Thompson sampling for combinatorial bandits Polynomial regret and mismatched sampling paradox

Raymond Zhang, joint work with Richard Combes

Université Paris Saclay, Centrale-Supélec / L2S, France

AEP 13, Tuesday, 3 December

Combinatorial bandits (10’)

Combinatorial semi-bandits

- A learner selects decision $A \in \mathcal{A} \subset \{0,1\}^d$ at time $t \in \{1, \dots, T\}$
- She obtains reward $A^\top X(t)$ with $(X(t))_t$ i.i.d. with mean $\mu^\star \in \mathbb{R}^d$ and independent entries
- She observes $A(t) \odot X(t) = (A_i(t) X_i(t))_{1 \le i \le d}$
- Goal: minimize regret

$$R(T) = \underbrace{\max_{A \in \mathcal{A}} \mathbb{E}\left[\sum_{t=1}^{T} A(t)^\top X(t)\right]}_{\text{oracle}} - \underbrace{\mathbb{E}\left[\sum_{t=1}^{T} A(t)^\top X(t)\right]}_{\text{your algorithm}} .$$

- Size $m = \max_{A \in \mathcal{A}} \|A\|_1$, gap $\Delta(A) = (\max_{a \in \mathcal{A}} A^\top \mu^\star) - A^\top \mu^\star$.

Cesa-Bianchi and Lugosi, 2011, "Combinatorial bandits"

Optimistic algorithms 1: CUCB

- Estimate of $\mu^\star$ at time t

$$\hat{\mu}_i(t) = \frac{1}{N_i(t)} \sum_{s<t} A_i(s) X_i(s) \text{ with } N_i(t) = \sum_{s<t} A_i(s)$$

- Optimistic algorithm, extension of UCB1, select

$$A(t) \in \arg\max_{A \in \mathcal{A}} \left[\sum_{i=1}^{d} A_i \hat{\mu}_i(t) + A_i \sqrt{\frac{2 \ln t}{N_i(t)}} \right]$$

- Implementable by linear programming over $\mathcal{X}$

Theorem (Kveton et al, 2014)

The regret of CUCB verifies $R(T) \le C_1 dm \ln T / \Delta_{\min}$ with C_1 a universal constant.

Kveton, Wen, Ashkan and Szepesvari, 2014, "Tight regret bounds for stochastic combinatorial semi-bandits"

Optimistic algorithms 2: ESCB

- Same idea as CUCB, with tighter confidence bounds

$$A(t) \in \arg\max_{A \in \mathcal{A}} \left[\sum_{i=1}^{d} A_i \hat{\mu}_i(t) + \sqrt{\sum_{i=1}^{d} A_i \frac{2 \ln t}{N_i(t)}} \right]$$

- Can be NP-Hard to implement, even if linear programming over $\mathcal{X}$ is polynomial

Theorem (Degenne et al, 2016)

The regret of ESCB verifies $R(T) \leq C_2 d (\ln m)^2 \ln T / \Delta_{\min} + P_2(m, d, 1/\Delta_{\min})$ *with* C_2 *a universal constant and* P_2 *a polynomial.*

Combes, Lelarge, Proutiere and Talebi, 2015, "Combinatorial bandits revisited"

Degenne and Perchet, 2016, "Combinatorial semi-bandit with known covariance"

Sampling algorithms: CTS

- Observations up to time t, $Y(t) = (A(s) \odot X(t))_{s<t}$, prior p_μ on $\mu^\star$
- Posterior sampling algorithm, select decision

$$A(t) \in \arg\max_{A \in \mathcal{X}} \left(A^\top \theta(t)\right) \text{ with } \theta(t) \sim p_{\mu | Y(t)}$$

- Example 1: (B-CTS) Bernoulli rewards and uniform priors, then

$$\theta_i(t) \sim \text{Beta}\left(N_i(t)(1-\hat{\theta}_i(t)), N_x(t)\hat{\theta}_i(t)\right)$$

  with independent entries.

Theorem (Perrault et al, 2012)

The regret of B-CTS verifies $R(T) \leq C_3 d (\ln m)^2 \ln T / \Delta_{\min} + Q_3(m, d, 1/\Delta_{\min})$ *with* C_3 *a universal constant and* Q *growing at least exponentially in* m, d.

Perrault, Boursier, Valko and Perchet, 2020, "Statistical efficiency of thompson sampling for combinatorial semi-bandits"

- Example 2: (G-CTS) Gaussian rewards and gaussian priors, then

$$A(t) \in \arg\max_{A \in \mathcal{A}} \left(A^\top \theta(t) \right) \text{ with } \theta(t) \sim N(\hat{\mu}(t), 2V(t))$$
$$V(t) = \text{diag}\left(\frac{1}{N_1(t)}, ..., \frac{1}{N_d(t)} \right)$$

Perrault, Boursier, Valko and Perchet, 2020, "Statistical efficiency of thompson sampling for combinatorial semi-bandits"

The BG-CTS Algorithm (5’)

BG-CTS algorithm

- Sampling algorithm

$$A(t) \in \arg\max \left(A^\top \theta(t) \right) \text{ with } \theta(t) \sim N(\hat{\mu}(t), 2g(t)V(t))$$
$$V(t) = \text{diag}\left(\frac{1}{N_1(t)}, ..., \frac{1}{N_d(t)} \right)$$

- Exploration boost

$$g(t) = (1+\lambda)\frac{\ln t + (m+2)\ln\ln t + (m/2)\ln(1+e/\lambda)}{\ln t}$$

- Similar to G-CTS for Gaussian rewards with a well chosen boost
- Implementable by linear programming over $\mathcal{X}$

Zhang and Combes, 2024, "Thompson sampling for combinatorial bandits: polynomial regret and mismatched sampling paradox"

Regret and complexity of BG-CTS

Theorem (Zhang et al, 2024)

Consider 1-subgaussian rewards. The regret of BG-CTS verifies $R(T) \leq C_4 d (\ln m) \ln T / \Delta_{\min} + P_4(m, d, 1/\Delta_{\min})$ *with* C_4 *a universal constant and* P_4 *a polynomial.*

- Valid for Gaussian, Bernoulli, bounded etc.
- If linear programming over $\mathcal{X}$ is polynomial then polynomial complexity
- Best known polynomial (complexity, regret) algorithm for asymptotic regret for general action set.
- Leads to an interesting paradox ...

Zhang and Combes, 2024, "Thompson sampling for combinatorial bandits: polynomial regret and mismatched sampling paradox"

Rationale: self normalized concentration inequalities

- Why is the "correct" confidence boost $g(t)$?
- Self-normalized concentration inequality, choose $g(t)$ such that

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

- The boost insure that : 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 $\beta>0$.

Degenne and Perchet, 2016, "Combinatorial semi-bandit with known covariance"

Mismatched sampling paradox (5')

Exponential regret of CTS

Theorem (Zhang et al, 2021)

Consider Bernoulli rewards. For any d there exists at least one θ and $\mathcal{X}$ such that the regret of B-CTS is greater than that of random choice for all $t \leq T_0(d)$ with T_0 growing at least exponentially in d. Also, B-CTS is not minimax optimal.

- CTS is too greedy, and can get "stuck" for exponentially long
- For $d = 20$, $T_0(d)$ is greater than the age of the universe (!)
- High dimensional phenomenon, when d is large enough, posterior is too concentrated around its mean

Zhang and Combes, 2021, "On the suboptimality of thompson sampling in high dimensions"

Mismatched sampling paradox

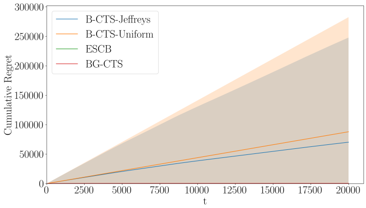
Consider a problem with Bernoulli rewards and parameters in $[0, 1]^d$.

- Learner 1 knows the rewards distribution and the support $[0, 1]^d$, uses a uniform (or Jeffrey's) prior over $[0, 1]^d$ and Bernoulli likelihood (B-CTS)
- Learner 2 does not know the rewards distribution and the support $[0, 1]^d$, uses a Gaussian prior and Gaussian likelihood over $\mathbb{R}^d$ and a boost (BG-CTS)

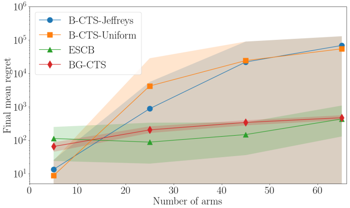
Paradox: Learner 1 performs <u>exponentially worse</u> than Learner 2

Zhang and Combes, 2024, "Thompson sampling for combinatorial bandits: polynomial regret and mismatched sampling paradox"

Performance comparison between Thompson Sampling and the Boosted Gaussian Thompson Sampling and ESCB.



(a) Average regret over time



(b) Average final regret as a function of m

Some reflections about sampling

- Sampling algorithms are fine, but posterior sampling sometimes does not work
- Putting mass outside of the parameter space can make things exponentially better (?!?)
- The Bayesian rationale of predicting using the posterior distribution is not universal for online problems
- Open problem: is there a simple rationale for designing efficient sampling algorithms for online problems ?

Zhang and Combes, 2024, "Thompson sampling for combinatorial bandits: polynomial regret and mismatched sampling paradox"

Thank you for your attention !

Paper here

https://arxiv.org/abs/2410.05441

Code here

https://github.com/RaymZhang/CTS-Mismatched-Paradox