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 general lower bound of $O(\sqrt{NT})$ for the multi-armed bandit problem. 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. [2012b].

Appearing in Proceedings of the 16^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2013, Scottsdale, AZ, USA. Volume 31 of JMLR: W&CP 31.

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., 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 Bayesian setting with known priors and geometric time-discounted rewards. In one of the earliest works on stochastic bandit problems, 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 very natural algorithm and 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 [2012a] and Kaufmann et al. [2012b]. In Agrawal and Goyal [2012a], the first logarithmic bound on expected regret of TS was proven. Kaufmann et al. [2012b] provided a bound that matches the asymptotic lower bound of Lai and Robbins [1985] for this problem. However, both 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 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 Chapelle and Li [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 (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, 2012b]. While one of the basic ideas for the analysis in the contextual bandits setting of Agrawal and Goyal [2012b] is similar to an idea in this paper, the details are substantially different.

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

The notation $P(\cdot)$ in above denotes probability density. TS maintains a posterior distribution for the underlying parameter 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 that 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 respectively.

We emphasize that the Beta priors and 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, i.e., the rewards are in the range $[0, 1]$. In description of Thompson Sampling using Beta priors and Bernoulli likelihood, we do start

with the description of the algorithm for 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 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 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 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 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 [2012a]. This extension, on observing a reward $r_t \in [0,1]$, tosses a coin with probability r_t, and uses the $\{0,1\}$ outcome to update the beta distribution in the same way as in the above algorithm. It is easy to show that any expected regret bounds produced for Algorithm 1 will also hold for this extension [Agrawal and Goyal, 2012a].

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 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 $\hat{\mu}_{i(t)} = \frac{(\hat{\mu}_{i(t)} k_{i(t)} + r_t)}{k_{i(t)}+2}$, $k_{i(t)} = k_{i(t)} + 1$.
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 Agrawal and Goyal [2012a].

Theorem 1. *For the N-armed stochastic bandit problem, TS 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.

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 [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 achieve the lower bounds of Lai and Robbins [1985], and match the upper bounds provided by Kaufmann et al. [2012b] for TS.

Theorem 2 and 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 multi-armed bandit problem within logarithmic factors. The problem-dependent bounds in the existing work on TS implied only suboptimal problem-independent bounds: The results of Agrawal and Goyal [2012a] 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 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 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 when T is larger than N. 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-inpdependent 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 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.

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 differences between the analysis for obtaining the logarithmic problem-dependent bound of Theorem 1, and the problem-independent bound of Theorem 2 and Theorem 3 are technical, and occur only towards the end of the proof.

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

Definition 1 ($F^B_{n,p}, f^B_{n,p}, F^{beta}_{\alpha,\beta}$)**.** *$F^B_{n,p}(\cdot)$ 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 2 (**Quantities** $k_i(t), i(t), S_i(t), \hat{\mu}_i(t)$)**.** *Let $i(t)$ denote the arm played at time t. $k_i(t)$ denotes the number of plays of arm i until time $t-1$. $S_i(t)$ is the number of successes among the plays of arm i until time $t-1$ for the Bernoulli bandit case. And, empirical*

mean $\hat{\mu}_i(t)$ is defined as $\hat{\mu}_i(t) = \frac{\sum_{\tau=1:i(\tau)=i}^{t-1} r_i(\tau)}{(k_i(t)+1)}$, where $r_i(t)$ is the reward for arm i at time t. (note that $\hat{\mu}_i(t) = 0$ when $k_i(t) = 0$).

Definition 3 (Quantities x_i, y_i). *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.*

Definition 4 (Events $E_i^\mu(t)$ and $E_i^\theta(t)$). *Define $E_i^\mu(t)$ as the event that $\hat{\mu}_i(t) \leq x_i$. Define $E_i^\theta(t)$ as the event that $\theta_i(t) \leq y_i$.*

Intuitively, $E_i^\mu(t)$, $E_i^\theta(t)$ 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.

Definition 5 (Filtration $\mathcal{F}_{t-1}$). *Define filtration $\mathcal{F}_{t-1}$ as the history of plays until time $t-1$, i.e.*

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

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

By definition, $\mathcal{F}_1 \subseteq \mathcal{F}_2 \cdots \subseteq \mathcal{F}_{T-1}$. Also by definition, for every arm i, the quantities $S_i(t)$, $k_i(t)$, $\hat{\mu}_i(t)$, the distribution of $\theta_i(t)$, and whether $E_i^\mu(t)$ is true or not, are determined by the history of plays until time $t-1$ and therefore are included in $\mathcal{F}_{t-1}$.

Definition 6. *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 $i \neq 1$,*

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

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

Proof. Note that whether $E_i^\mu(t)$ is true or not is included in $\mathcal{F}_{t-1}$ (refer to Defintion 5). Assume that filtration $\mathcal{F}_{t-1}$ 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

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

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

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

The first equality holds because because given $\mathcal{F}_{t-1}$ (and hence $S_i(t), k_j(t), \hat{\mu}_j(t)$ and the distributions of $\theta_j(t)$F 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,

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

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

2.1 Proof of Theorem 1

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

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

Let τ_k denote the time step at which arm 1 is played for the k^{th} time for $k \geq 1$, and let $\tau_0 = 0$. Note that for any i, for $k > k_i(T)$, $\tau_k > T$. Also, $\tau_T \geq T$. Then, using Lemma 1, we can bound the first term above as:

$$\begin{aligned} & \textstyle\sum_{t=1}^T \Pr(i(t) = i, E_i^\mu(t), E_i^\theta(t)) \\ =\ & \textstyle\sum_{t=1}^T \mathbb{E}\left[\Pr\left(i(t) = i, E_i^\mu(t), E_i^\theta(t) \,\middle|\, \mathcal{F}_{t-1}\right)\right] \\ \leq\ & \textstyle\sum_{t=1}^T \mathbb{E}\left[\frac{(1-p_{i,t})}{p_{i,t}} \Pr\left(i(t) = 1, E_i^\theta(t), E_i^\mu(t) \,\middle|\, \mathcal{F}_{t-1}\right)\right] \\ =\ & \textstyle\sum_{t=1}^T \mathbb{E}\left[\mathbb{E}\left[\frac{(1-p_{i,t})}{p_{i,t}} I(i(t) = 1, E_i^\theta(t), E_i^\mu(t)) \,\middle|\, \mathcal{F}_{t-1}\right]\right] \\ =\ & \textstyle\sum_{t=1}^T \mathbb{E}\left[\frac{(1-p_{i,t})}{p_{i,t}} I(i(t) = 1, E_i^\theta(t), E_i^\mu(t))\right] \\ \leq\ & \textstyle\sum_{k=0}^{T-1} \mathbb{E}\left[\frac{(1-p_{i,\tau_k+1})}{p_{i,\tau_k+1}} \sum_{t=\tau_k+1}^{\tau_{k+1}} I(i(t) = 1)\right] \\ =\ & \textstyle\sum_{k=0}^{T-1} \mathbb{E}\left[\frac{1}{p_{i,\tau_k+1}} - 1\right]. \quad (3) \end{aligned}$$

The second equality uses that $p_{i,t}$ is fixed given $\mathcal{F}_{t-1}$. The last inequality 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_k denote the time step at which k^{th} trial of first arm happens, then* [1]

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

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

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

Proof. The proof of this inequality uses careful numerical estimates; see Appendix B.3. □

For the remaining two terms in Equation (2), we prove the following lemmas.

Lemma 3.

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

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

Lemma 4.

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

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

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 $L_i(T)$; see Appendix B.2. □

For obtaining the problem-dependent bound of Theorem 1, for any $0 < \epsilon \leq 1$, 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 $d(x_i,y_i) = d(x_i,\mu_1)/(1+\epsilon) = d(\mu_i,\mu_1)/(1+\epsilon)^2$ ([2]). 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 μ_is and Δ_is. Substituting in Equation (2), and Equation (3), we get,

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

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

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

where $\epsilon' = 3\epsilon$, and the 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 (2), and Equation (3), we get,

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

[1] For any two functions $f(n), g(n)$, $f(n) = \Theta(g(n))$ if there exist constants b, c and n_0 such that for all $n \geq n_0$, $bg(n) \leq f(n) \leq cg(n)$.

[2] 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 Garivier and Cappé [2011], Maillard et al. [2011], Kaufmann et al. [2012a].

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 independent 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.

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

where $L_i(T) = \frac{18 \ln(T\Delta_i^2)}{\Delta_i^2}$.

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*

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

where $L_i(T) = \frac{18 \ln(T\Delta_i^2)}{\Delta_i^2}$.

Proof. See Appendix C.2. □

Now, substituting in Equation (2), and using Equation (3),

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

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

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

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 TS 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 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 or not is included in $\mathcal{F}_{t-1}$.

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,

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

We will show that given any filtration $\mathcal{F}_{t-1}$ and the event A_{t-1}, 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 filtration $\mathcal{F}_{t-1}$ 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 symmetry of Gaussian distribution,

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

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,

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

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

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

To summarize, for any t, the probability of playing a suboptimal arm at time t,

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

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

Conclusions. In this paper, we proved a 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, handle different prior distributions. (It can also be used for contextual bandits as described in Agrawal and Goyal [2012b] though that requires more work and hence is treated separately.)

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

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 *COLT*, 2012a.

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

J.-Y. Audibert and S. Bubeck. Minimax Policies for Adversarial and Stochastic Bandits. In *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.

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

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

A. Garivier and O. Cappé. The KL-UCB Algorithm for Bounded Stochastic Bandits and Beyond. In *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 *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 (AISTAT)*, 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. L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6:4–22, 1985.

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.

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.

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 *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.