,\middle|\, E_i^\theta(t), \mathcal{F}_{t-1}=F_{t-1}\right) \;\le\; \frac{(1-p_{i,t})}{p_{i,t}} \Pr\left(i(t)=1 \,\middle|\, E_i^\theta(t), \mathcal{F}_{t-1}=F_{t-1}\right). \quad (1)$$

We will use the observation that since $E_i^\theta(t)$ is the event that $\theta_i(t) \le y_i$, therefore, given $E_i^\theta(t)$, we have $i(t)=i$ only if $\theta_j(t) \le y_i, \forall j$. Therefore, for any $i \neq 1$,

$$\begin{aligned}
\Pr\left(i(t)=i \,\middle|\, E_i^\theta(t), \mathcal{F}_{t-1}=F_{t-1}\right) &\le \Pr\left(\theta_j(t)\le y_i, \forall j \,\middle|\, E_i^\theta(t), \mathcal{F}_{t-1}=F_{t-1}\right) \\
&= \Pr\left(\theta_1(t)\le y_i \,|\, \mathcal{F}_{t-1}=F_{t-1}\right) \\
&\qquad \cdot \Pr\left(\theta_j(t)\le y_i, \forall j\neq 1 \,\middle|\, E_i^\theta(t), \mathcal{F}_{t-1}=F_{t-1}\right) \\
&= (1-p_{i,t}) \cdot \Pr\left(\theta_j(t)\le y_i, \forall j\neq 1 \,\middle|\, E_i^\theta(t), \mathcal{F}_{t-1}=F_{t-1}\right).
\end{aligned}$$

The first equality holds because given $\mathcal{F}_{t-1}=F_{t-1}$ (and hence $S_j(t), k_j(t), \hat{\mu}_j(t)$ and the distributions of $\theta_j(t)$ for all j), $\theta_1(t)$ is independent of all the other $\theta_j(t)$ and events $E_j^\theta(t), j\neq 1$. Similarly,

$$\begin{aligned}
\Pr\left(i(t)=1 \,\middle|\, E_i^\theta(t), \mathcal{F}_{t-1}=F_{t-1}\right) &\ge \Pr\left(\theta_1(t) > y_i \ge \theta_j(t), \forall j\neq 1 \,\middle|\, E_i^\theta(t), \mathcal{F}_{t-1}=F_{t-1}\right) \\
&= \Pr\left(\theta_1(t) > y_i \,|\, \mathcal{F}_{t-1}=F_{t-1}\right) \\
&\qquad \cdot \Pr\left(\theta_j(t)\le y_i, \forall j\neq 1 \,\middle|\, E_i^\theta(t), \mathcal{F}_{t-1}=F_{t-1}\right) \\
&= p_{i,t} \cdot \Pr\left(\theta_j(t)\le y_i, \forall j\neq 1 \,\middle|\, E_i^\theta(t), \mathcal{F}_{t-1}=F_{t-1}\right).
\end{aligned}$$

Combining the above two inequalities, we get (1). □

Now we are ready to prove the upper bounds on regret in Theorems 1.1, 1.2, and 1.3.

2.1. Proof of Theorem 1.1

We can decompose the expected number of plays of a suboptimal arm $i \neq 1$ as follows.

$$\begin{aligned}
\mathbb{E}[k_i(T)] &= \sum_{t=1}^{T} \Pr(i(t)=i) \\
&= \sum_{t=1}^{T} \Pr(i(t)=i, E_i^\mu(t), E_i^\theta(t)) + \sum_{t=1}^{T} \Pr\left(i(t)=i, E_i^\mu(t), \overline{E_i^\theta(t)}\right) + \sum_{t=1}^{T} \Pr\left(i(t)=i, \overline{E_i^\mu(t)}\right).
\end{aligned} \quad (2)$$

Next, we bound each of the above terms. For the first term above, applying Lemma 2.8 and some algebraic manipulations using properties of conditional expectations, we

have

$$\begin{aligned}\sum_{t=1}^{T}\Pr(i(t)=i,E_i^{\mu}(t),E_i^{\theta}(t)) &= \sum_{t=1}^{T}\mathbb{E}\left[\Pr\left(i(t)=i,E_i^{\mu}(t),E_i^{\theta}(t)\,\middle|\,\mathcal{F}_{t-1}\right)\right]\\ &\le \sum_{t=1}^{T}\mathbb{E}\left[\frac{(1-p_{i,t})}{p_{i,t}}\Pr\left(i(t)=1,E_i^{\theta}(t),E_i^{\mu}(t)\,\middle|\,\mathcal{F}_{t-1}\right)\right]\\ &= \sum_{t=1}^{T}\mathbb{E}\left[\mathbb{E}\left[\frac{(1-p_{i,t})}{p_{i,t}}I(i(t)=1,E_i^{\theta}(t),E_i^{\mu}(t))\,\middle|\,\mathcal{F}_{t-1}\right]\right]\\ &= \sum_{t=1}^{T}\mathbb{E}\left[\frac{(1-p_{i,t})}{p_{i,t}}I(i(t)=1,E_i^{\theta}(t),E_i^{\mu}(t))\right]. \qquad (3)\end{aligned}$$

The second equality above uses that $p_{i,t}$ is fixed given $\mathcal{F}_{t-1}$. Now, let τ_k denote the time step at which arm 1 is played for the k^{th} time for $k \ge 1$, and let $\tau_0 = 0$. (Note that for any i, for $k > k_i(T)$, $\tau_k > T$. Also, $\tau_T \ge T$.) Observe that $p_{i,t} = \Pr(\theta_1(t) > y_i|\mathcal{F}_{t-1})$ changes only when the distribution of $\theta_1(t)$ changes, that is, only on the time step after each play of the first arm. Thus, $p_{i,t}$ is the same at all time steps $t \in \{\tau_k+1,\ldots,\tau_{k+1}\}$, for every k. Using this observation, we can decompose the above term in the following way.

$$\begin{aligned}\sum_{t=1}^{T}\mathbb{E}\left[\frac{(1-p_{i,t})}{p_{i,t}}I(i(t)=1,E_i^{\theta}(t),E_i^{\mu}(t))\right] &= \sum_{k=0}^{T-1}\mathbb{E}\left[\frac{(1-p_{i,\tau_k+1})}{p_{i,\tau_k+1}}\sum_{t=\tau_k+1}^{\tau_{k+1}}I(i(t)=1,E_i^{\theta}(t),E_i^{\mu}(t))\right]\\ &\le \sum_{k=0}^{T-1}\mathbb{E}\left[\frac{(1-p_{i,\tau_k+1})}{p_{i,\tau_k+1}}\right]. \qquad (4)\end{aligned}$$

We prove the following lemma to bound the sum of $\frac{1}{p_{i,\tau_k+1}}$.

LEMMA 2.9. *Let τ_k denote the time step at which k^{th} trial of the first arm happens, then for $i \neq 1$ we have* [3]

$$\mathbb{E}\left[\frac{1}{p_{i,\tau_k+1}}-1\right] \le \begin{cases}\frac{3}{\Delta_i'} & \text{for } k < \frac{8}{\Delta_i'},\\ \Theta(e^{-\Delta_i'^2 k/2}+\frac{1}{(k+1)\Delta_i'^2}e^{-D_i k}+\frac{1}{e^{\Delta_i'^2 k/4}-1}) & \text{for } k \ge \frac{8}{\Delta_i'},\end{cases}$$

where $\Delta_i' = \mu_1 - y_i$ and $D_i = y_i \ln\frac{y_i}{\mu_1} + (1-y_i)\ln\frac{1-y_i}{1-\mu_1}$.

PROOF. The proof of this inequality uses careful numerical estimates and appears in Appendix B. □

Substituting the bound from Lemma 2.9 into (4), we obtain the following bound on the first term on the right hand side of (2).

LEMMA 2.10. *For $i \neq 1$,*

$$\sum_{t=1}^{T}\Pr(i(t)=i,E_i^{\mu}(t),E_i^{\theta}(t)) \le \frac{24}{\Delta_i'^2}+\sum_{j\ge 8/\Delta_i'}\Theta\left(e^{-\Delta_i'^2 j/2}+\frac{1}{(j+1)\Delta_i'^2}e^{-D_i j}+\frac{1}{e^{\Delta_i'^2 j/4}-1}\right).$$

[3]Here for two functions $f(x), g(x)$ taking positive values and with the same domain we say that $f(x) = \Theta(g(x))$ if there exist absolute constants $b, c > 0$ such that for all x in the domains of f and g we have $bg(x) \le f(x) \le cg(x)$.

To bound the remaining two terms in (2), we use the fact that as the number of plays of arm i increases, the probability of violating the events $E_i^\mu(t)$ and $E_i^\theta(t)$ decreases exponentially. More precisely, we prove the following lemmas.

LEMMA 2.11. *For $i \neq 1$,*

$$\sum_{t=1}^{T} \Pr\left(i(t) = i, \overline{E_i^\mu(t)}\right) \leq \frac{1}{d(x_i,\mu_i)} + 1.$$

PROOF. Let τ_k denote the time at which the k^{th} trial of arm i happens. Set $\tau_0 = 0$. (Note that $\tau_k > T$ for $k > k_i(T)$. Note also that $T \leq \tau_T$.) Recall that the event $\overline{E_i^\mu(t)}$ was defined as $\hat{\mu}_i(t) > x_i$. We have

$$\sum_{t=1}^{T} \Pr(i(t) = i, \overline{E_i^\mu(t)}) \leq \sum_{k=0}^{T-1} \Pr(\overline{E_i^\mu(\tau_{k+1})}).$$

Each summand on the right hand side in the inequality above is a fixed number even though the random variables τ_{k+1} appear in it. This is because the distribution of $\hat{\mu}_i(\tau_{k+1})$ only depends on k and not on τ_{k+1}. At time τ_{k+1} for $k \geq 1$, $\hat{\mu}_i(\tau_{k+1}) = \frac{S_i(\tau_{k+1})}{k+1} \leq \frac{S_i(\tau_{k+1})}{k}$, where the latter is simply the average of the outcomes of k i.i.d. plays of arm i, each of which is a Bernoulli trial with mean μ_i. Using the Chernoff-Hoeffding bounds (Fact 1), we obtain $\Pr(\hat{\mu}_i(\tau_{k+1}) > x_i) \leq \Pr(\frac{S_i(\tau_{k+1})}{k} > x_i) \leq e^{-kd(x_i,\mu_i)}$. Substituting, we get,

$$\begin{aligned}\sum_{k=0}^{T-1} \Pr(\overline{E_i^\mu(\tau_{k+1})}) = \sum_{k=0}^{T-1} \Pr(\hat{\mu}_i(\tau_{k+1}) > x_i) &\leq 1 + \sum_{k=1}^{T-1} \exp(-kd(x_i,\mu_i)) \\ &\leq 1 + \frac{1}{d(x_i,\mu_i)}.\end{aligned}$$

The last inequality uses the fact that $d(x_i, \mu_i) > 0$. □

LEMMA 2.12. *For $i \neq 1$,*

$$\sum_{t=1}^{T} \Pr\left(i(t) = i, \overline{E_i^\theta(t)}, E_i^\mu(t)\right) \leq L_i(T) + 1,$$

where $L_i(T) = \frac{\ln T}{d(x_i,y_i)}$.

PROOF. We decompose the probability term into two parts, based on whether or not $k_i(T)$ is large ($k_i(t) > L_i(t)$).

$$\begin{aligned}\sum_{t=1}^{T} \Pr\left(i(t) = i, \overline{E_i^\theta(t)}, E_i^\mu(t)\right) &= \sum_{t=1}^{T} \Pr\left(i(t) = i, k_i(t) \leq L_i(T), \overline{E_i^\theta(t)}, E_i^\mu(t)\right) \\ &\quad + \sum_{t=1}^{T} \Pr\left(i(t) = i, k_i(t) > L_i(T), \overline{E_i^\theta(t)}, E_i^\mu(t)\right). \quad (5)\end{aligned}$$

The first term in the above decomposition is bounded by $\mathbb{E}[\sum_{t=1}^{T} I(i(t) = i, k_i(t) \leq L_i(T))]$, which is bounded trivially by $L_i(T)$. What remains is to bound the second term by 1. To bound the second term, we demonstrate that if $k_i(t)$ is large and the event $E_i^\mu(t)$ is satisfied, then the probability that the event $E_i^\theta(t)$ is violated, is small. Recall that $E_i^\theta(t)$ is defined as the event that $\theta_i(t) \leq y_i$. And, $E_i^\mu(t)$ is the event that $\hat{\mu}_i(t) \leq x_i$.

Then,

$$\begin{aligned}
&\sum_{t=1}^{T} \Pr\left(i(t)=i, k_i(t) > L_i(T), \overline{E_i^\theta(t)}, E_i^\mu(t)\right) \\
&= \sum_{t=1}^{T} \mathbb{E}\left[I\left(i(t)=i, k_i(t) > L_i(T), \overline{E_i^\theta(t)}, E_i^\mu(t)\right)\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{E}\left[I\left(i(t)=i, k_i(t) > L_i(T), \overline{E_i^\theta(t)}, E_i^\mu(t)\right) \,\middle|\, \mathcal{F}_{t-1}\right]\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T} I\left(k_i(t) > L_i(T), E_i^\mu(t)\right) \Pr\left(i(t)=i, \overline{E_i^\theta(t)} \,\middle|\, \mathcal{F}_{t-1}\right)\right] \\
&\leq \mathbb{E}\left[\sum_{t=1}^{T} I\left(k_i(t) > L_i(T), \hat{\mu}_i(t) \leq x_i\right) \Pr\left(\theta_i(t) > y_i \,|\, \mathcal{F}_{t-1}\right)\right]. \qquad (6)
\end{aligned}$$

The third equality above uses the fact that $k_i(t)$ and $E_i^\mu(t)$ are determined by the history $\mathcal{F}_{t-1}$.

Now, by definition, $S_i(t) = \hat{\mu}_i(t)(k_i(t)+1)$, and therefore, $\theta_i(t)$ is a Beta$(\hat{\mu}_i(t)(k_i(t)+1)+1, (1-\hat{\mu}_i(t))(k_i(t)+1))$ distributed random variable. A Beta(α,β) random variable is stochastically dominated by Beta(α',β') if $\alpha' \geq \alpha, \beta' \leq \beta$. Therefore, if $\hat{\mu}_i(t) \leq x_i$, the distribution of $\theta_i(t)$ is stochastically dominated by Beta$(x_i(k_i(t)+1)+1, (1-x_i)(k_i(t)+1))$. Therefore, given an instantiation F_{t-1} of $\mathcal{F}_{t-1}$ such that $\hat{\mu}_i(t) \leq x_i$ and $k_i(t) > L_i(T)$, we have

$$\Pr\left(\theta_i(t) > y_i \,|\, \mathcal{F}_{t-1} = F_{t-1}\right) \leq 1 - F^{beta}_{x_i(k_i(t)+1)+1,(1-x_i)(k_i(t)+1))}(y_i).$$

Now, using Fact 3 along with the Chernoff-Hoeffding bounds (Fact 1), we obtain that for any fixed $k_i(t) > L_i(T)$,

$$\begin{aligned}
1 - F^{beta}_{x_i(k_i(t)+1)+1,(1-x_i)(k_i(t)+1))}(y_i) &= F^{B}_{k_i(t)+1,y_i}(x_i(k_i(t)+1)) \\
&\leq e^{-(k_i(t)+1)d(x_i,y_i)} \\
&\leq e^{-L_i(T)d(x_i,y_i)},
\end{aligned}$$

which is smaller than $\frac{1}{T}$ because $L_i(T) = \frac{\ln T}{d(x_i,y_i)}$. Substituting, we get that for any instantiation F_{t-1} of $\mathcal{F}_{t-1}$ such that $\hat{\mu}_i(t) \leq x_i$ and $k_i(t) > L_i(T)$,

$$\Pr\left(\theta_i(t) > y_i \,|\, \mathcal{F}_{t-1} = F_{t-1}\right) \leq \frac{1}{T}.$$

For other instantiations of $\mathcal{F}_{t-1}$, the indicator term $I\left(k_i(t) > L_i(T), \hat{\mu}_i(t) \leq x_i\right)$ in (6) will be 0. Summing over t, this bounds the second term in (5) by 1 to complete the proof of the lemma. □

Putting it all together: Substituting the results from Lemma 2.10, Lemma 2.11 and Lemma 2.12 into (2), we obtain

$$\mathbb{E}[k_i(T)] \leq \frac{24}{\Delta_i'^2} + \sum_{j \geq 8/\Delta_i'} \Theta\left(e^{-\Delta_i'^2 j/2} + \frac{1}{(j+1)\Delta_i'^2} e^{-D_i j} + \frac{1}{e^{\Delta_i'^2 j/4} - 1}\right) + L_i(T) + 1 + \frac{1}{d(x_i,\mu_i)} + 1. \qquad (7)$$

To obtain the problem-dependent bound of Theorem 1.1, for $0 < \epsilon \leq 1$, we set $x_i \in (\mu_i, \mu_1)$ such that $d(x_i, \mu_1) = d(\mu_i, \mu_1)/(1+\epsilon)$, and set $y_i \in (x_i, \mu_1)$ such that $d(x_i, y_i) = d(x_i, \mu_1)/(1+\epsilon) = d(\mu_i, \mu_1)/(1+\epsilon)^2$ ([4]). This gives

$$L_i(T) = \frac{\ln T}{d(x_i, y_i)} = (1+\epsilon)^2 \frac{\ln T}{d(\mu_i, \mu_1)}.$$

Also, by some simple algebraic manipulations of the equality $d(x_i, \mu_1) = d(\mu_i, \mu_1)/(1+\epsilon)$, we can obtain

$$x_i - \mu_i \geq \frac{\epsilon}{(1+\epsilon)} \cdot \frac{d(\mu_i, \mu_1)}{\ln\left(\frac{\mu_1(1-\mu_i)}{\mu_i(1-\mu_1)}\right)},$$

giving $\frac{1}{d(x_i,\mu_i)} \leq \frac{1}{2(x_i-\mu_i)^2} = O(\frac{1}{\epsilon^2})$. Here big-Oh is hiding functions of the μ_i's and Δ_i's. Substituting in (7), we get

$$\begin{aligned}
\mathbb{E}[k_i(T)] &\leq \frac{24}{\Delta_i'^2} + \sum_{j \geq 8/\Delta_i'} \Theta\left(e^{-\Delta_i'^2 j/2} + \frac{1}{(j+1)\Delta_i'^2} e^{-D_i j} + \frac{1}{e^{\Delta_i'^2 j/4} - 1}\right) + L_i(T) + 1 + \frac{1}{d(x_i, \mu_i)} + 1 \\
&\leq \frac{24}{\Delta_i'^2} + \Theta\left(\frac{1}{\Delta_i'^2} + \frac{1}{\Delta_i'^2 D_i} + \frac{1}{\Delta_i'^4}\right) + (1+\epsilon)^2 \frac{\ln T}{d(\mu_i, \mu_1)} + O\left(\frac{1}{\epsilon^2}\right) \\
&= O(1) + (1+\epsilon)^2 \frac{\ln T}{d(\mu_i, \mu_1)} + O\left(\frac{1}{\epsilon^2}\right).
\end{aligned}$$

The big-Oh above hides dependence on the μ_i's and Δ_i's. This gives expected regret bound

$$\begin{aligned}
\mathbb{E}[\mathcal{R}(T)] &= \sum_i \Delta_i \mathbb{E}[k_i(T)] \\
&\leq \sum_i (1+\epsilon)^2 \frac{\ln T}{d(\mu_i, \mu_1)} \Delta_i + O(\frac{N}{\epsilon^2}) \\
&\leq \sum_i (1+\epsilon') \frac{\ln T}{d(\mu_i, \mu_1)} \Delta_i + O(\frac{N}{\epsilon'^2}),
\end{aligned}$$

where $\epsilon' = 3\epsilon$, and the big-Oh above hides μ_is and Δ_is in addition to the absolute constants. This completes the proof of Theorem 1.1. □

2.2. Proof of Theorem 1.2

The proof of $O(\sqrt{NT \ln T})$ problem-independent bound of Theorem 1.2 is basically the same as the proof of Theorem 1.1, except for the choice of the x_i's and y_i's. Here, we pick $x_i = \mu_i + \frac{\Delta_i}{3}, y_i = \mu_1 - \frac{\Delta_i}{3}$, so that $\Delta_i'^2 = (\mu_1 - y_i)^2 = \frac{\Delta_i^2}{9}$, and using Pinsker's inequality, $d(x_i, \mu_i) \geq 2(x_i - \mu_i)^2 = \frac{2\Delta_i^2}{9}$, $d(x_i, y_i) \geq 2(y_i - x_i)^2 \geq \frac{2\Delta_i^2}{9}$. Then, $L_i(T) = \frac{\ln T}{d(x_i, y_i)} \leq \frac{9 \ln T}{2\Delta_i^2}$,

[4]This way of choosing thresholds, in order to obtain bounds in terms of the KL-divergences $d(\mu_i, \mu_1)$ rather than the Δ_i's, is inspired by [Garivier and Cappé 2011; Maillard et al. 2011; Kaufmann et al. 2012a].

and $\frac{1}{d(x_i,\mu_i)} \leq \frac{9}{2\Delta_i^2}$. Then, substituting these bounds in (7), we get

$$\begin{aligned}
\mathbb{E}[k_i(T)] &\leq \frac{24}{\Delta_i'^2} + \sum_{j\geq 8/\Delta_i'}^{T-1} \Theta\left(e^{-\Delta_i'^2 j/2} + \frac{1}{(j+1)\Delta_i'^2} e^{-D_i j} + \frac{1}{e^{\Delta_i'^2 j/4}-1}\right) + L_i(T) + 1 + \frac{1}{d(x_i,\mu_i)} + 1 \\
&\leq \sum_{j\geq 8/\Delta_i'}^{T-1} \Theta\left(e^{-\Delta_i'^2 j/2} + \frac{1}{(j+1)\Delta_i'^2} + \frac{1}{j\Delta_i'^2}\right) + O\left(\frac{\ln T}{\Delta_i^2}\right) \\
&= \Theta\left(\frac{1}{\Delta_i'^2} + \frac{\ln T}{\Delta_i'^2}\right) + O\left(\frac{\ln T}{\Delta_i^2}\right) \\
&= O\left(\frac{\ln T}{\Delta_i^2}\right).
\end{aligned}$$

Therefore, for every arm i with $\Delta_i \geq \sqrt{\frac{N\ln T}{T}}$, expected regret is bounded by $\Delta_i \mathbb{E}[k_i(T)] = O(\sqrt{\frac{T\ln T}{N}})$. For arms with $\Delta_i \leq \sqrt{\frac{N\ln T}{T}}$, total expected regret is bounded by $\sqrt{NT\ln T}$. This gives a total regret bound of $O(\sqrt{NT\ln T})$, completing the proof of Theorem 1.2. □

2.3. Proof of Theorem 1.3

The regret analysis of TS with Gaussian priors follows essentially the same steps as in the analysis of the version with Beta priors. Here, we choose $x_i = \mu_i + \frac{\Delta_i}{3}$, $y_i = \mu_1 - \frac{\Delta_i}{3}$, $L_i(T) = \frac{32\ln(T\Delta_i^2+e^{32})}{(y_i-x_i)^2} = \frac{288\ln(T\Delta_i^2+e^{32})}{\Delta_i^2}$. We prove lemmas similar to Lemma 2.10-2.12 to bounds the three terms in (2).

To obtain bounds on the first term, observe that the derivation of Lemma 2.8 is independent of the type of priors used, therefore the derivation of (4) holds as is for Gaussian priors. We prove the following lemma corresponding to Lemma 2.9.

LEMMA 2.13. *Let τ_j denote the time of the j^{th} play of the first arm. Then*

$$\mathbb{E}\left[\frac{1}{p_{i,\tau_j+1}} - 1\right] \leq \begin{cases} e^{64}+5 & \forall j, \\ \frac{5}{T\Delta_i^2}, & j > L_i(T), \end{cases}$$

where $L_i(T) = \frac{288\ln(T\Delta_i^2+e^{32})}{\Delta_i^2}$.

PROOF. From Definitions 2.3 and 2.7, recall that $p_{i,t}$ denotes the probability of $\theta_i(t)$ exceeding y_i, given $\mathcal{F}_{t-1}$. And for the algorithm with Gaussian priors $\theta_i(t)$ has distribution $\mathcal{N}(\hat{\mu}_i(t), \frac{1}{k_i(t)+1})$.

Given $\mathcal{F}_{\tau_j}$, let Θ_j denote a $\mathcal{N}(\hat{\mu}_1(\tau_j+1), \frac{1}{j+1})$ distributed Gaussian random variable. Let G_j be the geometric random variable denoting the number of consecutive independent trials *until and including* the trial where a sample of Θ_j becomes greater than y_i. Then observe that $p_{i,\tau_j+1} = \Pr(\Theta_j > y_i|\mathcal{F}_{\tau_j})$ and

$$\mathbb{E}\left[\frac{1}{p_{i,\tau_j+1}}\right] = \mathbb{E}[\mathbb{E}[G_j|\mathcal{F}_{\tau_j}] = \mathbb{E}[G_j]$$

First, we will bound the expected value of G_j by a constant for all j.

Consider any integer $r \geq 1$. Let $z = \sqrt{\ln r}$ and let random variable MAX_r denote the maximum of r independent samples of Θ_j. We abbreviate $\hat{\mu}_1(\tau_j+1)$ to $\hat{\mu}_1$ in the

following. Then, for any integer $r \geq 1$,

$$\begin{aligned}
\Pr(G_j \leq r) &\geq \Pr(\mathbf{MAX}_r > y_i) \\
&\geq \Pr(\mathbf{MAX}_r > \hat{\mu}_1 + \frac{z}{\sqrt{j+1}} \geq y_i) \\
&= \mathbb{E}\left[\mathbb{E}\left[I\left(\mathbf{MAX}_r > \hat{\mu}_1 + \frac{z}{\sqrt{j+1}} \geq y_i\right) \middle| \mathcal{F}_{\tau_j}\right]\right] \\
&= \mathbb{E}\left[I\left(\hat{\mu}_1 + \frac{z}{\sqrt{j+1}} \geq y_i\right) \Pr\left(\mathbf{MAX}_r > \hat{\mu}_1 + \frac{z}{\sqrt{j+1}} \middle| \mathcal{F}_{\tau_j}\right)\right].
\end{aligned} \tag{8}$$

The following anti-concentration bound can be derived for the Gaussian r.v. Z with mean μ and std deviation σ, using Formula 7.1.13 from [Abramowitz and Stegun 1964].

$$\Pr(Z > \mu + x\sigma) \geq \frac{1}{\sqrt{2\pi}} \frac{x}{x^2+1} e^{-x^2/2}.$$

For any instantiation F_{τ_j} of $\mathcal{F}_{\tau_j}$, since Θ_j is Gaussian $\mathcal{N}(\hat{\mu}_1, \frac{1}{j+1})$ distributed r.v., this gives

$$\begin{aligned}
\Pr\left(\mathbf{MAX}_r > \hat{\mu}_1 + \frac{z}{\sqrt{j+1}} \middle| \mathcal{F}_{\tau_j} = F_{\tau_j}\right) &\geq 1 - \left(1 - \frac{1}{\sqrt{2\pi}} \frac{z}{(z^2+1)} e^{-z^2/2}\right)^r \\
&= 1 - \left(1 - \frac{1}{\sqrt{2\pi}} \frac{\sqrt{\ln r}}{(\ln r + 1)} \frac{1}{\sqrt{r}}\right)^r \\
&\geq 1 - e^{-\frac{r}{\sqrt{4\pi r \ln r}}}.
\end{aligned}$$

Observe that for large r (in particular for any $r \geq e^{11}$), $e^{-\sqrt{\frac{r}{2\pi \ln r}}} \leq \frac{1}{r^2}$. Therefore, for any $r \geq e^{11}$,

$$\Pr\left(\mathbf{MAX}_r > \hat{\mu}_1 + \frac{z}{\sqrt{j+1}} \middle| \mathcal{F}_{\tau_j} = F_{\tau_j}\right) \geq 1 - \frac{1}{r^2}. \tag{9}$$

Substituting back in (8), we have for any $r \geq e^{11}$,

$$\begin{aligned}
\Pr(G_j \leq r) &\geq \mathbb{E}\left[I\left(\hat{\mu}_1 + \frac{z}{\sqrt{j+1}} \geq y_i\right)\left(1 - \frac{1}{r^2}\right)\right] \\
&= \left(1 - \frac{1}{r^2}\right) \Pr\left(\hat{\mu}_1 + \frac{z}{\sqrt{j+1}} \geq y_i\right)
\end{aligned} \tag{10}$$

Next, we apply Chernoff-Hoeffding bounds to lower bound the probability $\Pr\left(\hat{\mu}_1 + \frac{z}{\sqrt{j+1}} \geq y_i\right)$. Recall from Definition 2.2 that $\hat{\mu}_1(t) = \frac{\sum_{s=1:i(s)=1}^{t-1} r_1(s)}{k_1(t)+1}$. Using Chernoff bounds for $t = \tau_j + 1, k_1(t) = j$:

$$\Pr(\hat{\mu}_1 + \tfrac{1}{j+1} + \tfrac{x}{\sqrt{j+1}} \geq \mu_1) \geq 1 - e^{-2x^2}.$$

Here, the term $1/(j+1)$ was added to $\hat{\mu}_1$ to adjust for the fact that $\hat{\mu}_1$ is not simply average of the past j observations, instead, it is the sum of past j observations divided

by $j+1$ (Definition 2.2). Now, we use $x := z - \frac{1}{\sqrt{j+1}} \geq z-1$ to get

$$\Pr\left(\hat{\mu}_1 + \frac{z}{\sqrt{j+1}} \geq \mu_1\right) \geq 1 - e^{-2(z-1)^2} \geq 1 - e^{-2z^2+4z} \geq 1 - \frac{1}{r^2}e^4\sqrt{\ln(r)}.$$

Using, $y_i \leq \mu_1$, this gives

$$\Pr\left(\hat{\mu}_1 + \frac{z}{\sqrt{j+1}} \geq y_i\right) \geq 1 - \frac{1}{r^2}e^4\sqrt{\ln(r)}. \tag{11}$$

Observe that for large r (in particular, for any $r \geq e^{64}$), $\frac{1}{r^2}e^{\sqrt{\ln(r)}} \leq \frac{1}{r^{1.5}}$.

Therefore, substituting, for any $r \geq e^{64}$,

$$\Pr(G_j \leq r) \;\geq\; 1 - \frac{1}{r^2} - \frac{1}{r^{1.5}}. \tag{12}$$

This gives,

$$\begin{aligned} E[G_j] \;&=\; \sum_{r=0}^{\infty} \Pr(G_j \geq r) \\ &=\; 1 + \sum_{r=1}^{\infty} \Pr(G_j \geq r) \\ &\leq\; 1 + e^{64} + \sum_{r\geq 1}\left(\frac{1}{r^2} + \frac{1}{r^{1.5}}\right) \\ &\leq\; 1 + e^{64} + 2 + 2.7. \end{aligned} \tag{13}$$

This proves a constant bound of $\mathbb{E}\left[\frac{1}{p_{i,\tau_j+1}} - 1\right] = E[G_j] - 1 \leq e^{64} + 5$ for all j.

Next, we derive a tighter bound for large j. Consider $j > L_i(T)$. Given any $r \geq 1$, define G_j, $\mathbf{MAX}_r$, and $z = \sqrt{\ln r}$ as defined earlier. Then,

$$\begin{aligned} \Pr(G_j \leq r) \;&\geq\; \Pr(\mathbf{MAX}_r > y_i) \\ &\geq\; \Pr(\mathbf{MAX}_r > \hat{\mu}_1 + \frac{z}{\sqrt{j+1}} - \frac{\Delta_i}{6} \geq y_i) \\ &=\; \mathbb{E}\left[\mathbb{E}\left[I\left(\mathbf{MAX}_r > \hat{\mu}_1 + \frac{z}{\sqrt{j+1}} - \frac{\Delta_i}{6} \geq y_i\right) \,\middle|\, \mathcal{F}_{\tau_j}\right]\right] \\ &=\; \mathbb{E}\left[I\left(\hat{\mu}_1 + \frac{z}{\sqrt{j+1}} + \frac{\Delta_i}{6} \geq \mu_1\right) \Pr\left(\mathbf{MAX}_r > \hat{\mu}_1 + \frac{z}{\sqrt{j+1}} - \frac{\Delta_i}{6} \,\middle|\, \mathcal{F}_{\tau_j}\right)\right]. \end{aligned} \tag{14}$$

where we used that $y_i = \mu_1 - \frac{\Delta_i}{3}$. Now, since $j+1 \geq L_i(T) = \frac{288\ln(T\Delta_i^2+e^{32})}{\Delta_i^2}$,

$$2\frac{\sqrt{2\ln(T\Delta_i^2+e^{32})}}{\sqrt{j+1}} \leq \frac{\Delta_i}{6}.$$

Therefore, for $r \leq (T\Delta_i^2+e^{32})^2$,

$$\frac{z}{\sqrt{j+1}} - \frac{\Delta_i}{6} = \frac{\sqrt{\ln(r)}}{\sqrt{j+1}} - \frac{\Delta_i}{6} \leq -\frac{\Delta_i}{12}$$

Then, since Θ_j is $\mathcal{N}(\hat{\mu}_1(\tau_j+1), \frac{1}{j+1})$ distributed random variable, using the upper bound in Fact 4, we obtain for any instantiation F_{τ_j} of history $\mathcal{F}_{\tau_j}$,

$$\Pr\left(\Theta_j > \hat{\mu}_1(\tau_j+1) - \frac{\Delta_i}{12} \middle| \mathcal{F}_{\tau_j} = F_{\tau_j}\right) \geq 1 - \frac{1}{2}e^{(j+1)\frac{\Delta_i^2}{288}} \geq 1 - \frac{1}{2(T\Delta_i^2 + e^{32})}$$

This implies

$$\Pr\left(\mathbf{MAX}_r > \hat{\mu}_1(\tau_j+1) + \frac{z}{\sqrt{j+1}} - \frac{\Delta_i}{6} \middle| \mathcal{F}_{\tau_j} = F_{\tau_j}\right) \geq 1 - \frac{1}{2^r(T\Delta_i^2 + e^{32})^r}$$

Also, for any $t \geq \tau_j + 1$, we have $k_1(t) \geq j$, and using Chernoff-Hoeffding bounds (Fact 2), we get

$$\Pr\left(\hat{\mu}_1(t) + \frac{z}{\sqrt{j+1}} - \frac{\Delta_i}{6} \geq y_i\right) \geq \Pr\left(\hat{\mu}_1(t) \geq \mu_1 - \frac{\Delta_i}{6}\right) \geq 1 - e^{-2k_1(t)\Delta_i^2/36} \geq 1 - \frac{1}{(T\Delta_i^2 + e^{32})^{16}}$$

Let $T' = (T\Delta_i^2 + e^{32})^2$. Therefore, for $1 \leq r \leq T'$

$$\Pr(G_j \leq r) \geq 1 - \frac{1}{2^r(T')^{r/2}} - \frac{1}{(T')^8}$$

When $r \geq T' \geq e^{64}$, we can use (12) to obtain

$$\Pr(G_j \leq r) \geq 1 - \frac{1}{r^2} - \frac{1}{r^{1.5}}.$$

Combining these bounds,

$$\begin{aligned}
E[G_j] &\leq \sum_{r=0}^{\infty} \Pr(G_j \geq r) \\
&\leq 1 + \sum_{r=1}^{T'} \Pr(G_j \geq r) + \sum_{r=T'}^{\infty} \Pr(G_j \geq r) \\
&\leq 1 + \sum_{r=1}^{T'} \frac{1}{(2\sqrt{T'})^r} + \frac{1}{(T')^7} + \sum_{r=T'}^{\infty} \frac{1}{r^2} + \frac{1}{r^{1.5}} \\
&\leq 1 + \frac{1}{\sqrt{T'}} + + \frac{1}{(T')^7} + \frac{2}{T'} + \frac{3}{\sqrt{T'}} \\
&\leq 1 + \frac{5}{T\Delta_i^2 + e^{32}}.
\end{aligned}$$

Above gives an upper bound of $\mathbb{E}\left[\frac{1}{p_{i,\tau_j+1}}\right] - 1 = \mathbb{E}[G_j] - 1 \leq \frac{5}{T\Delta_i^2}$ for $j > L_i(T)$.

□

Substituting the bound from Lemma 2.13 into (4), we obtain the following bound on the first term on the right hand side of (2).

LEMMA 2.14.

$$\sum_{t=1}^{T} \Pr(i(t) = i, E_i^{\mu}(t), E_i^{\theta}(t)) \leq (e^{64} + 5)(L_i(T)) + \frac{5}{\Delta_i^2}.$$

The proof of Lemma 2.11 can be easily adapted to Gaussian priors. So, this lemma holds as is for this case. Here, $x_i = \mu_i + \frac{\Delta_i}{3}$, therefore, using Pinsker's inequality $d(x_i, \mu_i) \geq 2(x_i - \mu_i)^2 = \frac{2\Delta_i^2}{9}$.

LEMMA 2.15.

$$\sum_{t=1}^{T} \Pr\left(i(t) = i, \overline{E_i^{\mu}(t)}\right) \;\leq\; \frac{1}{d(x_i, \mu_i)} + 1 \leq \frac{9}{2\Delta_i^2} + 1.$$

Corresponding to Lemma 2.12, we prove the following lemma.

LEMMA 2.16.

$$\textstyle\sum_{t=1}^{T} \Pr\left(i(t) = i, \overline{E_i^{\theta}(t)}, E_i^{\mu}(t)\right) \;\leq\; L_i(T) + \frac{1}{\Delta_i^2}.$$

where $L_i(T) \geq \frac{2\ln(T\Delta_i^2)}{(y_i - x_i)^2}$.

PROOF. The proof of this lemma is similar to the proof of Lemma 2.12. We decompose each summand into two parts, based on whether or not $k_i(T)$ is large ($k_i(t) > L_i(t)$).

$$\begin{aligned}\sum_{t=1}^{T} \Pr\left(i(t) = i, \overline{E_i^{\theta}(t)}, E_i^{\mu}(t)\right) \;=\; &\sum_{t=1}^{T} \Pr\left(i(t) = i, k_i(t) \leq L_i(T), \overline{E_i^{\theta}(t)}, E_i^{\mu}(t)\right) \\ &+ \sum_{t=1}^{T} \Pr\left(i(t) = i, k_i(t) > L_i(T), \overline{E_i^{\theta}(t)}, E_i^{\mu}(t)\right). \quad (15)\end{aligned}$$

The first term in the above decomposition is bounded by $\mathbb{E}[\sum_{t=1}^{T} I(i(t) = i, k_i(t) \leq L_i(T))]$, which is bounded trivially by $L_i(T)$. What remains is to bound the second term by $1/\Delta_i^2$. To this end, we show that if $k_i(t)$ is large and the event $E_i^{\mu}(t)$ is satisfied, then the probability that the event $E_i^{\theta}(t)$ is violated is small. Recall that $E_i^{\theta}(t)$ is defined as the event that $\theta_i(t) \leq y_i$. And, $E_i^{\mu}(t)$ is the event that $\hat{\mu}_i(t) \leq x_i$. Then,

$$\begin{aligned}\sum_{t=1}^{T} \Pr\left(i(t) = i, k_i(t) > L_i(T), \overline{E_i^{\theta}(t)}, E_i^{\mu}(t)\right) \;&\leq\; \mathbb{E}\left[\sum_{t=1}^{T} \Pr\left(i(t) = i, \overline{E_i^{\theta}(t)} \,\middle|\, k_i(t) > L_i(T), E_i^{\mu}(t), \mathcal{F}_{t-1}\right)\right] \\ &\leq\; \mathbb{E}\left[\sum_{t=1}^{T} \Pr\left(\theta_i(t) > y_i \,|\, k_i(t) > L_i(T), \hat{\mu}_i(t) \leq x_i, \mathcal{F}_{t-1}\right)\right]\end{aligned}$$

Now, $\theta_i(t)$ is a $\mathcal{N}\left(\hat{\mu}_i(t), \frac{1}{k_i(t)+1}\right)$ distributed Gaussian random variable. An $\mathcal{N}(m, \sigma^2)$ distributed r.v. (i.e., a Gaussian random variable with mean m and variance σ^2) is stochastically dominated by $\mathcal{N}(m', \sigma^2)$ distributed r.v. if $m' \geq m$. Therefore, given $\hat{\mu}_i(t) \leq x_i$, the distribution of $\theta_i(t)$ is stochastically dominated by $\mathcal{N}\left(\hat{\mu}_i(t), \frac{1}{k_i(t)+1}\right)$. That is,

$$\Pr\left(\theta_i(t) > y_i \,|\, k_i(t) > L_i(T), \hat{\mu}_i(t) \leq x_i, \mathcal{F}_{t-1}\right) \;\leq\; \Pr\left(\mathcal{N}\left(x_i, \tfrac{1}{k_i(t)+1}\right) > y_i \,\middle|\, \mathcal{F}_{t-1}, k_i(t) > L_i(T)\right).$$

Here, we slightly abused the notation for readability: $\Pr(\mathcal{N}(m, \sigma^2) > y_i)$ represents the probability that a random variable distributed as $\mathcal{N}(m, \sigma^2)$ takes value greater than y_i.

Now, using the concentration of Gaussian distribution (Fact 4), we obtain that for any fixed $k_i(t) > L_i(T)$,

$$\begin{aligned}\Pr\left(\mathcal{N}\left(x_i, \tfrac{1}{k_i(t)+1}\right) > y_i\right) &\leq \frac{1}{2}e^{-\frac{(k_i(t)+1)(y_i-x_i)^2}{2}} \\ &\leq \frac{1}{2}e^{-\frac{(L_i(t))(y_i-x_i)^2}{2}},\end{aligned}$$

which is smaller than $\frac{1}{T\Delta_i^2}$ because $L_i(T) \geq \frac{2\ln(T\Delta_i^2)}{(y_i-x_i)^2}$. Substituting, we get,

$$\Pr\left(\theta_i(t) > y_i \,|\, k_i(t) > L_i(T), \hat{\mu}_i(t) \leq x_i, \mathcal{F}_{t-1}\right) \leq \frac{1}{T\Delta_i^2}.$$

Summing over $t = 1, \ldots, T$, we get a bound of $\frac{1}{\Delta_i^2}$ on the second term in (5), completing the proof of the lemma. □

Substituting the bounds from Lemma 2.14-2.16 in (2), we get

$$\mathbb{E}[k_i(T)] \leq (e^{64}+5)\frac{288\ln(T\Delta_i^2+e^{32})}{\Delta_i^2} + \frac{5}{\Delta_i^2} + \frac{288\ln(T\Delta_i^2+e^{32})}{\Delta_i^2} + \frac{1}{\Delta_i^2} + \frac{9}{2\Delta_i^2} + 1.$$

Thus the expected regret due to arm i is upper bounded by

$$\Delta_i\mathbb{E}[k_i(T)] \leq \frac{21}{2\Delta_i} + (e^{64}+6)\frac{288\ln(T\Delta_i^2+e^{32})}{\Delta_i} + \Delta_i.$$

The above is decreasing in Δ_i for $\Delta_i \geq \frac{e}{\sqrt{T}}$. Therefore, for every arm i with $\Delta_i \geq e\sqrt{\frac{N\ln N}{T}}$, the expected regret is bounded by

$$O\left(\sqrt{\frac{T\ln N}{N}} + 1\right).$$

For arms with $\Delta_i \leq e\sqrt{\frac{N\ln N}{T}}$, the total regret is bounded by $e\sqrt{NT\ln N}$. This bounds the total regret by $O(N + \sqrt{NT\ln N})$, or $O(\sqrt{NT\ln N})$ assuming $T \geq N$. This proves Theorem 1.3. □

3. PROOF OF THE LOWER BOUND

In this section we prove Theorem 1.4. To this end, we construct a problem instance such that TS has regret $\Omega(\sqrt{NT\ln N})$ at time T. Let each arm i when played produce a reward of μ_i. That is, the reward distribution for every arm is a one point distribution. Set $\mu_1 := \Delta := \sqrt{\frac{N\ln N}{T}}$, and $\mu_2 := 0, \cdots, \mu_N := 0$.

Note that $\hat{\mu}_i(t), i \neq 1$, will always be 0, as $\hat{\mu}_i(1) = 0$, and these arms will always produce reward 0 when played. For arm 1, $\hat{\mu}_1(t) = \frac{k_1(t)\mu_1}{k_1(t)+1} \leq \mu_1$. Every time an arm other than arm 1 is played, there is a regret of Δ. Let $\mathcal{F}_{t-1}$ represent the history of plays and outcomes until time t as defined earlier, which includes $k_i(t), \hat{\mu}_i(t), i = 1, \ldots, N$. Define A_{t-1} as the event that $\sum_{i\neq 1} k_i(t) \leq \frac{c\sqrt{NT\ln N}}{\Delta}$ for a fixed constant c (to be specified later). Note that whether the event A_{t-1} is true, is determined by $\mathcal{F}_{t-1}$.

Now, if A_{t-1} is not true, then the regret until time t is at least $c\sqrt{NT\ln N}$. Therefore, for any $t \leq T$ we can assume that $\Pr(A_{t-1}) \geq \frac{1}{2}$. Otherwise, the expected regret until

time t,

$$\begin{aligned}\mathbb{E}\left[\mathcal{R}(t)\right] &\geq \mathbb{E}\left[\mathcal{R}(t)|\overline{A_{t-1}}\right]\cdot\tfrac{1}{2}\\ &\geq \tfrac{1}{2}c\sqrt{NT\ln N} = \Omega(\sqrt{NT\ln N}).\end{aligned}$$

We will show that given any instantiation of the history $\mathcal{F}_{t-1}$ such that the event A_{t-1} is true, the probability of playing a suboptimal arm is at least a constant, so that the regret is $\Omega(T\Delta) = \Omega(\sqrt{NT\ln N})$. For this, we show that with constant probability, $\theta_1(t)$ will be smaller than μ_1, and $\theta_i(t)$ for some suboptimal arm i will be larger than μ_1.

Now, given any history $\mathcal{F}_{t-1}$ with any value of $k_1(t)$, $\theta_1(t)$ is a Gaussian r.v. with mean $\hat{\mu}_1(t) = \frac{k_1(t)\mu_1}{k_1(t)+1} \leq \mu_1$, therefore, by the symmetry of the Gaussian distribution,

$$\Pr\left(\theta_1(t) \leq \mu_1 \,|\, \mathcal{F}_{t-1}\right) \geq \tfrac{1}{2}.$$

Also, given any instantiation $\mathcal{F}_{t-1} = F_{t-1}$, the $\theta_i(t)$'s for $i \neq 1$ are independent Gaussian distributed random variables with mean 0 and variance $\frac{1}{k_i(t)+1}$, therefore, using anti-concentration inequality provided by Fact 4 for Gaussian random variables we get

$$\begin{aligned}&\Pr\left(\exists i \neq 1, \theta_i(t) > \mu_1 \,|\, \mathcal{F}_{t-1} = F_{t-1}\right)\\ =\ &\Pr\left(\exists i \neq 1, (\theta_i(t)-0)\sqrt{k_i(t)+1} > \Delta\sqrt{k_i(t)+1}\,\middle|\, \mathcal{F}_{t-1} = F_{t-1}\right)\\ \geq\ &\left(1 - \prod_{i\neq 1}\left(1 - \tfrac{1}{8\sqrt{\pi}}e^{-(k_i(t)+1)\frac{7\Delta^2}{2}}\right)\right).\end{aligned}$$

Now, given an instantiation F_{t-1} of $\mathcal{F}_{t-1}$ such that A_{t-1} is true, we have $\sum_{i\neq 1} k_i(t) \leq \frac{c\sqrt{NT\ln N}}{\Delta}$, so that the right hand side in the above inequality is minimized when $k_i(t) = \frac{c\sqrt{NT\ln N}}{(N-1)\Delta}$ for all $i \neq 1$. Then, substituting $\Delta = \sqrt{\frac{N\ln N}{T}}$ and choosing the constant c appropriately, we get

$$\begin{aligned}\Pr\left(\exists i, \theta_i(t) > \mu_1 \,|\, \mathcal{F}_{t-1} = F_{t-1}\right) &\geq \left(1 - \textstyle\prod_{i\neq 1}\left(1 - e^{-\ln N}\right)\right)\\ &= 1 - \left(1 - \tfrac{1}{N}\right)^{N-1}.\end{aligned}$$

for any F_{t-1} such that A_{t-1} is true.

Let us use the notation $\mathcal{F}_{t-1}/_{A_{t-1}}$ to indicate the random variable $\mathcal{F}_{t-1}$ conditioned on A_{t-1} being true. Then, to summarize, for any t, the probability of playing a suboptimal arm at time t satisfies

$$\begin{aligned}\Pr\left(\exists i \neq 1, i(t) = i\right) &\geq \Pr\left(\theta_1(t) \leq \mu_1, \exists i, \theta_i(t) > \mu_1\right)\\ &= \mathbb{E}\left[\Pr\left(\theta_1(t) \leq \mu_1, \exists i, \theta_i(t) > \mu_1 \,|\, \mathcal{F}_{t-1}\right)\right]\\ &\geq \mathbb{E}\left[\Pr\left(\theta_1(t) \leq \mu_1, \exists i, \theta_i(t) > \mu_1 \,\middle|\, \mathcal{F}_{t-1}/_{A_{t-1}}\right)\right]\cdot\Pr\left(A_{t-1}\right)\\ &= \mathbb{E}\left[\Pr\left(\theta_1(t) \leq \mu_1 \,\middle|\, \mathcal{F}_{t-1}/_{A_{t-1}}\right)\cdot\Pr\left(\exists i, \theta_i(t) > \mu_1 \,\middle|\, \mathcal{F}_{t-1}/_{A_{t-1}}\right)\right]\cdot\Pr\left(A_{t-1}\right)\\ &\geq \tfrac{1}{2}\cdot\left(1 - \left(1 - \tfrac{1}{N}\right)^{N-1}\right)\cdot\tfrac{1}{2}\\ &\geq p,\end{aligned}$$

for some constant $p \in (0,1)$. Therefore the regret in time T is at least $Tp\Delta = \Omega(\sqrt{NT\ln N})$. This proves Theorem 1.4. □

Conclusions. In this paper, we proved optimal problem dependent regret bounds for Thompson Sampling for stochastic MAB problem with Bernoulli arms. Further, we provided near-optimal problem-dependent and problem-independent regret bounds for

the general MAB problem with bounded rewards. Specifically, our technique yields the first problem-independent regret upper bound of $O(\sqrt{NT \ln T})$ for the version of TS with Beta priors and an upper bound of $O(\sqrt{NT \ln N})$ for the version of TS with Gaussian priors along with a matching lower bound. The availability of strong anti-concentration bounds for Gaussian distribution allowed us to derive these tight upper and lower bounds for the version of TS with Gaussian priors. Similar lower bound may exist for TS with Beta priors.

In addition to near-optimal regret bounds, an important contribution of this paper is a simple proof technique that is easily adapted to provide optimal or near-optimal problem-dependent and problem independent bounds, and handle different prior distributions. The basic techniques presented in this work have also been adapted to prove Thompson Sampling regret bounds for the contextual bandits problem in subsequent work ([Agrawal and Goyal 2013b]).

Acknowledgement. We thank the anonymous referees for careful reading and suggestions that have improved the presentation.

REFERENCES

M. Abramowitz and I. A. Stegun. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Dover, New York, 1964.

S. Agrawal and N. Goyal. Analysis of Thompson Sampling for the Multi-armed Bandit Problem. In *Proceedings of the 25th Annual Conference on Learning Theory (COLT)*, 2012.

S. Agrawal and N. Goyal. Further optimal regret bounds for thompson sampling. In *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics, (AISTATS)*, 2013a.

S. Agrawal and N. Goyal. Thompson sampling for contextual bandits with linear payoffs. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*, 2013b.

J.-Y. Audibert and S. Bubeck. Minimax Policies for Adversarial and Stochastic Bandits. In *Proceedings of the 22th Annual Conference on Learning Theory (COLT)*, 2009.

P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2-3):235–256, 2002.

S. Bubeck and N. Cesa-Bianchi. Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems. *CoRR*, 2012.

S. Bubeck and C. Liu. Prior-free and prior-dependent regret bounds for Thompson Sampling. In *48th Annual Conference on Information Sciences and Systems, CISS 2014, Princeton, NJ, USA, March 19-21, 2014*, pages 1–9, 2014.

O. Chapelle and L. Li. An Empirical Evaluation of Thompson Sampling. In *Advances in Neural Information Processing Systems (NIPS) 24*, pages 2249–2257, 2011.

A. Garivier and O. Cappé. The KL-UCB Algorithm for Bounded Stochastic Bandits and Beyond. In *Proceedings of the 24th Annual Conference on Learning Theory (COLT)*, 2011.

J. C. Gittins. *Multi-armed Bandit Allocation Indices*. Wiley Interscience Series in Systems and Optimization. John Wiley and Son, 1989.

T. Graepel, J. Q. Candela, T. Borchert, and R. Herbrich. Web-Scale Bayesian Click-Through rate Prediction for Sponsored Search Advertising in Microsoft's Bing Search Engine. In *Proceedings of the 27th International Conference on Machine Learning (ICML)*, pages 13–20, 2010.

O.-C. Granmo. Solving Two-Armed Bernoulli Bandit Problems Using a Bayesian Learning Automaton. *International Journal of Intelligent Computing and Cybernetics (IJICC)*, 3(2):207–234, 2010.

E. Jeřábek. Dual weak pigeonhole principle, Boolean complexity, and derandomization. *Annals of Pure and Applied Logic*, 129(1-3):1–37, October 2004.

E. Kaufmann, O. Cappé, and A. Garivier. On Bayesian Upper Confidence Bounds for Bandit Problems. In *Fifteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2012a.

E. Kaufmann, N. Korda, and R. Munos. Thompson Sampling: An Optimal Finite Time Analysis. In *International Conference on Algorithmic Learning Theory (ALT)*, 2012b.

T. Kocák, M. Valko, R. Munos, and S. Agrawal. Spectral Thompson Sampling. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, pages 1911–1917, 2014.

N. Korda, E. Kaufmann, and R. Munos. Thompson sampling for 1-dimensional exponential family bandits. In *Advances in Neural Information Processing Systems (NIPS) 26*, pages 1448–1456, 2013.

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

L. Li. Generalized Thompson Sampling for Contextual Bandits. *CoRR*, abs/1310.7163, 2013. URL http://arxiv.org/abs/1310.7163.

L. Li and O. Chapelle. Open Problem: Regret Bounds for Thompson Sampling. In *Proceedings of the 25th Annual Conference on Learning Theory (COLT)*, 2012.

O.-A. Maillard, R. Munos, and G. Stoltz. Finite-time analysis of multi-armed bandits problems with Kullback-Leibler divergences. In *Proceedings of the 24th Annual Conference on Learning Theory (COLT)*, 2011.

B. C. May and D. S. Leslie. Simulation studies in optimistic Bayesian sampling in contextual-bandit problems. Technical Report 11:02, Statistics Group, Department of Mathematics, University of Bristol, 2011.

B. C. May, N. Korda, A. Lee, and D. S. Leslie. Optimistic Bayesian sampling in contextual-bandit problems. Technical Report 11:01, Statistics Group, Department of Mathematics, University of Bristol, 2011.

P. A. Ortega and D. A. Braun. Linearly parametrized bandits. *Journal of Artificial Intelligence Research*, 38:475–511, 2010.

D. Russo and B. Van Roy. Learning to Optimize Via Posterior Sampling. *Mathematics of Operations Research*, 39(4):1221–1243, 2014.

D. Russo and B. Van Roy. An Information-Theoretic Analysis of Thompson Sampling. *Journal of Machine Learning Research (to appear)*, 2015.

D. Russo, I. Osband, and B. Van Roy. (more) efficient reinforcement learning via posterior sampling. *Advances in Neural Information Processing Systems (NIPS) 26*, 2013.

S. Scott. A modern Bayesian look at the multi-armed bandit. *Applied Stochastic Models in Business and Industry*, 26:639–658, 2010.

M. J. A. Strens. A Bayesian Framework for Reinforcement Learning. In *Proceedings of the Seventeenth International Conference on Machine Learning (ICML)*, pages 943–950, 2000.

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

J. Wyatt. *Exploration and Inference in Learning from Reinforcement*. PhD thesis, Department of Artificial Intelligence, University of Edinburgh, 1997.

A. SOME WELL-KNOWN INEQUALITIES

FACT 1 (CHERNOFF-HOEFFDING BOUND). *Let $X_1, \ldots, X_n$ be independent $0-1$ r.v.s with $E[X_i] = p_i$ (not necessarily equal). Let $X = \frac{1}{n}\sum_i X_i$, $\mu = E[X] = \frac{1}{n}\sum_{i=1}^n p_i$. Then, for any $0 < \lambda < 1 - \mu$,*

$$\Pr(X \geq \mu + \lambda) \leq \exp\{-nd(\mu + \lambda, \mu)\},$$

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

$$\Pr(X \leq \mu - \lambda) \leq \exp\{-nd(\mu - \lambda, \mu)\},$$

where $d(a, b) = a \ln \frac{a}{b} + (1 - a) \ln \frac{(1-a)}{(1-b)}$.

FACT 2 (CHERNOFF–HOEFFDING BOUND). *Let $X_1, ..., X_n$ be random variables with common range $[0, 1]$ and such that $\mathbb{E}[X_t \mid X_1, ..., X_{t-1}] = \mu$. Let $S_n = X_1 + \ldots + X_n$. Then for all $a \geq 0$,*

$$\Pr(S_n \geq n\mu + a) \leq e^{-2a^2/n},$$

$$\Pr(S_n \leq n\mu - a) \leq e^{-2a^2/n}.$$

FACT 3.

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

for all positive integers α, β.

Formula 7.1.13 from [Abramowitz and Stegun 1964] can be used to derive the following concentration for Gaussian distributed random variables.

FACT 4. *[Abramowitz and Stegun 1964] For a Gaussian distributed random variable Z with mean m and variance σ^2, for any z,*

$$\frac{1}{4\sqrt{\pi}} \cdot e^{-7z^2/2} < \Pr(|Z - m| > z\sigma) \leq \frac{1}{2} e^{-z^2/2}.$$

B. THOMPSON SAMPLING WITH BETA DISTRIBUTION

Proof of Lemma 2.9

Let $k_1(t) = j, S_1(t) = s$. Let $y = y_i$. Then, $p_{i,t} = \Pr(\theta_1(t) > y) = F^B_{j+1,y}(s)$ (using Fact 3). Let $\tau_j + 1$ denote the time step after the j^{th} play of arm 1. Then, $k_1(\tau_j + 1) = j$, and

$$\mathbb{E}\left[\frac{1}{p_{i,\tau_j+1}}\right] = \sum_{s=0}^{j} \frac{f_{j,\mu_1}(s)}{F^B_{j+1,y}(s)}.$$

Let $\Delta' = \mu_1 - y$.

In the derivation below, we abbreviate $F^B_{j+1,y}(s)$ as $F_{j+1,y}(s)$.

Case $j < \frac{8}{\Delta'}$. Let $R = \frac{\mu_1(1-y)}{y(1-\mu_1)}$, $D = y \ln \frac{y}{\mu_1} + (1-y) \ln \frac{1-y}{1-\mu_1}$. Note that $\mu_1 \geq y$, so that $R \geq 1$.

$$\begin{aligned}
\sum_{s=0}^{j} \frac{f_{j,\mu_1}(s)}{F_{j+1,y}(s)} &\leq \frac{1}{1-y} \sum_{s=0}^{j} \frac{f_{j,\mu_1}(s)}{F_{j,y}(s)} \\
&\leq \frac{1}{1-y} \sum_{s=0}^{\lfloor yj \rfloor} \frac{f_{j,\mu_1}(s)}{f_{j,y}(s)} + \frac{1}{1-y} \sum_{s=\lceil yj \rceil}^{j} 2 f_{j,\mu_1}(s) \\
&= \frac{1}{1-y} \sum_{s=0}^{\lfloor yj \rfloor} R^s \frac{(1-\mu_1)^j}{(1-y)^j} + \frac{1}{1-y} \sum_{s=\lceil yj \rceil}^{j} 2 f_{j,\mu_1}(s) \\
&= \frac{1}{1-y} \left(\frac{R^{\lfloor yj \rfloor + 1} - 1}{R-1} \right) \frac{(1-\mu_1)^j}{(1-y)^j} \\
&\quad + \frac{1}{1-y} \sum_{s=\lceil yj \rceil}^{j} 2 f_{j,\mu_1}(s) \\
&\leq \frac{1}{1-y} \left(\frac{R}{R-1} \right) R^{yj} \frac{(1-\mu_1)^j}{(1-y)^j} + \frac{2}{\Delta'} \\
&= \frac{\mu_1}{\Delta'} e^{-Dj} + \frac{2}{\Delta'} \\
&\leq \frac{3}{\Delta'}. \qquad (16)
\end{aligned}$$

Case $j \geq \frac{8}{\Delta'}$. We will divide the sum $Sum(0,j) = \sum_{s=0}^{j} \frac{f_{j,\mu_1}(s)}{F_{j+1,y}(s)}$ into four partial sums and prove that

$$\begin{aligned}
Sum(0, \lfloor yj \rfloor - 1) &\leq \Theta\left(e^{-Dj} \tfrac{1}{(j+1)} \tfrac{1}{\Delta'^2} \right) \\
&\quad + \Theta(e^{-2\Delta'^2 j}), \\
Sum(\lfloor yj \rfloor, \lfloor yj \rfloor) &\leq 3e^{-Dj}, \\
Sum(\lceil yj \rceil, \lfloor \mu_1 j - \tfrac{\Delta'}{2} j \rfloor) &\leq \Theta(e^{-\Delta'^2 j/2}), \\
Sum(\lceil \mu_1 j - \tfrac{\Delta'}{2} j \rceil, j) &\leq 1 + \tfrac{1}{e^{\Delta'^2 j/4} - 1}.
\end{aligned}$$

Together, the above estimates will prove the required bound.

We use the following bounds on the cdf of Binomial distribution [Jeřábek 2004, Prop. A.4].

For $s \leq y(j+1) - \sqrt{(j+1)y(1-y)}$,

$$F_{j+1,y}(s) = \Theta\left(\frac{y(j+1-s)}{y(j+1)-s} \binom{j+1}{s} y^s (1-y)^{j+1-s} \right). \qquad (17)$$

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

$$F_{j+1,y}(s) = \Theta(1). \qquad (18)$$

Bounding $Sum(0, \lfloor yj \rfloor - 1)$. Using the bounds just given, for any s,

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

This gives

$$Sum(0,\lfloor yj\rfloor-1) \leq \Theta\left(\frac{(1-\mu_1)^j}{(1-y)^{j+1}}\sum_{s=0}^{\lfloor yj\rfloor-1}\left(1-\frac{s}{y(j+1)}\right)\cdot R^s\right)+\Theta(1)\sum_{s=0}^{\lfloor yj\rfloor-1}f_{j,\mu_1}(s). \tag{19}$$

We now bound the first expression on the RHS.

$$\begin{aligned}\frac{(1-\mu_1)^j}{(1-y)^{j+1}}\sum_{s=0}^{\lfloor yj\rfloor-1}\left(1-\frac{s}{y(j+1)}\right)\cdot R^s &= \frac{(1-\mu_1)^j}{(1-y)^{j+1}}\left(\frac{R^{\lfloor yj\rfloor}-1}{R-1}\right.\\ &\quad \left.-\frac{1}{y(j+1)}\left(\frac{(\lfloor yj\rfloor-1)R^{\lfloor yj\rfloor}}{R-1}-\frac{R^{\lfloor yj\rfloor}-R}{(R-1)^2}\right)\right)\\ &\leq \frac{(1-\mu_1)^j}{(1-y)^{j+1}}\left(\frac{1}{y(j+1)}\frac{R^{\lfloor yj\rfloor}}{(R-1)^2}\right.\\ &\quad \left.+\frac{(y(j+1)-\lfloor yj\rfloor+1)}{y(j+1)}\frac{R^{\lfloor yj\rfloor}}{(R-1)}\right)\\ &\leq \frac{(1-\mu_1)^j}{(1-y)^{j+1}}\frac{3}{y(j+1)}\frac{R^{\lfloor yj\rfloor+1}}{(R-1)^2}\\ &\leq e^{-Dj}\frac{3}{y(1-y)(j+1)}\frac{R}{(R-1)^2}\end{aligned}$$

The last inequality uses

$$\frac{(1-\mu_1)^j}{(1-y)^j}R^{\lfloor yj\rfloor} \leq \frac{(1-\mu_1)^j}{(1-y)^j}R^{yj} = e^{-Dj}.$$

Now, $R-1 = \frac{\mu_1(1-y)}{y(1-\mu_1)} - 1 = \frac{\mu_1-y}{y(1-\mu_1)}$. And, $\frac{R}{R-1} = \frac{\mu_1(1-y)}{\mu_1-y}$. Therefore,

$$\begin{aligned}\frac{1}{y(1-y)(j+1)}\frac{R}{(R-1)^2} &= \frac{1}{y(1-y)(j+1)}\cdot\frac{\mu_1(1-y)}{\mu_1-y}\cdot\frac{y(1-\mu_1)}{\mu_1-y}\\ &= \frac{1}{(j+1)}\frac{\mu_1(1-\mu_1)}{(\mu_1-y)^2}.\end{aligned}$$

Substituting, we get

$$\frac{(1-\mu_1)^j}{(1-y)^{j+1}}\sum_{s=0}^{\lfloor yj\rfloor}\left(1-\frac{s}{y(j+1)}\right)\cdot R^s \leq e^{-Dj}\frac{1}{(j+1)}\frac{\mu_1(1-\mu_1)}{(\mu_1-y)^2}.$$

Substituting in (19)

$$\begin{aligned} Sum(0, \lfloor yj \rfloor - 1) &\leq \Theta\left(e^{-Dj}\frac{1}{(j+1)}\frac{1}{\Delta'^2}\right) + \Theta(1)\sum_{s=0}^{\lfloor yj \rfloor - 1} f_{j,\mu_1}(s) \\ &\leq \Theta\left(e^{-Dj}\frac{1}{(j+1)}\frac{1}{\Delta'^2}\right) + \Theta(e^{-2(\mu_1 - y)^2 j}). \end{aligned}$$

Bounding $Sum(\lfloor yj \rfloor, \lfloor yj \rfloor)$. We use $\frac{f_{j,\mu_1}(s)}{F_{j+1,y}(s)} \leq \frac{f_{j,\mu_1}(s)}{f_{j+1,y}(s)} = \left(1 - \frac{s}{j+1}\right) R^s \frac{(1-\mu_1)^j}{(1-y)^{j+1}}$, to get

$$\begin{aligned} Sum(\lfloor yj \rfloor, \lfloor yj \rfloor) &= \frac{f_{j,\mu_1}(\lfloor yj \rfloor)}{F_{j+1,y}(\lfloor yj \rfloor)} \\ &\leq \left(1 - \frac{yj - 1}{j+1}\right) R^{yj} \frac{(1-\mu_1)^j}{(1-y)^{j+1}} \\ &\leq \frac{(1 - y + \frac{2}{j+1})}{1-y} R^{yj} \frac{(1-\mu_1)^j}{(1-y)^j} \\ &\leq 3e^{-Dj}. \end{aligned} \tag{20}$$

The last inequality uses $j \geq \frac{1}{\Delta'} \geq \frac{1}{1-y}$.

Bounding $Sum(\lceil yj \rceil, \lfloor \mu_1 j - \frac{\Delta'}{2} j \rfloor)$. Now, if $j > \frac{1}{\Delta'}$,then $\sqrt{(j+1)y(1-y)} > \sqrt{y} > y$, so $y(j+1) - \sqrt{(j+1)y(1-y)} < yj \leq \lceil yj \rceil$. Therefore, (using the bounds by [Jeřábek 2004] given in (18)) for $s \geq \lceil yj \rceil$, $F_{j+1,y}(s) = \Theta(1)$. Using this observation, we derive the following.

$$\begin{aligned} Sum(\lceil yj \rceil, \lfloor \mu_1 j - \frac{\Delta'}{2} j \rfloor) &= \sum_{s=\lceil yj \rceil}^{\lfloor \mu_1 j - \frac{\Delta'}{2} j \rfloor} \frac{f_{j,\mu_1}(s)}{F_{j+1,y}(s)} \\ &= \Theta\left(\sum_{s=\lceil yj \rceil}^{\lfloor \mu_1 j - \frac{\Delta'}{2} j \rfloor} f_{j,\mu_1}(s)\right) \\ &\leq \Theta(e^{-2\left(\mu_1 j - \lfloor \mu_1 j - \frac{\Delta'}{2} j \rfloor\right)^2 / j}) \\ &= \Theta(e^{-\Delta'^2 j/2}), \end{aligned} \tag{21}$$

where the inequality follows using the Chernoff-Hoeffding bounds (Fact 2).

Bounding $Sum(\lceil \mu_1 j - \frac{\Delta'}{2} j \rceil, j)$. For $s \geq \lceil \mu_1 j - \frac{\Delta'}{2} j \rceil = \lceil yj + \frac{\Delta'}{2} j \rceil$, again using the Chernoff-Hoeffding bounds from Fact 2,

$$\begin{aligned} F_{j+1,y}(s) &\geq 1 - e^{-2(yj + \frac{\Delta'}{2} j - y(j+1))^2/(j+1)} \\ &\geq 1 - e^{2\Delta'} e^{-\Delta'^2 j/2} \\ &\geq 1 - e^{\Delta'^2 j/4} e^{-\Delta'^2 j/2} \\ &= 1 - e^{-\Delta'^2 j/4}. \end{aligned}$$

The last inequality uses $j \geq \frac{8}{\Delta'}$.

$$\begin{aligned} Sum(\lceil \mu_1 j - \frac{\Delta'}{2} j \rceil, j) &= \sum_{s=\lceil \mu_1 j - \frac{\Delta'}{2} j \rceil}^{j} \frac{f_{j,\mu_1}(s)}{F_{j+1,y}(s)} \\ &\leq \frac{1}{1 - e^{-\Delta'^2 j/4}} \\ &= 1 + \frac{1}{e^{\Delta'^2 j/4} - 1}. \end{aligned} \tag{22}$$

Combining, we get for $j \geq \frac{8}{\Delta'}$,

$$\begin{aligned} &\mathbb{E}\left[\frac{1}{p_{i,\tau_{j+1}}}\right] \\ &\leq 1 + \Theta(e^{-\Delta'^2 j/2} + \frac{1}{(j+1)\Delta'^2} e^{-Dj} + \frac{1}{e^{\Delta'^2 j/4} - 1}) \end{aligned}$$

□