Further Optimal Regret Bounds for Thompson Sampling

Shipra Agrawal
Microsoft Research India

Navin Goyal
Microsoft Research India

Abstract

Thompson Sampling 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 it to have comparable or better empirical performance compared to the state of the art methods. In this paper, we provide a novel regret analysis for Thompson Sampling that proves the first near-optimal problem-independent bound of $O(\sqrt{NT \ln T})$ on the expected regret of this algorithm. Our novel martingale-based analysis techniques are conceptually simple, and easily extend to distributions other than the Beta distribution. For the version of Thompson Sampling that uses Gaussian priors, we prove a problem-independent bound of $O(\sqrt{NT \ln N})$ on the expected regret, and demonstrate the optimality of this bound by providing a matching lower bound. This lower bound of $\Omega(\sqrt{NT \ln N})$ is the first lower bound on the performance of a natural version of Thompson Sampling that is away from the optimal bound ($O(\sqrt{NT})$) achievable for the multi-armed bandit problem by another algorithm [4]. Our near-optimal problem-independent bounds for Thompson Sampling solve a COLT 2012 open problem of Chapelle and Li. Additionally, our techniques simultaneously provide 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. The optimal problem-dependent regret bound for this problem was first proven recently by Kaufmann et al. [15].

1 Introduction

Multi-armed bandit problem models the exploration/exploitation trade-off inherent in sequential decision problems. Many versions and generalizations of the multi-armed bandit problem 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 the stochastic bandit problem, some popular ones include Upper Confidence Bound (UCB) family of algorithms, (e.g., [16, 5], and more recently [4, 9, 17, 14]), which have good theoretical guarantees, and the algorithm by [10], which gives optimal strategy under Bayesian setting with known priors and geometric time-discounted rewards. In one of the earliest works on stochastic bandit problems, [23] 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 very natural algorithm and the same idea has been rediscovered many times independently in the context of reinforcement learning, e.g., in [24, 20, 22].

Recently, TS has attracted considerable attention. Several studies (e.g., [12, 21, 11, 7, 19, 15]) 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. [12, 18] provide weak guarantees, namely, a bound of $o(T)$ on expected regret in time T. Significant progress was made in more recent work [2, 15]. In [2], the first logarithmic bound on expected regret of TS algorithm were proven.

Preliminary work. Under review by AISTATS 2013. Do not distribute.

[15] provided a bound that matches the asymptotic lower bound of [16] for this problem. However, both these bounds were problem dependent, i.e. the regret bounds are logarithmic in 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 existing 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 Chapaelle and Li [8].

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 (arguably simpler than the previous work). Our technique easily extends to distributions other than Beta distribution, and it also extends to the more general contextual bandits setting [3]. While one of the basic ideas for the analysis in the contextual bandits setting of [3] is similar to an idea in this paper, the details are substantially different.

Before stating our results, we describe the MAB problem and the TS algorithm formally.

1.1 The multi-armed bandit problem

We consider the stochastic multi-armed bandit (MAB) 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 reward 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 the MAB problem 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 is to maximize the expected total reward in 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 more convenient 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 $\Delta_i := \mu^* - \mu_i$. Also, let $k_i(t)$ denote the number of times arm i has been played up to step $t-1$. Then the expected total regret in time T is given by

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

Other performance measures include PAC-style guarantees; we do not consider those measures here.

1.2 Thompson Sampling

The basic idea 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 Thompson Sampling is a specific algorithm due to Thompson, in this paper we will use Thompson Sampling (TS) to refer to a class of algorithms that have a similar structure. The general structure of TS for the contextual bandits problem involves the following elements (this description of TS follows closely that of [7]):

1. a set Θ 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 past time steps;
4. an assumed likelihood function $P(r|\tilde{\mu})$, which gives the probability of reward given a context b and a parameter $\tilde{\mu}$;
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.

In each round, TS plays an arm according to its posterior probability of maximizing the expected reward. A simple way to achieve that is to produce a sample of reward for each arm, using the posterior distributions, 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 respectively.

Thompson Sampling using Beta priors 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 (reward $=1$) is μ_i. Beta priors is useful 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 $(0,1)$. 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 as $\text{Beta}(S_i(t)+1, F_i(t)+1)$. The algorithm then samples from these posterior distributions of the μ_i's, and plays an arm according to the probability of its mean being the largest.

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 $\text{Beta}(S_i + 1, F_i + 1)$ 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 [2], which seamlessly extends results for Bernoulli bandits to general stochastic bandit problem.

Thompson Sampling using Gaussian priors As before, let $k_i(t)$ denote the number of plays of arm i until time $t-1$, $i(t)$ denote the arm played at time t. Let $r_i(t)$ denote the reward of arm i at time t, and define $\hat{\mu}_i(t)$ as:

$$\hat{\mu}_i(t) = \frac{\sum_{w=1:i(w)=i}^{t-1} r_i(t)}{k_i(t)+1}.$$

Note that $\hat{\mu}_i(1) = 0$. To derive TS algorithm 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.

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 $\hat{\mu}_{i(t)} = \frac{(\hat{\mu}_{i(t)} k_{i(t)} + r_t)}{k_{i(t)}+1}$, $k_{i(t)} = k_{i(t)} + 1$.
end
```

1.3 Our results

In this article, we bound the *finite time* expected regret of TS. From now on 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 and of course the algorithm does not use this assumption. The assumption of *unique* optimal arm is also without loss of generality, since adding more arms with $\mu_i = \mu^*$ can only decrease the expected regret; details of this argument were provided in [2].

Upper bounds

Theorem 1. *For the N-armed stochastic bandit problem, TS algorithm, using Beta priors has expected regret*

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

in 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 2. *For the N-armed stochastic bandit problem, TS using Beta priors, has expected regret*

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

in time T, where the big-Oh notation hides only the absolute constants.

Theorem 3. *For the N-armed stochastic bandit problem, TS using Gaussian priors, has expected regret*

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

in time $T \geq N$, where the big-Oh notation hides only the absolute constants.

Lower bound

Theorem 4. *There exists an instance of N-armed stochastic bandit problem, for which TS, using Gaussian priors, has expected regret*

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

in time $T \geq N$.

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 problem parameters $\mu_i, \Delta_i, i = 1, \ldots, N$. Lai and Robbins [16] essentially proved the following lower bound on the regret of any bandit algorithm (see [16] for a precise statement): $\mathbb{E}[\mathcal{R}(T)] \geq \left[\sum_{i=2}^{N} \frac{\Delta_i}{d(\mu_i,\mu_1)} + o(1)\right] \ln T$. They also gave algorithms asymptotically achieving this guarantee, though unfortunately their algorithms are not efficient. Auer et al. [5] gave the UCB1 algorithm, which is efficient and achieves the following bound: $\mathbb{E}[\mathcal{R}(T)] \leq \left[8 \sum_{i=2}^{N} \frac{1}{\Delta_i}\right] \ln T + (1 + \pi^2/3) \left(\sum_{i=2}^{N} \Delta_i\right)$. More recently, Kaufmann et al. [14] gave Bayes-UCB algorithm which achieves the lower bound of [16] for Bernoulli rewards. Bayes-UCB is a UCB-like algorithm, where the upper confidence bounds are based on the quantiles of Beta posterior distributions. Interestingly, these upper confidence bounds turn out to be similar to those used by algorithms in [9] and [17]; these latter papers also achieve the lower bound of [16] using UCB-like algorithms. Our bounds in Theorem 1 achieve the asymptotic lower bounds of [16], and match those provided by [15] for TS.

Theorem 2 and 3 shows that TS with Beta and Gaussian distribution 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-inpdependent lower bound (see Section 3.3 of [6]) for the multi-armed bandit problem within logarithmic factors. The problem-dependent bounds in the existing work implied only suboptimal problem-independent bounds: [2] implied a problem independent bound of $\tilde{O}(N^{1/5}T^{4/5})$. In [15], the additive problem dependent term was not explicitly calculated, which makes it difficult to derive the corresponding problem independent bound, but on a preliminary examination, it appears 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 [6]). More recently, Audibert and Bubeck [4] gave an algorithm MOSS, inspired by UCB1, with regret $O(\sqrt{NT})$ that matches the $\Omega(\sqrt{NT})$ problem-inpdependent lower bound for the multi-armed bandit problem. Interestingly, Theorem 4 shows that this is unachievable for TS with Gaussian priors, as there is a lower bound of $\Omega(\sqrt{NT \ln N})$ on expected regret. This is the first lower bound for TS that differs from the general lower bound for the problem, and demonstrates a slight limitation of TS, although for only a specifically designed problem instance.

2 Proofs of upper bounds

In this section, we prove Theorems 1, 2 and 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 1, which forms the core of our analysis. Further, we show that the coefficient in this linear function decreases exponentially fast with the increase in the number of plays of the optimal arm (Lemma 2), this allows us to bound the total number of plays of every suboptimal arm, to bound the regret as desired. The difference between the analysis for obtaining the logarithmic problem-dependent bound of Theorem 1, and the problem-independent bound of Theorem 2 is technical, and occurs only towards the end of the proof.

We recall some of the definitions introduced earlier, and introduce some new notations used in the proof. $F^B_{n,p}(\cdot)$ denotes the cdf and $f^B_{n,p}(\cdot)$ denotes the probability mass function of the binomial distribution with parameters n, p. Let $F^{beta}_{\alpha,\beta}(\cdot)$ denote the cdf of the beta distribution with parameters α, β.

Definition 1. *$k_i(t)$ is defined as the number of plays of arm i until time $t-1$, and $S_i(t)$ as the number of successes among the plays of arm i until time $t-1$. Also, $i(t)$ denotes the arm played at time t.*

Definition 2. *For each arm i, 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 approporiate points in the proof. Define* $L_i(T) = \frac{\ln T}{d(x_i,y_i)}$, *and* $\hat{\mu}_i(t) = S_i(t)/(k_i(t)+1)$ *(note that* $\hat{\mu}_i(t) = 0$ *when* $k_i(t) = 0$*). Define* $E^{\mu}_i(t)$ *as the event that* $\hat{\mu}_i(t) \leq x_i$*. Define* $E^{\theta}_i(t)$ *as the event that* $\theta_i(t) \leq y_i$*.*

Intuitively, $E^{\mu}_i(t)$*,* $E^{\theta}_i(t)$ *are the events that* $\hat{\mu}_i(t)$ *and* $\theta_i(t)$*, respectively, are not too far from the mean* μ_i*. As we show later, these events will hold with high probability for most time steps.*

Definition 3. *Define filtration* $\mathcal{F}_{t-1}$ *as the history of plays until time* $t-1$*, i.e.*

$$\mathcal{F}_{t-1} = \{i(w), r_{i(w)}(w), i = 1, \ldots, N, w = 1, \ldots, t-1\},$$

where $i(t)$ *denotes the arm played at time* t*, and* $r_i(t)$ *denotes the reward observed for arm* i *at time* t*.*

Definition 4. *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 determined by* $\mathcal{F}_{t-1}$*.*

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

Lemma 1. *For all* $t \in [1, T]$*, and* $i \neq 1$*,*

$$\begin{aligned} &\Pr\left(i(t) = i, E^{\mu}_i(t), E^{\theta}_i(t)) \,\middle|\, \mathcal{F}_{t-1}\right) \\ \leq\ & \frac{(1-p_{i,t})}{p_{i,t}} \Pr\left(i(t) = 1, E^{\mu}_i(t), E^{\theta}_i(t) \,\middle|\, \mathcal{F}_{t-1}\right), \end{aligned}$$

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

Proof. Note that whether $E^{\mu}_i(t)$ is true or not is determined by $\mathcal{F}_{t-1}$. Assume that filtration $\mathcal{F}_{t-1}$ is such that $E^{\mu}_i(t)$ is true (otherwise the probability on the left hand side is 0 and the inequality is trivially true). It then suffices to prove that

$$\begin{aligned} &\Pr\left(i(t) = i \,\middle|\, E^{\theta}_i(t), \mathcal{F}_{t-1}\right) \\ \leq\ & \frac{(1-p_{i,t})}{p_{i,t}} \Pr\left(i(t) = 1 \,\middle|\, E^{\theta}_i(t), \mathcal{F}_{t-1}\right). \end{aligned} \tag{1}$$

Let $M_i(t)$ denote the event that arm i exceeds all the suboptimal arms at time t. That is,

$$M_i(t) : \theta_i(t) \geq \theta_j(t), \forall j \neq 1.$$

We will prove the following two inequalities which immediately give (1).

$$\begin{aligned} &\Pr\left(i(t) = 1 \,\middle|\, E^{\theta}_i(t), \mathcal{F}_{t-1}\right) \\ \geq\ & p_{i,t} \cdot \Pr\left(M_i(t) \,\middle|\, E^{\theta}_i(t), \mathcal{F}_{t-1}\right), \end{aligned} \tag{2}$$

$$\begin{aligned} &\Pr\left(i(t) = i \,\middle|\, E^{\theta}_i(t), \mathcal{F}_{t-1}\right) \\ \leq\ & (1-p_{i,t}) \cdot \Pr\left(M_i(t) \,\middle|\, E^{\theta}_i(t), \mathcal{F}_{t-1}\right). \end{aligned} \tag{3}$$

We have

$$\begin{aligned} &\Pr\left(i(t) = 1 \,\middle|\, E^{\theta}_i(t), \mathcal{F}_{t-1}\right) \\ \geq\ & \Pr\left(i(t) = 1, M_i(t) \,\middle|\, E^{\theta}_i(t), \mathcal{F}_{t-1}\right) \\ =\ & \Pr\left(M_i(t) \,\middle|\, E^{\theta}_i(t), \mathcal{F}_{t-1}\right) \cdot \Pr\left(i(t) = 1 \,\middle|\, M_i(t), E^{\theta}_i(t), \mathcal{F}_{t-1}\right). \end{aligned} \tag{4}$$

Now, given $M_i(t), E^{\theta}_i(t)$, it holds that for all $j \neq i, j \neq 1$,

$$\theta_j(t) \leq \theta_i(t) \leq y_i,$$

and so

$$\begin{aligned} &\Pr(i(t) = 1 \,|\, M_i(t), E^{\theta}_i(t), \mathcal{F}_{t-1}) \\ \geq\ & \Pr(\theta_1(t) > y_i \,|\, M_i(t), E^{\theta}_i(t), \mathcal{F}_{t-1}) \\ =\ & \Pr(\theta_1(t) > y_i \,|\, \mathcal{F}_{t-1}) \\ =\ & p_{i,t}. \end{aligned}$$

The second last equality follows because the events $M_i(t)$ and $E^{\theta}_i(t), \forall i \neq 1$ involve conditions on only $\theta_j(t)$ for $j \neq 1$, and given $\mathcal{F}_{t-1}$ (and hence $\hat{\mu}_j(t), k_j(t), \forall j$), $\theta_1(t)$ is independent of all the other $\theta_j(t), j \neq 1$, and hence independent of these events. This together with (4) gives (2).

Since $E^{\theta}_i(t)$ is the event that $\theta_i(t) \leq y_i$, therefore, given $E^{\theta}_i(t)$, $i(t) = i$ only if $\theta_1(t) < y_i$. This gives (3):

$$\begin{aligned} &\Pr\left(i(t) = i \,\middle|\, E^{\theta}_i(t), \mathcal{F}_{t-1}\right) \\ \leq\ & \Pr\left(\theta_1(t) \leq y_i, \theta_i(t) \geq \theta_j(t), \forall j \neq 1, \,\middle|\, E^{\theta}_i(t), \mathcal{F}_{t-1}\right) \\ =\ & \Pr\left(\theta_1(t) \leq y_i \,\middle|\, \mathcal{F}_{t-1}\right) \cdot \Pr\left(\theta_i(t) \geq \theta_j(t), \forall j \neq 1 \,\middle|\, E^{\theta}_i(t), \mathcal{F}_{t-1}\right) \\ =\ & (1-p_{i,t}) \cdot \Pr(M_i(t) \,|\, E^{\theta}_i(t), \mathcal{F}_{t-1}). \end{aligned}$$

□

2.1 Proof of Theorem 1

We can bound the expected number of plays of a suboptimal arm i as follows:

$$\begin{aligned} \mathbb{E}[k_i(T)] &= \sum_{t=1}^{T} \Pr(i(t) = i) \\ &= \sum_{t=1}^{T} \Pr(i(t) = i, E^{\mu}_i(t), E^{\theta}_i(t)) \\ &\quad + \sum_{t=1}^{T} \Pr\left(i(t) = i, E^{\mu}_i(t), \overline{E^{\theta}_i(t)}\right) \\ &\quad + \sum_{t=1}^{T} \Pr\left(i(t) = i, \overline{E^{\mu}_i(t)}\right) \end{aligned} \tag{5}$$

Let τ_k denote the time step at which arm 1 is played for the k^{th} time for $k \geq 1$, and let $\tau_0 = 0$. Then, using Lemma 1, we can bound the first term above as:

$$
\begin{aligned}
& \sum_{t=1}^{T} \Pr(i(t)=i, E_i^{\mu}(t), E_i^{\theta}(t)) \\
& \leq \sum_{t=1}^{T} \mathbb{E}\left[\frac{(1-p_{i,t})}{p_{i,t}} I(i(t)=1, E_i^{\theta}(t), E_i^{\mu}(t))\right] \\
(*) & \leq \sum_{k=0}^{T-1} \mathbb{E}\left[\frac{(1-p_{i,\tau_k+1})}{p_{i,\tau_k+1}} \sum_{t=\tau_k+1}^{\tau_{k+1}} I(i(t)=1)\right] \\
& = \sum_{k=0}^{T-1} \mathbb{E}\left[\frac{1}{p_{i,\tau_k+1}} - 1\right]. \qquad (6)
\end{aligned}
$$

The inequality marked $(*)$ uses the observation 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 first arm. Thus, $p_{i,t}$ is same at all time steps $t \in \{\tau_k + 1, \ldots, \tau_{k+1}\}$, for every k. We prove the following lemma to bound the sum of $\frac{1}{p_{i,\tau_k+1}}$.

Lemma 2. *Let τ_j denote the time step at which j^{th} trial of first arm happens, then*

$$\mathbb{E}[\frac{1}{p_{i,\tau_k+1}}] \leq$$

$$
\begin{cases}
1 + \frac{3}{\Delta_i'}, & \textit{for } k < \frac{8}{\Delta_i'}, \\
1 + \Theta(e^{-\Delta_i'^2 k/2} + \frac{1}{(k+1)\Delta_i'^2} e^{-D_i k} & \\
\qquad + \frac{1}{e^{\Delta_i'^2 k/4} - 1}), & \textit{for } k \geq \frac{8}{\Delta_i'},
\end{cases}
$$

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

Proof. The proof of this inequality involves some careful algebraic manipulations using tight estimates for partial Binomial sums provided by [13]. Refer to Appendix B.3 for details. □

For the remaining two terms in Equation (5), we prove the following lemmas.

Lemma 3.

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

Proof. This follows from the Chernoff-Hoeffding bounds for concentration of $\hat{\mu}_i(t)$. Appendix B.1 has details. □

Lemma 4.

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

Proof. This follows from the observation that $\theta_i(t)$ is well-concentrated around its mean when $k_i(t)$ is large, that is, larger than $L_i(T)$. Appendix B.2 has details. □

For obtaining the problem-dependent bound of Theorem 1, for some $0 < \epsilon \leq 1$, 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 $d(x_i, y_i) = d(x_i, \mu_1)/(1+\epsilon) = d(\mu_i, \mu_1)/(1+\epsilon)^2$ [1]. 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

$$\frac{1}{d(x_i, \mu_i)} \leq \frac{2}{(x_i - \mu_i)^2} = O(\frac{1}{\epsilon^2}).$$

Here order notation is hiding functions of μ_is and Δ_is, since they are assumed to be constants. Substituing in Equation (5), and Equation (6), we get,

$$
\begin{aligned}
& \mathbb{E}[k_i(T)] \\
\leq \quad & \frac{24}{\Delta_i'^2} + \sum_{j=0}^{T-1} \Theta\left(e^{-\Delta_i'^2 j/2} + \frac{1}{(j+1)\Delta_i'^2} e^{-D_i j} + \right. \\
& \left. \frac{1}{e^{\Delta_i'^2 j/4} - 1}\right) + L_i(T) + 1 + \frac{1}{d(x_i, \mu_i)} + 1 \\
\leq \quad & \frac{24}{\Delta_i'^2} + \Theta\left(\frac{1}{\Delta_i'^2} + \frac{1}{\Delta_i'^2 D} + \frac{1}{\Delta_i'^4} + \frac{1}{\Delta_i'^2}\right) \\
& + (1+\epsilon)^2 \frac{\ln T}{d(\mu_i, \mu_1)} + O(\frac{1}{\epsilon^2}) \\
= \quad & O(1) + (1+\epsilon)^2 \frac{\ln T}{d(\mu_i, \mu_1)} + O(\frac{1}{\epsilon^2}).
\end{aligned}
$$

The order notation above hides dependence on μ_is and Δ_is. This gives expected regret bound

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

[1] This way of choosing thresholds, in order to obtain bounds in terms of KL-divergences $d(\mu_i, \mu_1)$ rather than Δ_is, is inspired by [9, 17, 14].

where $\epsilon' = 3\epsilon$, and the order notation in above hides μ_is and Δ_is in addition to the absolute constants.

2.2 Proof of Theorem 2

The proof of $O(\sqrt{NT\ln T})$ problem-independent bound of Theorem 2 is basically the same as the proof of Theorem 1, except for the choice of x_i and y_i. Here, we pick $x_i = \mu_i + \frac{\Delta_i}{3}, y_i = \mu_1 - \frac{\Delta_i}{3}$, so that $\Delta'^2 = (\mu_1 - y_i)^2 = \frac{\Delta_i^2}{9}$, 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, $L_i(T) = \frac{\ln T}{d(x_i, y_i)} \leq \frac{9 \ln T}{2\Delta_i^2}$, and $\frac{1}{d(x_i,\mu_i)} \leq \frac{9}{2\Delta_i^2}$. Then, as in previous subsection, substituting in Equation (5), and Equation (6), we get,

$$\begin{aligned}
& \mathbb{E}[k_i(T)] \\
\leq \;& \frac{24}{\Delta_i'^2} + \sum_{j=0}^{T-1} \Theta\left(e^{-\Delta_i'^2 j/2} + \frac{1}{(j+1)\Delta_i'^2} e^{-D_i j} + \right. \\
& \left. \frac{1}{e^{\Delta_i'^2 j/4} - 1}\right) + L_i(T) + 1 + \frac{1}{d(x_i, \mu_i)} + 1 \\
\leq \;& \sum_{j=0}^{T-1} \Theta\left(e^{-\Delta_i'^2 j/2} + \frac{1}{(j+1)\Delta_i'^2} + \frac{4}{\Delta_i'^2 j}\right) \\
& +O\left(\frac{\ln T}{\Delta_i^2}\right) \\
= \;& \Theta\left(\frac{1}{\Delta_i'^2} + \frac{\ln T}{\Delta_i'^2}\right) + O\left(\frac{\ln T}{\Delta_i^2}\right) = O\left(\frac{\ln T}{\Delta_i^2}\right).
\end{aligned}$$

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 $O(\sqrt{NT\ln T})$. □

2.3 Proof of Theorem 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 $x_i = \mu_i + \frac{\Delta_i}{3}$, $y_i = \mu_1 - \frac{\Delta_i}{3}$, $L_i(T) = \frac{2\ln(T\Delta_i^2)}{(y_i - x_i)^2} = \frac{18\ln(T\Delta_i^2)}{\Delta_i^2}$. Lemma 1 is indepndent of the type of priors used, and the proof of Lemma 3 can be easily adapted to Gaussian priors. So, both these lemmas hold as it is for this case. Corresponding to Lemma 4, and Lemma 2, we prove the following for the Gaussian distribution case.

Lemma 5.

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

Proof. The proof of this lemma follows from the concentration of the Gaussian distribution (Fact 4); see Appendix C.1. □

Lemma 6. *Let τ_j denote the time of the j^{th} play of the first arm. Then*

$$\mathbb{E}\left[\frac{1}{p_{i,\tau_j+1}} - 1\right] \leq \begin{cases} e^{11} + 4 & j \leq L_i(T) \\ \frac{1}{T\Delta_i^2} & j > L_i(T) \end{cases}$$

Proof. See Appendix C.2. □

Now, substituting in Equation (5), and using Equation (6),

$$\begin{aligned}
& \mathbb{E}[k_i(T)] \\
\leq \;& \sum_{k=0}^{T-1} \mathbb{E}\left[\frac{1}{p_{i,\tau_k+1}} - 1\right] + L_i(T) + \frac{1}{\Delta_i^2} \\
& + \frac{1}{d(x_i, \mu_i)} + 1 \\
\leq \;& (e^{11} + 5) + \frac{1}{\Delta_i^2} + \frac{18\ln(T\Delta_i^2)}{\Delta_i} + \frac{1}{\Delta_i^2} + \frac{9}{2\Delta_i^2}.
\end{aligned}$$

This gives a bound on expected regret due to arm i as

$$\Delta_i \mathbb{E}[k_i(T)] \leq (e^{11} + 5) + \frac{13}{2\Delta_i} + \frac{18\ln(T\Delta_i^2)}{\Delta_i}$$

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

$$\begin{aligned}
& (e^{11} + 5) + 18\ln(N\ln N)\sqrt{\frac{T}{N\ln N}} + 39\sqrt{\frac{T}{N\ln N}} \\
& \leq (e^{11} + 5) + 75\sqrt{\frac{T\ln N}{N}}.
\end{aligned}$$

For arms with $\Delta_i \leq e\sqrt{\frac{N\ln N}{T}}$, 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 $T \geq N$.

3 Proof of the lower lound

In this section, we prove Theorem 4. We construct a problem instance such that the TS algorithm has regret of $\Omega(\sqrt{NT\ln N})$ in time T. Let each arm i when played produces 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 $\mu_2 = \cdots = \mu_N = 0$.

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 until time t, which consists of $k_i(t), \hat{\mu}_i(t), i = 1, \ldots, N$. We say that $\mathcal{F}_{t-1} \in A_{t-1}$ if $\sum_{i\neq 1} k_i(t) \leq \frac{c\sqrt{NT\ln N}}{\Delta}$ for a fixed constant c (to be specified later), i.e. A_{t-1} is the set of histories which satisfy the given condition.

Note that if $\mathcal{F}_{t-1} \notin A_{t-1}$ then the regret until time t is at least $\sqrt{NT\ln N}$. Using this observation we show that for any $t \leq T$, we can assume that $\Pr(\mathcal{F}_{t-1} \in A_{t-1}) \geq \frac{1}{2}$. This is because otherwise the expected regret until time t

$$\begin{aligned}\mathbb{E}[\mathcal{R}(t)] &\geq \mathbb{E}[\mathcal{R}(t)|\mathcal{F}_{t-1} \notin A_{t-1}] \cdot \frac{1}{2} \\ &\geq \frac{1}{2}c\sqrt{NT\ln N} = \Omega(\sqrt{NT\ln N}),\end{aligned}$$

which would mean $\mathbb{E}[\mathcal{R}(T)] \geq \mathbb{E}[\mathcal{R}(t)] = \Omega(\sqrt{NT\ln N})$.

Now, given any history $\mathcal{F}_{t-1}$, $\theta_1(t)$ 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 symmetry of Gaussian distribution,

$$\Pr\left(\theta_1(t) \leq \mu_1 \mid \mathcal{F}_{t-1} \in A_{t-1}\right) \geq \frac{1}{2}. \qquad (7)$$

Also, given any $\mathcal{F}_{t-1}$, $\theta_i(t)$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,

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

Given $\sum_{i\neq 1} k_i(t) \leq \frac{c\sqrt{NT\log N}}{\Delta}$, 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

$$\begin{aligned}&\Pr\left(\exists i, \theta_i(t) > \mu_1 \mid \mathcal{F}_{t-1} \in A_{t-1}\right) \\ \geq\ &\Pr\left(\exists i, \theta_i(t) > \mu_1 \;\middle|\; k_i(t), \forall i, \sum_{i\neq 1} k_i(t) \leq \frac{\sqrt{NT\log N}}{\Delta}\right) \\ \geq\ &\left(1 - \prod_i \left(1 - e^{-\ln N}\right)\right) \\ =\ &1 - \left(1 - \frac{1}{N}\right)^{N-1}\end{aligned}$$

To summarize, for any t,

$$\begin{aligned}&\Pr\left(\exists i \neq 1, i(t) = i\right) \\ \geq\ &\Pr\left(\theta_1(t) \leq \mu_1, \exists i, \theta_i(t) > \mu_1\right) \\ \geq\ &\Pr\left(\theta_1(t) \leq \mu_1, \exists i, \theta_i(t) > \mu_1 \mid \mathcal{F}_{t-1} \in A_{t-1}\right) \\ &\qquad \cdot \Pr\left(\mathcal{F}_{t-1} \in A_{t-1}\right) \\ =\ &\Pr\left(\theta_1(t) \leq \mu_1 \mid \mathcal{F}_{t-1} \in A_{t-1}\right) \\ &\cdot \Pr\left(\exists i, \theta_i(t) > \mu_1 \mid \mathcal{F}_{t-1} \in A_{t-1}\right) \cdot \Pr\left(\mathcal{F}_{t-1} \in A_{t-1}\right) \\ \geq\ &\frac{1}{2} \cdot \left(1 - \left(1 - \frac{1}{N}\right)^{N-1}\right) \cdot \frac{1}{2} \\ \geq\ &p\end{aligned}$$

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

Acknowledgement. We thank Emil Jeřábek for telling us about his estimates of partial binomial sums. We also thank MathOverflow for connecting us with Emil.

References

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

[2] S. Agrawal and N. Goyal. Analysis of Thompson Sampling for the Multi-armed Bandit Problem. In *COLT*, 2012.

[3] S. Agrawal and N. Goyal. Thompson sampling for contextual bandits with linear payoffs. *Manuscript*, 2012.

[4] J.-Y. Audibert and S. Bubeck. Minimax Policies for Adversarial and Stochastic Bandits. In *COLT*, 2009.

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

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

[7] O. Chapelle and L. Li. An Empirical Evaluation of Thompson Sampling. In *NIPS*, pages 2249–2257, 2011.

[8] O. Chapelle and L. Li. Open Problem: Regret Bounds for Thompson Sampling. In *COLT*, 2012.

[9] A. Garivier and O. Cappé. The KL-UCB Algorithm for Bounded Stochastic Bandits and Beyond. In *Conference on Learning Theory (COLT)*, 2011.

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

[11] 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 *ICML*, pages 13–20, 2010.

[12] 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.

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

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

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

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

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

[18] B. C. May, N. Korda, A. Lee, and D. S. Leslie. Optimistic Bayesian sampling in contextual-bandit problems. Technical report, Statistics Group, Department of Mathematics, University of Bristol.

[19] B. C. May and D. S. Leslie. Simulation studies in optimistic Bayesian sampling in contextual-bandit problems. Technical report, Statistics Group, Department of Mathematics, University of Bristol.

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

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

[22] M. J. A. Strens. A Bayesian Framework for Reinforcement Learning. In *ICML*, pages 943–950, 2000.

[23] 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.

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

A Some results used in the proofs

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* $\mathbb{E}\left[X_t \mid X_1, ..., X_{t-1}\right] = \mu$. *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 [1] can be used to derive the following concentration for Gaussian distributed random variables.

Fact 4. *[1] 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

B.1 Proof of Lemma 3

Let τ_k denote the time at which k^{th} trial of arm i happens. Let $\tau_0 = 0$; Then,

$$\begin{aligned}
& \sum_{t=1}^{T} \Pr(i(t) = i, \overline{E^{\mu}_i(t)}) \\
\leq \quad & \mathbb{E}\left[\sum_{k=1}^{T} \sum_{t=\tau_k+1}^{\tau_{s+1}} I(i(t) = i) I(\overline{E^{\mu}_i(t)})\right] \\
= \quad & \mathbb{E}\left[\sum_{k=0}^{T-1} I(\overline{E^{\mu}_i(\tau_k + 1)}) \sum_{t=\tau_k+1}^{\tau_{k+1}} I(i(t) = i)\right] \\
= \quad & \mathbb{E}\left[\sum_{k=0}^{T-1} I(\overline{E^{\mu}_i(\tau_k + 1)})\right] \\
\leq \quad & \mathbb{E}\left[\sum_{k=0}^{T-1} I(\overline{E^{\mu}_i(\tau_k + 1)})\right] \\
\leq \quad & 1 + \mathbb{E}\left[\sum_{k=1}^{T-1} I(\overline{E^{\mu}_i(\tau_k + 1)})\right] \\
\leq \quad & 1 + \sum_{k=1}^{T-1} \exp(-kd(x_i, \mu_i)) \\
\leq \quad & 1 + \frac{1}{d(x_i, \mu_i)}
\end{aligned}$$

The second last inequality follows from the observation that the event $\overline{E^{\mu}_i(t)}$ was defined as $\hat{\mu}_i(t) > x_i$. At time $\tau_k + 1$ 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 latter is simply the average of the outcomes observed from k i.i.d. plays of arm i, each of which is a Bernoulli trial with mean μ_i. Using Chernoff-Hoeffding bounds (Fact 1), we obtain that $\Pr(\hat{\mu}_i(\tau_k + 1) > x_i) \leq \Pr(\frac{S_i(\tau_k+1)}{k} > x_i) \leq e^{-kd(x_i,\mu_i)}$. □

B.2 Proof of Lemma 4

$$\begin{aligned}
& \Pr\left(i(t) = i, \overline{E^{\theta}_i(t)} \,\middle|\, E^{\mu}_i(t), \mathcal{F}_{t-1}\right) \\
\leq \quad & \Pr\left(\theta_i(t) > y_i \mid \hat{\mu}_i(t) \leq x_i, \mathcal{F}_{t-1}\right) \\
= \quad & \Pr\left(Beta(\hat{\mu}_i(t)(k_i(t) + 1) + 1, (1 - \hat{\mu}_i(t))(k_i(t) + 1)) > y_i \right. \\
& \qquad\qquad \left. \mid \hat{\mu}_i(t) \leq x_i\right) \\
\leq \quad & \Pr\left(Beta(x_i(k_i(t) + 1) + 1, (1 - x_i)(k_i(t) + 1)) > y_i\right) \\
= \quad & F^{B}_{k_i(t)+1,y_i}(x_i(k_i(t) + 1)) \\
\leq \quad & e^{-(k_i(t)+1)d(x_i,y_i)},
\end{aligned}$$

where the second last equality follows from (Fact 3), and the last inequality follows from Chernoff-Hoeffding bounds (refer to Fact 1). In above, we slightly abused the notation for readability – the notation $\Pr(Beta(\alpha, \beta) > y_i)$ represented the probability that a random variable distributed as $Beta(\alpha, \beta)$ takes a value greater than y_i.

This implies that for t such that $k_i(t) > L_i(T)$,

$$\Pr\left(i(t) = i, \overline{E^{\theta}_i(t)} \,\middle|\, E^{\mu}_i(t), \mathcal{F}_{t-1}\right) \leq \frac{1}{T}.$$

Let τ be the largest time step until $k_i(t) \leq L_i(T)$, then,

$$
\begin{aligned}
& \sum_{t=1}^{T} \Pr\left(i(t)=i, \overline{E_i^{\theta}(t)}, E_i^{\mu}(t)\right) \\
\leq\ & \mathbb{E}\left[\sum_{t=1}^{\tau} \Pr\left(i(t)=i, \overline{E_i^{\theta}(t)}, E_i^{\mu}(t)\right)\right. \\
& \qquad \left. + \sum_{t=\tau+1}^{T} \Pr\left(i(t)=i, \overline{E_i^{\theta}(t)}, E_i^{\mu}(t)\right)\right] \\
\leq\ & \mathbb{E}\left[\sum_{t=1}^{\tau} \Pr\left(i(t)=i\right)\right. \\
& \qquad \left. + \sum_{t=\tau+1}^{T} \Pr\left(i(t)=i, \overline{E_i^{\theta}(t)} \,\middle|\, E_i^{\mu}(t)\right)\right] \\
\leq\ & \mathbb{E}\left[\sum_{t=1}^{\tau} \Pr\left(i(t)=i\right) + \sum_{t=\tau+1}^{T} \frac{1}{T}\right] \\
\leq\ & \mathbb{E}\left[\sum_{t=1}^{\tau} I(i(t)=i)\right] + 1 \\
\leq\ & L_i(T) + 1.
\end{aligned}
$$

□

B.3 Proof of Lemma 2

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

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

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

For $j < \frac{8}{\Delta'}$: Let $R = \frac{\mu_1(1-y)}{y(1-\mu_1)}$, $D = y \log \frac{y}{\mu_1} + (1-y)\log\frac{1-y}{1-\mu_1}$.

$$
\begin{aligned}
& \sum_{s=0}^{j} \frac{f_{j,\mu_1}(s)}{F_{j+1,y}(s)} \\
\leq\ & \frac{1}{1-y} \sum_{s=0}^{j} \frac{f_{j,\mu_1}(s)}{F_{j,y}(s)} \\
\leq\ & \frac{1}{1-y} \sum_{s=0}^{\lfloor yj \rfloor} \frac{f_{j,\mu_1}(s)}{f_{j,y}(s)} + \frac{1}{1-y} \sum_{s=\lceil yj \rceil}^{j} 2 f_{j,\mu_1}(s) \\
=\ & \frac{1}{1-y} \sum_{s=0}^{\lfloor yj \rfloor} R^s \frac{(1-\mu_1)^j}{(1-y)^j} + \frac{1}{1-y} \sum_{s=\lceil yj \rceil}^{j} 2 f_{j,\mu_1}(s) \\
=\ & \frac{1}{1-y} \left(\frac{R^{\lfloor yj \rfloor + 1} - 1}{R-1}\right) \frac{(1-\mu_1)^j}{(1-y)^j} + \frac{1}{1-y} \sum_{s=\lceil yj \rceil}^{j} 2 f_{j,\mu_1}(s) && (8) \\
\leq\ & \frac{1}{1-y} \left(\frac{R}{R-1}\right) R^{yj} \frac{(1-\mu_1)^j}{(1-y)^j} + \frac{2}{\Delta'} && (9) \\
=\ & \frac{\mu_1}{\Delta'} e^{-Dj} + \frac{2}{\Delta'} \\
\leq\ & \frac{3}{\Delta'}. && (10)
\end{aligned}
$$

For $j \geq \frac{8}{\Delta'}$: 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

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

Together, the above estimates will prove the required bound.

We use the following bounds on the cdf of Binomial distribution [13, Prop. A.4].
For $s \leq y(j+1) - \sqrt{(j+1)y(1-y)}$,

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

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

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

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

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

This gives

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

We now bound the first expression on the RHS.

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

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

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

Substituting in (11)

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

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

$$
\begin{aligned}
Sum(\lfloor yj \rfloor, \lfloor yj \rfloor) \quad & = \quad \frac{f_{j,\mu_1}(\lfloor yj \rfloor)}{F_{j+1,y}(\lfloor yj \rfloor)} \\
& \leq \quad \left(1 - \frac{yj - 1}{j+1}\right) R^{yj} \frac{(1-\mu_1)^j}{(1-y)^{j+1}} \\
& \leq \quad \frac{(1 - y + \frac{2}{j+1})}{1-y} R^{yj} \frac{(1-\mu_1)^j}{(1-y)^j} \\
& \leq \quad 3e^{-Dj}. \qquad (11)
\end{aligned}
$$

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, for $s \geq \lceil yj \rceil$, $F_{j+1,y}(s) = \Theta(1)$. Using this observation, we derive the following.

$$
\begin{aligned}
Sum(\lceil yj \rceil, \lfloor \mu_1 j - \frac{\Delta'}{2} j \rfloor) \quad & = \quad \sum_{s=\lceil yj \rceil}^{\lfloor \mu_1 j - \frac{\Delta'}{2} j \rfloor} \frac{f_{j,\mu_1}(s)}{F_{j+1,y}(s)} \\
& = \quad \Theta\left(\sum_{s=\lceil yj \rceil}^{\lfloor \mu_1 j - \frac{\Delta'}{2} j \rfloor} f_{j,\mu_1}(s)\right) \\
& \leq \quad \Theta(e^{-2\left(\mu_1 j - \lfloor \mu_1 j - \frac{\Delta'}{2} j \rfloor\right)^2 / j}) \\
& = \quad \Theta(e^{-\Delta'^2 j / 2}), \qquad (12)
\end{aligned}
$$

where the inequality follows using Chernoff-Hoeffding bounds (refer to 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 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 $j \geq \frac{8}{\Delta'}$.

$$\begin{aligned} Sum(\lceil \mu_1 j - \frac{\Delta'}{2} j\rceil, j) &= \sum_{s=\lceil \mu_1 j - \frac{\Delta'}{2} j\rceil}^{j} \frac{f_{j,\mu_1}(s)}{F_{j+1,y}(s)} \\ &\leq \frac{1}{1 - e^{-\Delta'^2 j/4}} \\ &= 1 + \frac{1}{e^{\Delta'^2 j/4} - 1}. \quad (13) \end{aligned}$$

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

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

□

C Thompson Sampling with Gaussian Distribution

C.1 Proof of Lemma 5

Proof.

$$\begin{aligned} &\Pr\left(i(t) = i, \overline{E_i^\theta(t)} \,\middle|\, E_i^\mu(t), \mathcal{F}_{t-1}\right) \\ \leq\ & \Pr\left(\theta_i(t) > y_i \mid \hat{\mu}_i(t) \leq x_i, \mathcal{F}_{t-1}\right) \\ =\ & \Pr\left(\mathcal{N}(\hat{\mu}_i(t), \frac{1}{k_i(t)+1}) > y_i \,\middle|\, \hat{\mu}_i(t) \leq x_i\right) \\ \leq\ & \Pr\left(\mathcal{N}(x_i, \frac{1}{k_i(t)+1}) > y_i\right) \\ \leq\ & \frac{1}{2} e^{-\frac{(k_i(t)+1)(y_i - x_i)^2}{2}}, \end{aligned}$$

where the last inequality follows from the concentration of Gaussian distribution (refer to Fact 4). Therefore, for t such that $k_i(t) > L_i(T) = \frac{2\ln(T\Delta_i^2)}{(y_i - x_i)^2}$,

$$\Pr\left(i(t) = i, \overline{E_i^\theta(t)} \,\middle|\, E_i^\mu(t), \mathcal{F}_{t-1}\right) \leq \frac{1}{T\Delta_i^2}.$$

Let τ be the largest time step until $k_i(t) \leq L_i(T)$, then,

$$\begin{aligned} &\sum_{t=1}^{T} \Pr\left(i(t) = i, \overline{E_i^\theta(t)}, E_i^\mu(t)\right) \\ \leq\ & \mathbb{E}\left[\sum_{t=1}^{\tau} \Pr\left(i(t) = i, \overline{E_i^\theta(t)}, E_i^\mu(t)\right) \right. \\ & \left. + \sum_{t=\tau+1}^{T} \Pr\left(i(t) = i, \overline{E_i^\theta(t)}, E_i^\mu(t)\right)\right] \\ \leq\ & \mathbb{E}\left\{\sum_{t=1}^{\tau} \Pr\left(i(t) = i\right) \right. \\ & \left. + \sum_{t=\tau+1}^{T} \Pr\left(i(t) = i, \overline{E_i^\theta(t)} \,\middle|\, E_i^\mu(t)\right)\right\} \\ \leq\ & \mathbb{E}\left[\sum_{t=1}^{\tau} \Pr\left(i(t) = i\right) + \sum_{t=\tau+1}^{T} \frac{1}{T\Delta_i^2}\right] \\ \leq\ & \mathbb{E}\left[\sum_{t=1}^{\tau} I(i(t) = i)\right] + \frac{1}{\Delta_i^2} \\ \leq\ & L_i(T) + \frac{1}{\Delta_i^2}. \end{aligned}$$

□

C.2 Proof of Lemma 6

Let Θ_j denote a $\mathcal{N}(\hat{\mu}_1(\tau_j + 1), \frac{1}{k_1(\tau_j+1)})$ distributed Gaussian random variable. Let G_j be a geomertic random variable denoting the number of consecutive independent trials until $\Theta_j > y_i$. Then, observe that

$$\frac{1}{p_{i,\tau_j+1}} - 1 = \mathbb{E}[G_j] = \sum_{r=1}^{\infty} \Pr(G_j \geq r)$$

We will bound the expected value of G_j by a constant for all j. Consider any integer $r \geq 1$. Let $z = \sqrt{\ln r}$, let random variable MAX_r denote the maximum of r independent samples of Θ_j. We abbreviate $\hat{\mu}_1(\tau_j + 1)$ as $\hat{\mu}_1$ and $k_1(\tau_j + 1)$ as k_1 in the following. Then

$$\begin{aligned} \Pr(G_j < r) &\geq \Pr(MAX_r > y_i) \\ &\geq \Pr\left(MAX_r > \hat{\mu}_1 + \frac{z}{\sqrt{k_1}} \,\middle|\, \hat{\mu}_1 + \frac{z}{\sqrt{k_1}} \geq y_i\right) \\ &\qquad \cdot \Pr\left(\hat{\mu}_1 + \frac{z}{\sqrt{k_1}} \geq y_i\right) \end{aligned}$$

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 [1]

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

This gives

$$\begin{aligned}
& \Pr\left(MAX_r > \hat{\mu}_1 + \frac{z}{\sqrt{k_1}} \,\middle|\, \hat{\mu}_1 + \frac{z}{\sqrt{k_1}} \geq y_i \right) \\
\geq \quad & 1 - \left(1 - \frac{1}{\sqrt{2\pi}} \frac{z}{(z^2+1)} e^{-z^2/2} \right)^r \\
= \quad & 1 - \left(1 - \frac{1}{\sqrt{2\pi}} \frac{\sqrt{\ln r}}{(\ln r + 1)} \frac{1}{\sqrt{r}} \right)^r \\
\geq \quad & 1 - e^{-\frac{r}{\sqrt{4\pi r \ln r}}}.
\end{aligned}$$

Also, using Chernoff-Hoeffding bounds (refer to Fact 2),

$$\Pr(\hat{\mu}_1 + \frac{z}{\sqrt{k_1}} \geq \mu_1) \geq 1 - e^{-2z^2} = 1 - \frac{1}{r^2}.$$

Therefore, substituting,

$$\begin{aligned}
\Pr(G_j < r) \quad & \geq \quad (1 - e^{-\sqrt{\frac{r}{4\pi \ln r}}}) \cdot (1 - \frac{1}{r^2}) \\
& \geq \quad 1 - \frac{1}{r^2} - e^{-\sqrt{\frac{r}{4\pi \ln r}}}.
\end{aligned}$$

$$\begin{aligned}
E[G_j] \quad & = \quad \sum_{r \geq 1} (1 - \Pr(G_j < r)) \\
& \leq \quad \sum_{r \geq 1} \frac{1}{r^2} + e^{-\sqrt{\frac{r}{2\pi \ln r}}} \\
& \leq \quad e^{11} + \sum_r 2 \frac{1}{r^2} \\
& \leq \quad e^{11} + 4,
\end{aligned}$$

The second last inequality in above uses the fact that for $r \geq e^{11}$, $e^{-\sqrt{\frac{r}{2\pi \ln r}}} \leq \frac{1}{r^2}$.