Thompson Sampling For Combinatorial Bandits

Polynomial Regret and Mismatched Sampling Paradox

Raymond Zhang Richard Combes

Laboratoire des signaux et systèmes, CentraleSupélec, CNRS, Université Paris-Saclay

Abstract

We consider Thompson Sampling (TS) for linear combinatorial semi-bandits and subgaussian rewards. We propose the first known TS whose finite-time regret does not scale exponentially with the dimension of the problem. We further describe what we call the "mismatched sampling paradox". The code used to generate the experiments is available at https://github.com/RaymZhang/CTS-Mismatched-Paradox

The stochastic combinatorial bandit model

At time $t = 1, 2, ..., T$,

1. A decision maker selects a decision $A(t) \in \mathcal{A}$ where $\mathcal{A} \subset \{0,1\}^d$
2. The environment draws a random vector $X(t) \in \mathbb{R}^d$
3. The learner observes $Y(t) = A(t) \odot X(t)$ (Semi bandit feedback).
4. The learner receives a Linear reward $r(t) = A(t)^\top X(t)$

Minimize :

$$R(T,\mu^*) := T \max_{A\in\mathcal{A}} \left\{ \mathbb{E}\left[A^\top X(t)\right] \right\} - \sum_{t=1}^{T} \mathbb{E}\left[A(t)^\top X(t)\right]$$
$$= T \max_{A\in\mathcal{A}} \left\{ A^\top \mu^\star \right\} - \sum_{t=1}^{T} \mathbb{E}\left[A(t)^\top X(t)\right].$$

Assumption: *The random variables $(X(t))_{t\in[T]}$ are independent and identically distributed through time with independent entries, with mean $\mathbb{E}\left[X(t)\right] = \mu^\star$ that we suppose unknown. The random variables are subgaussian with variance proxy σ^2 i.e. $\forall \lambda \in \mathbb{R}^d$ and $\forall t \in [T]$ we have :*

$$\mathbb{E}\left[\exp\left(\lambda^\top (X(t) - \mu^\star)\right)\right] \leqslant \exp\left(\frac{\|\lambda\|^2\sigma^2}{2}\right).$$

The Boosted variance Gaussian Combinatorial Thompson Sampling (BG-CTS)

Input: $\lambda > 0, \sigma > 0$
Initialization : $\forall i \in [d], N_i(0) = 0, \widehat{\mu}_i(0) = 0$ select decisions until $\min_{i\in[d]} N_i(t) > 0$. Update $\forall i \in [d], N_i(0), \mu_i(0)$ accordingly. Generate $\forall t \in [T], \forall i \in [d], Z_i(t) \sim \mathcal{N}(0,1)$ i.i.d.
for $t = 1, ..., T$ **do**

Compute $\theta_i(t) = \widehat{\mu}_i(t-1) + \sigma\sqrt{\frac{2g(t)}{N_i(t-1)}} Z_i(t)$.
Compute $A(t) = \arg\max_{A\in\mathcal{A}} \{A^T\theta(t)\}$.
The environment draws $\forall i \in [d], X_i(t)$.
Observe $X(t) \odot A(t)$, Receive reward $A(t)^T X(t)$.
Update $\forall i \in A(t), N_i(t) = N_i(t-1) + 1, \widehat{\mu}_i(t) = \frac{N_i(t)-1}{N_i(t)}\mu_i(t-1) + \frac{X_i(t)}{N_i(t)}$.

end

Here $g(t)$ is a boost function defined as :

$$g(t) := \frac{(1+\lambda)\left(\ln t + (m+2)\ln\ln t + \frac{m}{2}\ln\left(1+\frac{1}{e}\right)\right)}{\ln(t)} \qquad (1)$$

Main Theorem

Theorem: *For $\lambda = 1$, and σ^2 subgaussian rewards, the regret of BG-CTS is upper bounded by:*

$$R(T,\mu^\star) \leqslant C\frac{\sigma^2 d \ln m}{\Delta_{\min}} \ln T + C'\frac{\sigma^2 d^2 m \ln m}{\Delta_{\min}} \ln\ln T + P\left(m, d, \frac{1}{\Delta_{\min}}, \Delta_{\max}, \sigma\right) \qquad (2)$$

with $m := \max_A \|A\|_1$, $\Delta_A = A^\top\mu^\star - A^\top\mu^\star$, $\Delta_{\min} = \min_{A,\Delta_A>0} \Delta_A$, C, C' universal constants and P a polynomial in $m, d, \frac{1}{\Delta_{\min}}, \sigma, \Delta_{\max}$.

Reminder of the vanilla Combinatorial Thompson Sampling

We set a prior $\pi(\mu)$ on $\mu^\star$, then at each time select decision

$$A(t) \in \arg\max_{A\in\mathcal{A}} \left(A^\top\theta(t)\right) \text{ with } \theta(t) \sim p(\mu|(A(s), Y(s))_{s\in[t-1]})$$

- Example 1: (B-CTS) Bernoulli rewards and uniform priors on $[0,1]^d$, then
$$\theta_i(t) \sim \text{Beta}\left(N_i(t-1)\widehat{\mu}_i(t-1), N_i(t-1)(1-\widehat{\mu}_i(t-1))\right)$$
- Example 2: (G-CTS) Gaussian rewards with variance σ^2 and uniform priors on $\mathbb{R}^d$, then
$$\theta_i(t) \sim \mathcal{N}\left(\widehat{\mu}_i(t-1), (1+\lambda)\frac{\sigma^2}{N_i(t-1)}\right)$$

Note that BG-CTS is G-CTS with an enlarged variance of $g(t)$.

Upper bound regret comparison

The best known upper bound of vanilla Thompson Sampling for combinatorial bandits are given in theorem 3 and 1 in [3], they are respectively :

- For example 2 :
$$O\left(\frac{\sigma^2 d(\ln m)^2}{\Delta_{\min}} \ln T + \frac{dm^3}{\Delta_{\min}^2} + m\left(\sigma\frac{m^2+1}{\Delta_{min}}\right)^{2+4m}\right).$$
With a universal constant C.
- For example 1 : It is the same upper bound with $\sigma = 1$ and another universal constant C'.

Lower bound of Thompson Sampling on a Bernoulli case

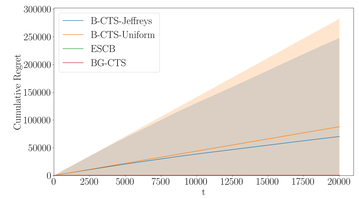
In the paper [4], we showed that if Thompson sampling is applied to the following combinatorial set : $\mathcal{A} = \{A^1, A^2\}$ with action of size $m = d/2$ with $A^1 = (1, ..., 1, 0, ..., 0)$ and $A^2 = (0, ..., 0, 1, ..., 1)$ and parameters $\mu^\star = (\mu_1, ..., \mu_1, \mu_2, ..., \mu_2)$ $\mu_1, \mu_2 > 0.5$ with a uniform prior and Bernoulli likelihood then :

$$R(T,\mu^\star) \geqslant \frac{\Delta_{\min}}{4p_{\Delta_{\min}}}(1-(1-p_{\Delta_{\min}})^{T-1}), \text{ with } p_{\Delta_{\min}} = \exp\left\{-\frac{2m}{9}\left(\frac{1}{2} - (\frac{\Delta_{\min}}{m} + \frac{1}{\sqrt{m}})\right)^2\right\}.$$

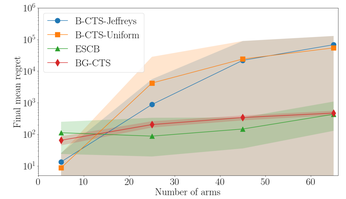
Here $p_{\Delta_{\min}}$ increases exponentially with m and for large T, the regret is exponential in m.

Numerical experiments

We showcase a performance comparison between Thompson Sampling and the Boosted Gaussian Thompson Sampling and ESCB [1] on the previous combinatorial set that showed exponential regret.



(a) Average regret over time



(b) Average final regret as a function of m

Rationale: self normalized concentration inequalities [2]

This function $g(t)$ is chosen to ensure that :

$$\mathbb{P}\left(\sup_{s\leqslant t} \frac{|A^{\star\top}(\widehat{\mu}(s) - \mu^\star)|}{\sqrt{A^{\star\top}V(s)A^\star}} \geqslant \sigma\sqrt{2\ln(t)g(t)}\right) \approx \frac{1}{t(\ln t)^2}$$

where $V(t) = \text{diag}((N_i(t)^{-1})_{i\in[d]})$

Thanks to this boost the proof relies on showing that with high probability :

$$\forall t \in [T], \sum_s^t \mathbf{1}\left\{A^{\star\top}\theta(s) > A^{\star\top}\mu^\star\right\} > ct^\beta$$

with a constant $\beta > 0$.

Some thoughts on this work

- Putting (infinite !) mass outside the possible range of parameters can be beneficial and even exponentially better. This is the Mismatched Sampling Paradox.
- The prior on the rewards of each action is actually not uniform but can be very concentrated if you put a uniform prior on a bounded set for each parameter.
- There still seem to be a trade-off between the complexity of the methods and the regret.

References

[1] Richard Combes, Stefan Magureanu, and Alexandre Proutiere. "Minimal Exploration in Structured Stochastic Bandits". In: *Proc. of NeurIPS*. 2017.
[2] Remy Degenne and Vianney Perchet. "Combinatorial semi-bandit with known covariance". In: *Proc. of NeurIPS*. 2016.
[3] Pierre Perrault et al. "Statistical Efficiency of Thompson Sampling for Combinatorial Semi-Bandits". In: *Proc. of NeurIPS*. 2020.
[4] Raymond Zhang and Richard Combes. "On the Suboptimality of Thompson Sampling in High Dimensions". In: *Proc. of NeurIPS*. 2021.

raymond.zhang@centralesupelec.fr

richard.combes.centralesupelec.fr