Spectral Thompson Sampling

Tomáš Kocák, Michal Valko, Rémi Munos, Shipra Agrawal

Spectral Thompson Sampling

Tomáš Kocák
SequeL team
INRIA Lille - Nord Europe
France

Michal Valko
SequeL team
INRIA Lille - Nord Europe
France

Rémi Munos
SequeL team
INRIA Lille, France
Microsoft Research NE, USA

Shipra Agrawal
ML and Optimization Group
Microsoft Research
Bangalore, India

Abstract

Thompson Sampling (TS) has surged a lot of interest due to its good empirical performance, in particular in the computational advertising. Though successful, the tools for its performance analysis appeared only recently. In this paper, we describe and analyze SpectralTS algorithm for a bandit problem, where the payoffs of the choices are *smooth* given an underlying graph. In this setting, each choice is a node of a graph and the expected payoffs of the neighboring nodes are assumed to be similar. Although the setting has application both in recommender systems and advertising, the traditional algorithms would scale poorly with the number of choices. For that purpose we consider an *effective dimension* d, which is small in real-world graphs. We deliver the analysis showing that the regret of SpectralTS scales as $d\sqrt{T \ln N}$ with high probability, where T is the time horizon and N is the number of choices. Since a $d\sqrt{T \ln N}$ regret is comparable to the known results, SpectralTS offers a computationally more efficient alternative. We also show that our algorithm is competitive on both synthetic and real-world data.

1 Introduction

Thompson Sampling (Thompson, 1933) is one of the oldest heuristics for sequential problems with limited feedback, also known as *bandit problems*. It solves the exploration-exploitation dilemma by a simple and intuitive rule: when choosing the next action to play, *choose it according to probability that it is the best one*; that is the one that *maximizes the expected payoff*. Using this heuristic, it is straightforward to design many bandit algorithms, such as the SpectralTS algorithm presented in this paper.

What is challenging though, is to provide the analysis and prove *performance guarantees* for TS algorithms. This may be the reason, why TS has not been in the center of interest of sequential machine learning, where mostly optimistic algorithms were studied (Auer, Cesa-Bianchi, and Fischer, 2002; Auer, 2002). Nevertheless, the past few years witnessed the rise of interest in TS due to its empirical performance, in particular in the computational advertising (Chapelle and Li, 2011), a major source of income for Internet companies. This motivated the researchers to explain the success of TS.

A major breakthrough in this aspect was the work of Agrawal and Goyal (2012a), who provided the first finite-time analysis of TS. It was shortly after followed by a refined analysis (Kaufmann, Korda, and Munos, 2012) showing the optimal performance of TS for Bernoulli distributions. Some of the asymptotic results for TS were proved by May et al. (2012). Agrawal and Goyal (2013a) then provided distribution-independent analysis of TS for the multi-arm bandit. The most relevant results for our work are by Agrawal and Goyal (2013b), who bring a new martingale technique, enabling us to analyze cases where the payoffs of the actions are linear in some basis. Later, Korda, Kaufmann, and Munos (2013) extended the known optimal TS analysis to 1-dimensional exponential family distributions. Finally, in the Bayesian setting, Russo and Van Roy (2013, 2014) analyzed TS with respect to the *Bayesian risk*.

In our prior work (Valko et al., 2014), we introduced a *spectral bandit* setting, relevant for *content-based recommender systems* (Pazzani and Billsus, 2007), where the payoff function is expressed as a linear combination of a *smooth basis*. In such systems, we aim to recommend a content to a user, based on her personal preferences. Recommender systems take advantage of the *content similarity* in order to offer relevant suggestions. In other words, we assume that the user preferences are smooth on the similarity graph of content items. The results from *manifold learning* (Belkin, Niyogi, and Sindhwani, 2006) show that the eigenvectors related to the smallest eigenvalues of the similarity graph Laplacian offer a useful smooth basis. Another example of leveraging useful structural property present in the real-world data is to take advantage of the hierarchy of the content features (Yue, Hong, and Guestrin, 2012).

Although LinUCB (Li et al., 2010), GP-UCB (Srinivas et al., 2010), and LinearTS (Agrawal and Goyal, 2013b) could be used for the spectral bandit setting, they would not scale well with the number of possible items to recommend. This is why we defined (Valko et al., 2014) the *effective dimension* d, likely to be small in real-world graphs, and provided an algorithm based on the optimistic principle: SpectralUCB. In this paper, we focus on the TS alternative: *Spectral Thompson Sampling* (Table 1). The algorithm is easy to obtain, since there is no need to derive the upper confidence bounds. Furthermore, one of the main benefits of SpectralTS is its computational efficiency.

	Linear	*Spectral*
Optimistic Approach D^2N per step update	**LinUCB** $D\sqrt{T\ln T}$	**SpectralUCB** $d\sqrt{T\ln T}$
Thompson Sampling $D^2 + DN$ per step update	**LinearTS** $D\sqrt{T\ln N}$	***SpectralTS*** $d\sqrt{T\ln N}$

Table 1: Linear vs. Spectral Bandits

The main contribution of this paper is the finite-time analysis of SpectralTS. We prove that the regret of SpectralTS scales as $d\sqrt{T\ln N}$, which is comparable to the known results. Although the regret is $\sqrt{\ln N}$ away from the one of SpectralUCB, this factor is negligible for the relevant applications, e.g., movie recommendation. Interestingly, even for the linear case, there is no polynomial time algorithm for linear contextual bandits with better than $D\sqrt{T\ln N}$ regret (Agrawal and Goyal, 2012b), where D is the dimension of the context vector. Optimistic approach (UCB) for linear contextual bandits is not polynomially implementable, where the numbers of choices are at least exponential in D (e.g., when set of arms is all the vectors in a polytope) and the approximation given by Dani, Hayes, and Kakade (2008) achieves only $D^{3/2}\sqrt{T}$ regret. Similarly for the spectral bandit case, SpectralTS offers a computationally attractive alternative to SpectralUCB (Table 1, left). Instead of computing the upper confidence bounds for all the arms in each step, we only need to sample from our current belief of the true model and perform the maximization given this belief. We support this claim with an empirical evaluation.

2 Setting

In this section, we formally define the *spectral bandit* setting. The most important notation is summarized in Table 2. Let $\mathcal{G}$ be the given graph with the set $\mathcal{V}$ of N nodes. Let $\mathcal{W}$ be the $N \times N$ matrix of edge weights and $\mathcal{D}$ is a $N \times N$ diagonal matrix with the entries $d_{ii} = \sum_j w_{ij}$. The graph Laplacian of $\mathcal{G}$ is a $N \times N$ matrix defined as $\mathcal{L} = \mathcal{D} - \mathcal{W}$. Let $\mathcal{L} = \mathbf{Q}\mathbf{\Lambda}_{\mathcal{L}}\mathbf{Q}^\intercal$ be the eigendecomposition of $\mathcal{L}$, where $\mathbf{Q}$ is $N \times N$ orthogonal matrix with eigenvectors of $\mathcal{L}$ in columns. Let $\{\mathbf{b}_i\}_{1 \le i \le N}$ be the N rows of $\mathbf{Q}$. This way, each $\mathbf{b}_i$ corresponds to the features of action i (commonly referred to as the *arm* i) in the *spectral basis*. The reason for spectral basis comes from manifold learning (Belkin, Niyogi, and Sindhwani, 2006). In our setting, we assume that the neighboring content has similar payoffs, which means that the payoff function is smooth on $\mathcal{G}$. Belkin, Niyogi, and Sindhwani (2006) showed that that smooth functions can be represented as a linear combination of eigenvectors with small eigenvalues. This explains the choice of regularizer in Section 3.

We now describe the learning setting. Each time t, the recommender chooses an action $a(t)$ and receives a payoff that is in expectation *linear* in the associated features $\mathbf{b}_{a(t)}$,

$$r(t) = \mathbf{b}_{a(t)}^\intercal \boldsymbol{\mu} + \varepsilon_t,$$

where $\boldsymbol{\mu}$ encodes the (unknown) vector of user preferences and ε_t is R-sub-Gaussian noise, i.e.,

$$\forall \xi \in \mathbb{R},\ \mathbb{E}[e^{\xi\varepsilon_t} \mid \{\mathbf{b}_i\}_{i=1}^N, \mathcal{H}_{t-1}] \le \exp\left(\frac{\xi^2 R^2}{2}\right).$$

- N – number of arms
- $\mathbf{b}_i$ – feature vector of arm i
- d – effective dimension
- $a(t)$ – arm played at time t
- a^* – optimal arm
- λ – regularization parameter
- C – upperbound on $\|\boldsymbol{\mu}\|_{\mathbf{\Lambda}}$
- $\boldsymbol{\mu}$ – true (unknown) vector of weights
- $v = R\sqrt{6d\ln((\lambda+T)/(\delta\lambda)} + C$
- $p = 1/(4e\sqrt{\pi})$
- $l = R\sqrt{2d\ln((\lambda+T)T^2/(\delta\lambda))} + C$
- $g = v\sqrt{4\ln TN} + l$
- $\Delta_i = \mathbf{b}_{a^*}^\intercal \mu - \mathbf{b}_i^\intercal \mu$

Table 2: Overview of the notation.

The history $\mathcal{H}_{t-1}$ is defined as:

$$\mathcal{H}_{t-1} = \{a(\tau),\, r(\tau),\, \tau = 1, \ldots, t-1\}.$$

We now define the performance metric of any algorithm for the setting above. The instantaneous (pseudo)-regret in time t is defined as the difference between the mean payoff (reward) of the optimal arm a^* and the arm $a(t)$ chosen by the recommender,

$$\text{regret}(t) = \Delta_{a(t)} = \mathbf{b}_{a^*}^\intercal \boldsymbol{\mu} - \mathbf{b}_{a(t)}^\intercal \boldsymbol{\mu}.$$

The performance of any algorithm is measured in terms of cumulative regret, which is the sum of regrets over time,

$$\mathcal{R}(T) = \sum_{t=1}^{T} \text{regret}(t).$$

3 Algorithm

In this paper, we use TS to decide which arm to play. Specifically, we represent our current knowledge about $\boldsymbol{\mu}$ as the normal distribution $\mathcal{N}(\hat{\boldsymbol{\mu}}(t), v^2\mathbf{B}_t^{-1})$, where $\hat{\boldsymbol{\mu}}(t)$ is our actual approximation of the unknown parameter $\boldsymbol{\mu}$ and $v^2\mathbf{B}_t^{-1}$ reflects our uncertainty about it. As mentioned, we assume that the reward function is a linear combination of eigenvectors of $\mathcal{L}$ with large coefficients corresponding to the eigenvectors with small eigenvalues. We encode this assumption into our initial confidence ellipsoid by setting $\mathbf{B}_1 = \mathbf{\Lambda} = \mathbf{\Lambda}_{\mathcal{L}} + \lambda\mathbf{I}_N$, where λ is a regularization parameter.

After that, every time step t we generate a sample $\tilde{\boldsymbol{\mu}}(t)$ from the distribution $\mathcal{N}(\hat{\boldsymbol{\mu}}(t), v^2\mathbf{B}_t^{-1})$ and chose an arm $a(t)$ that maximizes $\mathbf{b}_i^\intercal\tilde{\boldsymbol{\mu}}(t)$. After receiving a reward, we update our estimate of $\boldsymbol{\mu}$ and the confidence of it, i.e., we compute $\hat{\boldsymbol{\mu}}(t+1)$ and $\mathbf{B}(t+1)$,

$$\mathbf{B}_{t+1} = \mathbf{B}_t + \mathbf{b}_{a(t)}\mathbf{b}_{a(t)}^\intercal$$

$$\hat{\boldsymbol{\mu}}(t+1) = \mathbf{B}_{t+1}^{-1}\left(\sum_{i=1}^{t} \mathbf{b}_{a(i)} r(i)\right).$$

Remark 1. *Since TS is a Bayesian approach, it requires a prior to run and we choose it here to be a Gaussian. However, this does not pose any assumption whatsoever about*

the actual data both for the algorithm and the analysis. The only assumptions we make about the data are: (a) that the mean payoff is linear in the features, (b) that the noise is sub-Gaussian, and (c) that we know a bound on the Laplacian norm of the mean reward function. We provide a frequentist bound on the regret (and not an average over the prior) which is a much stronger worst case result.

The computational advantage of SpectralTS in Algorithm 1, compared to SpectralUCB, is that we do not need to compute the confidence bound for each arm. Indeed, in SpectralTS we need to sample $\tilde{\boldsymbol{\mu}}$ which can be done in N^2 time (note that $\mathbf{B}_t$ is only changing by a rank one update) and a maximum of $\mathbf{b}_i^\top \tilde{\boldsymbol{\mu}}$ which can be also done in N^2 time. On the other hand, in SpectralUCB, we need to compute a $\mathbf{B}_t^{-1}$ norm for each of N feature vectors which amounts to a ND^2 time. Table 1 (left) summarizes the computational complexity of the two approaches. Notice that in our setting $D = N$, which comes to a N^2 vs. N^3 time per step. We support this argument in Section 5. Finally note, that the eigendecomposition needs to be done only once in the beginning and since $\mathcal{L}$ is diagonally dominant, this can be done for N in millions (Koutis, Miller, and Peng, 2010).

Algorithm 1 Spectral Thompson Sampling

Input:
N: number of arms, T: number of pulls
$\{\boldsymbol{\Lambda}_{\mathcal{L}}, \mathbf{Q}\}$: spectral basis of graph Laplacian $\mathcal{L}$
λ, δ: regularization and confidence parameters
R, C: upper bounds on noise and $\|\boldsymbol{\mu}\|_{\boldsymbol{\Lambda}}$
Initialization:
$v = R\sqrt{6d\ln((\lambda+T)/\delta\lambda)} + C$
$\hat{\boldsymbol{\mu}} = 0_N$, $\boldsymbol{f} = 0_N$, $\mathbf{B} = \boldsymbol{\Lambda}_{\mathcal{L}} + \lambda \mathbf{I}_N$
Run:
for $t = 1$ **to** T **do**
Sample $\tilde{\boldsymbol{\mu}} \sim \mathcal{N}(\hat{\boldsymbol{\mu}}, v^2\mathbf{B}^{-1})$
$a(t) \leftarrow \arg\max_a \mathbf{b}_a^\top \tilde{\boldsymbol{\mu}}$
Observe a noisy reward $r(t) = \mathbf{b}_{a(t)}^\top \boldsymbol{\mu} + \varepsilon_t$
$\boldsymbol{f} \leftarrow \boldsymbol{f} + \mathbf{b}_{a(t)} r(t)$
Update $\mathbf{B} \leftarrow \mathbf{B} + \mathbf{b}_{a(t)}\mathbf{b}_{a(t)}^\top$
Update $\hat{\boldsymbol{\mu}} \leftarrow \mathbf{B}^{-1}\boldsymbol{f}$
end for

In order to state our result, we first define the *effective dimension*, a quantity shown to be small in the real-world graphs (Valko et al., 2014).

Definition 1. *Let the **effective dimension** d be the largest d such that*

$$(d-1)\lambda_d \le \frac{T}{\ln(1+T/\lambda)}.$$

We would like to stress that we consider the regime when $T < N$, because we aim for applications with a large set of arms and we are interested in a satisfactory performance after just a few iterations. For instance, when we aim to recommend N movies, we would like to have useful recommendations in the time $T < N$, i.e., before the user saw all of them. In the typical $T > N$ setting, d can be of the order of N and our approach does not bring an improvement over linear bandit algorithms. The following theorem upper-bounds the cumulative regret of SpectralTS in terms of d.

Theorem 1. *Let d be the effective dimension and λ be the minimum eigenvalue of $\boldsymbol{\Lambda}$. If $\|\boldsymbol{\mu}\|_{\boldsymbol{\Lambda}} \le C$ and for all $\mathbf{b}_i$, $|\mathbf{b}_i^\top \boldsymbol{\mu}| \le 1$, then the cumulative regret of Spectral Thompson Sampling is with probability at least $1-\delta$ bounded as*

$$\mathcal{R}(T) \le \frac{11g}{p}\sqrt{\frac{4+4\lambda}{\lambda} dT \ln\frac{\lambda+T}{\lambda}} + \frac{1}{T} + \frac{g}{p}\left(\frac{11}{\sqrt{\lambda}} + 2\right)\sqrt{2T\ln\frac{2}{\delta}},$$

where $p = 1/(4e\sqrt{\pi})$ and

$$g = \sqrt{4\ln TN}\left(R\sqrt{6d\ln\left(\frac{\lambda+T}{\delta\lambda}\right)} + C\right) + R\sqrt{2d\ln\left(\frac{(\lambda+T)T^2}{\delta\lambda}\right)} + C.$$

Remark 2. *Substituting g and p we see that regret bound scales as $d\sqrt{T\ln N}$. Note that $N = D$ could be exponential in d and we need to consider factor $\sqrt{\ln N}$ in our bound. On the other hand, if N is indeed exponential in d, then our algorithm scales with $\ln D\sqrt{T\ln D} = \ln(D)^{3/2}\sqrt{T}$ which is even better.*

4 Analysis

Preliminaries In the first five lemmas we state the known results on which we build in our analysis.

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

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

Multiple use of Sylvester's determinant theorem gives:

Lemma 2. *Let $\mathbf{B}_t = \boldsymbol{\Lambda} + \sum_{\tau=1}^{t-1}\mathbf{b}_\tau\mathbf{b}_\tau^\top$, then*

$$\ln\frac{|\mathbf{B}_t|}{|\boldsymbol{\Lambda}|} = \sum_{\tau=1}^{t}\ln(1+\|\mathbf{b}_\tau\|_{\mathbf{B}_\tau^{-1}})$$

Lemma 3. *(Abbasi-Yadkori, Pál, and Szepesvári, 2011). Let $\mathbf{B}_t = \boldsymbol{\Lambda} + \sum_{\tau=1}^{t-1}\mathbf{b}_\tau\mathbf{b}_\tau^\top$ and define $\boldsymbol{\xi}_t = \sum_{\tau=1}^{t}\varepsilon_\tau\mathbf{b}_\tau$. With probability at least $1-\delta$, $\forall t \ge 1$:*

$$\|\boldsymbol{\xi}_t\|^2_{\mathbf{B}_t^{-1}} \le 2R^2\ln\left(\frac{|\mathbf{B}_t|^{1/2}}{\delta|\boldsymbol{\Lambda}|^{1/2}}\right).$$

The next lemma is a generalization of Theorem 2 in Abbasi-Yadkori, Pál, and Szepesvári (2011) for any $\boldsymbol{\Lambda}$.

Lemma 4. *(Lemma 3 by Valko et al. (2014)). Let $\|\boldsymbol{\mu}\|_{\boldsymbol{\Lambda}} \le C$ and $\mathbf{B}_t$ is as above. Then for any $\mathbf{x}$ with probability at least $1-\delta$, $\forall t \ge 1$:*

$$|\mathbf{x}^\top(\hat{\boldsymbol{\mu}}(t) - \boldsymbol{\mu})| \le \|\mathbf{x}\|_{\mathbf{B}_t^{-1}}\left(R\sqrt{2\ln\left(\frac{|\mathbf{B}_t|^{1/2}}{\delta|\boldsymbol{\Lambda}|^{1/2}}\right)} + C\right)$$

Lemma 5. *(Lemma 7 by Valko et al. (2014)). Let d be the effective dimension. Then:*

$$\ln\frac{|\mathbf{B}_t|}{|\boldsymbol{\Lambda}|} \le 2d\ln\left(1+\frac{T}{\lambda}\right).$$

Cumulative Regret analysis Our analysis is based on the proof technique of Agrawal and Goyal (2013b). The summary of the technique follows. Each time an arm is played, our algorithm improves the confidence about our actual estimate of $\boldsymbol{\mu}$ via update of $\mathbf{B}_t$ and thus the update of confidence ellipsoid. However, when we play a suboptimal arm, the regret we obtain can be much higher than the improvement of our knowledge. To overcome this difficulty, the arms are divided into two groups of *saturated* and *unsaturated* arms, based on whether the standard deviation for an arm is smaller than the standard deviation of the optimal arm (Definition 3) or not. Consequently, the optimal arm is in group of unsaturated arms. The idea is to bound the regret of playing an unsaturated arm in terms of standard deviation and to show that the probability that the saturated arm is played is small enough. This way we overcome the difficulty of high regret and small knowledge obtained by playing an arm. In the following we use the notation from Table 2.

Definition 2. *We define $E^{\hat{\mu}}(t)$ as the event that for all i,*

$$|\mathbf{b}_i^\intercal \hat{\boldsymbol{\mu}}(t) - \mathbf{b}_i^\intercal \boldsymbol{\mu}| \leq l \|\mathbf{b}_i\|_{\mathbf{B}_t^{-1}}$$

and $E^{\tilde{\mu}}(t)$ as the event that for all i,

$$|\mathbf{b}_i^\intercal \tilde{\boldsymbol{\mu}}(t) - \mathbf{b}_i^\intercal \hat{\boldsymbol{\mu}}(t)| \leq v \|\mathbf{b}_i\|_{\mathbf{B}_t^{-1}} \sqrt{4 \ln(TN)}.$$

Definition 3. *We say that an arm i is **saturated** at time t if $\Delta_i > g\|\mathbf{b}_i\|_{\mathbf{B}_t^{-1}}$, and **unsaturated** otherwise (including a^*). Let $C(t)$ denote the set of saturated arms at time t.*

Definition 4. *We define filtration $\mathcal{F}_{t-1}$ as the union of the history until time $t-1$ and features, i.e.,*

$$\mathcal{F}_{t-1} = \{\mathcal{H}_{t-1}\} \cup \{\mathbf{b}_i, i = 1, \ldots, N\}$$

By definition, $\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \cdots \subseteq \mathcal{F}_{T-1}$.

Lemma 6. *For all t, $0 < \delta < 1$, $Pr(E^{\hat{\mu}}(t)) \geq 1 - \delta/T^2$ and for all possible filtrations $\mathcal{F}_{t-1}$,*

$$Pr(E^{\tilde{\mu}}(t) \,|\, \mathcal{F}_{t-1}) \geq 1 - 1/T^2.$$

Proof. **Bounding the probability of event $E^{\hat{\mu}}(t)$:** Using Lemma 4, where C is such that $\|\boldsymbol{\mu}\|_{\boldsymbol{\Lambda}} \leq C$, for all i with probability at least $1 - \delta'$ we have

$$\begin{aligned}|\mathbf{b}_i^\intercal(\hat{\boldsymbol{\mu}}(t) - \boldsymbol{\mu})| &\leq \|\mathbf{b}_i\|_{\mathbf{B}_t^{-1}} \left(R\sqrt{2 \ln \left(\frac{|\mathbf{B}_t|^{1/2}}{\delta' |\boldsymbol{\Lambda}|^{1/2}} \right)} + C \right) \\ &= \|\mathbf{b}_i\|_{\mathbf{B}_t^{-1}} \left(R\sqrt{\ln \frac{|\mathbf{B}_t|}{|\boldsymbol{\Lambda}|} + 2 \ln \frac{1}{\delta'}} + C \right).\end{aligned}$$

Therefore, using Lemma 5 and substituting $\delta' = \delta/T^2$, we get that with probability at least $1 - \delta/T^2$, for all i,

$$\begin{aligned}&|\mathbf{b}_i^\intercal(\hat{\boldsymbol{\mu}}(t) - \boldsymbol{\mu})| \\ &\leq \|\mathbf{b}_i\|_{\mathbf{B}_t^{-1}} \left(R\sqrt{2d \ln \frac{\lambda + T}{\lambda} + 2d \ln \frac{T^2}{\delta}} + C \right) \\ &= \|\mathbf{b}_i\|_{\mathbf{B}_t^{-1}} \left(R\sqrt{2d \ln \left(\frac{(\lambda + T)T^2}{\delta \lambda} \right)} + C \right) \\ &= l \|\mathbf{b}_i\|_{\mathbf{B}_t^{-1}}.\end{aligned}$$

Bounding the probability of event $E^{\tilde{\mu}}(t)$: The probability of each individual term $|\mathbf{b}_i^\intercal(\tilde{\boldsymbol{\mu}}(t) - \hat{\boldsymbol{\mu}}(t))| < \sqrt{4 \ln(TN)}$ can be bounded using Lemma 1 to get

$$\begin{aligned}&Pr\left(|\mathbf{b}_i^\intercal(\tilde{\boldsymbol{\mu}}(t) - \hat{\boldsymbol{\mu}}(t))| \geq v \|\mathbf{b}_i\|_{\mathbf{B}_t^{-1}} \sqrt{4 \ln(TN)} \right) \\ &\leq \frac{e^{-2 \ln TN}}{\sqrt{\pi 4 \ln(TN)}} \leq \frac{1}{T^2 N}.\end{aligned}$$

We complete the proof by taking a union bound over all N vectors $\mathbf{b}_i$. Notice that we took a different approach than Agrawal and Goyal (2013b) to avoid the dependence on the ambient dimension D. □

Lemma 7. *For any filtration $\mathcal{F}_{t-1}$ such that $E^{\hat{\mu}}(t)$ is true,*

$$Pr\left(\mathbf{b}_{a^*}^\intercal \tilde{\boldsymbol{\mu}}(t) > \mathbf{b}_{a^*}^\intercal \boldsymbol{\mu} \,|\, \mathcal{F}_{t-1}\right) \geq \frac{1}{4e\sqrt{\pi}}.$$

Proof. Since $\mathbf{b}_{a^*}^\intercal \tilde{\boldsymbol{\mu}}(t)$ is a Gaussian random variable with the mean $\mathbf{b}_{a^*}^\intercal \hat{\boldsymbol{\mu}}(t)$ and the standard deviation $v\|\mathbf{b}_{a^*}\|_{B_t^{-1}}$, we can use the anti-concentration inequality in Lemma 1,

$$\begin{aligned}&Pr\left(\mathbf{b}_{a^*}^\intercal \tilde{\boldsymbol{\mu}}(t) \geq \mathbf{b}_{a^*}^\intercal \boldsymbol{\mu} \,|\, \mathcal{F}_{t-1}\right) \\ &= Pr\left(\frac{\mathbf{b}_{a^*}^\intercal \tilde{\boldsymbol{\mu}}(t) - \mathbf{b}_{a^*}^\intercal \hat{\boldsymbol{\mu}}(t)}{v\|\mathbf{b}_{a^*}\|_{\mathbf{B}_t^{-1}}} \geq \frac{\mathbf{b}_{a^*}^\intercal \boldsymbol{\mu} - \mathbf{b}_{a^*}^\intercal \hat{\boldsymbol{\mu}}(t)}{v\|\mathbf{b}_{a^*}\|_{\mathbf{B}_t^{-1}}} \,|\, \mathcal{F}_{t-1} \right) \\ &\geq \frac{1}{4\sqrt{\pi} Z_t} e^{-Z_t^2},\end{aligned}$$

$$\text{where} \quad |Z_t| = \left| \frac{\mathbf{b}_{a^*}^\intercal \boldsymbol{\mu} - \mathbf{b}_{a^*}^\intercal \hat{\boldsymbol{\mu}}(t)}{v\|\mathbf{b}_{a^*}\|_{\mathbf{B}_t^{-1}}} \right|.$$

Since we consider a filtration $\mathcal{F}_{t-1}$ such that $E^{\hat{\mu}}(t)$ is true, we can upperbound the numerator to get

$$|Z_t| \leq \frac{l\|\mathbf{b}_{a^*}\|_{\mathbf{B}_t^{-1}}}{v\|\mathbf{b}_{a^*}\|_{\mathbf{B}_t^{-1}}} = \frac{l}{v} \leq 1.$$

Finally, $Pr\left(\mathbf{b}_{a^*}^\intercal \tilde{\boldsymbol{\mu}}(t) > \mathbf{b}_{a^*}^\intercal \boldsymbol{\mu} \,|\, \mathcal{F}_{t-1}\right) \geq \frac{1}{4e\sqrt{\pi}}.$ □

Lemma 8. *For any filtration $\mathcal{F}_{t-1}$ such that $E^{\hat{\mu}}(t)$ is true,*

$$Pr(a(t) \notin C(t) \,|\, \mathcal{F}_{t-1}) \geq \frac{1}{4e\sqrt{\pi}} - \frac{1}{T^2}.$$

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

$$\begin{aligned}&Pr(a(t) \notin C(t) \,|\, \mathcal{F}_{t-1}) \\ &\quad \geq Pr(\mathbf{b}_{a^*}^\intercal \tilde{\boldsymbol{\mu}}(t) > \mathbf{b}_j^\intercal \tilde{\boldsymbol{\mu}}(t), \forall j \in C(t) \,|\, \mathcal{F}_{t-1}).\end{aligned}$$

By definition, for all saturated arms, i.e., for all $j \in C(t)$, $\Delta_j > g\|\mathbf{b}_j\|_{\mathbf{B}_t^{-1}}$. Now if both of the events $E^{\hat{\mu}(t)}$ and $E^{\tilde{\mu}(t)}$ are true then, by definition of these events, for all $j \in C(t)$, $\mathbf{b}_j^\intercal \tilde{\boldsymbol{\mu}}(t) \leq \mathbf{b}_j^\intercal \boldsymbol{\mu}(t) + g\|\mathbf{b}_j\|_{\mathbf{B}_t^{-1}}$. Therefore, given the filtration $\mathcal{F}_{t-1}$, such that $E^{\hat{\mu}}(t)$ is true, either $E^{\tilde{\mu}}(t)$ is false, or else for all $j \in C(t)$,

$$\mathbf{b}_j^\intercal \tilde{\boldsymbol{\mu}}(t) \leq \mathbf{b}_j^\intercal \boldsymbol{\mu} + g\|\mathbf{b}_j\|_{\mathbf{B}_t^{-1}} \leq \mathbf{b}_{a^*}^\intercal \boldsymbol{\mu}.$$

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

$$\begin{aligned}
&Pr(\mathbf{b}_{a^*}{}^{\mathsf{T}}\tilde{\boldsymbol{\mu}}(t) > \mathbf{b}_j^{\mathsf{T}}\tilde{\boldsymbol{\mu}}(t),\, \forall j \in C(t) \,|\, \mathcal{F}_{t-1}) \\
&\quad \geq Pr(\mathbf{b}_{a^*}^{\mathsf{T}}\tilde{\boldsymbol{\mu}}(t) > \mathbf{b}_{a^*}^{\mathsf{T}}\boldsymbol{\mu} \,|\, \mathcal{F}_{t-1}) - Pr\left(\overline{E^{\hat{\mu}}(t)} \,|\, \mathcal{F}_{t-1}\right) \\
&\quad \geq \frac{1}{4e\sqrt{\pi}} - \frac{1}{T^2}.
\end{aligned}$$

In the last inequality we used Lemma 6 and Lemma 7. □

Lemma 9. *For any filtration $\mathcal{F}_{t-1}$ such that $E^{\hat{\mu}}(t)$ is true,*

$$\mathbb{E}[\Delta_{a(t)} \,|\, \mathcal{F}_{t-1}] \leq \frac{11g}{p}\mathbb{E}[\|\mathbf{b}_{a(t)}\|_{\mathbf{B}_t^{-1}} \,|\, \mathcal{F}_{t-1}] + \frac{1}{T^2}$$

Proof. Let $\overline{a}(t)$ denote the unsaturated arm with the smallest norm $\|\mathbf{b}_i\|_{\mathbf{B}_t^{-1}}$, i.e.,

$$\overline{a}(t) = \operatorname*{arg\,min}_{i \notin C(t)} \|\mathbf{b}_i\|_{\mathbf{B}_t^{-1}}.$$

Notice that since $C(t)$ and $\|\mathbf{b}_i\|_{\mathbf{B}_t^{-1}}$ for all i, are fixed on fixing $\mathcal{F}_{t-1}$, so is $\overline{a}(t)$. Now, using Lemma 8, for any $\mathcal{F}_{t-1}$ such that $E^{\hat{\mu}}(t)$ is true,

$$\begin{aligned}
&\mathbb{E}[\|\mathbf{b}_{a(t)}\|_{\mathbf{B}_t^{-1}} \,|\, \mathcal{F}_{t-1}] \\
&\quad \geq \mathbb{E}[\|\mathbf{b}_{a(t)}\|_{\mathbf{B}_t^{-1}} \,|\, \mathcal{F}_{t-1}, a(t) \notin C(t)] \\
&\qquad \cdot Pr(a(t) \notin C(t) \,|\, \mathcal{F}_{t-1}) \\
&\quad \geq \|\mathbf{b}_{\overline{a}(t)}\|_{\mathbf{B}_t^{-1}} \left(\frac{1}{4e\sqrt{\pi}} - \frac{1}{T^2}\right).
\end{aligned}$$

Now, if the events $E^{\hat{\mu}}(t)$ and $E^{\tilde{\mu}}(t)$ are true, then for all i, by definition, $\mathbf{b}_i^{\mathsf{T}}\tilde{\boldsymbol{\mu}}(t) \leq \mathbf{b}_i^{\mathsf{T}}\boldsymbol{\mu} + g\|\mathbf{b}_i\|_{\mathbf{B}_t^{-1}}$. Using this observation along with $\mathbf{b}_{a(t)}^{\mathsf{T}}\tilde{\boldsymbol{\mu}}(t) \geq \mathbf{b}_i^{\mathsf{T}}\tilde{\boldsymbol{\mu}}(t)$ for all i,

$$\begin{aligned}
\Delta_{a(t)} &= \Delta_{\overline{a}(t)} + (\mathbf{b}_{\overline{a}(t)}^{\mathsf{T}}\boldsymbol{\mu} - \mathbf{b}_{a(t)}^{\mathsf{T}}\boldsymbol{\mu}) \\
&\leq \Delta_{\overline{a}(t)} + (\mathbf{b}_{\overline{a}(t)}^{\mathsf{T}}\tilde{\boldsymbol{\mu}}(t) - \mathbf{b}_{a(t)}^{\mathsf{T}}\tilde{\boldsymbol{\mu}}(t)) \\
&\quad + g\|\mathbf{b}_{\overline{a}(t)}\|_{\mathbf{B}_t^{-1}} + g\|\mathbf{b}_{a(t)}\|_{\mathbf{B}_t^{-1}} \\
&\leq \Delta_{\overline{a}(t)} + g\|\mathbf{b}_{\overline{a}(t)}\|_{\mathbf{B}_t^{-1}} + g\|\mathbf{b}_{a(t)}\|_{\mathbf{B}_t^{-1}} \\
&\leq g\|\mathbf{b}_{\overline{a}(t)}\|_{\mathbf{B}_t^{-1}} + g\|\mathbf{b}_{\overline{a}(t)}\|_{\mathbf{B}_t^{-1}} + g\|\mathbf{b}_{a(t)}\|_{\mathbf{B}_t^{-1}}.
\end{aligned}$$

Therefore, for any $\mathcal{F}_{t-1}$ such that $E^{\hat{\mu}}(t)$ is true, either $\Delta_{a(t)} \leq 2g\|\mathbf{b}_{\overline{a}(t)}\|_{\mathbf{B}_t^{-1}} + g\|\mathbf{b}_{a(t)}\|_{\mathbf{B}_t^{-1}}$, or $E^{\tilde{\mu}}(t)$ is false. We can deduce that

$$\begin{aligned}
&\mathbb{E}[\Delta_{a(t)} \,|\, \mathcal{F}_{t-1}] \\
&\quad \leq \mathbb{E}\left[2g\|\mathbf{b}_{\overline{a}(t)}\|_{\mathbf{B}_t^{-1}} + g\|\mathbf{b}_{a(t)}\|_{\mathbf{B}_t^{-1}} \,|\, \mathcal{F}_{t-1}\right] \\
&\qquad + Pr\left(\overline{E^{\tilde{\mu}}(t)}\right) \\
&\quad \leq \frac{2g}{p - \frac{1}{T^2}}\mathbb{E}\left[\|\mathbf{b}_{a(t)}\|_{\mathbf{B}_t^{-1}} \,|\, \mathcal{F}_{t-1}\right] \\
&\qquad + g\mathbb{E}\left[\|\mathbf{b}_{a(t)}\|_{\mathbf{B}_t^{-1}} \,|\, \mathcal{F}_{t-1}\right] + \frac{1}{T^2} \\
&\quad \leq \frac{11g}{p}\mathbb{E}[\|\mathbf{b}_{a(t)}\|_{\mathbf{B}_t^{-1}} \,|\, \mathcal{F}_{t-1}] + \frac{1}{T^2}.
\end{aligned}$$

In the last inequality we used that $1/(p - 1/T^2) \leq 5/p$, which holds trivially for $T \leq 4$. For $T \geq 5$, we get that $T^2 \geq 5e\sqrt{\pi}$, which holds for $T \geq 5$. □

Definition 5. *We define* $\mathrm{regret}'(t) = \mathrm{regret}(t) \cdot I(E^{\hat{\mu}}(t))$.

Definition 6. *A sequence of random variables $(Y_t;\, t \geq 0)$ is called a* ***super-martingale*** *corresponding to a filtration $\mathcal{F}_t$, if for all t, Y_t is $\mathcal{F}_t$-measurable, and for $t \geq 1$,*

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

Next, following Agrawal and Goyal (2013b), we establish a super-martingale process that will form the basis of our proof of the high-probability regret bound.

Definition 7. *Let*

$$\begin{aligned}
X_t &= \mathrm{regret}'(t) - \frac{11g}{p}\|\mathbf{b}_{a(t)}\|_{\mathbf{B}_t^{-1}} - \frac{1}{T^2} \\
Y_t &= \sum_{w=1}^{t} X_w.
\end{aligned}$$

Lemma 10. *$(Y_t;\, t = 0, \ldots, T)$ is a super-martingale process with respect to filtration $\mathcal{F}_t$.*

Proof. We need to prove that for all $t \in \{1, \ldots, T\}$, and any possible $\mathcal{F}_{t-1}$, $\mathbb{E}[Y_t - Y_{t-1} \,|\, \mathcal{F}_{t-1}] \leq 0$, i.e.

$$\mathbb{E}[\mathrm{regret}'(t) \,|\, \mathcal{F}_{t-1}] \leq \frac{11g}{p}\|\mathbf{b}_{a(t)}\|_{\mathbf{B}_t^{-1}} + \frac{1}{T^2}.$$

Note that whether $E^{\hat{\mu}}(t)$ is true or not, is completely determined by $\mathcal{F}_{t-1}$. If $\mathcal{F}_{t-1}$ is such that $E^{\hat{\mu}}(t)$ is not true, then $\mathrm{regret}'(t) = \mathrm{regret}(t) \cdot I(E^{\hat{\mu}}(t)) = 0$, and the above inequality holds trivially. Moreover, for $\mathcal{F}_{t-1}$ such that $E^{\hat{\mu}}(t)$ holds, the inequality follows from Lemma 9. □

Unlike (Agrawal and Goyal, 2013b; Abbasi-Yadkori, Pál, and Szepesvári, 2011), we do not want to require $\lambda \geq 1$. Therefore, we provide the following lemma that shows the dependence of $\|\mathbf{b}_{a(t)}\|^2_{\mathbf{B}_t^{-1}}$ on λ.

Lemma 11. *For all t,*

$$\|\mathbf{b}_{a(t)}\|^2_{\mathbf{B}_t^{-1}} \leq \left(2 + \frac{2}{\lambda}\right) \ln\left(1 + \|\mathbf{b}_{a(t)}\|^2_{\mathbf{B}_t^{-1}}\right).$$

Proof. Note, that $\|\mathbf{b}_{a(t)}\|_{\mathbf{B}_t^{-1}} \leq (1/\sqrt{\lambda})\|\mathbf{b}_{a(t)}\| \leq (1/\sqrt{\lambda})$ and for all $0 \leq x \leq 1$ we have

$$x \leq 2\ln(1 + x). \tag{1}$$

Now we consider two cases depending on λ. If $\lambda \geq 1$, we know that $0 \leq \|\mathbf{b}_{a(t)}\|_{\mathbf{B}_t^{-1}} \leq 1$ and therefore by (1),

$$\|\mathbf{b}_{a(t)}\|^2_{\mathbf{B}_t^{-1}} \leq 2\ln\left(1 + \|\mathbf{b}_{a(t)}\|^2_{\mathbf{B}_t^{-1}}\right).$$

Similarly, if $\lambda < 1$, then $0 \leq \lambda\|\mathbf{b}_{a(t)}\|_{\mathbf{B}_t^{-1}} \leq 1$ and we get

$$\begin{aligned}
\|\mathbf{b}_{a(t)}\|^2_{\mathbf{B}_t^{-1}} &\leq \frac{2}{\lambda}\ln\left(1 + \lambda\|\mathbf{b}_{a(t)}\|^2_{\mathbf{B}_t^{-1}}\right) \\
&\leq \frac{2}{\lambda}\ln\left(1 + \|\mathbf{b}_{a(t)}\|^2_{\mathbf{B}_t^{-1}}\right).
\end{aligned}$$

Combining the two, we get that for all $\lambda \geq 0$,

$$\begin{aligned}
\|\mathbf{b}_{a(t)}\|^2_{\mathbf{B}_t^{-1}} &\leq \max\left(2, \frac{2}{\lambda}\right)\ln\left(1 + \|\mathbf{b}_{a(t)}\|^2_{\mathbf{B}_t^{-1}}\right) \\
&\leq \left(2 + \frac{2}{\lambda}\right)\ln\left(1 + \|\mathbf{b}_{a(t)}\|^2_{\mathbf{B}_t^{-1}}\right).
\end{aligned}$$

□

Proof of Theorem 1. First, notice that X_t is bounded as $|X_t| \leq 1+11g/(p\sqrt{\lambda})+1/T^2 \leq (11/\sqrt{\lambda}+2)g/p$. Thus, we can apply Azuma-Hoeffding inequality to obtain that with probability at least $1-\delta/2$,

$$\sum_{t=1}^{T} \text{regret}'(t) \leq \sum_{t=1}^{T} \frac{11g}{p} \|\mathbf{b}_{a(t)}\|_{\mathbf{B}_t^{-1}} + \sum_{t=1}^{T} \frac{1}{T^2} + \sqrt{2\left(\sum_{t=1}^{T} \frac{g^2}{p^2}\left(\frac{11}{\sqrt{\lambda}}+2\right)^2\right) \ln \frac{2}{\delta}}.$$

Since p and g are constants, then with probability $1-\delta/2$,

$$\sum_{t=1}^{T} \text{regret}'(t) \leq \frac{11g}{p} \sum_{t=1}^{T} \|\mathbf{b}_{a(t)}\|_{\mathbf{B}_t^{-1}} + \frac{1}{T} + \frac{g}{p}\left(\frac{11}{\sqrt{\lambda}}+2\right)\sqrt{2T \ln \frac{2}{\delta}}.$$

The last step is to upperbound $\sum_{t=1}^{T} \|\mathbf{b}_{a(t)}\|_{\mathbf{B}_t^{-1}}$. For this purpose, Agrawal and Goyal (2013b) rely on the analysis of Auer (2002) and the assumption that $\lambda \geq 1$. We provide an alternative approach using Cauchy-Schwartz inequality, Lemma 2, and Lemma 11 to get

$$\sum_{t=1}^{T} \|\mathbf{b}_{a(t)}\|_{\mathbf{B}_t^{-1}} \leq \sqrt{T \sum_{t=1}^{T} \|\mathbf{b}_{a(t)}\|^2_{\mathbf{B}_t^{-1}}} \leq \sqrt{T\left(2+\frac{2}{\lambda}\right) \ln \frac{|\mathbf{B}_T|}{|\mathbf{\Lambda}|}} \leq \sqrt{\frac{4+4\lambda}{\lambda} dT \ln \frac{\lambda+T}{\lambda}}.$$

Finally, we know that $E^{\hat{\mu}}(t)$ holds for all t with probability at least $1-\frac{\delta}{2}$ and $\text{regret}'(t) = \text{regret}(t)$ for all t with probability at least $1-\frac{\delta}{2}$. Hence, with probability $1-\delta$,

$$\mathcal{R}(T) \leq \frac{11g}{p}\sqrt{\frac{4+4\lambda}{\lambda} dT \ln \frac{\lambda+T}{\lambda}} + \frac{1}{T} + \frac{g}{p}\left(\frac{11}{\sqrt{\lambda}}+2\right)\sqrt{2T \ln \frac{2}{\delta}}.$$

□

5 Experiments

The aim of this section is to give empirical evidence that SpectralTS – a faster and simpler algorithm than SpectralUCB – also delivers comparable or better empirical performance. For the synthetic experiment, we considered a *Barabási-Albert* (BA) model (1999), known for its preferential attachment property, common in real-world graphs. We generated a random graph using BA model with $N = 250$ nodes and the degree parameter 3. For each run, we generated the weights of the edges uniformly at random. Then we generated $\boldsymbol{\mu}$, a random vector of weights (unknown to the algorithms) in order to compute the payoffs and evaluated the cumulative regret. The $\boldsymbol{\mu}$ in each simulation was a random linear combination of the first 3 smoothest eigenvectors of the graph Laplacian. In all experiments, we had $\delta = 0.001$, $\lambda = 1$, and $R = 0.01$. We evaluated the algorithms in $T < N$ regime, where the linear bandit algorithms are not expected to perform well. Figure 1 shows the results averaged over 10 simulations. Notice that while the result of SpectralTS is comparable to SpectralUCB, its computational time is much faster due the reasons discussed in Section 3. Recall that while both algorithms compute the least-square problem of the same size, SpectralUCB has then to compute the confidence interval for each arm.

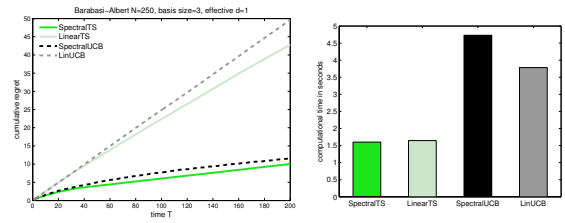

Figure 1: Barabási-Albert random graph results

Furthermore, we performed the comparison of the algorithms on the MovieLens dataset (Lam and Herlocker, 2012) of the movie ratings. The graph in this dataset is the graph of 2019 movies with edges corresponding to the movie similarities. For each user we have a graph function, unknown to the algorithms, that assigns to each node, the rating of the particular user. A detailed description on the preprocessing is deferred to (Valko et al., 2014). Our goal is then to recommend the most highly rated content. Figure 2 shows the MovieLens data results averaged over 10 randomly sampled users. Notice that the results follow the same trends as for the synthetic data.

Our results show that the empirical performance of the computationally more efficient SpectralTS is similar or slightly better than the one of SpectralUCB. The main contribution of this paper is the analysis that backs up this evidence with a theoretical justification.

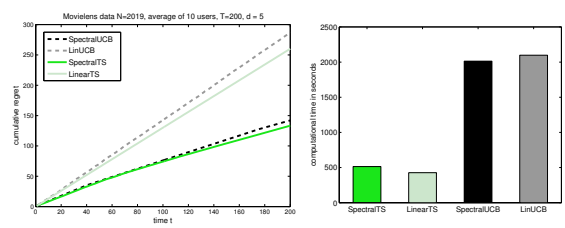

Figure 2: MovieLens data results

6 Conclusion

We considered the spectral bandit setting with a reward function assumed to be smooth on a given graph and proposed Spectral Thompson Sampling (TS) for it. Our main contribution is stated in Theorem 1 where we prove that the regret bound scales with effective dimension d, typically much smaller than the ambient dimension D, which is the case of linear bandit algorithms. In our experiments, we showed that SpectralTS outperforms LinearUCB and LinearTS, and provides similar results to SpectralUCB in the regime when $T < N$. Moreover, we showed that SpectralTS is computationally less expensive than SpectralUCB.

7 Acknowledgements

The research presented in this paper was supported by French Ministry of Higher Education and Research, by European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n°270327 (project CompLACS), and by Intel Corporation.

References

Abbasi-Yadkori, Y.; Pál, D.; and Szepesvári, C. 2011. Improved Algorithms for Linear Stochastic Bandits. In *Neural Information Processing Systems*.

Agrawal, S., and Goyal, N. 2012a. Analysis of Thompson Sampling for the multi-armed bandit problem. *Proceedings of the 25th Annual Conference on Learning Theory (COLT)*.

Agrawal, S., and Goyal, N. 2012b. Thompson Sampling for Contextual Bandits with Linear Payoffs. *CoRR, abs/1209.3352, http://arxiv.org/abs/1209.3352*.

Agrawal, S., and Goyal, N. 2013a. Further Optimal Regret Bounds for Thompson Sampling. In *Proceedings of the 16th International Conference on Articial Intelligence and Statistics (AISTATS)*, volume 31, 99–107.

Agrawal, S., and Goyal, N. 2013b. Thompson Sampling for Contextual Bandits with Linear Payoffs. In *International Conference on Machine Learning*.

Auer, P.; Cesa-Bianchi, N.; and Fischer, P. 2002. Finite-time Analysis of the Multiarmed Bandit Problem. *Machine Learning* 47(2-3):235–256.

Auer, P. 2002. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research* 3:397–422.

Barabási, A.-L., and Albert, R. 1999. Emergence of scaling in random networks. *Science* 286:11.

Belkin, M.; Niyogi, P.; and Sindhwani, V. 2006. Manifold Regularization: A Geometric Framework for Learning from Labeled and Unlabeled Examples. *Journal of Machine Learning Research* 7:2399–2434.

Chapelle, O., and Li, L. 2011. An Empirical Evaluation of Thompson Sampling. In *Neural Information Processing Systems*.

Dani, V.; Hayes, T. P.; and Kakade, S. M. 2008. The Price of Bandit Information for Online Optimization. In *Neural Information Processing Systems*. MIT Press.

Kaufmann, E.; Korda, N.; and Munos, R. 2012. Thompson Sampling: An Asymptotically Optimal Finite Time Analysis. *Algorithmic Learning Theory*.

Korda, N.; Kaufmann, E.; and Munos, R. 2013. Thompson Sampling for 1-Dimensional Exponential Family Bandits. In *Neural Information Processing Systems*.

Koutis, I.; Miller, G. L.; and Peng, R. 2010. Approaching Optimality for Solving SDD Linear Systems. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, 235–244. IEEE.

Lam, S., and Herlocker, J. 2012. MovieLens 1M Dataset. http://www.grouplens.org/node/12.

Li, L.; Chu, W.; Langford, J.; and Schapire, R. E. 2010. A Contextual-Bandit Approach to Personalized News Article Recommendation. *WWW 10*.

May, B. C.; Korda, N.; Lee, A.; and Leslie, D. S. 2012. Optimistic Bayesian sampling in contextual-bandit problems. *The Journal of Machine Learning Research* 13(1):2069–2069–2106–2106.

Pazzani, M. J., and Billsus, D. 2007. Content-Based Recommendation Systems. *The adaptive web*.

Russo, D., and Van Roy, B. 2013. Eluder Dimension and the Sample Complexity of Optimistic Exploration. In *Neural Information Processing Systems*.

Russo, D., and Van Roy, B. 2014. Learning to Optimize Via Posterior Sampling. *Mathematics of Operations Research*.

Srinivas, N.; Krause, A.; Kakade, S.; and Seeger, M. 2010. Gaussian Process Optimization in the Bandit Setting: No Regret and Experimental Design. *Proceedings of International Conference on Machine Learning* 1015–1022.

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

Valko, M.; Munos, R.; Kveton, B.; and Kocák, T. 2014. Spectral Bandits for Smooth Graph Functions. In *31th International Conference on Machine Learning*.

Yue, Y.; Hong, S. A.; and Guestrin, C. 2012. Hierarchical Exploration for Accelerating Contextual Bandits. In Langford, J., and Pineau, J., eds., *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, 1895–1902. New York, NY, USA: ACM.