On the Suboptimality of Thompson Sampling in High Dimensions

Anonymous Author(s)
Affiliation
Address
email

Abstract

In this paper we consider Thompson Sampling for combinatorial semi-bandits. We demonstrate that, perhaps surprisingly, Thompson Sampling is sub-optimal for this problem in the sense that its regret scales exponentially in the ambient dimension, and its minimax regret scales almost linearly. This phenomenon occurs under a wide variety of assumptions including both non-linear and linear reward functions in the Bernoulli distribution setting. We also show that including a fixed amount of forced exploration to Thompson Sampling does not alleviate the problem. We complement our theoretical results with numerical results and show that in practice Thompson Sampling indeed can perform very poorly in some high dimensional situations.

1 Introduction

We consider the problem of combinatorial bandits with semi-bandit feedback. At time t a learner selects a decision $x(t) \in \mathcal{X}$ where $\mathcal{X} \subset \{0,1\}^d$ is the set of available decisions. The environment then draws a random vector $Z(t) \in \mathbb{R}^d$. The learner then observes $Y(t) = x(t) \odot Z(t)$, where $\odot$ denotes the Hadamard (elementwise) product. This is called the semi bandit feedback. Then the learner receives the reward $f(x(t), Z(t))$ where f is a known function. We assume that $(Z(t))_{t\geq 1}$ are i.i.d., and that $Z_1(t), ..., Z_d(t)$ are independent and distributed as $Z_i(t) \sim \text{Bernoulli}(\theta_i)$ for all t,i.

The goal is to minimize the regret:

$$R(T, \theta) = T \max_{x\in\mathcal{X}}\{f(x, \theta)\} - \sum_{t=1}^{T} \mathbb{E}(f(x(t), \theta))$$

Initially θ is unknown to the learner and minimizing regret involves exploring suboptimal decisions just enough in order to identify the optimal decision. For any decision $x \in \mathcal{X}$ we define the reward gap

$$\Delta_x = \max_{x\in\mathcal{X}}\{f(x, \theta)\} - f(x, \theta)$$

which is the amount of regret incurred by choosing x instead of an optimal decision

$$x^\star \in \arg\max_{x\in\mathcal{X}}\{f(x, \theta)\}$$

and $\Delta_{\min} = \min_{x\in\mathcal{X}:\Delta_x>0} \Delta_x$ the minimal gap. We define $m \triangleq \max_{x\in\mathcal{X}} \sum_{i=1}^{d} |x_i|$ the size of the maximal decision.

For this problem, an algorithm which has attracted a lot of interest is Thompson Sampling, which at time t selects the decision maximizing $f(x(t), V(t))$ where $V(t)$ is a random variable distributed as

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

the posterior distribution of θ knowing the information available at time t, which is $Y(1), ..., Y(t-1)$. This strategy is usually computationally simple (for instance when f is linear) since it involves maximizing a function f over $\mathcal{X}$. Also, for some problem instances it tends to perform well numerically. A particular case of interest is *linear combinatorial semi-bandits* where $f(x, \theta) = \theta^\top x$ so that the reward is a linear function of the decision.

Our contribution. We show that the regret of Thompson sampling in general does not scale polynomially in the ambient dimension d.

(i) We provide several examples, both for linear and non-linear combinatorial bandits, where the regret of Thompson sampling does not scale polynomially in the dimension d (in fact in some cases it may scale even faster than exponentially in the dimension). In some cases, we show that one must wait for an amount of time greater than $\Omega(d^d)$ for Thompson Sampling to perform at least as well as random choice where one simply chooses $x(t)$ uniformly distributed in $\mathcal{X}$ at every round. Therefore, in high dimensions, in some instances, Thompson Sampling in general can perform strictly worse than random choice for all practically relevant time horizons.

(ii) We show that the minimax regret of Thompson Sampling scales at least as $\Omega(T^{1-\frac{1}{d}})$ so that it is not minimax optimal, as there exists algorithms such as CUCB and ESCB with minimax regret $O(\textbf{poly}(d)\sqrt{T(\ln T)})$. In fact, in high dimensions, the minimax regret of Thomson sampling is almost linear.

(iii) We further show that adding forced exploration before Thompson Sampling does not correct for this problem, so that this is not an artifact due to initialization.

(iv) Using numerical experiments, we show that indeed, for reasonable time horizons, Thompson sampling performs very poorly in high dimensions in some instances.

We believe that our results highlight two general characteristics of Thompson sampling. First Thompson Sampling tends to be much more greedy than optimistic algorithms such as ESCB, CUCB. This greedy behavior explains why the regret of Thompson Sampling is, in some instances, much smaller than that of optimistic algorithms. In fact it is sometimes so greedy that it misses the optimal decision. Second Thompson Sampling tends to be by nature a "risky" algorithm so that its regret exhibits very large fluctuations across runs. In some cases it finds the optimal arm very quickly and with little to no regret, while in other cases it simply misses the optimal decision and performs worse than random choice.

Related work. Combinatorial bandits are a generalization of classical bandits studied in [12]. Several asymptotically optimal algorithms are known for classical bandits, including the algorithm of [13], KL-UCB [2], DMED [9] and Thompson Sampling [15, 10]. Other algorithms include the celebrated UCB1 [1]. A large number of algorithms for combinatorial semi-bandits have been proposed, many of which naturally extend algorithms for classical bandits to the combinatorial setting. CUCB [3, 11] is a natural extension of UCB1 to the combinatorial setting. ESCB [5, 7] is an improvement of CUCB which leverages the independence of rewards between items. AESCB [6] is an approximate version of ESCB with roughly the same performance guarantees and reduced computational complexity. Thompson Sampling for combinatorial bandits was considered in [8, 16, 14]. Also, combinatorial semi bandits are a particular case of structured bandits, for which there exists asymptotically optimal algorithms such as OSSB [4]. Table 1 presents the best known regret upper bounds for CUCB, ESCB and TS.

We provide two types of bounds: problem dependent bounds (sometimes called gap-dependent bounds) and minimax bounds (or gap-free bounds). The former involves T, d, m and $\Delta_{\min}$, while the latter hold for any value of $\Delta_{\min}$.

An important observation is that all of the known regret upper bounds for Thompson sampling [8], [16], [14] feature at least one term that does not scale polynomially with the dimension. In particular, the paper [14] shows that the regret of Thompson Sampling is upper bounded by

$$R(T) \leq C(d(\ln m)^2 \ln(|\mathcal{X}|T)/\Delta_{\min} + dm^3/\Delta_{\min}^2 + m((m^2+1)/\Delta_{min})^{2+4m})$$

This is a general bound for all combinatorial set of interest. There is a super exponential constant (that does not depend on T) term : $m((m^2+1)/\Delta_{min})^{2+4m})$. Regret bounds for CUCB and ESCB do not feature this exponential dependency in the dimension. However, since Thompson Sampling tends to perform very well in all of the numerical experiments presented in the literature, a natural intuition would be that all known upper bounds are simply not sharp, and that the true regret of

Table 1: Algorithms and best known regret bounds.

Algo.	Regret ((i) problem dependent and (ii) minimax)
CUCB	(i) $O(dm(\ln T)/\Delta_{min})$ [11][Theorem 4] (ii) $O(\sqrt{dmT(\ln T)} + dm)$ [11][Theorem 6]
ESCB	(i) $O(d(\ln m)^2(\ln T)/\Delta_{min} + dm^3/\Delta^2_{\min})$ [7][Theorem 2] (ii) $O(\sqrt{d(\ln m)^2 T(\ln T)} + dm)$ [7][Corollary 1]
TS	(i) $O(d(\ln m)^2 \ln(\lvert\mathcal{X}\rvert T)/\Delta_{\min} + dm^3/\Delta^2_{\min}\ m((m^2+1)/\Delta_{min})^{2+4m})$ [14][Theorem 1] (ii) not available

Thompson Sampling does not really grow exponentially with the dimension. We show in this work that this intuition is incorrect, and that the regret of Thompson Sampling really does scale (at least) exponentially in the dimension i.e. it suffers from the "curse of dimensionality". This directly implies that, in high dimensions, CUCB and ESCB perform much better than Thompson Sampling in general.

Prior knowledge to the learner: combinatorial set $\mathcal{X} \subset \{0,1\}^d$, function $f : [0,1]^d \times \mathcal{X} \to \mathbb{R}$
For $t = 1, ..., T$:

1. The learner computes statistics $A(t) = \sum_{s=1}^{t-1} Z(s) \odot x(s)$ and $B(t) = \sum_{s=1}^{t-1} x(s) - Z(s) \odot x(s)$
2. The learner draws $V(t)$ with $V_1(t), ..., V_d(t)$ independent and $V_i(t) \sim \text{Beta}(A_i(t), B_i(t))$
3. The learner chooses decision $x(t) \in \arg\max_{x\in\mathcal{X}}\{f(x, V(t))\}$
4. The environment draws $Z(t)$ with $Z_1(t), ..., Z_d(t)$ independent and $Z_i(t) \sim \text{Ber}(\theta_i)$
5. The learner observes $Z(t) \odot x(t)$ and receive the reward $f(x(t), Z(t))$

Performance metric: expected regret $R(T,\theta) = T(\max_{x\in\mathcal{X}}\{f(x,\theta)\}) - \sum_{t=1}^{T} \mathbb{E}(f(\theta, x(t)))$

Figure 1: Thompson sampling for combinatorial semi-bandits with Bernoulli rewards and uniform prior.

2 Model

2.1 Problem Dependent Regret and Minimax Regret

In order to evaluate the performance of an algorithm over the set of instances $\theta \in [0,1]^d$, there are two main figures of merit that we study in this paper. The first is the problem-dependent regret which is $R(T,\theta)$ when $\theta \in [0,1]^d$ is fixed. The second is the minimax regret which is the worse case over θ for T fixed: $\max_{\theta\in[0,1]^d} R(T,\theta)$.

2.2 Thompson Sampling

The basic Thompson sampling algorithm works as follows. For $i = 1, ..., d$, define $A_i(t) = \sum_{s=1}^{t-1} Z_i(s)x_i(s)$ and $B_i(t) = \sum_{s=1}^{t-1}(1 - Z_i(s))x_i(s)$ which represent the number of successes and failures observed when getting a sample to estimate θ_i. We define $N_i(t) = A_i(t) + B_i(t) = \sum_{s=1}^{t-1} x_i(s)$ the number of samples available at time t to estimate θ_i, and

$$\hat{\theta}_i(t) = \frac{A_i(t)}{\max(N_i(t), 1)}$$

the corresponding estimate of θ_i which is simply the empirical mean.

The Thompson Sampling algorithm selects decision

$$x(t) \in \arg\max_{x\in\mathcal{X}} f(x, V(t)) \text{ where } V_i(t) \sim \text{Beta}(A_i(t)+1, B_i(t)+1)$$

and $V_1(t), ..., V_d(t)$ are independent. Vector $V(t)$ is called the *Thompson sample* at time t. Thompson sampling is based on a Bayesian argument, $V(t)$ is drawn according to the posterior distribution of θ

knowing the information available at time t, where the prior distribution for θ is uniform over $[0,1]^d$. Choosing decision $x(t)$ as done above should ensure that one explores just enough to find the optimal decision. In the linear case, the decision can be computed by linear maximization over $\mathcal{X}$

$$x(t) \in \arg\max_{x \in \mathcal{X}} \{V(t)^\top x\}$$

This explains the practical appeal of Thompson sampling, since whenever linear maximization over $\mathcal{X}$ can be implemented efficiently, the algorithm has low computational complexity.

2.3 Thompson Sampling with Forced Exploration

A natural extension of Thompson sampling is to add ℓ forced exploration rounds, where ℓ is a fixed number, in order to avoid some artifacts that could possibly occur due to the prior distribution. The algorithm operates as follows. At time $1 \leq t \leq \ell$, one selects $x(t) \in \mathcal{X}$ such that $x_{i(t)}(t) = 1$ with

$$i(t) \in \min_{i=1,\dots,d} N_i(t).$$

Otherwise for $t \geq \ell + 1$ one selects

$$x(t) \in \arg\max_{x \in \mathcal{X}} f(x, V(t))$$

with $V(t)$ the Thompson sample defined above. Namely, one first perform a forced exploration during ℓ rounds then apply Thompson sampling. This guarantees that $N_i(t) \geq \lfloor \ell/d \rfloor$ samples are available to estimate θ_i for all $i = 1, ..., d$, and one subsequently applies Thompson sampling. We call this variant Thompson Sampling with ℓ forced exploration rounds.

3 Main Results

We now state our main theoretical results. All proofs are found in the supplementary material.

3.1 Some combinatorial sets of interest

We will provide several examples of combinatorial sets where the regret of Thompson Sampling indeed scales exponentially with the dimension, so that this phenomenon is quite general and is not an artifact that only occurs for one particular family of combinatorial structures. We define $\mathcal{X}^p$ the set of paths of the directed acyclic graph depicted in figure 2. This set has two disjoint decisions $(1, ..., 1, 0, ..., 0)$ and $(0, ..., 0, 1, ..., 1)$ of equal size $m = \frac{d}{2}$. We define $\mathcal{X}^m$ the set of matchings of the bipartite graph depicted in figure 2. This graph has d vertices and d edges.

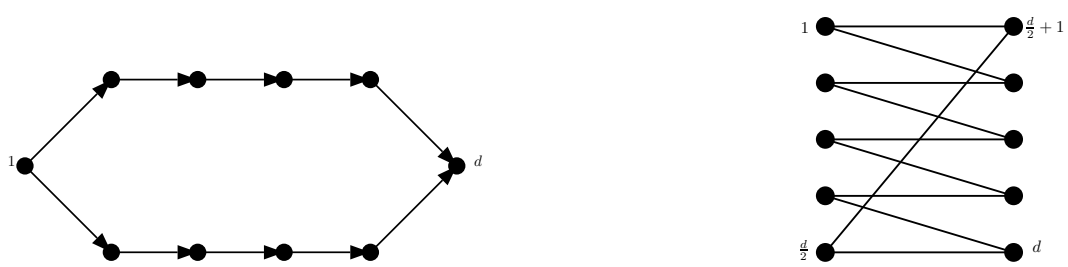


Figure 2: Paths in a directed acyclic graph (left) and matchings of the Z graph (right)

The combinatorial sets presented are very simple. However our results can by generalised for more complex set of interest without losing the exponential nature of the regret. For example the two path environment can be generalised to $k \geq 2 \in \mathbb{N}$ paths. It can also be generalised for non disjoint paths if the optimal path does not share "a lot" of edges with all the other paths. This could be the case for real life applications like shortest path routing or in medical trials where treatments cannot be associated with each other. With those simple examples in mind many other more complex sets that exhibit exponential regret can be found. However we do not provide formal proof for those more complex examples as the simple examples are sufficient to prove the suboptimality of Thompson sampling here.

3.2 Linear Combinatorial Bandits

Due to space constrain we only consider linear bandits, where the expected reward function is linear so that $f(x, \theta) = x^\top \theta$. For an analysis of the non linear case see the appendix **??** in the supplementary material. In this appendix we have only considered non linear problems inspired by theorem 3 of [16], and those problems had a distinctive feature which was that the Thompson sample of sub optimal decisions had no variance. One could be lead to think that the exponential dependency of the regret on the dimension could be caused by this feature, and it is hence natural to investigate the linear case, which is not only more common, but also where the Thompson sample of any decision always has a non-null variance.

In Theorem 1 we consider a linear problem over the combinatorial set $\mathcal{X}^p$ which is formed of two disjoint paths. We show that the regret of Thompson Sampling does scale exponentially in the dimension for this problem. Therefore this phenomenon is not linked to a particular, well chosen, non-linear reward function, but also occurs for the classical case of linear reward functions.

Theorem 1. *Consider a linear combinatorial bandit problem over combinatorial set $\mathcal{X}^p$ and parameter $\theta_i = 1$ if $1 \leq i \leq d/2$ and $\theta_i = 1 - \frac{\Delta}{m}$ otherwise. If $\frac{\Delta}{m} + \frac{1}{\sqrt{m}} < \frac{1}{2}$. Then the regret of Thompson Sampling is lower bounded by*

$$R(T, \theta) \geq \frac{\Delta}{4p_\Delta}(1 - (1 - p_\Delta)^{T-1}) \text{ with } p_\Delta = \exp\left\{-\frac{2m}{9}(\frac{1}{2} - (\frac{\Delta}{m} + \frac{1}{\sqrt{m}}))^2\right\}$$

Theorem 1 is proven by controlling the expectation of the first time that the optimal decision is selected, and showing that it is exponentially large in general. The central argument can be summarized as follows. Consider t such that at times $1, ..., t$ only the suboptimal decision has been selected. The probability of selecting the optimal decision is $\mathbb{P}\left(\sum_{i=1}^m V_i(t) \geq \sum_{i=m+1}^d V_i(t) | A(t), B(t)\right)$ where $V_1(t), ..., V_d(t)$ are independent, distributed in $[0, 1]$, and their respective expectations are $\mathbb{E}\Big(V_i(t)|A(t), B(t)\Big) = \begin{cases} \frac{1}{2} & \text{if } 1 \leq i \leq m \\ \frac{A_i(t)+1}{t+2} & \text{if } m+1 \leq i \leq d \end{cases}$ Furthermore, from the law of large numbers, when t is large, $\sum_{i=m+1}^d A_i(t) \approx (1 - \frac{\Delta}{m})t$ since we have sampled the suboptimal decision t times and therefore $\sum_{i=1}^m \mathbb{E}(V_i(t)|A(t), B(t)) = \frac{m}{2}$ and $\sum_{i=m+1}^d \mathbb{E}(V_i(t)|A(t), B(t)) \approx m - \Delta$ Since $V_1(t), ..., V_d(t)$ are independent and distributed in $[0, 1]$, their sums must concentrate around their expectation, and from Hoeffding's inequality: $\mathbb{P}(\sum_{i=1}^m V_i(t) \geq \sum_{i=m+1}^d V_i(t)|A(t), B(t)) \leq O(e^{-um((\frac{1}{2} - \frac{\Delta}{m}))^2})$ where $u > 0$ is some positive exponent related to how concentrated the Thompson samples are. This implies that, for large t, the probability of selecting the optimal decision is exponentially small if it has never been selected previously.

Also, we see that this phenomenon of lack of exploration by Thompson Sampling is a typically high dimensional phenomenon. In short, the Thompson samples of decisions will tend to concentrate around their expectation, so that Thompson Sampling will, most of the time, act greedily and simply select the decision maximizing the empirical reward.

We can also emphasize the fact that when m grows, we can have an arbitrary large gap Δ and still have exponential regret. This is a very unusual behavior the difficulty of a bandit problem should be inversely proportional to its gap Δ.

Furthermore another version of theorem 1 which exhibits exponential behavior can be shown with parameters $\theta_i = u$ if $1 \leq i \leq d/2$ and $\theta_i = u - \frac{\Delta}{m}$ otherwise, under the condition $\frac{\Delta}{m} + \frac{1}{\sqrt{m}} < u - \frac{1}{2}$ where $u \in]\frac{1}{2}, 1]$. We chose the parameters of theorem 1 for the sake of clarity and being at the edge of the parameter space is not a necessary condition to have exponential regret.

3.3 Linear Combinatorial Bandits: Small Gap Regime

We now derive Theorem 2, which is another regret bound for the regret of Thompson Sampling, however this bound is more accurate in the regime where Δ is small, and allows to deduce a lower bound for the minimax regret as well. The proof is a more intricate version of the proof of Theorem 2 which we briefly explained above. Corollary 3 states that the minimax regret of Thompson Sampling scales as $O(T)$ for $T \leq m!$, and at least as $O(T^{1-\frac{2}{d}})$ for $T \geq m!$. Of course, in practice, when m is

large we have $T \leq m!$ for any reasonable time horizon, so that the regret of Thompson Sampling is linear in this regime. Also, as stated by corollary 3, Thompson Sampling is provably not minimax optimal, as there exist algorithms with much smaller regret scaling as $O(d^{1/4}\sqrt{T \ln T})$.

Theorem 2. *Consider a linear combinatorial bandit problem over combinatorial set $\mathcal{X}^p$ and parameter $\theta_i = 1$ if $1 \leq i \leq d/2$ and $\theta_i = 1 - \frac{\Delta}{m}$ otherwise.*

Then there exists universal constants C_2, C_3 such that for all $T \geq T_0 \equiv C_2 m^2 \ln m$, and $\Delta \leq 1/6$ and $m \geq 3$, the regret of Thompson Sampling is lower bounded by

$$R(T,\theta) \geq \frac{\Delta}{4p_\Delta}(1-(1-p_\Delta)^{T_0-1}) + C_3\Delta\frac{(1-g_\Delta)^{T_0} - (1-g_\Delta)^{T-T_0+1}}{g_\Delta}$$

with

$$g_\Delta = \frac{(2\Delta)^m}{m!} \text{ and } p_\Delta = \exp\left\{-\frac{2m}{9}(\frac{1}{2} - (\frac{\Delta}{m} + \frac{1}{\sqrt{m}}))^2\right\}$$

Corollary 3. *Consider a linear combinatorial bandit problem over combinatorial set $\mathcal{X}^p$ with $m \geq 3$. If $T > 2^m m!$ then the minimax regret of Thompson sampling is lower bounded by*

$$\max_{\theta\in[0,1]^d} R(T,\theta) \geq C_4 m T^{1-\frac{2}{d}}$$

Otherwise it is bounded by

$$\max_{\theta\in[0,1]^d} R(T,\theta) \geq C_4' T$$

with $C_4 > 0$ and $C_4' > 0$ universal constants. The minimax regret of ESCB for this set $\mathcal{X}^p$ is upper bounded by:

$$\max_{\theta\in[0,1]^d} R(T,\theta) \leq C_5 d^{\frac{1}{4}}\sqrt{T(\ln T)}.$$

with $C_5 \geq 0$ a universal constant. Therefore Thompson Sampling is not *minimax optimal.*

3.4 Linear Combinatorial Bandits with Forced Exploration

We finally extend our results to show that, even when forced exploration is added, Thompson Sampling still provably incurs a regret growing exponentially with the dimension, as stated by Theorem 4 and Corollary 5. In particular, if the number of forced exploration rounds ℓ satisfies

$$\ln\left(\frac{1}{1-\frac{\Delta}{m}}\right) \leq \frac{2\left(\frac{1}{\frac{\ell}{2}+2} - (\frac{\Delta}{m} + \frac{1}{\sqrt{m}})\right)^2}{(\frac{\ell}{2}+3)^2\frac{\ell}{2}}$$

then Theorem 4 implies that the regret of Thompson Sampling still increases exponentially in m, in spite of the forced exploration added to the algorithm. In fact it is impossible to set ℓ to prevent exponential regret from happening, unless the learner knows the value of the gap Δ in advance. Indeed, for any fixed ℓ, the above inequality always holds providing that Δ is small enough.

Theorem 4. *Consider a linear combinatorial bandit problem over combinatorial set $\mathcal{X}^p$ and parameter $\theta_i = 1$ if $1 \leq i \leq d/2$ and $\theta_i = 1 - \frac{\Delta}{m}$ otherwise.*

Then if $\frac{\Delta}{m} + \frac{1}{\sqrt{m}} < \frac{1}{\frac{\ell}{2}+2}$ the regret of Thompson Sampling with ℓ forced exploration rounds is lower bounded by

$$R^{TSFE}(T,\theta) \geq (1-\frac{\Delta}{m})^{\frac{m\ell}{2}}\frac{\Delta}{4p^\ell}(1-(1-p^\ell)^{T-1}) \text{ with}$$

$$p^\ell = \exp\{-2m\Big(\frac{1}{\frac{\ell}{2}+2} - (\frac{\Delta}{m} + \frac{1}{\sqrt{m}})\Big)^2/(\frac{\ell}{2}+3)^2\}$$

Theorem 5. *Consider a linear combinatorial bandit problem over combinatorial set $\mathcal{X}^p$ with $m \geq 3$ and parameter $\theta_i = 1$ if $1 \leq i \leq d/2$ and $\theta_i = 1 - \frac{\Delta}{m}$ otherwise.*

Then if $T > 4^m m!$ and $T > T_0(m,l) \equiv C8\ell^2 m^2 \ln m$ the minimax regret of Thompson sampling with ℓ forced round of exploration is lower bounded by :

$$\max_{\theta\in[0,1]^d} R(T,\theta) \geq C_7 C(\ell,m)\frac{m}{\ell}T^{1-\frac{1}{m}}$$

Otherwise it is bounded by

$$\max_{\theta \in [0,1]^d} R(T, \theta) \geq C_7' C(\ell, m) \frac{T}{\ell}$$

with $C_7 > 0$ *and* $C_7' > 0$ *universal constants and* C *a function such that* $\forall \ell \in \mathbb{N}, \lim_{m \to \infty} C(\ell, m) = 1$

4 Numerical Experiments

We now illustrate the exponential regret of Thompson sampling in practical settings using numerical experiments. Due to this exponential nature, some of those experiments involve a significant amount of computing time in high dimensions. Due to limited space, we solely consider the linear case, which is most often considered in the literature. Unless specified otherwise we use 1000 independent sample paths for averaging, and 95% confidence intervals are presented on the plots.

First selection of the optimal decision As shown by our theoretical results, the first time that the optimal decision is selected $\tau = \min\{t \geq 1 : x(t) = x^\star\}$ can be exponentially large, and this is what causes exponential regret. On Figure 3 , we present the c.d.f. (cumulative distribution function) of τ as a function of m for combinatorial sets $\mathcal{X}^p$ and $\mathcal{X}^m$ introduced above. The parameter values are chosen as in the previous sections $\theta_i = 1$ if $1 \leq i \leq d/2$ and $\theta_i = \frac{\Delta}{m}$ otherwise. For each sample path we generate τ by simulating Thompson Sampling until the optimal decision is played for the first time.

Some quantiles of τ indeed seem to increase exponentially as m grows and for reasonable values of m, τ can be very large with appreciable probability, for instance on Fig. 3 for $m = 14$, $\tau \geq 5.10^4$ with probability greater than 0.1. Clearly, on sample paths where this happens, Thompson sampling performs worse than random choice for the first 5.10^4 time steps which is surprisingly poor behaviour, especially on such a simple problem. This also showcases the fact that those sample paths happen relatively often. Thus the regret of Thompson sampling is not only due to very rare occasion with high regret but those poor behavior of Thompson sampling can indeed happen quite often.

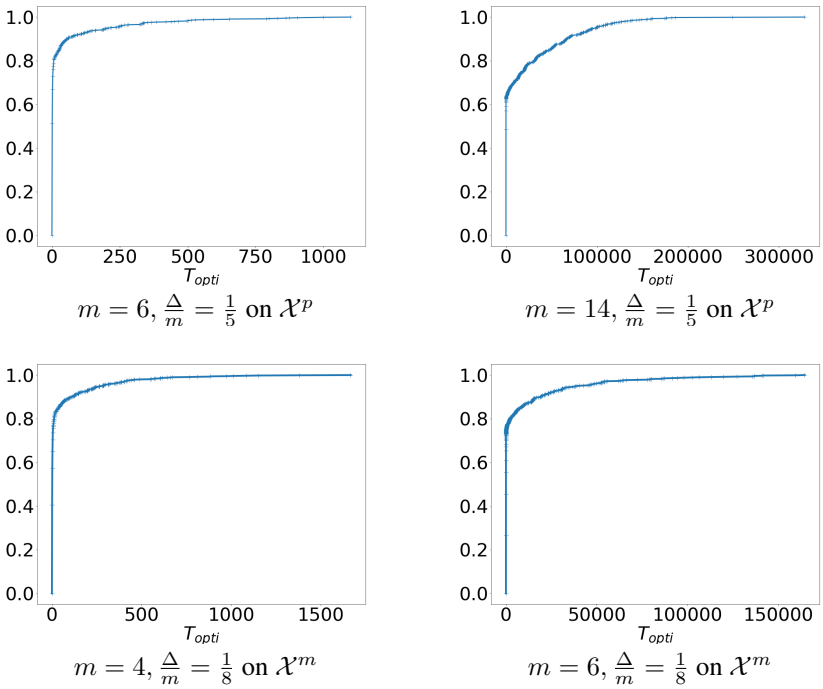
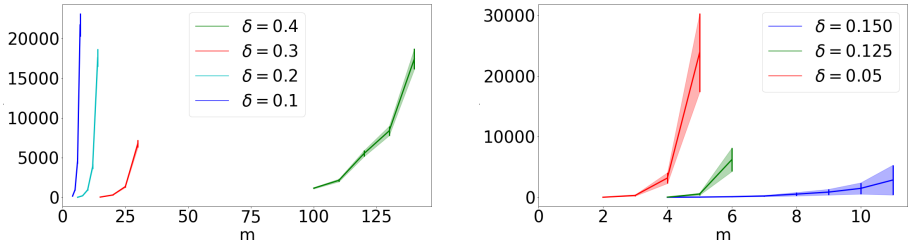


Figure 3: C.d.f. of the first time the optimal decision is played τ as a function of m for set of paths and matchings $\mathcal{X}^p$, $\mathcal{X}^m$

To investigate the impact of the gap Δ, on Figure 4a and Figure 4b, we plot the expected first time the optimal decision is selected $\mathbb{E}(\tau)$ as a function of m for various $\delta = \frac{\Delta}{m}$, for the sets of paths $\mathcal{X}^p$ and matchings $\mathcal{X}^m$. Once again we observe an exponential growth in both figures, and this growth is particularly fast for small values of δ. When the gap gets small, the exponential growth of regret is exacerbated, leading to an even worse performance.

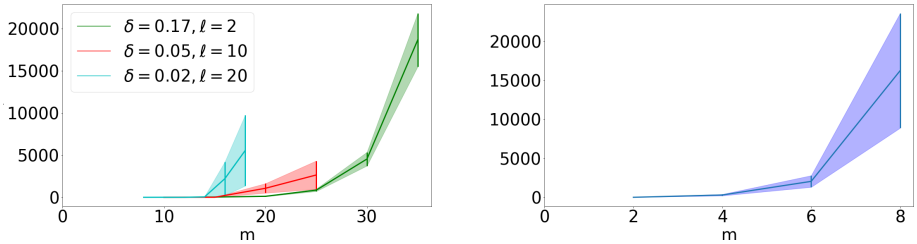


(a) Set of paths $\mathcal{X}^p$ combinatorial set (b) Set of matchings $\mathcal{X}^m$ combinatorial set

Figure 4: Expectation of the first time the optimal decision is played $\mathbb{E}(\tau)$ as a function of m and various values of $\frac{\Delta}{m} = \delta$

Impact of forced exploration We now investigate if forced exploration alleviates the problem in practice, and consider ℓ forced exploration rounds. On figures 5a we plot the expected first time the optimal decision is selected $\mathbb{E}(\tau)$ as a function of m for various $\delta = \frac{\Delta}{m}$ and ℓ, for the sets of paths $\mathcal{X}^p$. As predicted by theorem 4, $\mathbb{E}(\tau)$ still seems to scale exponentially in m and induce en exponential regret.

Theorem 4 states that if ℓ is chosen such that $\frac{\Delta}{m} < \frac{2}{\ell+1}$ then regret scales exponentially. On the other hand one could think that when ℓ is chosen large enough to violate this condition then regret does not grow as rapidly. Figure 5b shows that, at least numerically, this does not appear to be the case, indeed, we choose $\ell = m$, and there are values of δ such that the regret still seem to scale exponentially in m.

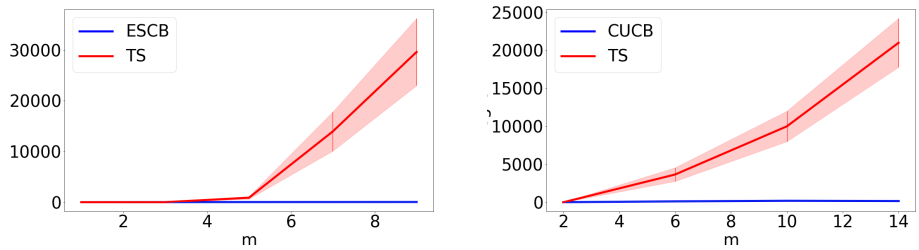


(a) Set of paths $\mathcal{X}^p$ combinatorial set (b) Set of matchings $\mathcal{X}^m$ combinatorial set $\frac{\Delta}{m} = \frac{1}{100}$ and $\ell = m$

Figure 5: Expectation of the first time the optimal decision is played $\mathbb{E}(\tau)$ as a function of m and various values of $\frac{\Delta}{m} = \delta$ and ℓ

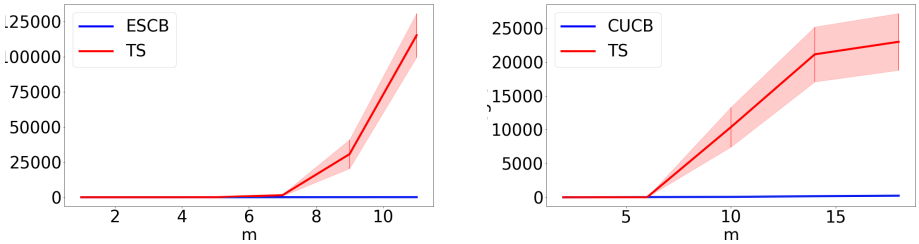
Comparison with Optimistic Algorithms We now compare Thompson sampling with respectively the state-of-the-art frequentist algorithms ESCB and/or CUCB. These experiments are averaged over 40 sample path due to computational limits. On figure 6a and figure 6b we present the regret as a function of m for the set of paths $\mathcal{X}^p$, $\frac{\Delta}{m} = 0.1$, $T = 4.10^4$ and the set of matchings $\mathcal{X}^m$, $\frac{\Delta}{m} = 0.05$, $T = 4.10^4$ respectively. The results show that the regret of Thompson Sampling is larger than that of CUCB and/or ESCB by several orders of magnitude in high dimensions, as predicted by our theoretical results. In fact the regret of Thompson Sampling is so overwhelmingly large that, due to the scale of the figure, it looks like the regret of ESCB and/or CUCB does not increase with the dimension (this is of course not the case).

On figures 7a and 7b we perform similar experiments but with smaller gaps, $\frac{\Delta}{m} = \frac{1}{m}$ and the same behaviour arises.



(a) $\delta = 0.1$ and $T = 4.10^4$, Set of paths $\mathcal{X}^m$ (b) $\delta = 0.05$ and $T = 10^5$, set of matchings $\mathcal{X}^m$

Figure 6: Regret comparison between ESCB and TS for set of paths $\mathcal{X}^p$ and matching $\mathcal{X}^m$ (Averaged over 40 experiences)



(a) $\Delta = 1$ and $T = 4.10^5$, set of paths $\mathcal{X}^m$ (b) $\Delta = 1$ and $T = 10^5$, set of matchings $\mathcal{X}^m$

Figure 7: Regret comparison between ESCB and TS for set of paths $\mathcal{X}^p$ and matching $\mathcal{X}^m$ For a fixed Δ (Averaged over 20 and 40 experiences)

5 Conclusion

We have shown through both theoretical analysis as well as numerical experiments that Thompson Sampling can perform very poorly in high dimensions, both for both linear and non linear problems, and for various combinatorial structures such as sets of paths and matchings (one could consider more complex combinatorial set including multiple non disjoint paths). Introducing forced exploration does not alleviate the problem either. Therefore, this is not an artifact, but rather a general problem. In essence, Thompson performs poorly because it has a tendency to play much too greedily, and in high dimensions this sometimes leads to a complete lack of exploration and missing the optimal arm. Our work points towards a new challenging open problem which is to design better Thompson-like algorithms for regret minimization that can deal with high-dimensional problems, while retaining the computational efficiency of Thompson Sampling. Our work concern the Bernoulli setting but we believe that our results can be generalized to bounded distributions. It is not obvious whether or not our results still hold for Gaussian distributions, and seems like an interesting open problem.

References

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

[2] O. Cappé, A. Garivier, O. Maillard, R. Munos, and G. Stoltz. Kullback-leibler upper confidence bounds for optimal sequential allocation. *Annals of Statistics*, 41(3):516–541, June 2013.

[3] Wei Chen, Yajun Wang, and Yang Yuan. Combinatorial multi-armed bandit: General framework and applications. In *Proc. of ICML*, 2013.

[4] Richard Combes, Stefan Magureanu, and Alexandre Proutiere. Minimal exploration in structured stochastic bandits. In *Proc. of NIPS*, 2017.

[5] Richard Combes, M. Sadegh Talebi, Alexandre Proutiere, and Marc Lelarge. Combinatorial bandits revisited. In *Proc. of NIPS*, 2015.

[6] Thibaut Cuvelier, Richard Combes, and Eric Gourdin. Statistically efficient, polynomial time algorithms for combinatorial semi bandits. In *Arxiv 2002.07258*, 2020.

[7] Remy Degenne and Vianney Perchet. Combinatorial semi-bandit with known covariance. In *Proc. of NIPS*, 2016.

[8] Aditya Gopalan, Shie Mannor, and Yishay Mansour. Thompson sampling for complex online problems. In *Proc. of ICML*, 2014.

[9] Junya Honda and Akimichi Takemura. An asymptotically optimal bandit algorithm for bounded support models. In *Proc. of COLT*, pages 67–79, 01 2010.

[10] Emilie Kaufmann, Nathaniel Korda, and Rémi Munos. Thompson sampling: An asymptotically optimal finite-time analysis. In *Proc. of ALT*, 2012.

[11] Branislav Kveton, Zheng Wen, Azin Ashkan, and Csaba Szepesvari. Tight regret bounds for stochastic combinatorial semi-bandits. In *Proc. of AISTATS*, 2015.

[12] T.L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–2, 1985.

[13] Tze Leung Lai. Adaptive treatment allocation and the multi-armed bandit problem. *Ann. Statist.*, 15(3):1091–1114, 09 1987.

[14] Pierre Perrault, Etienne Boursier, Michal Valko, and Vianney Perchet. Statistical efficiency of thompson sampling for combinatorial semi-bandits. In *Proc. of NIPS*, 2020.

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

[16] Siwei Wang and Wei Chen. Thompson sampling for combinatorial semi-bandits. In *Proc. of ICML*, 2018.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] See our contribution part in the introduction, i.e. section 1
   (b) Did you describe the limitations of your work? [Yes] See subsection 3.1. Also, the assumptions of each theorem are clearly stated.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] We do not believe that our work may have a negative societal impact.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] The assumptions for each theorem are clearly stated.
   (b) Did you include complete proofs of all theoretical results? [Yes] All the lemmas and proofs used are in the supplementary material.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] All code used is provided as supplementary material. See the jupyter notebook in the supplementary material. It is recommended to open it with "Collapsible Headings" option on because of the length of the notebook.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] The parameters chosen are in the jupyter notebook provided as well as in the "Numerical Experiments" section. We mainly chose these parameters so that computation times remain reasonable.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] All figures show error bars which represent the 95% confidence intervals.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] While our experiments are not extremely computationally intensive, we still also picked parameters so that the computation does not take to long, because of the exponential nature of the problem.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]