Thompson Sampling for Combinatorial Bandits

Polynomial Regret and Mismatched Sampling Paradox

Raymond Zhang Richard Combes
name.surname@centralesupelec.fr
November 12, 2024

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

Outline

1. Combinatorial Bandits
2. Thompson Sampling for Combinatorial Bandits

Combinatorial Bandits

Combinatorial Bandits model

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

Combinatorial Bandits model

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

1. A decision maker selects a decision $A(t) \in \mathcal{A}$ where $\mathcal{A} \subset \{0, 1\}^d$

Combinatorial Bandits 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 then draws a random vector $X(t) \in \mathbb{R}^d$ where the $(X(t))_{t\in[T]}$ are i.i.d. with $\mathbb{E}[X(t)] := \mu^\star$.

We also assume that the entries of X are subgaussian of parameter σ, $\forall \lambda \in \mathbb{R}^d, \mathbb{E}\left[\exp(\lambda^\top(X(t) - \mu^\star))\right] < \exp\left(\frac{||\lambda||^2\sigma^2}{2}\right)$

Combinatorial Bandits 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 then draws a random vector $X(t) \in \mathbb{R}^d$ where the $(X(t))_{t \in [T]}$ are i.i.d. with $\mathbb{E}[X(t)] := \mu^\star$.

We also assume that the entries of X are subgaussian of parameter σ, $\forall \lambda \in \mathbb{R}^d, \mathbb{E}\left[\exp(\lambda^\top(X(t) - \mu^\star))\right] < \exp\left(\frac{||\lambda||^2 \sigma^2}{2}\right)$

3. The learner then observes $Y(t) = A(t) \odot X(t)$ (Semi bandit feedback).

Combinatorial Bandits 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 then draws a random vector $X(t) \in \mathbb{R}^d$ where the $(X(t))_{t \in [T]}$ are i.i.d. with $\mathbb{E}[X(t)] := \mu^\star$.

We also assume that the entries of X are subgaussian of parameter σ, $\forall \lambda \in \mathbb{R}^d, \mathbb{E}\left[\exp(\lambda^\top (X(t) - \mu^\star))\right] < \exp\left(\frac{||\lambda||^2 \sigma^2}{2}\right)$

3. The learner then observes $Y(t) = A(t) \odot X(t)$ (Semi bandit feedback).

4. Receives a Linear reward $r(t) = A(t)^\top X(t)$

Goal

Minimize :

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

Thompson Sampling for Combinatorial Bandits

Thompson Sampling[1]

Given a prior on the parameter μ^* : $\pi(\mu)$

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

1. Thompson Sampling draws $\theta(t)$ from the posterior distribution $\pi_{t-1}(t) := \pi(\mu | X(t-1), ..., X(0), A(t-1), ..., A(0))$ and selects :

$$A(t) \in \arg\max_{A \in \mathcal{A}} \{A^\top \theta(t)\}$$

2. The environment then draws a random vector $X(t) \in \mathbb{R}^d$. The learner then observes :

$$Y(t) = A(t) \odot X(t)$$

3. Receives a Linear reward $r(t) = A(t)^\top X(t)$

[1] Wang and Chen 2020.

Thompson Sampling

4. Update the posterior $\pi_t(\mu)$ using the Bayes rule.

If we suppose $X(t)$ to be Gaussian with variance $\sigma^2 I_d$ and mean $\mu^\star$. It is reasonable to give ourselves a prior $\pi_0(\mu)$ uniform on $\mathbb{R}^d$ and a Gaussian likelihood with variance σ^2.

The posterior can therefore be written :

$$\forall i \in [d], \theta_i(t) \sim \mathcal{N}\left(\frac{\sum_s^t Y_i(s)}{N_i(t)}, \frac{\sigma^2}{N_i(t)}\right) \quad (1)$$

With $N_i(t) := \sum_s^t A_i(s)$ the number of time item i has been selected.

Our proposed version of TS[2]

We propose to draw :

$$\forall i \in [d], \theta_i(t) \sim \mathcal{N}\left(\frac{\sum_s^t Y_i(s)}{N_i(t)}, \frac{2g(t)\sigma^2}{N_i(t)}\right) \tag{2}$$

With :

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

With $m := \max_{A \in \mathcal{A}} \|A\|_1$. Note that $g(t) \to 2$

[2]Zhang and Combes 2024.

Regret of Thompson Sampling

Upper bound of algorithm the first version (1) for subgaussian rewards :

$$O\left(\frac{\sigma^2 d(\ln m)^2}{\Delta_{\min}}\ln T + \frac{dm^3}{\Delta_{\min}^2} + m\left(\sigma\frac{m^2+1}{\Delta_{min}}\right)^{2+4m}\right).$$

Upper bound of algorithm the second version (2) for subgaussian rewards :

$$O\left(\frac{\sigma^2 d\ln m}{\Delta_{\min}}\ln T + \frac{\sigma^2 d^2 m\ln m}{\Delta_{\min}}\ln\ln T + P\left(m, d, \frac{1}{\Delta_{\min}}, \Delta_{\max}, \sigma\right)\right)$$

The degrees of the polynomial in $m, d, 1/\Delta_{\min}, \sigma$ are respectively $30, 10, 20, 20$.

Lower bound of TS for Bernoulli rewards

In our paper[3] we proved a lower bound for the regret of Thompson Sampling for Bernoulli rewards and Bernoulli likelihood and Beta prior:

$$R(T,\theta) \geqslant \frac{\Delta_{\min}}{4p_{\Delta_{\min}}}(1-(1-p_{\Delta_{\min}})^{T-1})$$

With : $p_{\Delta_{\min}} = \exp\left\{-\frac{2m}{9}\left(\frac{1}{2}-(\frac{\Delta_{\min}}{m}+\frac{1}{\sqrt{m}})\right)^2\right\}$

[3] Zhang and Combes 2021.

References

Wang, Siwei and Wei Chen (Mar. 2020). **"Thompson Sampling for Combinatorial Semi-Bandits".** In: *arXiv:1803.04623 [cs]*. arXiv: 1803.04623. URL: http://arxiv.org/abs/1803.04623 (visited on 05/26/2020).

Zhang, Raymond and Richard Combes (Oct. 2021). **"On the Suboptimality of Thompson Sampling in High Dimensions".** In: *arXiv:2102.05502 [cs, stat]*. URL: http://arxiv.org/abs/2102.05502 (visited on 04/19/2023).

— (Oct. 2024). ***Thompson Sampling For Combinatorial Bandits: Polynomial Regret and Mismatched Sampling Paradox.*** arXiv:2410.05441. URL: http://arxiv.org/abs/2410.05441 (visited on 10/17/2024).