Thompson Sampling for Contextual Bandits with Linear Payoffs

Shipra Agrawal
shipra@microsoft.com
Microsoft Research India

Navin Goyal
navingo@microsoft.com
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 better empirical performance compared to the state of the art methods. However, many questions regarding its theoretical performance remained open. In this paper, we design and analyze Thompson Sampling algorithm for the contextual multi-armed bandit problem with linear payoff functions, when the contexts are provided by an adaptive adversary. This is perhaps the most important and widely studied version of the contextual bandits problem. We prove a high probability regret bound of $\tilde{O}(\frac{1}{\sqrt{\epsilon}}\sqrt{T^{1+\epsilon}}d)$ in time T for any $0 < \epsilon < 1$, where d is the dimension of each context vector and ϵ is a parameter used by the algorithm. Our results provide the first theoretical guarantees for the contextual version of Thompson Sampling, and are close to the lower bound of $\Omega(\sqrt{Td})$ for this problem. This essentially solves the COLT open problem of Chapelle and Li [COLT 2012] regarding regret bounds for Thompson Sampling for contextual bandits problem.

Our version of Thompson sampling uses Gaussian prior and Gaussian likelihood function. Our novel martingale-based analysis techniques also allow easy extensions to the use of more general distributions, satisfying certain general conditions.

1 Introduction

Multi-armed bandit (MAB) problems model the exploration/exploitation trade-off inherent in many sequential decision problems. There are many versions of multi-armed bandit problems; a particularly useful version is the contextual multi-armed bandit problem. In this problem, on each of T rounds, a learner is presented with the choice of taking one of N actions, referred to as N arms. Before making the choice of which arm to play, the learner sees a d-dimensional feature vector b_i, referred to as "context", associated with each arm i. The learner uses these feature vectors along with the feature vectors and rewards of arms played by her in the past to make the choice of arm. Over time, the learner's aim is gather enough information about how the feature vectors and rewards are related to each other, so that she can predict, with some certainty, which arm will give the best reward by looking at the feature vectors. The learner competes with a class of predictors, in which each predictor takes in the feature vectors and predicts which arm will give the best reward. If the learner can guarantee to do nearly as well as the predictions of the best predictor in hindsight (to have low regret), the learner is said to successfully compete with that class.

In the contextual bandits setting with *linear payoff functions*, the learner competes with the class of all "linear" predictors on the feature vectors. That is, a predictor is defined by N d-dimensional parameters $\{\tilde{\mu}_i\}_{i=1}^N$, and the predictor ranks the arms according to $b_i^T\tilde{\mu}_i$. We consider the contextual bandit problem under the linear realizability assumption, that is, we assume that there are unknown underlying parameters $\{\mu_i\}_{i=1}^N$ such that the expected reward for each arm i, given context b_i, is $b_i^T\mu_i$. Under this realizability assumption, the linear predictor corresponding to $\{\mu_i\}_{i=1}^N$ is in fact the best predictor and the learner's aim is to learn these underlying parameters. This realizability assumption is standard in the existing literature on contextual multi-armed bandits [4, 11, 9, 1].

In this paper, we analyze Thompson Sampling (TS) algorithm for the contextual bandits problem with linear payoffs. Thompson Sampling is one the earliest heuristics for the multi-armed bandit problems. The first version of this Bayesian heuristic is around 80 years old, dating to Thompson (1933) [25]. Since then, it got rediscovered numerous times independently in the context of reinforcement learning, e.g., in [27, 20, 24]. It is a member of the family of *randomized probability matching* algorithms. 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. The general structure of Thompson sampling for the contextual bandits problem involves the following elements:

1. a set Θ of parameters $\tilde{\mu}$;
2. an assumed prior distribution $P(\tilde{\mu})$ on these parameters;
3. past observation $\mathcal{D}$ consisting of (context b, reward r) for the past time steps;
4. an assumed likelihood function $P(r|b,\tilde{\mu})$, 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. We emphasize that although TS algorithm is a Bayesian approach, the description of the algorithm and our analysis apply to the prior-free stochastic MAB model, and are directly comparable to the UCB family of algorithms which are a frequentist approach to the same problem. One could interpret the Bayesian priors used by the TS algorithm as a way of capturing the current knowledge about the arms.

Recently, TS has attracted considerable attention. Several studies (e.g., [13, 22, 12, 7, 19, 15]) have empirically demonstrated the efficacy of TS: Scott [22] provides a detailed discussion of probability matching techniques in many general settings along with favorable empirical comparisons with other techniques. Chapelle and Li [7] demonstrate that for the basic stochastic MAB problem, empirically TS achieves regret comparable to the lower bound of [16]; and in applications like display advertising and news article recommendation modeled by the contextual bandits problem, it is competitive to or better than the other methods such as UCB. In their experiments, TS is also more robust to delayed or batched feedback than the other methods. TS has also been used in an industrial scale application for CTR prediction of search ads on search engines [12]. Kaufmann et al. [15] do a thorough comparison of TS with the best known versions of UCB, and show that TS has the lowest regret in the long run.

Despite being easy to implement and being competitive to the state of the art methods, the theoretical understanding of TS algorithm is limited. [13, 18] provided weak guarantees, namely, a bound of $o(T)$ on expected regret in time T. More recently, some significant progress was made by [3, 15], who provided near-optimal problem-dependent bounds on the expected regret of TS for the basic (i.e. without contexts) version of the stochastic MAB problem. However, many questions regarding theoretical analysis of TS remained open, including near-optimal problem-independent regret bounds, high probability regret bounds, and regret bounds for the more general contextual bandits setting. Some of these questions were formally raised as a COLT 2012 open problem [8]. In this paper, we use novel and simple martingale-based analysis techniques to demonstrate that TS achieves high probability, near-optimal problem independent regret bounds for contextual bandits with linear payoffs. To our knowledge, ours are the first non-trivial regret bounds for TS for the contextual bandits problem. Additionally, our results are the first high probability regret bounds for TS, even in the case of basic MAB problem. This essentially solves the COLT 2012 open problem [8] for linear contextual bandits.

The contextual MAB problem does not seem easily amenable to the techniques used so far for analyzing the basic MAB problem by [3, 15]. In Section 2.3, we describe some of the challenges, and our martingale-based solution ideas to handle them. Our version of Thompson Sampling algorithm, described formally in Section 2.2, uses Gaussian prior and Gaussian likelihood functions. As we discuss towards the end of the paper in Section 4, our techniques are easily extensible to the use of other prior distributions, satisfying certain conditions.

1.1 Our Results

The formal problem statement appears in Sec. 2.1.

Theorem 1. *For the contextual bandit problem with linear payoffs, with probability* $1-\delta$*, the total regret in time* T *for Thompson Sampling is bounded by* $O\left(d\sqrt{\frac{NT^{1+\epsilon}}{\epsilon}\ln N}\ln T\ln\frac{1}{\delta}\right)$ *, for any* $0<\epsilon<1$*. Here,* ϵ *is a parameter used by the Thompson Sampling algorithm.*

Theorem 2. *When* $\mu_1=\mu_2\cdots=\mu_N=\mu$*, i.e. there is a single underlying* d*-dimensional parameter* μ*, then with probability* $1-\delta$*, the total regret in time* T *for Thompson Sampling is bounded by* $O\left(d\sqrt{\frac{T^{1+\epsilon}}{\epsilon}\ln N}\ln T\ln\frac{1}{\delta}\right)$.

Remark 1. *Here $0 < \epsilon < 1$ can be chosen to be any constant. If T is known, one could choose $\epsilon = \frac{1}{\ln T}$, to get $\tilde{O}(d\sqrt{NT})$ (and $\tilde{O}(d\sqrt{T})$) regret bound.*

Remark 2. *Note that Theorem 2 has only a logarithmic dependence on the number of arms N, which makes it particularly useful when the number of arms N is very large, but there is a single underlying d-dimensional parameter μ, with d being much smaller than N. One could also recover the setting with different μ_is from the setting with a single μ, by letting μ be an Nd-dimensional vector formed by appending all the μ_is, and letting context $b_i(t)$ be an Nd-dimensional vector which is 0 in all but d positions corresponding to arm i. However, a direct application of Theorem 2 would then give a slightly weaker bound of $\tilde{O}(\frac{1}{\sqrt{\epsilon}}Nd\sqrt{T^{1+\epsilon}})$ compared to Theorem 1.*

We will mainly describe the algorithm and regret analysis for the setting of single parameter $\mu = \mu_1 = \cdots \mu_N$, i.e. the proof of Theorem 2. The algorithm and analysis for the setting with different μ_is is similar. In Section 4, we describe the changes required for the latter setting in order to get the result of Theorem 1.

1.2 Related Work

The contextual bandit problem with linear payoffs is a widely studied problem in statistics and machine learning often under different names as mentioned by Chu et al. [9]: bandit problems with covariates [26, 21], associative reinforcement learning [14], associative bandit problems [4, 23], and bandit problems with expert advice [5]. The name contextual bandits was coined in Langford and Zhang [17].

Chu et al. [9] show that for any algorithm the regret is $\Omega(\sqrt{Td})$ for $d^2 \leq T$ for the N-armed contextual bandits problem with linear payoffs and single parameter. Auer [4] and Chu et al. [9] SupLinUCB, a complicated algorithm using UCB as a subroutine, for this problem. Chu et al. achieve a regret bound of $O(\sqrt{Td\ln^3(NT\ln(T)/\delta)})$ with probability at least $1-\delta$ (Auer [4] proves similar results). Let us compare these results with ours. Our bounds have a factor of d compared to $\sqrt{d}$ in the bounds just mentioned. As can be observed in our regret analysis, the extra $\sqrt{d}$ factor in our bounds appears because we will use a concentration inequality that provides only a concentration of $O(\sqrt{d\ln\frac{1}{\delta}})$ for the empirical estimate of the mean rewards around the actual mean. The advantage of using this (weaker though more generally applicable) concentration inequality is that it allows for statistical dependence between the samples used in the estimates of the mean rewards, which could be due to the dependence between the past rewards and the future choice of arms, or because the contexts are generated by an *adaptive adversary*. By contrast, in the analysis of SupLinUCB, [4, 9] consider only *oblivious adversary*, and achieve statistical independence of samples by using a complicated master procedure SupLin on top of the basic UCB style algorithm. This allows them to use a stronger concentration of $O(\sqrt{\ln\frac{1}{\delta}})$ given by the Azuma-Hoeffding inequality. We do not use any such master algorithm.

A closely related setting is that of *linear stochastic bandits problem*, e.g. [10, 1]. In linear stochastic bandits problem, every arm i is associated with a known fixed vector b_i, and the expected reward of the arm, when played, is $b_i^T\mu$ for some common unknown underlying parameter μ. Abbasi-Yadkori et al. [1] analyze a UCB-style algorithm for that problem. When adapted to our setting, their regret bound is $O(d\log(T)\sqrt{T} + \sqrt{dT\log(T/\delta)})$. Note that their regret bound does not depend on N, and thus can even be applied to infinite set of arms, for example, when the set of arms is specified as those corresponding to all vectors in a d-dimensional polytope. The lower bound for this setting was given by Dani et al. [10] as $\Omega(d\sqrt{T})$. The state-of-the-art bounds for linear bandits problem in case of finite N are given by [6]. They provide an algorithm based on exponential weights, with regret of order $\sqrt{dT \log N}$ for any finite set of N actions. However, their setting is slightly different from ours. They consider a non-stochastic (adversarial) bandit setting where the reward at time t for arm i is $b_i^T \mu_t$ with μ_t chosen by an adversary. The set of arms and the associated b_i vectors are *non-adaptive* and fixed in advance.

While the results in this paper do not claim to provide regret bounds for Thompson Sampling algorithm that match or better the best available bounds of this extensively studied problem, our bounds for this natural and efficient heuristic are close to the best bounds. Our bounds are essentially within a factor of $\sqrt{d} \ln T$ of the best bounds for finite N (those for UCB1 by [9, 4], and for Exp2 algorithm by [6]), and within $\sqrt{\ln N}$ factor of the best bounds that do not depend on N (by Abbasi-Yadkori et al. [1]). The main contribution of this paper is to provide tools for the analysis of Thompson Sampling algorithm for contextual bandits, which despite of being popular and empirically attractive, has eluded theoretical analysis. While significant recent progress was made in analyzing it for basic MAB [3, 15], it was not clear how to extend that to contextual bandits problem, for which no regret bounds were available. There were considerable difficulties in extending the existing techniques to this case, some of which were also pointed out in [8]. We believe the techniques used in this paper will provide useful insights into the workings of this Bayesian algorithm, and may be useful for further improvements and extensions.

2 Problem setting and algorithm description

2.1 Problem setting

There are N arms. At time $t = 1, 2, \ldots$, a context vector $b_i(t) \in \mathbb{R}^d$, $||b_i(t)|| \le 1$, is revealed for every arm i. These context vectors are chosen by an adversary in an adaptive manner after observing the arms played and their rewards up to time $t-1$, i.e. history $\mathcal{H}_{t-1}$,

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

where $i(t)$ denotes the arm played at time t.

Given $b_i(t)$, reward for arm i at time t is generated from an (unknown) distribution with mean $b_i(t)^T \mu_i$, where $\mu_i \in \mathbb{R}^d$, $||\mu_i|| \le 1$ are fixed but unknown parameters. Also, given history $\mathcal{H}_{t-1}$, and $b_i(t), i = 1, \ldots, N$, reward for arms $i, i', i \neq i'$ are independent of each other.

$$\mathbb{E}\left[r_i(t) \,\middle|\, \{b_i(t)\}_{i=1}^N, \mathcal{H}_{t-1}\right] = \mathbb{E}\left[r_i(t) \,|\, b_i(t)\right] = b_i(t)^T \mu_i$$

Furthermore, we assume that $\eta_{i,t} = r_i(t) - b_i(t)^T \mu_i$ is conditionally R-sub-Gaussian for constant $R \ge 0$, i.e.,

$$\forall \lambda \in \mathbb{R}, \mathbb{E}[e^{\lambda \eta_{i,t}} | \{b_i(t)\}_{i=1}^N, \mathcal{H}_{t-1}] \le \exp\left(\frac{\lambda^2 R^2}{2}\right).$$

This assumption is satisfied if $r_i(t) \in [b_i(t)^T \mu_i - R, b_i(t)^T \mu_i + R]$ (refer to Remark 1 in Appendix A.1 of [11]). Note that this assumption is weaker than assuming $r_i(t)$ is bounded.

An algorithm for the *contextual bandit problem* needs to choose, at every time t, an arm $i(t)$ to play, using history $\mathcal{H}_{t-1}$ and current contexts $b_i(t), i = 1, \ldots, N$. Let $i^*(t)$ denote the optimal arm at time t, i.e. $i^*(t) = \arg\max_i b_i(t)^T \mu_i$. Then the regret at time t,

$$\text{regret}(t) = b_{i^*(t)}(t)^T \mu_{i^*(t)} - b_{i(t)}(t)^T \mu_{i(t)}.$$

The objective is to minimize the total regret $\mathcal{R}(T) = \sum_{t=1}^{T} \text{regret}(t)$ in time T. The time horizon T is finite but possibly unknown.

Remark 3. *An alternative definition of regret that appears in literature is*

$$\mathcal{R}(T) = \textstyle\sum_{t=1}^{T} r_{i^*(t)}(t) - r_{i(t)}(t).$$

When the reward $r_i(t)$ at all time steps is bounded as $|r_i(t)| \leq R$ for some constant R, for all i, then we can obtain the same results as in Theorem 1 and Theorem 2 for this definition of regret. The details are provided in Section 3.1.

2.2 Thompson Sampling algorithm

Here, we describe the algorithm for the setting of single parameter $\mu = \mu_1 = \cdots \mu_N$. (For the case of N different parameters, see Section 4.) Since there is a single underlying parameter, TS will maintain a common prior distribution over this parameter.

We use Gaussian likelihood function and Gaussian prior in our version of Thompson Sampling algorithm. More precisely, we assume that the **likelihood** of reward $r_i(t)$ at time t, given context $b_i(t)$ and parameter μ, is given by the pdf of Gaussian distribution $\mathcal{N}(b_i(t)^T\mu, v^2)$. Here, $v = R\sqrt{\frac{6}{\epsilon} d \ln(\frac{1}{\delta})}$, with $\epsilon \in (0,1)$ which parameterizes our algorithm. Let

$$B(t) = I_d + \textstyle\sum_{w=1}^{t-1} b_{i(w)}(w) b_{i(w)}(w)^T, \quad \hat{\mu}(t) = B(t)^{-1}\left(\sum_{w=1}^{t-1} b_{i(w)}(w) r_{i(w)}(w)\right).$$

Then, assuming that the **prior** for μ at time t is given by $\mathcal{N}(\hat{\mu}(t), v^2 B(t)^{-1})$, it easy to compute the **posterior** distribution

$$\Pr(\tilde{\mu} | r_i(t)) \propto \Pr(r_i(t) | b_i(t)^T \tilde{\mu}) \Pr(\tilde{\mu})$$

as $\mathcal{N}(\hat{\mu}(t+1), v^2 B(t+1)^{-1})$ (details of this computation are in Appendix A). Or, equivalently, the posterior distribution of the mean reward $b_i(t+1)^T\hat{\mu}(t+1)$ for arm i is $\mathcal{N}(b_i(t+1)^T\hat{\mu}(t+1), v^2 b_i(t+1)^T B(t+1)^{-1} b_i(t+1))$. In our Thompson Sampling algorithm, for each arm i, we will generate an independent sample $\theta_i(t)$ from the distribution $\mathcal{N}(b_i(t)^T\hat{\mu}(t), v^2 b_i(t)^T B(t)^{-1} b_i(t))$ at time t. The arm with maximum value of $\theta_i(t)$ will be played.

```
Algorithm 1: Thompson Sampling for Contextual bandits
  Set $B = I_d, \hat{\mu} = 0_d,\ f = 0_d$.
  foreach $t = 1, 2, \ldots,$ do
      For each arm $i = 1, \ldots, N$, sample $\theta_i(t)$ independently from distribution
      $\mathcal{N}(b_i(t)^T\hat{\mu}, v^2 b_i(t)^T B^{-1} b_i(t))$.
      Play arm $i(t) := \arg\max_i \theta_i(t)$ and observe reward $r_t$.
      Update $B = B + b_{i(t)}(t) b_{i(t)}(t)^T$, $f = f + b_{i(t)}(t) r_t$, $\hat{\mu} = B^{-1} f$.
  end
```

Remark 4. *Note that in the case of a single underlying parameter μ, one could alternatively first generate a single $\tilde{\mu}$ from distribution $\mathcal{N}(\hat{\mu}(t+1), v^2 B(t+1)^{-1})$, and then generate $\theta_i(t)$ as $b_i(t)^T\tilde{\mu}$. This alternative algorithm could be more efficient if N is large, but there is an efficient way to compute $\max_i b_i(t)^T\tilde{\mu}$. While in this alternative algorithm, the marginal distribution of each $\theta_i(t)$ remains $\mathcal{N}(b_i(t+1)^T\hat{\mu}(t+1), v^2 b_i(t+1)^T B(t+1)^{-1} b_i(t+1))$, $\theta_i(t)$s are not independent anymore. In our proof, we utilize the independence of $\theta_i(t)$s (in Lemma 2), and it is not clear to us at this point whether the algorithm with dependent θ_is will have the same regret.*

For the case of N different parameters, this distinction is not important, as a separate $\tilde{\mu}_i$ has to be generated for every i.

2.3 Challenges and solution outline

The contextual version of the multi-armed bandit problem presents new challenges for the analysis of TS algorithm, and the techniques used so far for analyzing the basic multi-armed bandit problem by [3, 15] do not seem directly applicable. Let us describe some of these difficulties and our novel solution ideas to resolve them.

In the basic MAB problem there are N arms, each with mean reward μ_i, and the regret for playing a suboptimal arm i is $\mu_{i^*} - \mu_i$, where i^* is the arm with highest mean. Let us compare this to a 1-dimensional contextual MAB problem, where each arm i is associated with a parameter μ_i, but in addition, at every time t, it is associated with a context $b_i(t)$, so that mean reward is $b_i(t)\mu_i$, the best arm $i^*(t)$ at time t is the arm with the highest mean at time t, and the regret for playing arm i is $b_{i^*(t)}(t)\mu_{i^*(t)} - b_i(t)\mu_i$.

In general, the basis of regret analysis for stochastic MAB is to prove that the variance of empirical estimates for all arms decreases fast enough, so that the regret incurred until the variance becomes small enough, is small. In the basic MAB, the variance of the empirical mean is inversely proportional to the number of plays $k_i(t)$ of arm i at time t. Thus, every time a suboptimal arm is played, we know that even though a regret of $\mu_{i^*} - \mu_i \leq 1$ in incurred, there is also an improvement of exactly 1 in the number of plays of that arm, and hence, corresponding decrease in the variance. The techniques for analyzing basic MAB rely on this observation to precisely quantify the exploration-exploitation tradeoff. On the other hand, the variance of empirical mean for contextual case is given by inverse of $B_i(t) = \sum_{u=1}^{t} b_u(t)^2$. When a suboptimal arm i is played, if $b_i(t)$ is small, the regret $b_{i^*(t)}(t)\mu_{i^*(t)} - b_i(t)\mu_i$ could be much higher than the improvement $b_i(t)$ in $B_i(t)$.

In our solution, we overcome this difficulty by bounding the expected regret at every step by a function of the probability of playing the optimal arm at that step. So, a high expected regret would mean large expected number of plays of optimal arm, in turn implying that regret is small. More precisely, we prove that, for "most histories" $\mathcal{F}_{t-1}$,

$$\frac{1}{(\sqrt{4\ln(NT)}\ v+\ell(T))}\mathbb{E}[\text{regret}(t)|\mathcal{F}_{t-1}] \lesssim \frac{1}{p}\Pr\left(i(t) = i^*(t) \mid \mathcal{F}_{t-1}\right) s_{t,i^*(t)} + s_{t,i(t)}.$$

This inequality will form the basis for establishing our super-martingale process. Here filtration $\mathcal{F}_{t-1}$ will be defined as the union of history $\mathcal{H}_{t-1}$ and the contexts $b_i(t), i = 1, \ldots, N$ at time t. And, $p = \frac{1}{2e\sqrt{\pi T^{\epsilon}}}$, $\ell(T) = R\sqrt{d\ln(\frac{T^3}{\delta})} + 1$, $s_{t,i} = \sqrt{b_i(t)^T B(t)^{-1} b_i(t)}$.

The main idea behind proving the above inequality is to divide the arms into two groups at any given time :

- *unsaturated arms* defined as those with $\Delta_i(t) := b_{i^*(t)}(t)^T\mu - b_i(t)^T\mu \leq (\sqrt{4\ln(NT)}\ v + \ell(T))s_{t,i}$,
- *saturated arms* defined as those with $\Delta_i(t) > (\sqrt{4\ln(NT)}\ v + \ell(T))s_{t,i}$.

Note that $s_{t,i}$ gives the standard deviation of the estimate $b_i(t)^T\hat{\mu}(t)$ and of $\theta_i(t)$. Thus, intuitively saturated arms are arms with the property that the estimates of the means constructed so far in the direction of their contexts are good, making the deviations $s_{t,i}$s small enough—significantly smaller than their current $\Delta_i(t)$.

If an unsaturated arm is played at time t, then regret is at most $\Delta_{i(t)}(t) \leq (\sqrt{4\ln(NT)}\ v + \ell(T))s_{t,i(t)}$. For saturated arms, the regret can be large, but on the other hand, since their deviation $s_{t,i}$ is small, the concentration of $\theta_i(t)$ and $b_i(t)^T\hat{\mu}(t)$ will ensure that with reasonable probability, the algorithm is able to distinguish between them and the optimal arm. In particular, we prove that the probability of playing a saturated arm is within p of the probability of playing the optimal arm. Further, using concentration bounds for $\theta_i(t)$ and $\hat{\mu}(t)$, the regret at any time can be bounded by $(\sqrt{4\ln(NT)}\ v + \ell(T))(s_{t,i^*(t)} + s_{t,i(t)})$ to get the desired inequality.

Then, using the Azuma-Hoeffding inequality for super-martingales, it will follow that with high probability,

$$\begin{aligned}\mathcal{R}(T) = \sum_{t=1}^T \text{regret}(t) &\lesssim (\sqrt{4\ln(NT)}\ v + \ell(T))\left(\tfrac{1}{p}\sum_{t=1}^T I(i(t) = i^*(t)) s_{t,i^*(t)} + \sum_t s_{t,i(t)}\right)\\ &\leq (\sqrt{4\ln(NT)}\ v + \ell(T))\left(\tfrac{1}{p}\sum_t s_{t,i(t)} + \sum_t s_{t,i(t)}\right)\end{aligned}$$

Then, we will use the inequality $\sum_t s_{t,i(t)} = O(\sqrt{Td})$ (derived along the lines of [4]), to get the desired regret bound.

3 Regret Analysis: Proof of Theorem 2

Definition 1. *Define* $\ell(T) = R\sqrt{d\ln(\frac{T^3}{\delta})} + 1$, $v = R\sqrt{\frac{6}{\epsilon}d\ln(\frac{1}{\delta})}$. *And for all* i, *define* $s_{t,i} = \sqrt{b_i(t)^T B(t)^{-1} b_i(t)}$, $\Delta_i(t) = b_{i^*(t)}(t)^T\mu - b_i(t)^T\mu$. *Also define filtration* $\mathcal{F}_t$ *as the union of history until time* t, *and the contexts at time* $t+1$, *i.e.,* $\mathcal{F}_t = \{\mathcal{H}_t, b_i(t+1), i = 1, \ldots, N\}$.

Definition 2. *An arm* i *is called* saturated *at time* t *if* $\Delta_i(t) > (\sqrt{4\ln(NT)}\ v + \ell(T))s_{t,i}$, *and* unsaturated *otherwise. Let* $C(t)$ *denote the set of saturated arms at time* t. *Note that the optimal arm at time* t *is always unsaturated at time* t, *i.e.,* $i^*(t) \notin C(t)$, *and an arm may keep shifting from saturated to unsaturated and vice-versa over time.*

Definition 3. *Define* $E(t)$ *and* $\tilde{E}(t)$ *as the events that* $b_i(t)^T\hat{\mu}(t)$ *and* $\theta_i(t)$ *are concentrated around their respective means. More precisely, define* $E(t)$ *as the event that*

$$\forall i : |b_i(t)^T\hat{\mu}(t) - b_i(t)^T\mu| \leq \ell(T)s_{t,i}.$$

Define $\tilde{E}_i(t)$ *as the event that*

$$|\theta_i(t) - b_i(t)^T\hat{\mu}(t)| \leq \sqrt{4\ln(NT)}\ v s_{t,i},$$

and $\tilde{E}(t)$ *as the event that* $\forall i, \tilde{E}_i(t)$ *holds.*

Lemma 1. *For all* t, $0 < \delta < 1$, $\Pr(E(t)) \geq 1 - \frac{\delta}{T^2}$. *And, for all possible filtrations* $\mathcal{F}_{t-1}$, $\forall i, \Pr(\tilde{E}_i(t)|\mathcal{F}_{t-1}) \geq 1 - \frac{1}{NT^2}$, *and* $\Pr(\tilde{E}(t)|\mathcal{F}_{t-1}) \geq 1 - \frac{1}{T^2}$.

Proof. The complete proof of this lemma appears in Appendix B.2. The probability bound for $E(t)$ will be proven using the concentration inequality given by Theorem 1 in [1]). The probability bound for $\tilde{E}_i(t)$ will be proven using a concentration inequality for Gaussian random variables from [2] stated as Lemma 4 in Appendix B.1 . □

Definition 4. *Recall that regret*(t) *was defined as the regret at time* t, *regret*$(t) = b_{i^*(t)}(t)^T\mu - b_{i(t)}(t)^T\mu$. *Define regret*$'(t) = $ *regret*$(t) - I(\overline{E(t)})$.

Next, we establish a super-martingale process that will form the basis of proving our high-probability regret bound.

Definition 5. *Let*

$$X_t := \frac{1}{(\sqrt{4\ln(NT)}\ v + \ell(T))}\textit{regret}'(t) - \tfrac{1}{p}I(i(t) = i^*(t))s_{t,i^*(t)} - s_{t,i(t)} - \tfrac{5}{pT^2},$$

and $Y_t := \sum_{w=1}^{t} X_w$*, where* $p = \frac{1}{2e\sqrt{\pi T^{\epsilon}}}$.

Lemma 2. $(Y_t; t \geq 0)$ *is a super-martingale process with respect to filtration* $\mathcal{F}_t$.

Proof. We need to prove that for all $t \in [1, T]$,

$$\frac{1}{(\sqrt{4\ln(NT)}\ v + \ell(T))}\mathbb{E}[\text{regret}'(t)|\mathcal{F}_{t-1}] \leq \frac{\Pr\left(i(t) = i^*(t) \,\middle|\, \mathcal{F}_{t-1}\right)}{p} s_{t,i^*(t)} + \mathbb{E}\left[s_{t,i(t)} \,\middle|\, \mathcal{F}_{t-1}\right] + \frac{5}{pT^2}.$$

Let $g(T) = (\sqrt{4\ln(NT)}\ v + \ell(T))$. Note that whether $E(t)$ is true or not is completely determined by $\mathcal{F}_{t-1}$. If $\mathcal{F}_{t-1}$ is such that $E(t)$ does not hold, then $\text{regret}'(t) = \text{regret}(t) - I(\overline{E(t)}) \leq 0$, and the lemma holds trivially. So, we will prove the above lemma while assuming we are given an $\mathcal{F}_{t-1}$ such that $E(t)$ holds.

Let $E_s(t)$ denote the event that some (by definition suboptimal) saturated arm i in $C(t)$ exceeds all the suboptimal unsaturated arms at time t, i.e.,

$$E_s(t) : \exists i \in C(t), \text{ such that } \forall j \notin C(t), j \neq i^*(t), \quad \theta_i(t) \geq \theta_j(t).$$

We prove the following lower bound on the probability of playing the optimal arm,

$$\Pr\left(i(t) = i^*(t) \,\middle|\, \mathcal{F}_{t-1}\right) \geq p \Pr\left(E_s(t) \,\middle|\, \tilde{E}(t), \mathcal{F}_{t-1}\right) - \tfrac{2}{T^2}.$$

And we prove that

$$\tfrac{1}{g(T)}\mathbb{E}[\text{regret}'(t)|\mathcal{F}_{t-1}] \leq \Pr\left(E_s(t) \,\middle|\, \tilde{E}(t), \mathcal{F}_{t-1}\right) s_{t,i^*(t)} + \mathbb{E}\left[s_{t,i(t)} \,\middle|\, \mathcal{F}_{t-1}\right] + \tfrac{3}{T^2},$$

to get the desired inequality.

For the lower bound on $\Pr\left(i(t) = i^*(t) \,\middle|\, \mathcal{F}_{t-1}\right)$,

$$\begin{aligned} \Pr\left(i(t) = i^*(t) \,\middle|\, \mathcal{F}_{t-1}\right) &\geq \Pr(i(t) = i^*(t), E_s(t), \tilde{E}(t)|\mathcal{F}_{t-1}) \\ &= \Pr(i(t) = i^*(t) \,|\, E_s(t), \tilde{E}(t), \mathcal{F}_{t-1}) \cdot \Pr(E_s(t) \,|\, \tilde{E}(t), \mathcal{F}_{t-1}) \cdot \Pr(\tilde{E}(t) \,|\, \mathcal{F}_{t-1}) \end{aligned} \tag{1}$$

Now, given $\tilde{E}(t)$, and $\mathcal{F}_{t-1}$ such that $E(t)$ is true, using the definition of saturated arms, it holds that for all $i \in C(t)$,

$$\theta_i(t) \leq b_i(t)^T\mu + g(T)s_{t,i} \leq b_i(t)^T\mu + \Delta_i(t) \leq b_{i^*(t)}(t)^T\mu,$$

and given $E_s(t)$, it holds that there exists $j \in C(t)$ with

$$\max_{i \notin C(t)} \theta_i(t) \leq \theta_j(t) \leq b_j(t)^T\mu + \Delta_j(t) \leq b_{i^*(t)}(t)^T\mu,$$

so that for all arms i, $\theta_i(t) \leq b_{i^*(t)}(t)^T\mu$. Therefore,

$$\begin{aligned} \Pr(i(t) = i^*(t) \,|\, E_s(t), \tilde{E}(t), \mathcal{F}_{t-1}) &\geq \Pr(\theta_{i^*(t)}(t) \geq b_{i^*(t)}(t)^T\mu \,|\, E_s(t), \tilde{E}(t), \mathcal{F}_{t-1}) \\ &= \Pr(\theta_{i^*(t)}(t) \geq b_{i^*(t)}(t)^T\mu \,|\, \tilde{E}_{i^*(t)}(t), \mathcal{F}_{t-1}) \\ &\geq \Pr(\theta_{i^*(t)}(t) \geq b_{i^*(t)}(t)^T\mu \,|\, \mathcal{F}_{t-1}) - \tfrac{1}{T^2} \end{aligned}$$

The equality in above holds because events $E_s(t)$ and $\tilde{E}_i(t), \forall i \neq i^*(t)$ do not concern the optimal arm, and given $\mathcal{F}_{t-1}$ (and hence $i^*(t)$, $b_{i^*(t)}(t)$, $\hat{\mu}(t)$, and $B(t)$), $\theta_{i^*(t)}(t)$ is independent of these events. For the last inequality, we use that for any two events A, B, $\Pr(A) \leq \Pr(A|B) + \Pr(B)$.

In Lemma 3, we prove a lower bound of p on the probability of $\theta_{i^*(t)}(t)$ to exceed the optimal mean reward $b_{i^*(t)}(t)^T\mu_{i^*(t)}$ given $\mathcal{F}_{t-1}$ such that $E(t)$ holds. This will be proven using concentration provided by $E(t)$ and anti-concentration of Gaussian random variable $\theta_{i^*(t)}(t)$. Using Lemma 3,

$$\Pr(i(t) = i^*(t) \,\Big|\, E_s(t), \tilde{E}(t), \mathcal{F}_{t-1}) \geq p - \tfrac{1}{T^2}$$

Substituting this along with $\Pr\left(\tilde{E}(t) \,\Big|\, \mathcal{F}_{t-1}\right) \geq 1 - \frac{1}{T^2}$ in Equation (1), we get

$$\Pr\left(i(t) = i^*(t) \mid \mathcal{F}_{t-1}\right) \geq p \Pr\left(E_s(t) \,\Big|\, \tilde{E}_i(t), \mathcal{F}_{t-1}\right) - \frac{2}{T^2}. \tag{2}$$

For the regret upper bound, we observe that given $\tilde{E}(t)$, and $\mathcal{F}_{t-1}$ such that $E(t)$ holds, if an arm i is played at time t, then $\Delta_i(t) \leq g(T)(s_{t,i} + s_{t,i^*(t)})$. This holds because if an arm i is played at time t, then it must be true that $\theta_i(t) \geq \theta_{i^*(t)}(t)$. And, given $\tilde{E}(t)$ and $E(t)$,

$$\begin{array}{rcl} b_i(t)^T \mu & \geq & \theta_i(t) - g(T) s_{t,i} \\ & \geq & \theta_{i^*(t)}(t) - g(T) s_{t,i} \\ & \geq & b_{i^*(t)}(t)^T \mu - g(T) s_{t,i^*(t)} - g(T) s_{t,i}. \end{array}$$

Also, by definition of unsaturated arms, for every unsaturated arm i, $\Delta_i(t) \leq g(T) s_{t,i}$. Therefore,

$$\begin{array}{rcl} \mathbb{E}\left[\text{regret}'(t) \mid \mathcal{F}_{t-1}\right] & \leq & \mathbb{E}\left[\sum_{i \in C(t)} \Delta_i(t) I(i = i(t)) \,\Big|\, \mathcal{F}_{t-1}\right] + \mathbb{E}\left[\sum_{i \notin C(t), i \neq i^*(t)} g(T) s_{t,i} I(i = i(t)) \,\Big|\, \mathcal{F}_{t-1}\right] \\ (*) & \leq & \mathbb{E}\left[\sum_{i \in C(t)} \Delta_i(t) I(i = i(t)) \,\Big|\, \tilde{E}(t), \mathcal{F}_{t-1}\right] + \frac{1}{T^2} + g(T)\mathbb{E}\left[s_{t,i(t)} I(i(t) \notin C(t)) \,\big|\, \mathcal{F}_{t-1}\right] \\ & \leq & \mathbb{E}\left[\left(g(T) s_{t,i^*(t)} + g(T) s_{t,i(t)}\right) I(i(t) \in C(t)) \,\Big|\, \tilde{E}(t), \mathcal{F}_{t-1}\right] + \frac{1}{T^2} \\ & & +g(T)\mathbb{E}\left[s_{t,i(t)} I(i(t) \notin C(t)) \,\big|\, \mathcal{F}_{t-1}\right] \\ (**) & \leq & g(T) s_{t,i^*(t)} \cdot \mathbb{E}\left[I(i(t) \in C(t)) \,\Big|\, \tilde{E}(t), \mathcal{F}_{t-1}\right] + \frac{1}{T^2} + g(T)\mathbb{E}\left[s_{t,i(t)} \,\big|\, \mathcal{F}_{t-1}\right] \left(1 - \frac{1}{T^2}\right)^{-1} \\ & \leq & \left(g(T) s_{t,i^*(t)}\right) \Pr\left(E_s \,\Big|\, \tilde{E}(t), \mathcal{F}_{t-1}\right) + \frac{1}{T^2} + g(T)\mathbb{E}\left[s_{t,i(t)} \,\big|\, \mathcal{F}_{t-1}\right] (1 + \frac{2}{T^2}) \\ & \leq & \left(g(T) s_{t,i^*(t)}\right) \Pr\left(E_s \,\Big|\, \tilde{E}(t), \mathcal{F}_{t-1}\right) + \frac{1}{T^2} + g(T)\mathbb{E}\left[s_{t,i(t)} \,\big|\, \mathcal{F}_{t-1}\right] + \frac{2g(T)}{T^2} \end{array} \tag{3}$$

For the inequality marked $(*)$, we use that for any random variable $A \leq 1$, event B, and F, $\mathbb{E}[A|F] \leq \mathbb{E}[A|B, F] + 1 - \Pr(B|F)$. We use this with $A = \sum_{i \in C(t)} \Delta_i(t) I(i = i(t)) \leq \sum_{i \in C(t)} ||b_{i^*(t)}(t)|| \cdot ||\mu_{i^*(t)}|| I(i = i(t)) \leq 1$, $B = \tilde{E}(t)$,$F = \mathcal{F}_{t-1}$. For the inequality marked $(**)$, we use that for any events A, B, F, $\Pr(A|F) \geq \Pr(A|B, F)\Pr(B|F)$. We use this with $A = I(i(t) \notin C(t))$,$B = \tilde{E}(t)$, $F = \mathcal{F}_{t-1}$. For the last inequality, we use that $s_{t,i(t)} \leq ||b_{i(t)}(t)|| \leq 1$. □

The next lemma lower bounds the probability that the sample $\theta_{i^*(t)}(t)$ of the optimal arm at time t will exceed its mean reward.

Lemma 3. *For any filtration $\mathcal{F}_{t-1}$ such that $E(t)$ is true,*

$$\Pr\left(\theta_{i^*(t)}(t) \geq b_{i^*(t)}(t)^T \mu \,\big|\, \mathcal{F}_{t-1}\right) \geq \tfrac{1}{2e\sqrt{\pi T^{\epsilon}}}.$$

Proof. Given event $E(t)$, $|b_{i^*(t)}(t)^T \hat{\mu}(t) - b_{i^*(t)}(t)^T \mu| \leq \ell(T) s_{t,i^*(t)}$. And, since Gaussian random variable $\theta_{i^*(t)}(t)$ has mean $b_{i^*(t)}(t)^T \hat{\mu}(t)$ and standard deviation $v s_{t,i^*(t)}$, using anti-concentration inequality in Lemma 4, we will prove that with probability at least $\frac{1}{2e\sqrt{\pi T^{\epsilon}}}$, $\theta_{i^*(t)}(t)$ will exceed $b_{i^*(t)}(t)^T \hat{\mu}(t) + \left(\ln \frac{T}{\epsilon}\right) v s_{t,i^*(t)}$. Then, the proof will follow from observing that $b_{i^*(t)}(t)^T \hat{\mu}(t) + \left(\ln \frac{T}{\epsilon}\right) v s_{t,i^*(t)} \geq b_{i^*(t)}(t)^T \hat{\mu}(t) + \ell(T) s_{t,i^*(t)} \geq b_{i^*(t)}(t)^T \mu$. The details of the proof are in Appendix C. □

Now, we are ready to prove Theorem 2.

3.1 Proof of Theorem 2

Note that for super-martingale Y_t,

$$|Y_t - Y_{t-1}| = |X_t| = |\tfrac{1}{(\sqrt{4\ln(NT)}\ v+\ell(T))}\text{regret}'(t) - \tfrac{1}{p}I(i(t) = i^*(t))s_{t,i^*(t)} - s_{t,i(t)} - \tfrac{5}{pT^2}| \leq \tfrac{7}{p}.$$

The last inequality holds because for any i, $s_{t,i} = \sqrt{b_i(t)^T B^{-1}(t) b_i(t)} \leq ||b_i(t)||_2 \leq 1$. Therefore, by Azuma-Hoeffding inequality,

$$\Pr\left(Y_T - Y_0 > \tfrac{7}{p}\sqrt{T\ln(2/\delta)}\right) \leq \exp\left(\tfrac{-\ln(2/\delta)T}{T}\right) \leq \delta/2.$$

Therefore with probability $1 - \frac{\delta}{2}$,

$$\sum_{t=1}^{T}\left(\tfrac{1}{(\sqrt{4\ln(NT)}\ v+\ell(T))}\text{regret}'(t) - \tfrac{1}{p}I(i(t) = i^*(t))s_{t,i^*(t)} - s_{t,i(t)} - \tfrac{5}{pT^2}\right) \leq \tfrac{7}{p}\sqrt{T\ln(2/\delta)}.$$

Also,

$$\begin{array}{rcl}\frac{1}{p}\sum_{t=1}^{T} I(i(t) = i^*(t))s_{t,i^*(t)} + \sum_{t=1}^{T} s_{t,i(t)} + \frac{5}{pT} & = & \frac{1}{p}\sum_{t:i(t)=i^*(t)} s_{t,i^*(t)} + \sum_{t=1}^{T} s_{t,i(t)} + \frac{5}{pT} \\ & \leq & \frac{1}{p}\sum_{t=1}^{T} s_{t,i(t)} + \sum_{t=1}^{T} s_{t,i(t)} + \frac{5}{pT} \\ & = & O(\sqrt{T^\epsilon}\sqrt{Td\ln T})\end{array}$$

For the last inequality, we use that $\sum_{t=1}^{T} s_{t,i(t)} \leq 5\sqrt{dT\ln T}$, which can be derived along the lines of Lemma 3 of [9] using Lemma 11 of [4]. Details are in Appendix B.3. Therefore, with probability $1 - \frac{\delta}{2}$,

$$\sum_{t=1}^{T} \text{regret}'(t) \leq (\sqrt{4\ln(NT)}\ v + \ell(T)) \cdot \left(O(\sqrt{T^\epsilon}\sqrt{Td\ln T}) + \tfrac{7}{p}\sqrt{T\ln(\tfrac{2}{\delta})}\right) =$$
$$O\left(d\sqrt{\tfrac{T^{1+\epsilon}}{\epsilon}\ln N}\ln T\ln\tfrac{1}{\delta}\right)$$

Also, because $E(t)$ holds for all t with probability at least $1 - \frac{\delta}{2}$ (refer to Lemma 1), $\text{regret}'(t) = \text{regret}(t)$ for all t with probability at least $1 - \frac{\delta}{2}$. Hence, with probability $1 - \delta$,

$$\mathcal{R}(T) = \sum_{t=1}^{T} \text{regret}(t) = O\left(d\sqrt{\tfrac{T^{1+\epsilon}}{\epsilon}\ln N}\ln T\ln\tfrac{1}{\delta}\right).$$

To obtain bounds for the other definition of regret in Remark 3, observe that the expected regret for this definition is the same as before,

$$\mathbb{E}[\text{regret}(t)] = \mathbb{E}[r_{i^*(t)}(t) - r_{i(t)}(t)] = \mathbb{E}[\mathbb{E}[r_{i^*(t)}(t)|i^*(t)]] - \mathbb{E}[\mathbb{E}[r_{i(t)}(t)|i(t)]] = \mathbb{E}[b_{i^*(t)}(t)^T\mu - b_{i(t)}(t)^T\mu].$$

Therefore, Lemma 2 holds as it is, and Y_t defined in Definition 5 is a super-martingale with respect to this new definition of regret(t) as well. Now, if $|r_i(t)| \leq R$ for all i, then $|\text{regret}'(t)| \leq R$ and $|Y_t - Y_{t-1}| \leq \frac{7}{p} + R$, and we can apply Azuma-Hoeffding inequality exactly as in this subsection to obtain regret bounds of the same order as Theorem 2 for the new definition.

4 Extensions

4.1 N different parameters

Theorem 1 considers the setting where each arm i is associated with a parameter μ_i, where possibly $\mu_i \neq \mu_{i'}$ for two different arms i and i'. In this case, Thompson Sampling would maintain a separate estimate of mean $\hat{\mu}_i(t)$, and $B_i(t)$ for each arm i which would be updated only at the time instances when i is played. The statements of Lemma 1, and the super-martingale property established by Lemma 2 will hold as it is for the new definitions. The only difference will appear in the bound for $\sum_t s_{t,i(t)}$ used in the proof of Theorem 2. For the case of N different parameters, we will get a bound of $O(\sqrt{NTd\ln T})$ on this quantity instead of $O(\sqrt{Td\ln T})$, leading to the extra $\sqrt{N}$ factor in the bound in Theorem 1 compared to Theorem 2. The details of the algorithm for the case of N different parameters, and the changes in the analysis required for proving Theorem 1 are provided in Appendix D.

4.2 General distributions

In the algorithm in this paper, $\theta_i(t)$ is generated from a Gaussian distribution. However, the analysis techniques in this paper are easily extendable to an algorithm that uses a posterior distribution other than the Gaussian distribution. The only distribution specific properties we have used in the analysis are the concentration and anti-concentration inequalities for Gaussian distributed random variables mentioned in Lemma 4. The concentration inequality was used to prove that $\tilde{E}(t)$ happens with high probability in Lemma 1, and the anti-concentration inequality was used to lower bound the probability that Gaussian distributed random variable $\theta_{i*(t)}(t)$ exceeds its mean by some factors of its standard deviation in Lemma 3. If any other distribution provides similar tail inequalities, these inequalities can be used as a black box in the analysis, and the regret bounds can be reproduced for that distribution.

5 Conclusions

We provided a theoretical analysis of Thompson Sampling for the contextual bandits problem with linear payoffs. Our results resolve many open questions regarding the theoretical guarantees for Thompson Sampling, and establish that even for the contextual version of the stochastic MAB problem, TS achieves regret bounds comparable to the state-of-the-art methods. We used novel martingale-based analysis techniques which are simpler than those in the past work on TS [3, 15], and amenable to extensions. In fact, the techniques introduced in this paper could also be used to provide a simpler proof for the optimal expected regret bounds for TS for the basic MAB problem studied in [3, 15]. The proof of this claim will appear elsewhere.

Several questions remain open. A tighter analysis that can remove the dependence on ϵ is desirable. We believe that our techniques would adapt to provide such bounds for the *expected regret*. Other avenues to explore are contextual bandits with *generalized* linear models considered in [11], the setting with delayed and batched feedbacks, and the *agnostic* case of contextual bandits with linear payoffs. The agnostic case refers to the setting which does not make the realizability assumption that there exists a vector μ_i for each i for which $\mathbb{E}[r_i(t)|b_i(t)] = b_i(t)^T\mu_i$. To our knowledge, no existing algorithm has been shown to have non-trivial regret bounds for the agnostic case.

References

[1] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved Algorithms for Linear Stochastic Bandits. In *NIPS*, pages 2312–2320, 2011.

[2] Milton Abramowitz and Irene A. Stegun. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables.* Dover, New York, 1964.

[3] Shipra Agrawal and Navin Goyal. Analysis of Thompson Sampling for the Multi-armed Bandit Problem. In *COLT*, 2012.

[4] Peter Auer. Using Confidence Bounds for Exploitation-Exploration Trade-offs. *Journal of Machine Learning Research*, 3:397–422, 2002.

[5] Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The Nonstochastic Multiarmed Bandit Problem. *SIAM J. Comput.*, 32(1):48–77, 2002.

[6] Sébastien Bubeck, Nicolò Cesa-Bianchi, and Sham M. Kakade. Towards minimax policies for online linear optimization with bandit feedback. *Proceedings of the 25th Conference on Learning Theory (COLT)*, pages 1–14, 2012.

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

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

[9] Wei Chu, Lihong Li, Lev Reyzin, and Robert E. Schapire. Contextual Bandits with Linear Payoff Functions. *Journal of Machine Learning Research - Proceedings Track*, 15:208–214, 2011.

[10] Varsha Dani, Thomas P. Hayes, and Sham M. Kakade. Stochastic Linear Optimization under Bandit Feedback. In *COLT*, pages 355–366, 2008.

[11] Sarah Filippi, Olivier Cappé, Aurélien Garivier, and Csaba Szepesvári. Parametric Bandits: The Generalized Linear Case. In *NIPS*, pages 586–594, 2010.

[12] Thore Graepel, Joaquin Quiñonero Candela, Thomas Borchert, and Ralf Herbrich. Web-Scale Bayesian Click-Through rate Prediction for Sponsored Search Advertising in Microsoft's Bing Search Engine. In *ICML*, pages 13–20, 2010.

[13] 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.

[14] Leslie Pack Kaelbling. Associative Reinforcement Learning: Functions in k-DNF. *Machine Learning*, 15(3):279–298, 1994.

[15] Emilie Kaufmann, Nathaniel Korda, and Rémi Munos. Thompson Sampling: An Optimal Finite Time Analysis. *ALT*, 2012.

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

[17] John Langford and Tong Zhang. The Epoch-Greedy Algorithm for Multi-armed Bandits with Side Information. In *NIPS*, 2007.

[18] Benedict C. May, Nathan Korda, Anthony Lee, and David S. Leslie. Optimistic Bayesian sampling in contextual-bandit problems. Technical report, Statistics Group, Department of Mathematics, University of Bristol.

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

[20] Pedro A. Ortega and Daniel A. Braun. Linearly Parametrized Bandits. *Journal of Artificial Intelligence Research*, 38:475–511, 2010.

[21] Jyotirmoy Sarkar. One-armed badit problem with covariates. *The Annals of Statistics*, 19(4):1978–2002, 1991.

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

[23] Alexander L. Strehl, Chris Mesterharm, Michael L. Littman, and Haym Hirsh. Experience-efficient learning in associative bandit problems. In *ICML*, pages 889–896, 2006.

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

[25] William 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.

[26] Michael Woodroofe. A one-armed bandit problem with a concomitant variable. *Journal of the American Statistics Association*, 74(368):799–806, 1979.

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

A Posterior distribution computation

$$\begin{aligned}
\Pr(\tilde{\mu}|r_i(t)) &\propto \Pr(r_i(t)|b_i(t)^T\tilde{\mu})\Pr(\tilde{\mu}) \\
&\propto \exp\{-\frac{1}{2v^2}((r_i(t)-\tilde{\mu}^T b_i(t))^2 + (\tilde{\mu}-\hat{\mu}(t))^T B(t)(\tilde{\mu}-\hat{\mu}(t))\} \\
&\propto \exp\{-\frac{1}{2v^2}(r_i(t)^2 + \tilde{\mu}^T b_i(t) b_i(t)^T \tilde{\mu} + \tilde{\mu}^T B(t)\tilde{\mu} - 2\tilde{\mu}^T b_i(t) r_i(t) - 2\tilde{\mu}^T B(t)\hat{\mu}(t))\} \\
&\propto \exp\{-\frac{1}{2v^2}(\tilde{\mu}^T B(t+1)\tilde{\mu} - 2\tilde{\mu}^T B(t+1)\hat{\mu}(t+1))\} \\
&\propto \exp\{-\frac{1}{2v^2}(\tilde{\mu}-\hat{\mu}(t+1))^T B(t+1)(\tilde{\mu}-\hat{\mu}(t+1))\} \\
&\propto \mathcal{N}(\hat{\mu}(t+1), v^2 B(t+1)^{-1})
\end{aligned}$$

Therefore, the posterior distribution of μ at time $t+1$ is $\mathcal{N}(\hat{\mu}(t+1), v^2 B(t+1)^{-1})$,

B Proof of Theorem 2

B.1 Gaussian concentration

Formula 7.1.13 from [2] can be used to derive the following concentration and anti-concentration inequalities for Gaussian distributed random variables.

Lemma 4. *[2] For a Gaussian distributed random variable Z with mean m and variance σ^2, for any $z \geq 1$,*

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

B.2 Proof of Lemma 1

We will use the following lemma (implied by Theorem 1 in [1]):

Lemma 5. *[1] Let $(\mathcal{F}'_t; t \geq 0)$ be a filtration, $(m_t; t \geq 1)$ be an $\mathbb{R}^d$-valued stochastic process such that m_t is $(\mathcal{F}'_{t-1})$ measurable, $(\eta_t; t \geq 1)$ be a real-valued martingale difference process such that η_t is $(\mathcal{F}'_t)$ measurable and For $t \geq 0$, define $\xi_t = \sum_{u=1}^t m_u \eta_u$ and $M_t = I_d + \sum_{u=1}^t m_u m_u^T$, where I_d is the d-dimensional identity matrix. Assume η_t is conditionally R-sub-Gaussian.*

Then, for any $\delta' > 0$, $t \geq 0$, with probability at least $1-\delta'$,

$$||\xi_t||_{M_t^{-1}} \leq R\sqrt{d \ln\left(\frac{t+1}{\delta'}\right)}.$$

We use the above lemma with $m_t = b_{i(t)}(t)$, $\eta_t = r_{i(t)} - b_{i(t)}(t)^T\mu$, $\mathcal{F}'_t = (m_{u+1}, \eta_u : u \leq t)$. (Note that effectively, $\mathcal{F}'_t$ can be imagined to have all the information including the arms played until time $t+1$, except for the reward of the arm played at time $t+1$). By definition of $\mathcal{F}'_t$, m_t is $\mathcal{F}'_{t-1}$ measurable, and η_t is $\mathcal{F}'_t$ measurable. And, η_t is a martingale difference process:

$$\mathbb{E}\left[\eta_t | \mathcal{F}'_{t-1}\right] = \mathbb{E}[r_{i(t)} | b_{i(t)}(t), i(t)] - b_{i(t)}(t)^T \mu = 0.$$

Also, this makes

$$M_t = I_d + \sum_{u=1}^t m_u m_u^T = I_d + \sum_{u=1}^t b_{i(u)}(u) b_{i(u)}(u)^T,$$

$$\xi_t = \sum_{u=1}^t m_u \eta_u = \sum_{u=1}^t b_{i(u)}(u)(r_{i(u)} - b_{i(u)}(u)^T \mu).$$

Note that $B(t) = M_{t-1}$, and $\hat{\mu}(t) - \mu = M_{t-1}^{-1}(\xi_{t-1} - \mu)$, so that

$$|b_i(t)^T\hat{\mu}(t) - b_i(t)^T\mu| = |b_i(t)^T M_{t-1}^{-1}(\xi_{t-1}-\mu)| \leq ||b_i(t)||_{M_{t-1}^{-1}} ||\xi_{t-1}-\mu||_{M_{t-1}^{-1}} = ||b_i(t)||_{B(t)^{-1}} ||\xi_{t-1}-\mu||_{M_{t-1}^{-1}}.$$

The inequality holds because M_{t-1}^{-1} is a positive definite matrix. Using the above lemma, for any $\delta' > 0$, $t \geq 1$, with probability at least $1-\delta'$,

$$||\xi_{t-1}||_{M_{t-1}^{-1}} \leq R\sqrt{d \ln\left(\frac{t}{\delta'}\right)}.$$

Therefore, $||\xi_{t-1} - \mu||_{M_{t-1}^{-1}} \leq R\sqrt{d \ln\left(\frac{t}{\delta'}\right)} + ||\mu||_{M_{t-1}^{-1}} \leq R\sqrt{d \ln\left(\frac{t}{\delta'}\right)} + 1$. Substituting $\delta' = \frac{\delta}{T^2}$, we get that with probability $1 - \frac{\delta}{T^2}$, for all i,

$$|b_i(t)^T \hat{\mu}(t) - b_i(t)^T \mu| \leq s_{t,i} \cdot \left(R\sqrt{d \ln(\frac{T^3}{\delta})} + 1 \right) \leq \ell(T) s_{t,i}$$

This proves the bound on the probability of $E(t)$. To prove bound on probability of $\tilde{E}_i(t)$, we use the fact that since $\theta_i(t)$ is distributed as $\mathcal{N}(b_i(t)^T \hat{\mu}(t), v^2 b_i(t)^T B(t)^{-1} b_i(t))$, therefore, using concentration inequalities for Gaussian random variables,

$$\Pr(|\theta_i(t) - b_i(t)^T \hat{\mu}(t)| > z\ v\sqrt{b_i(t)^T B(t)^{-1} b_i(t)} | \mathcal{F}_{t-1}) \leq \frac{1}{\sqrt{\pi} z} e^{-z^2/2}$$

Substituting $z = \sqrt{4 \ln(NT)}$, we get the desired bound.

B.3 Bound on the sum of $s_{t,i(t)}$

We will use the following result, implied by the referred lemma in [4]

Lemma 6. [[4], Lemma 11]. *Let $A' = A + xx^T$, where $x \in \mathbb{R}^d, A, A' \in \mathbb{R}^{d \times d}$, and all the eigenvalues $\lambda_j, j = 1, \ldots, d$ of A are greater than or equal to 1. Then, the eigenvalues $\lambda'_j, j = 1, \ldots, d$ of A' can be arranged so that $\lambda_j \leq \lambda'_j$ for all j, and*

$$x^T A^{-1} x \leq 10 \sum_{j=1}^{d} \frac{\lambda'_j - \lambda_j}{\lambda_j}$$

Let $\lambda_{j,t}$ denote the eigenvalues of $B(t)$. Note that $B(t+1) = B(t) + b_{i(t)}(t) b_{i(t)}(t)^T$, and $\lambda_{j,t} \geq 1, \forall j$. Therefore, above implies

$$s_{t,i(t)}^2 \leq 10 \sum_{j=1}^{d} \frac{\lambda_{j,t+1} - \lambda_{j,t}}{\lambda_{j,t}}.$$

This allows us to derive the following along the lines of Lemma 3 of [9].

$$\sum_{t=1}^{T} s_{t,i(t)} \leq 5\sqrt{dT \ln T}.$$

C Proof of Lemma 3

Given event $E(t)$, $|b_{i^*(t)}(t)^T \hat{\mu}(t) - b_{i^*(t)}(t)^T \mu| \leq \ell(T) s_{t,i^*(t)}$. And, since Gaussian random variable $\theta_{i^*(t)}(t)$ has mean $b_{i^*(t)}(t)^T \hat{\mu}(t)$ and standard deviation $v s_{t,i^*(t)}$, using anti-concentration inequality in Lemma 4,

$$\begin{aligned} \Pr\left(\theta_{i^*(t)}(t) \geq b_{i^*(t)}(t)^T \mu \,\middle|\, \mathcal{F}_{t-1}\right) &= \Pr\left(\frac{\theta_{i^*(t)}(t) - b_{i^*(t)}(t)^T \hat{\mu}(t)}{v s_{t,i^*(t)}} \geq \frac{b_{i^*(t)}(t)^T \mu - b_{i^*(t)}(t)^T \hat{\mu}(t)}{v s_{t,i^*(t)}} \,\middle|\, \mathcal{F}_{t-1} \right) \\ &\geq \frac{1}{2\sqrt{\pi}} e^{-Z_t^2} \end{aligned}$$

where

$$\begin{aligned}
|Z_t| &= \left|\frac{b_{i^*(t)}(t)^T\mu - b_{i^*(t)}(t)^T\hat{\mu}(t)}{v s_{t,i^*(t)}}\right| \\
&\le \frac{\ell(T)s_{t,i^*(t)}}{v s_{t,i^*(t)}} \\
&\le \frac{R\sqrt{d\ln(\frac{T^3}{\delta})}+1}{R\sqrt{\frac{6}{\epsilon}d\ln(\frac{1}{\delta})}} \\
&\le \sqrt{\frac{\epsilon}{2}(\ln T+1)}
\end{aligned}$$

$$\Pr\left(\theta_{i^*(t)}(t) \ge b_{i^*(t)}(t)^T\mu \,\middle|\, \mathcal{F}_{t-1}\right) \ge \frac{1}{2\sqrt{\pi}}e^{-\frac{\epsilon}{2}(\ln T+1)} = \frac{1}{2e\sqrt{\pi}T^{\frac{\epsilon}{2}}}$$

D N different parameters: Proof of Theorem 1

Theorem 1 considers the setting where each arm i is associated with a parameter μ_i, where possibly $\mu_i \neq \mu_{i'}$ for two different arms i and i'. In this case, Thompson Sampling would maintain a separate estimate of mean $\hat{\mu}_i(t)$, and $B_i(t)$ for each arm i which would be updated only at the time instances when i is played.

$$B_i(t) = I_d + \sum_{u=1:i(u)=i}^{t-1} b_i(u)b_i(u)^T$$

$$\hat{\mu}_i(t) = B_i(t)^{-1}\left(\sum_{u=1:i(u)=i}^{t-1} b_i(u)r_i(u)\right)$$

$$s_{t,i} = \sqrt{b_i(t)^T B_i(t)^{-1} b_i(t)}$$

The posterior distribution for each arm i at time t would be $\mathcal{N}(b_i(t)^T\hat{\mu}_i(t), v^2\, b_i(t)^T B_i(t)^{-1}b_i(t))$.

```
Algorithm 2: Thompson Sampling for Contextual bandits with $N$ parameters
Set $B_i = I_d, \hat{\mu}_i = 0_d,\ i = 1,\ldots,N,\ f_i = 0_d$.
foreach $t = 1, 2, \ldots,$ do
    For each arm $i = 1,\ldots,N$, sample $\theta_i(t)$ independently from distribution
    $\mathcal{N}(b_i(t)^T\hat{\mu}_i, v^2\ b_i(t)^T B_i^{-1} b_i(t))$.
    Play arm $i(t) := \arg\max_i \theta_i(t)$ and observe reward $r_t$.
    Update $B_{i(t)} = B_{i(t)} + b_{i(t)}(t)b_{i(t)}(t)^T$, $f_{i(t)} = f_{i(t)} + b_{i(t)}(t)r_t$, $\hat{\mu}_{i(t)} = B_{i(t)}^{-1} f_{i(t)}$.
end
```

In the regret analysis, the events $E(t)$ will now be defined with respect to concentration of all $\hat{\mu}_i(t)$ around their respective means. That is,

$$E(t) : \forall i, b_i(t)^T\hat{\mu}_i(t) \in [b_i(t)^T\mu_i - \ell(T)s_{t,i},\ \ b_i(t)^T\mu_i + \ell(T)s_{t,i}]$$

Similarly, $\tilde{E}_i(t)$ will be the event that

$$\theta_i(t) \in [b_i(t)^T \hat{\mu}_i(t) - \sqrt{4\ln(NT)}\ v s_{t,i},\ \ b_i(t)^T \hat{\mu}_i(t) + \sqrt{4\ln(NT)}\ v s_{t,i}],$$

and $\tilde{E}(t)$ will be the event that $\forall i, \tilde{E}_i(t)$ holds. It is easy to observe that the statements of Lemma 1 and the super-martingale property established by Lemma 2 will hold as it is for these new definitions. The only difference will appear in the bound for $\sum_t s_{t,i(t)}$ used in the proof of Theorem 2. For the case of N different parameters, we will get a bound of $O(\sqrt{NTd\ln T})$ on this quantity.

Let $n_i(T)$ be the number of times arm i is played by time T. Then using Lemma 6, for two consequent time steps t, t' at which arm i is played

$$s^2_{t,i(t)} \leq 10 \sum_{j=1}^{d} \frac{\lambda_{j,t'} - \lambda_{j,t}}{\lambda_{j,t}}.$$

This allows us to derive the following lemma along the lines of Lemma 3 of [9].

Lemma 7. [[9], Lemma 3] *For $i = 1, \ldots, N$,*

$$\sum_{t=1:i(t)=i}^{T} s_{t,i(t)} \leq 5\sqrt{dn_i(T)\ln(n_i(T))}.$$

Using above lemma,

$$\sum_{t=1}^{T} s_{t,i(t)} = \sum_{i=1}^{N} \sum_{t=1:i(t)=i}^{T} s_{t,i(t)} \leq \sum_{i=1}^{N} 5\sqrt{n_i(T)d\ln T} \leq 5\sqrt{N}\sqrt{\sum_i n_i(T)}\sqrt{d\ln T} = 5\sqrt{NTd\ln T}.$$

Therefore, following the same lines as proof of Theorem 2, we will get a regret bound of $O(d\sqrt{\frac{T^\epsilon}{\epsilon}}\sqrt{NT\ln N}\ln T\ln\frac{1}{\delta})$