Thompson Sampling for Contextual Bandits with Linear Payoffs

Shipra Agrawal
Microsoft Research

Navin Goyal
Microsoft Research

February 4, 2014

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 a generalization of Thompson Sampling algorithm for the stochastic contextual multi-armed bandit problem with linear payoff functions, when the contexts are provided by an adaptive adversary. This is among the most important and widely studied version of the contextual bandits problem. We provide the first theoretical guarantees for the contextual version of Thompson Sampling. We prove a high probability regret bound of $\tilde{O}(d^{3/2}\sqrt{T})$ (or $\tilde{O}(d\sqrt{T\log(N)})$), which is the best regret bound achieved by any computationally efficient algorithm for this problem, and is within a factor of $\sqrt{d}$ (or $\sqrt{\log(N)}$) of the information-theoretic lower bound for this problem.

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, in each of T rounds, a learner is presented with the choice of taking one out of N actions, referred to as N arms. Before making the choice of which arm to play, the learner sees d-dimensional feature vectors 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 the arms played by her in the past to make the choice of the arm to play in the current round. Over time, the learner's aim is to gather enough information about how the feature vectors and rewards relate to each other, so that she can predict, with some certainty, which arm is likely to 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 (i.e., have low regret), then 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 a d-dimensional parameter $\overline{\mu} \in \mathbb{R}^d$, and the predictor ranks the arms according to $b_i^T \overline{\mu}$. We consider stochastic contextual bandit problem under linear realizability assumption, that is, we assume that there is an unknown underlying parameter $\mu \in \mathbb{R}^d$ such that the expected reward for each arm i, given context b_i, is $b_i^T \mu$. Under this realizability assumption, the linear predictor corresponding to μ is in fact the best predictor and the learner's aim is to learn this underlying parameter. This realizability assumption is standard in the existing literature on contextual multi-armed bandits, e.g. (Auer, 2002; Filippi et al., 2010; Chu et al., 2011; Abbasi-Yadkori et al., 2011).

Thompson Sampling (TS) is one of the earliest heuristics for multi-armed bandit problems. The first version of this Bayesian heuristic is around 80 years old, dating to Thompson (1933). Since then, it has been rediscovered numerous times independently in the context of reinforcement learning, e.g., in Wyatt (1997); Ortega & Braun (2010); Strens (2000). 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 TS for the contextual bandits problem involves the following elements:

1. a set Θ of parameters $\tilde{\mu}$;
2. a prior distribution $P(\tilde{\mu})$ on these parameters;
3. past observations $\mathcal{D}$ consisting of (context b, reward r) for the past time steps;

4. a 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 having the best parameter. A simple way to achieve this is to produce a sample of parameter for each arm, using the posterior distributions, and play the arm that produces the best sample. In this paper, we design and analyze a natural generalization of Thompson Sampling (TS) for contextual bandits; this generalization fits the above general structure, and uses Gaussian prior and Gaussian likelihood function. We emphasize that although TS is a Bayesian approach, the description of the algorithm and our analysis apply to the prior-free stochastic MAB model, and our regret bounds will hold irrespective of whether or not the actual reward distribution matches the Gaussian likelihood function used to derive this Bayesian heuristic. Thus, our bounds for TS algorithm are directly comparable to the UCB family of algorithms which form a frequentist approach to the same problem. One could interpret the priors used by TS as a way of capturing the current knowledge about the arms.

Recently, TS has attracted considerable attention. Several studies (e.g., Granmo (2010); Scott (2010); Graepel et al. (2010); Chapelle & Li (2011); May & Leslie (2011); Kaufmann et al. (2012)) have empirically demonstrated the efficacy of TS: Scott (2010) provides a detailed discussion of probability matching techniques in many general settings along with favorable empirical comparisons with other techniques. Chapelle & Li (2011) demonstrate that for the basic stochastic MAB problem, empirically TS achieves regret comparable to the lower bound of Lai & Robbins (1985); 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 been used in an industrial-scale application for CTR prediction of search ads on search engines (Graepel et al., 2010). Kaufmann et al. (2012) 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.

However, the theoretical understanding of TS is limited. Granmo (2010) and May et al. (2011) provided weak guarantees, namely, a bound of $o(T)$ on the expected regret in time T. For the the basic (i.e. without contexts) version of the stochastic MAB problem, some significant progress was made by Agrawal & Goyal (2012), Kaufmann et al. (2012) and, more recently, by Agrawal & Goyal (2013b), who provided optimal regret bounds on the expected regret. But, many questions regarding theoretical analysis of TS remained open, including high probability regret bounds, and regret bounds for the more general contextual bandits setting. In particular, the contextual MAB problem does not seem easily amenable to the techniques used so far for analyzing TS for the basic MAB problem. In Section 3.1, we describe some of these challenges. Some of these questions and difficulties were also formally raised as a COLT 2012 open problem (Chapelle & Li, 2012).

In this paper, we use novel martingale-based analysis techniques to demonstrate that TS (i.e., our Gaussian prior based generalization of TS for contextual bandits) achieves high probability, near-optimal regret bounds for stochastic contextual bandits with linear payoff functions. 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 by(Chapelle & Li, 2012) for contextual bandits with linear payoffs.

We provide a regret bound of $\tilde{O}(d^{3/2}\sqrt{T})$, or $\tilde{O}(d\sqrt{T\log(N)})$ (whichever is smaller), upper bound on the regret for Thompson Sampling algorithm. Moreover, the Thomspon Sampling algorithm we propose is efficient (runs in time polynomial in d) to implement as long as it is efficient to optimize a linear function over the set of arms (see Section 2.2 paragraph "Computational efficiency" for further discussion). Although the information theoretic lower bound for this problem is $\Omega(d\sqrt{T})$, an upper bound of $\tilde{O}(d^{3/2}\sqrt{T})$ is in fact the best achieved by any computationally efficient algorithm in the literature when number of arms N is large (see the related work section 2.4 for a detailed discussion). To determine whether there is a gap between computational and information theoretic lower bound for this problem is an intriguing open question.

Our version of Thompson Sampling algorithm for the contextual MAB problem, described formally in Section 2.2, uses Gaussian prior and Gaussian likelihood functions. Our techniques can be extended to the use of other prior distributions, satisfying certain conditions, as discussed in Section 4.

2 Problem setting and algorithm description

2.1 Problem setting

There are N arms. At time $t = 1, 2, \ldots$, a context vector $b_i(t) \in \mathbb{R}^d$, 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 $\mathcal{H}_{t-1}$,

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

where $a(\tau)$ denotes the arm played at time τ. Given $b_i(t)$, the reward for arm i at time t is generated from an (unknown) distribution with mean $b_i(t)^T\mu$, where $\mu \in \mathbb{R}^d$ is a fixed but unknown parameter.

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

An algorithm for the *contextual bandit problem* needs to choose, at every time t, an arm $a(t)$ to play, using history $\mathcal{H}_{t-1}$ and current contexts $b_i(t), i = 1, \ldots, N$. Let $a^*(t)$ denote the optimal arm at time t, i.e. $a^*(t) = \arg\max_i b_i(t)^T\mu$. And let $\Delta_i(t)$ be the difference between the mean rewards of the optimal arm and of arm i at time t, i.e.,

$$\Delta_i(t) = b_{a^*(t)}(t)^T\mu - b_i(t)^T\mu.$$

Then, the regret at time t is defined as

$$\text{regret}(t) = \Delta_{a(t)}(t).$$

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.

We assume that $\eta_{i,t} = r_i(t) - b_i(t)^T\mu$ is conditionally R-sub-Gaussian for a constant $R \geq 0$, i.e.,

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

This assumption is satisfied whenever $r_i(t) \in [b_i(t)^T\mu - R, b_i(t)^T\mu + R]$ (see Remark 1 in Appendix A.1 of Filippi et al. (2010)). We will also assume that $||b_i(t)|| \leq 1$, $||\mu|| \leq 1$, and $\Delta_i(t) \leq 1$ for all i, t (the norms, unless otherwise indicated, are ℓ_2-norms). These assumptions are required to make the regret bounds scale-free, and are standard in the literature on this problem. If $||\mu|| \leq c, ||b_i(t)|| \leq c, \Delta_i(t) \leq c$ instead, then our regret bounds would increase by a factor of c.

Remark 1. *An alternative definition of regret that appears in the literature is*

$$\textit{regret}(t) = r_{a^*(t)}(t) - r_{a(t)}(t).$$

We can obtain the same regret bounds for this alternative definition of regret. The details are provided in the supplementary material in Appendix A.5.

2.2 Thompson Sampling algorithm

We use Gaussian likelihood function and Gaussian prior to design our version of Thompson Sampling algorithm. More precisely, suppose that the **likelihood** of reward $r_i(t)$ at time t, given context $b_i(t)$ and parameter μ, were given by the pdf of Gaussian distribution $\mathcal{N}(b_i(t)^T\mu, v^2)$. Here, $v = R\sqrt{9d\ln(\frac{T}{\delta})}$.Let

$$B(t) = I_d + \sum_{\tau=1}^{t-1} b_{a(\tau)}(\tau) b_{a(\tau)}(\tau)^T$$

$$\hat{\mu}(t) = B(t)^{-1}\left(\sum_{\tau=1}^{t-1} b_{a(\tau)}(\tau) r_{a(\tau)}(\tau)\right).$$

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

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

as $\mathcal{N}(\hat{\mu}(t+1), v^2 B(t+1)^{-1})$ (details of this computation are in Appendix A.1). In our Thompson Sampling algorithm, at every time step t, we will simply generate a sample $\tilde{\mu}(t)$ from the distribution $\mathcal{N}(\hat{\mu}(t), v^2 B(t)^{-1})$, and play the arm i that maximizes $b_i(t)^T\tilde{\mu}(t)$.

We emphasize that the Gaussian priors and the Gaussian likelihood model for rewards are only used above to design the Thompson Sampling algorithm for contextual bandits. Our analysis of the algorithm 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 2.1, i.e., the R-sub-Gaussian assumption.

Algorithm 1 Thompson Sampling for Contextual bandits

for all $t = 1, 2, \ldots,$ **do**
  Sample $\tilde{\mu}(t)$ from distribution $\mathcal{N}(\hat{\mu}(t), v^2 B(t)^{-1})$.
  Play arm $a(t) := \arg\max_i b_i(t)^T \tilde{\mu}(t)$, and observe reward $r_{a(t)}(t)$.
end for

Knowledge of time horizon T: The parameter $v = R\sqrt{9d\ln(\frac{T}{\delta})}$ can be replaced by $v_t = R\sqrt{9d\ln(\frac{t}{\delta})}$ at time t, if the time horizon T is not known. In fact, this is the version of Thompson Sampling that we will analyze. The analysis we provide can be applied as it is (with only notational changes) to the version using the fixed value of v for all time steps, to get the same regret upper bound.

Computational efficiency: Every step t of Thompson Sampling (both algorithms) consists of generating a d-dimensional sample $\tilde{\mu}(t)$ from a multi-variate Gaussian distribution, and solving the problem $\arg\max_i b_i(t)^T \tilde{\mu}(t)$. Therefore, even if the number of arms N is large (or infinite), the above algorithms are efficient as long as the problem $\arg\max_i b_i(t)^T \tilde{\mu}(t)$ is efficiently solvable. This is the case, for example, when the set of arms at time t is given by a d-dimensional convex set $\mathcal{K}_t$ (every vector in $\mathcal{K}_t$ is a context vector, and thus corresponds to an arm). The problem to be solved at time step t is then $\max_{b \in \mathcal{K}_t} b^T \tilde{\mu}(t)$, where $\mathcal{K}_t$.

2.3 Our Results

Theorem 1. *With probability $1 - \delta$, the total regret for Thompson Sampling algorithm in time T is bounded as*

$$\mathcal{R}(T) = O\left(d^{3/2}\sqrt{T}\left(\ln(T) + \sqrt{\ln(T)\ln(\frac{1}{\delta})}\right)\right), \tag{1}$$

or,

$$\mathcal{R}(T) = O\left(d\sqrt{T\log(N)}\left(\ln(T) + \sqrt{\ln(T)\ln(\frac{1}{\delta})}\right)\right), \tag{2}$$

whichever is smaller, for any $0 < \delta < 1$, where δ is a parameter used by the algorithm.

Remark 2. *The regret bound in Equation* (1) *does not depend on N, and are applicable to the case of infinite arms, with only notational changes required in the analysis.*

2.4 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. (2011): bandit problems with co-variates (Woodroofe, 1979; Sarkar, 1991), associative reinforcement learning (Kaelbling, 1994), associative bandit problems (Auer, 2002; Strehl et al., 2006), bandit problems with expert advice (Auer et al., 2002), and linear bandits (Dani et al., 2008; Abbasi-Yadkori et al., 2011; Bubeck et al., 2012). The name *contextual bandits* was coined in Langford & Zhang (2007).

A lower bound of $\Omega(d\sqrt{T})$ for this problem was given by Dani et al. (2008), when the number of arms is allowed to be infinite. In particular, they prove their lower bound using an example where the set of arms correspond to all vectors in the intersection of a d-dimensional sphere and a cube. They also provide an upper bound of $\tilde{O}(d\sqrt{T})$, although their setting is slightly restrictive in the sense that the context vector for every arm is fixed in advanced and is not allowed to change with time. Abbasi-Yadkori et al. (2011) analyze a UCB-style algorithm and provide a regret upper bound of $O(d\log(T)\sqrt{T} + \sqrt{dT\log(T/\delta)})$.

For finite N, Chu et al. (2011) show a lower bound of $\Omega(\sqrt{Td})$ for $d^2 \leq T$. Auer (2002) and Chu et al. (2011) analyze SupLinUCB, a complicated algorithm using UCB as a subroutine, for this problem. Chu et al. (2011) achieve a regret bound of $O(\sqrt{Td\ln^3(NT\ln(T)/\delta)})$ with probability at least $1-\delta$ (Auer (2002) proves similar results). This regret bound is not applicable to the case of infinite arms, and assumes that context vectors are generated by an *oblivious* adversary. Also, this regret bound would give $O(d^2\sqrt{T})$ regret if N is exponential in d. The state-of-the-art bounds for linear bandits problem in case of finite N are given by Bubeck et al. (2012). They provide an algorithm based on exponential weights, with regret of order $\sqrt{dT\log N}$ for any finite set of N actions. This also gives $O(d\sqrt{T})$ regret when N is exponential in d.

However, none of the above algorithms is efficient when N is large, in particular, when the arms are given by all points in a continuous set of dimension d. The algorithm of Bubeck et al. (2012) requires to maintain a distribution of $O(N)$ support, and those of Chu et al. (2011), Dani et al. (2008), Abbasi-Yadkori et al. (2011) will need to solve an NP-hard problem at every step, even when the set of arms is given by a polytope of d-dimensions. In contrast, the Thompson Sampling algorithm we propose will run in time polynomial in d, as long as the one can efficiently optimize a linear function over the set of arms (maximize $b^T\tilde{\mu}(t)$ for $b \in \mathcal{K}$, where $\mathcal{K}$ is the set of arms). This can be done efficiently, for example, when the set of arms forms a convex set, and even for some combinatorial set of arms. We pay for this efficiency in terms of regret - our regret bounds are $\tilde{O}(d^{3/2}\sqrt{T})$ when N is large or infinite, which is a factor of $\sqrt{d}$ away from the information theoretic lower bound. The only other efficient algorithm for this problem that we are aware of was provided by Dani et al. (2008) (Algorithm 3.2), which also achieves a regret bound of $O(d^{3/2}\sqrt{T})$. Thus, Thompson Sampling achieves the best regret upper bound

achieved by an efficient algorithm in the literature. It is open problem to find a computationally efficient algorithm when N is large or infinite, that achieves the information theoretic lower bound of $O(d\sqrt{T})$ on regret.

Our results demonstrate that the natural and efficient heuristic of Thompson Sampling can achieve theoretical bounds that are close to the best bounds. The main contribution of this paper is to provide new tools for analysis of Thompson Sampling algorithm for contextual bandits, which despite being popular and empirically attractive, has eluded theoretical analysis. 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.

3 Regret Analysis: Proof of Theorem 1

3.1 Challenges and proof 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 Agrawal & Goyal (2012); Kaufmann et al. (2012) do not seem directly applicable. Let us describe some of these difficulties and our novel ideas to resolve them.

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

In general, the basis of regret analysis for stochastic MAB is to prove that the variances of empirical estimates for all arms decrease fast enough, so that the regret incurred until the variances become 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 the suboptimal arm i is played, we know that even though a regret of $\mu_{i^*} - \mu_i \leq 1$ is 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 the empirical mean for the contextual case is given by inverse of $B_i(t) = \sum_{\tau=1:a(\tau)=i}^{t} b_i(\tau)^2$. When a suboptimal arm i is played, if $b_i(t)$ is small, the regret $b_{a^*(t)}(t)\mu_{a^*(t)} - b_i(t)\mu_i$ could be much higher than the improvement $b_i(t)^2$ in $B_i(t)$.

In our proof, we overcome this difficulty by dividing the arms into two groups at any time: saturated and unsaturated arms, based on whether the standard deviation of the estimates for an arm is smaller or larger compared to the standard deviation for the optimal arm. The optimal arm is included in the group of unsaturated arms. We show that for the unsaturated arms, the regret on

playing the arm can be bounded by a factor of the standard deviation, which improves every time the arm is played. This allows us to bound the total regret due to unsaturated arms. For the saturated arms, standard deviation is small, or in other words, the estimates of the means constructed so far are quite accurate in the direction of the current contexts of these arms, so that the algorithm is able to distinguish between them and the optimal arm. We utilize this observation to show that the probability of playing such arms is small, and at every time step an unsaturated arm will be played with some constant probability.

Below is a more technical outline of the proof of Theorem 1. At any time step t, we divide the arms into two groups:

- *saturated arms* defined as those with $\Delta_i(t) > g_t \; s_i(t)$,
- *unsaturated arms* defined as those with $\Delta_i(t) \leq g_t \; s_i(t)$,

where $s_i(t) = \sqrt{b_i(t)^T B(t)^{-1} b_i(t)}$ and g_t, ℓ_t $(g_t > \ell_t)$ are deterministic functions of t, d, δ, defined later. Note that $s_i(t)$ is the standard deviation of the estimate $b_i(t)^T \hat{\mu}(t)$ and $v_t s_i(t)$ is the standard deviation of the random variable $b_i(t)^T \tilde{\mu}(t)$.

We use concentration bounds for $\tilde{\mu}(t)$ and $\hat{\mu}(t)$ to bound the regret at any time t by $g_t(s_{t,a^*(t)} + s_{a(t)}(t))$. Now, if an unsaturated arm is played at time t, then using the definition of unsaturated arms, the regret is at most $g_t s_{a(t)}(t)$. This is useful because of the inequality $\sum_{t=1}^{T} s_{a(t)}(t) = O(\sqrt{Td \ln T})$ (derived along the lines of Auer (2002)), which allows us to bound the total regret due to unsaturated arms.

To bound the regret irrespective of whether a saturated or unsaturated arm is played at time t, we lower bound the probability of playing an unsaturated arm at any time t. More precisely, we define $\mathcal{F}_{t-1}$ as the union of history $\mathcal{H}_{t-1}$ and the contexts $b_i(t), i = 1, \ldots, N$ at time t, and prove that for "most" (in a high probability sense) $\mathcal{F}_{t-1}$,

$$\Pr\left(a(t) \text{ is a unsaturated arm} \,\middle|\, \mathcal{F}_{t-1}\right) \geq p - \tfrac{1}{t^2},$$

where $p = \frac{1}{4e\sqrt{\pi}}$. Note that for p is constant for $\epsilon_t = 1/\ln(t)$. This observation allows us to establish that the expected regret at any time step t is upper bounded in terms of regret due to playing an unsaturated arm at that time, i.e. in terms of $s_{a(t)}(t)$. More precisely, we prove that for "most" $\mathcal{F}_{t-1}$

$$\mathbb{E}\left[\text{regret}(t) \,\middle|\, \mathcal{F}_{t-1}\right] \leq \tfrac{3g_t}{p} \mathbb{E}\left[s_{a(t)}(t) \,\middle|\, \mathcal{F}_{t-1}\right] + \tfrac{2g_t}{pt^2}.$$

We use these observations to establish that $(X_t; t \geq 0)$, where

$$X_t \simeq \text{regret}(t) - \tfrac{3g_t}{p} s_{a(t)}(t) - \tfrac{2g_t}{pt^2},$$

is a super-martingale difference process adapted to filtration $\mathcal{F}_t$. Then, using the Azuma-Hoeffding inequality for super-martingales, along with the inequality $\sum_t s_{a(t)}(t) = O(\sqrt{Td \ln T})$, we will obtain the desired high probability regret bound.

3.2 Formal proof

As mentioned earlier, we will analyze the version of Algorithm 1 that uses $v_t = R\sqrt{9d\ln(\frac{t}{\delta})}$ instead of $v = R\sqrt{9d\ln(\frac{T}{\delta})}$ at time t.

We start with introducing some notations. For quick reference, the notations introduced below also appear in a table of notations at the beginning of the supplementary material.

Definition 1. *For all i, define $\theta_i(t) = b_i(t)^T\tilde{\mu}(t)$, and $s_i(t) = \sqrt{b_i(t)^T B(t)^{-1} b_i(t)}$. By definition of $\tilde{\mu}(t)$ in Algorithm 2, marginal distribution of each $\theta_i(t)$ is Gaussian with mean $b_i(t)^T\hat{\mu}(t)$ and standard deviation $v_t s_i(t)$.*

Definition 2. *Recall that $\Delta_i(t) = b_{a^*(t)}(t)^T\mu - b_i(t)^T\mu$, the difference between the mean reward of optimal arm and arm i at time t.*

Definition 3. *Define $\ell_t = R\sqrt{d\ln\left(\frac{t^3}{\delta}\right)}+1$, $v_t = R\sqrt{9d\ln(\frac{t}{\delta})}$, $g_t = \min\{\sqrt{4d\ln(t)}, \sqrt{4\log(tN)}\,\}v_t + \ell_t$, and $p = \frac{1}{4e\sqrt{\pi}}$.*

Definition 4. *Define $E^\mu(t)$ and $E^\theta(t)$ as the events that $b_i(t)^T\hat{\mu}(t)$ and $\theta_i(t)$ are concentrated around their respective means. More precisely, define $E^\mu(t)$ as the event that*

$$\forall i : |b_i(t)^T\hat{\mu}(t) - b_i(t)^T\mu| \le \ell_t\ s_i(t).$$

Define $E^\theta(t)$ as the event that

$$\forall i : |\theta_i(t) - b_i(t)^T\hat{\mu}(t)| \le \min\{\sqrt{4d\ln(t)}, \sqrt{4\log(tN)}\,\}v_t\ s_i(t).$$

Definition 5. *An arm i is called* saturated *at time t if $\Delta_i(t) > g_t s_i(t)$, and* unsaturated *otherwise. Let $C(t)$ denote the set of saturated arms at time t. Note that the optimal arm is always unsaturated at time t, i.e., $a^*(t) \notin C(t)$. An arm may keep shifting from saturated to unsaturated and vice-versa over time.*

Definition 6. *Define filtration $\mathcal{F}_{t-1}$ as the union of history until time $t-1$, and the contexts at time t, i.e., $\mathcal{F}_{t-1} = \{\mathcal{H}_{t-1}, b_i(t), i = 1, \ldots, N\}$.*

By definition, $\mathcal{F}_1 \subseteq \mathcal{F}_2 \cdots \subseteq \mathcal{F}_{T-1}$. Observe that the following quantities are determined by the history $\mathcal{H}_{t-1}$ and the contexts $b_i(t)$ at time t, and hence are included in $\mathcal{F}_{t-1}$,

- *$\hat{\mu}(t), B(t)$,*
- *$s_i(t)$, for all i,*
- *the identity of the optimal arm $a^*(t)$ and the set of saturated arms $C(t)$,*
- *whether $E^\mu(t)$ is true or not,*
- *the distribution $\mathcal{N}(\hat{\mu}(t), v_t^2 B(t)^{-1})$ of $\tilde{\mu}(t)$, and hence the joint distribution of $\theta_i(t) = b_i(t)^T\tilde{\mu}(t), i = 1, \ldots, N$.*

Lemma 1. *For all t, $0 < \delta < 1$, $\Pr(E^\mu(t)) \ge 1 - \frac{\delta}{t^2}$. And, for all possible filtrations $\mathcal{F}_{t-1}$, $\Pr(E^\theta(t)|\mathcal{F}_{t-1}) \ge 1 - \frac{1}{t^2}$.*

Proof. The complete proof of this lemma appears in Appendix A.3. The probability bound for $E^{\mu}(t)$ will be proven using a concentration inequality given by Abbasi-Yadkori et al. (2011), stated as Lemma 8 in Appendix A.2. The R-sub-Gaussian assumption on rewards will be utilized here. The probability bound for $E^{\theta}(t)$ will be proven using a concentration inequality for Gaussian random variables from Abramowitz & Stegun (1964) stated as Lemma 6 in Appendix A.2 . □

The next lemma lower bounds the probability that $\theta_{a^*(t)}(t) = b_{a^*(t)}(t)^T\tilde{\mu}(t)$ for the optimal arm at time t will exceed its mean reward $b_{a^*(t)}(t)^T\mu$.

Lemma 2. *For any filtration $\mathcal{F}_{t-1}$ such that $E^{\mu}(t)$ is true,*

$$\Pr\left(\theta_{a^*(t)}(t) > b_{a^*(t)}(t)^T\mu \,\middle|\, \mathcal{F}_{t-1}\right) \geq p.$$

Proof. The proof uses anti-concentration of Gaussian random variable $\theta_{a^*(t)}(t) = b_{a^*(t)}(t)^T\tilde{\mu}(t)$, which has mean $b_{a^*(t)}(t)^T\hat{\mu}(t)$ and standard deviation $v_t s_{t,a^*(t)}$, provided by Lemma 6 in Appendix A.2, and the concentration of $b_{a^*(t)}(t)^T\hat{\mu}(t)$ around $b_{a^*(t)}(t)^T\mu$ provided by the event $E^{\mu}(t)$. The details of the proof are in Appendix A.4. □

The following lemma bounds the probability of playing saturated arms in terms of the probability of playing unsaturated arms.

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

$$\Pr\left(a(t) \notin C(t) \,\middle|\, \mathcal{F}_{t-1}\right) \geq p - \frac{1}{t^2}.$$

Proof. The algorithm chooses the arm with the highest value of $\theta_i(t) = b_i(t)^T\tilde{\mu}(t)$ to be played at time t. Therefore, if $\theta_{a^*(t)}(t)$ is greater than $\theta_j(t)$ for all saturated arms, i.e., $\theta_{a^*(t)}(t) > \theta_j(t), \forall j \in C(t)$, then one of the unsaturated arms (which include the optimal arm and other suboptimal unsaturated arms) must be played. Therefore,

$$\begin{aligned}&\Pr\left(a(t) \notin C(t) \,\middle|\, \mathcal{F}_{t-1}\right)\\ &\geq \Pr\left(\theta_{a^*(t)}(t) > \theta_j(t), \forall j \in C(t) \,\middle|\, \mathcal{F}_{t-1}\right). \qquad (3)\end{aligned}$$

By definition, for all saturated arms, i.e. for all $j \in C(t)$, $\Delta_j(t) > g_t s_{t,j}$. Also, if both the events $E^{\mu}(t)$ and $E^{\theta}(t)$ are true then, by the definitions of these events, for all $j \in C(t)$, $\theta_j(t) \leq b_j(t)^T\mu + g_t s_{t,j}$. Therefore, given an $\mathcal{F}_{t-1}$ such that $E^{\mu}(t)$ is true, either $E^{\theta}(t)$ is false, or else for all $j \in C(t)$,

$$\theta_j(t) \leq b_j(t)^T\mu + g_t s_{t,j} \leq b_{a^*(t)}(t)^T\mu.$$

Hence, for any $\mathcal{F}_{t-1}$ such that $E^{\mu}(t)$ is true,

$$\begin{aligned}&\Pr\left(\theta_{a^*(t)}(t) > \theta_j(t), \forall j \in C(t) \,\middle|\, \mathcal{F}_{t-1}\right)\\ \geq\ &\Pr\left(\theta_{a^*(t)}(t) > b_{a^*(t)}(t)^T\mu \,\middle|\, \mathcal{F}_{t-1}\right)\\ &\quad - \Pr\left(\overline{E^{\theta}(t)} \,\middle|\, \mathcal{F}_{t-1}\right)\\ \geq\ &p - \frac{1}{t^2}.\end{aligned}$$

The last inequality uses Lemma 2 and Lemma 1. □

Lemma 4. *For any filtration $\mathcal{F}_{t-1}$ such that $E^{\mu}(t)$ is true,*

$$\mathbb{E}\left[\Delta_{a(t)}(t) \,\middle|\, \mathcal{F}_{t-1}\right] \leq \frac{3g_t}{p}\ \mathbb{E}\left[s_{a(t)}(t) \,\middle|\, \mathcal{F}_{t-1}\right] + \frac{2g_t}{pt^2}.$$

Proof. Let $\bar{a}(t)$ denote the unsaturated arm with smallest $s_i(t)$, i.e.

$$\bar{a}(t) = \arg\min_{i \notin C(t)} s_i(t)$$

Note that since $C(t)$ and $s_i(t)$ for all i are fixed on fixing $\mathcal{F}_{t-1}$, so is $\bar{a}(t)$.

Now, using Lemma 3, for any $\mathcal{F}_{t-1}$ such that $E^{\mu}(\theta)$ is true,

$$\begin{aligned}
\mathbb{E}\left[s_{a(t)}(t) \,\middle|\, \mathcal{F}_{t-1}\right] &\geq \mathbb{E}\left[s_{a(t)}(t) \,\middle|\, \mathcal{F}_{t-1}, a(t) \notin C(t)\right] \\
&\quad \cdot \Pr\left(a(t) \notin C(t) \,\middle|\, \mathcal{F}_{T-1}\right) \\
&\geq s_{t,\bar{a}(t)}\left(p - \frac{1}{t^2}\right).
\end{aligned}$$

Now, if events $E^{\mu}(t)$ and $E^{\theta}(t)$ are true, then for all i, by definition, $\theta_i(t) \leq b_i(t)^T\mu + g_t s_i(t)$. Using this observation along with the fact that $\theta_{a(t)}(t) \geq \theta_i(t)$ for all i,

$$\begin{aligned}
\Delta_{a(t)}(t) &= \Delta_{\bar{a}(t)}(t) + (b_{\bar{a}(t)}(t)^T\mu - b_{a(t)}(t)^T\mu) \\
&\leq \Delta_{\bar{a}(t)}(t) + (\theta_{\bar{a}(t)}(t) - \theta_{a(t)}(t)) \\
&\quad + g_t s_{t,\bar{a}(t)} + g_t s_{a(t)}(t) \\
&\leq \Delta_{\bar{a}(t)}(t) + g_t s_{t,\bar{a}(t)} + g_t s_{a(t)}(t) \\
&\leq g_t s_{t,\bar{a}}(t) + g_t s_{t,\bar{a}(t)} + g_t s_{a(t)}(t)
\end{aligned}$$

Therefore, for any $\mathcal{F}_{t-1}$ such that $E^{\mu}(\theta)$ is true either $\Delta_{a(t)}(t) \leq 2g_t s_{t,\bar{a}}(t) + g_t s_{a(t)}(t)$ or $E^{\theta}(t)$ is false. Therefore,

$$\begin{aligned}
\mathbb{E}\left[\Delta_{a(t)}(t) \,\middle|\, \mathcal{F}_{t-1}\right] &\leq \mathbb{E}\left[2\ g_t s_{t,\bar{a}}(t) + g_t s_{a(t)}(t) \,\middle|\, \mathcal{F}_{t-1}\right] \\
&\quad + \Pr\left(\overline{E^{\theta}(t)}\right) \\
&\leq \frac{2\ g_t}{\left(p - \frac{1}{t^2}\right)}\ \mathbb{E}\left[s_{t,a(t)} \,\middle|\, \mathcal{F}_{t-1}\right] \\
&\quad + g_t \mathbb{E}\left[s_{t,a(t)} \,\middle|\, \mathcal{F}_{t-1}\right] + \frac{1}{t^2} \\
&\leq \frac{3}{p}\ g_t\ \mathbb{E}\left[s_{t,a(t)} \,\middle|\, \mathcal{F}_{t-1}\right] + \frac{2g_t}{pt^2}.
\end{aligned}$$

In the first inequality we used that for all i, $\Delta_i(t) \leq 1$. The second inequality used the inequality derived in the beginning of this proof, and Lemma 1 to apply $\Pr\left(\overline{E^{\theta}(t)}\right) \leq \frac{1}{t^2}$. The third inequality used the observation that $0 \leq s_{t,a(t)} \leq ||b_{a(t)}(t)|| \leq 1$.

□

Definition 7. *Recall that regret(t) was defined as, regret(t) = $\Delta_{a(t)}(t) = b_{a^*(t)}(t)^T\mu - b_{a(t)}(t)^T\mu$. Define regret$'(t)$ = regret$(t) \cdot I(E^\mu(t))$.*

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

Definition 8. *Let*

$$\begin{aligned} X_t &= \text{regret}'(t) - \tfrac{3g_t}{p} s_{a(t)}(t) - \frac{2g_t}{pt^2} \\ Y_t &= \textstyle\sum_{w=1}^{t} X_w. \end{aligned}$$

Lemma 5. *$(Y_t; t = 0, \ldots, T)$ is a super-martingale process with respect to filtration $\mathcal{F}_t$.*

Proof. See Definition 9 in Appendix A.2 for the definition of super-martingales. We need to prove that for all $t \in [1, T]$, and any $\mathcal{F}_{t-1}$, $\mathbb{E}[Y_t - Y_{t-1}|\mathcal{F}_{t-1}] \leq 0$, i.e.

$$\mathbb{E}\left[\text{regret}'(t) \,\middle|\, \mathcal{F}_{t-1}\right] \leq \frac{3g_t}{p}\mathbb{E}\left[s_{a(t)}(t) \,\middle|\, \mathcal{F}_{t-1}\right] + \frac{2g_t}{pt^2}.$$

Note that whether $E^\mu(t)$ is true or not is completely determined by $\mathcal{F}_{t-1}$. If $\mathcal{F}_{t-1}$ is such that $E^\mu(t)$ is not true, then regret$'(t) =$ regret$(t) \cdot I(E^\mu(t)) = 0$, and the above inequality holds trivially. And, for $\mathcal{F}_{t-1}$ such that $E^\mu(t)$ holds, the inequality follows from Lemma 4. □

Now, we are ready to prove Theorem 1.

Proof of Theorem 1 Note that X_t is bounded, $|X_t| \leq 1 + \frac{3}{p}g_t + \frac{2}{pt^2}g_t \leq \frac{6}{p}g_t$. Thus, we can apply Azuma-Hoeffding inequality (see Lemma 7 in Appendix A.2), to obtain that with probability $1 - \frac{\delta}{2}$,

$$\textstyle\sum_{t=1}^{T} \text{regret}'(t) \;\leq\; \sum_{t=1}^{T} \frac{3g_t}{p} s_{a(t)}(t) + \sum_{t=1}^{T} \frac{2g_t}{pt^2} + \sqrt{2\left(\sum_t \frac{36g_t^2}{p^2}\right)\ln(\tfrac{2}{\delta})} \tag{4}$$

Note that p is a constant. Also, by definition, $g_t \leq g_T$. Therefore, from above equation, with probability $1 - \frac{\delta}{2}$,

$$\textstyle\sum_{t=1}^{T} \text{regret}'(t) \;\leq\; \frac{3g_T}{p}\sum_{t=1}^{T} s_{a(t)}(t) + \frac{2g_T}{p}\sum_{t=1}^{T}\frac{1}{t^2} + \frac{6g_T}{p}\sqrt{2T\ln(\tfrac{2}{\delta})}$$

Now, we can use $\sum_{t=1}^{T} s_{a(t)}(t) \leq 5\sqrt{dT\ln T}$, which can be derived along the lines of Lemma 3 of Chu et al. (2011) using Lemma 11 of Auer (2002) (see Appendix A.5 for details). Also, by definition $g_T = O(\sqrt{d\ln(\frac{T}{\delta})}\cdot(\min\{\sqrt{d}, \sqrt{\log(N)}\}))$

(see the Table of notations in the beginning of the supplementary material). Substituting in above, we get

$$\begin{aligned}\sum_{t=1}^{T} \text{regret}'(t) &= O\left(d\sqrt{\ln(\tfrac{T}{\delta})} \cdot (\min\{\sqrt{d}, \sqrt{\log(N)}\})) \cdot \sqrt{dT \ln T}\right) \\ &= O\left(d\sqrt{T} \cdot (\min\{\sqrt{d}, \sqrt{\log(N)}\})) \cdot \left(\ln(T) + \sqrt{\ln(T)\ln(\tfrac{1}{\delta})}\right)\right).\end{aligned}$$

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

$$\begin{aligned}\mathcal{R}(T) &= \textstyle\sum_{t=1}^{T} \text{regret}(t) = \sum_{t=1}^{T} \text{regret}'(t) \\ &= O\left(d\sqrt{T} \cdot (\min\{\sqrt{d}, \sqrt{\log(N)}\})) \cdot \left(\ln(T) + \sqrt{\ln(T)\ln(\tfrac{1}{\delta})}\right)\right).\end{aligned}$$

The proof for the alternate definition of regret mentioned in Remark 1 is provided in Appendix A.5.

4 Conclusions

We provided a theoretical analysis of Thompson Sampling for the stochastic contextual bandits problem with linear payoffs. Our results resolve some 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 close to the state-of-the-art methods. We used a novel martingale-based analysis technique which is arguably simpler than the techniques in the past work on TS (Agrawal & Goyal, 2012; Kaufmann et al., 2012), and is amenable to extensions.

In the algorithm in this paper, Gaussian priors were used, so that $\tilde{\mu}(t)$ was generated from a Gaussian distribution. However, the analysis techniques in this paper are extendable to an algorithm that uses a prior 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 (Lemma 6), which were used to prove Lemma 1 and Lemma 2 respectively. If any other distribution provides similar tail inequalities, to allow us proving these lemmas, these can be used as a black box in the analysis, and the regret bounds can be reproduced for that distribution.

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 Filippi et al. (2010), the setting with delayed and batched feedback, 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.

Acknowledgements This is an extended and slightly modified version of Agrawal & Goyal (2013a).

References

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

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

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

Agrawal, Shipra and Goyal, Navin. Thompson Sampling for Contextual Bandits with Linear Payoffs. *ICML*, 2013a.

Agrawal, Shipra and Goyal, Navin. Further Optimal Regret Bounds for Thompson Sampling. *AISTATS*, 2013b.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

May, Benedict C., Korda, Nathan, Lee, Anthony, and Leslie, David S. Optimistic Bayesian sampling in contextual-bandit problems. Technical Report 11:01, Statistics Group, Department of Mathematics, University of Bristol, 2011.

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

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

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

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

Strens, Malcolm J. A. A Bayesian Framework for Reinforcement Learning. In *ICML*, pp. 943–950, 2000.

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

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

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

Nomenclature

$a(t)$ The arm played at time t

$a^*(t)$ The optimal arm at time t

$B(t)$ $= I_d + \sum_{\tau=1}^{t-1} b_{a(\tau)}(\tau) b_{a(\tau)}(\tau)^T$

$b_i(t)$ context vector for arm i at time t

$C(t)$ The set of saturated arms at time t.

d The dimension of context vectors

$\Delta_i(t)$ $= b_{a^*(t)}(t)^T \mu - b_i(t)^T \mu$

$E^\mu(t)$ Event $\forall i : |b_i(t)^T \hat{\mu}(t) - b_i(t)^T \mu| \leq \ell_t s_i(t)$

$E^\theta(t)$ Event $\forall i : |\theta_i(t) - b_i(t)^T \hat{\mu}(t)| \leq \min\{\sqrt{4d\ln(t)}, \sqrt{4\log(tN)}\ \} v_t s_i(t)$

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

g_t $= \min\{\sqrt{4d\ln(t)}, \sqrt{4\log(tN)}\ \} v_t + \ell_t$

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

ℓ_t $= R\sqrt{d \ln\left(\frac{t^3}{\delta}\right)} + 1$

μ The unknown d-dimensional parameter

$\hat{\mu}(t)$ $= B(t)^{-1} \left(\sum_{\tau=1}^{t-1} b_{a(\tau)}(\tau) r_{a(\tau)}(\tau) \right)$ (Empirical estimate of mean at time t)

$\tilde{\mu}(t)$ d-dimensional sample generated by from distribution $\mathcal{N}(\hat{\mu}(t), v_t^2 B(t)^{-1})$.

N number of arms

p_t $= \frac{1}{4e\sqrt{\pi}}$

$r_i(t)$ Reward for arm i at time t

regret(t) Regret at time t

$s_i(t)$ $= \sqrt{b_i(t)^T B(t)^{-1} b_i(t)}$

saturated arm any arm i with $\Delta_i(t) > g_t s_i(t)$.

$\theta_i(t)$ $= b_i(t)^T \tilde{\mu}(t)$

v_t $= R\sqrt{9d \ln(\frac{t}{\delta})}$

A

A.1 Posterior distribution computation

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

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

A.2 Some concentration inequalities

Formula 7.1.13 from Abramowitz & Stegun (1964) can be used to derive the following concentration and anti-concentration inequalities for Gaussian distributed random variables.

Lemma 6. *(Abramowitz & Stegun, 1964) For a Gaussian distributed random variable Z with mean m and variance σ^2, for any $z \geq 1$,*

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

Definition 9 (Super-martingale)**.** *A sequence of random variables $(Y_t; t \geq 0)$ is called a super-martingale corresponding to filtration $\mathcal{F}_t$, if for all t, Y_t is $\mathcal{F}_t$-measurable, and for $t \geq 1$,*

$$\mathbb{E}\left[Y_t - Y_{t-1} \,\middle|\, \mathcal{F}_{t-1}\right] \leq 0.$$

Lemma 7 (Azuma-Hoeffding inequality)**.** *If a super-martingale $(Y_t; t \geq 0)$, corresponding to filtration $\mathcal{F}_t$, satisfies $|Y_t - Y_{t-1}| \leq c_t$ for some constant c_t, for all $t = 1, \ldots, T$, then for any $a \geq 0$,*

$$\Pr(Y_T - Y_0 \geq a) \leq e^{-\frac{a^2}{2\sum_{t=1}^{T} c_t^2}}.$$

The following lemma is implied by Theorem 1 in Abbasi-Yadkori et al. (2011):

Lemma 8. *(Abbasi-Yadkori et al., 2011) 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. For $t \geq 0$, define $\xi_t = \sum_{\tau=1}^{t} m_\tau \eta_\tau$ and $M_t = I_d + \sum_{\tau=1}^{t} m_\tau m_\tau^T$, 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 $1 - \delta'$,

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

where $||\xi_t||_{M_t^{-1}} = \sqrt{\xi_t^T M_t^{-1} \xi_t}$.

A.3 Proof of Lemma 1

Bounding the probability of event $E^\mu(t)$: We use Lemma 8 with $m_t = b_{a(t)}(t)$, $\eta_t = r_{a(t)}(t) - b_{a(t)}(t)^T\mu$, $\mathcal{F}'_t = (a(\tau+1), m_{\tau+1}, \eta_\tau : \tau \leq t)$. (Note that effectively, $\mathcal{F}'_t$ has all the information, including the arms played, until time $t+1$, except for the reward of the arm played at time $t+1$). By the definition of $\mathcal{F}'_t$, m_t is $\mathcal{F}'_{t-1}$-measurable, and η_t is $\mathcal{F}'_t$-measurable. Also, η_t is conditionally R-sub-Gaussian due to the assumption mentioned in the problem settings (refer to Section 2.1), and is a martingale difference process:

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

Also, this makes

$$M_t = I_d + \sum_{\tau=1}^{t} m_\tau m_\tau^T = I_d + \sum_{\tau=1}^{t} b_{a(\tau)}(\tau) b_{a(\tau)}(\tau)^T,$$

$$\xi_t = \sum_{\tau=1}^{t} m_\tau \eta_\tau = \sum_{\tau=1}^{t} b_{a(\tau)}(\tau)(r_{a(\tau)} - b_{a(\tau)}(\tau)^T\mu).$$

Note that $B(t) = M_{t-1}$, and $\hat{\mu}(t) - \mu = M_{t-1}^{-1}(\xi_{t-1} - \mu)$. Let for any vector $y \in \mathbb{R}$ and matrix $A \in \mathbb{R}^{d \times d}$, $||y||_A$ denote $\sqrt{y^T A y}$. Then, for all i,

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

The inequality holds because M_{t-1}^{-1} is a positive definite matrix. Using Lemma 8, 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,

$$\begin{aligned} & |b_i(t)^T \hat{\mu}(t) - b_i(t)^T \mu| \\ \leq \; & ||b_i(t)||_{B(t)^{-1}} \cdot \left(R\sqrt{d \ln\left(\frac{t}{\delta'}\right)} + 1 \right) \\ \leq \; & ||b_i(t)||_{B(t)^{-1}} \cdot \left(R\sqrt{d \ln\left(\frac{t^3}{\delta}\right)} + 1 \right) \\ = \; & \ell_t \; s_i(t). \end{aligned}$$

This proves the bound on the probability of $E^{\mu}(t)$.

Bounding the probability of event $E^{\theta}(t)$: Given any filtration $\mathcal{F}_{t-1}$, $b_i(t), B(t)$ are fixed. Then,

$$\begin{aligned} |\theta_i(t) - b_i(t)^T \hat{\mu}(t)| &= |b_i(t)^T(\tilde{\mu}(t) - \hat{\mu}(t))| \\ &= |b_i(t)^T B(t)^{-1/2} B(t)^{1/2}(\tilde{\mu}(t) - \hat{\mu}(t))| \\ &\leq v_t \sqrt{b_i(t)^T B(t)^{-1} b_i(t)} \cdot \left\| \left(\frac{1}{v_t} B(t)^{1/2}(\tilde{\mu}(t) - \hat{\mu}(t)) \right) \right\|_2 \\ &= v_t s_i(t) ||\zeta||_2 \\ &\leq v_t s_i(t) \sqrt{4d \ln t} \end{aligned}$$

with probability $1 - \frac{1}{t^2}$. Here, $\zeta_k, k = 1, \ldots, d$ denotes standard univariate normal random variable (mean 0 and variance 1).

Alternatively, we can bound $|\theta_i(t) - b_i(t)^T \hat{\mu}(t)|$ for every i by considering that $\theta_i(t)$ Gaussian random variable with mean $b_i(t)^T \hat{\mu}(t)$ and variance $v_t^2 s_i(t)^2$. Therefore, using Lemma 6, for every i

$$|\theta_i(t) - b_i(t)^T \hat{\mu}(t)| \;=\; \sqrt{4 \ln(Nt)} s_i(t)$$

with probability $1 - \frac{1}{Nt^2}$. Taking union bound over $i = 1, \ldots, N$, we obtain that $|\theta_i(t) - b_i(t)^T \hat{\mu}(t)| \leq \sqrt{4 \ln(Nt)} s_i(t)$ holds for all arms with probability $1 - \frac{1}{t^2}$.

Combined, the two bounds give that $E^{\theta}(t)$ holds with probability $1 - \frac{1}{t^2}$.

A.4 Proof of Lemma 2

Given event $E^{\mu}(t)$, $|b_{a^*(t)}(t)^T \hat{\mu}(t) - b_{a^*(t)}(t)^T \mu| \leq \ell_t s_{a^*(t)}(t)$. And, since Gaussian random variable $\theta_{a^*(t)}(t)$ has mean $b_{a^*(t)}(t)^T \hat{\mu}(t)$ and standard deviation

$v_t s_{a^*(t)}(t)$, using anti-concentration inequality in Lemma 6,

$$\begin{aligned} & \Pr\left(\theta_{a^*(t)}(t) \geq b_{a^*(t)}(t)^T\mu \,\middle|\, \mathcal{F}_{t-1}\right) \\ = \; & \Pr\Big(\tfrac{\theta_{a^*(t)}(t)-b_{a^*(t)}(t)^T\hat{\mu}(t)}{v_t s_{t,a^*(t)}} \\ & \qquad \geq \tfrac{b_{a^*(t)}(t)^T\mu-b_{a^*(t)}(t)^T\hat{\mu}(t)}{v_t s_{t,a^*(t)}} \,\Big|\, \mathcal{F}_{t-1}\Big) \\ \geq \; & \tfrac{1}{4\sqrt{\pi}} e^{-Z_t^2}. \end{aligned}$$

where

$$\begin{aligned} |Z_t| &= \left|\frac{b_{a^*(t)}(t)^T\mu - b_{a^*(t)}(t)^T\hat{\mu}(t)}{v_t s_{a^*(t)}(t)}\right| \\ &\leq \frac{\ell_t s_{a^*(t)}(t)}{v_t s_{a^*(t)}(t)} \\ &= \frac{\left(R\sqrt{d\ln\left(\frac{t^3}{\delta}\right)}+1\right)}{R\sqrt{9d\ln(\frac{t}{\delta})}} \\ &\leq 1. \end{aligned}$$

So

$$\Pr\left(\theta_{a^*(t)}(t) \geq b_{a^*(t)}(t)^T\mu \,\middle|\, \mathcal{F}_{t-1}\right) \geq \tfrac{1}{4e\sqrt{\pi}}.$$

A.5 Missing details from Section 3.2

To derive the inequality $\sum_{t=1}^{T} s_{a(t)}(t) \leq 5\sqrt{dT\ln T}$, we use the following result, implied by the referred lemma in Auer (2002).

Lemma 9. *(Auer, 2002, 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 $B(t+1) = B(t) + b_{a(t)}(t)b_{a(t)}(t)^T$, and $\lambda_{j,t} \geq 1, \forall j$. Therefore, above implies

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

This allows us to derive the given inequality after some algebraic computations following along the lines of Lemma 3 of Chu et al. (2011).

To obtain bounds for the other definition of regret in Remark 1, we observe that because $\mathbb{E}[r_i(t)|\mathcal{F}_{t-1}] = b_i(t)^T\mu$ for all i, the expected value of regret$'(t)$ given $\mathcal{F}_{t-1}$ for this definition of regret(t) is same as before. More precisely, for $\mathcal{F}_{t-1}$ such that $E^\mu(t)$ holds,

$$\begin{aligned} & \mathbb{E}\left[\text{regret}'(t) \,\middle|\, \mathcal{F}_{t-1}\right] \\ = \; & \mathbb{E}\left[\text{regret}(t) \,\middle|\, \mathcal{F}_{t-1}\right] \\ = \; & \mathbb{E}\left[r_{a^*(t)}(t) - r_{a(t)}(t) \,\middle|\, \mathcal{F}_{t-1}\right] \\ = \; & \mathbb{E}\left[b_{a^*(t)}(t)^T\mu - b_{a(t)}(t)^T\mu \,\middle|\, \mathcal{F}_{t-1}\right]. \end{aligned}$$

And, $\mathbb{E}\left[\text{regret}'(t) \,\middle|\, \mathcal{F}_{t-1}\right] = 0$ for other $\mathcal{F}_{t-1}$. Therefore, Lemma 5 holds as it is, and Y_t defined in Definition 8 is a super-martingale with respect to this new definition of regret(t) as well. Now, if $|r_i(t) - b_i(t)^T\mu| \leq R$, for all i, then $|\text{regret}'(t)| \leq 2R$ and $|Y_t - Y_{t-1}| \leq \frac{6}{p}\frac{g_t^2}{\ell_t} + 2R$, and we can apply Azuma-Hoeffding inequality exactly as in the proof of Theorem 1 to obtain regret bounds of the same order as Theorem 1 for the new definition. The results extend to the more general R-sub-Gaussian condition on $r_i(t)$, using a simple extension of Azuma-Hoeffding inequality; we omit the proof of that extension.