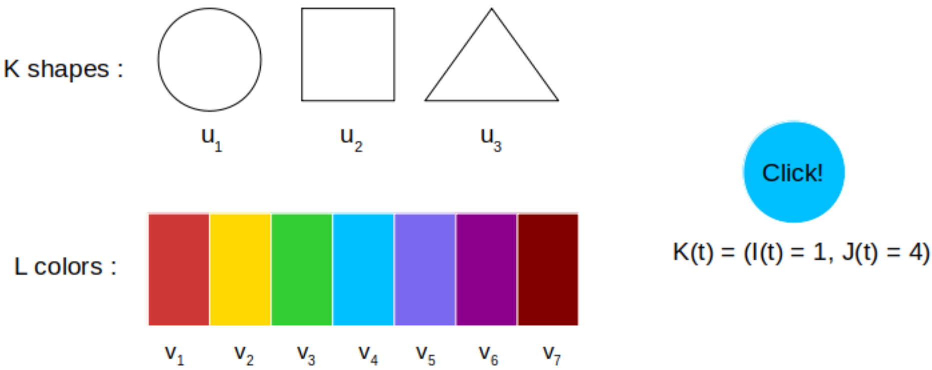
Solving Bernoulli Rank-One Bandits with Unimodal Thompson Sampling

Cindy Trinh, Émilie Kaufmann, Claire Vernade, Richard Combes

ALT 2020

Let us choose the design of a button for our website!



The user clicks if they are attracted both by the shape i and the color j.
If they click, the reward is 1.

The Rank-one bandit problem

Rank-one model Katariya et al. (2017b,a)

- $\mathbf{u} = (u_1, u_2, \ldots, u_K) \in [0,1]^K$ $\mathbf{v} = (v_1, v_2, \ldots, v_L) \in [0,1]^L$
- Arm $(i,j) \to$ Bernoulli distribution with mean $\mu_{ij} = u_i v_j$
- $\boldsymbol{\mu} = \boldsymbol{u}\boldsymbol{v}^T$ is a rank one matrix

At each time step $t = 1, \ldots, T$,

- The learner chooses an arm $K(t) = (I(t), J(t))$,
- And receives $r(t) \sim \mathcal{B}(\mu_{I(t),J(t)})$, where $\mu_{i,j} = u_i v_j$

Goal: minimize the expected cumulative regret

$$R_{\boldsymbol{\mu}}(T, \mathcal{A}) = \sum_{t=1}^{T} \left[\max_{(i,j)\in[K]\times[L]} \mu_{(i,j)} - \mathbb{E}_{\boldsymbol{\mu}}[\mu_{(I(t),J(t))}] \right]$$

Lower bound on the regret

Lower bound on the regret for Bernoulli Rank-one bandits, Katariya et al. (2017a). For any algorithm $\mathcal{A}$ which is uniformly efficient,

$$\liminf_{T\to\infty} \frac{R_{\boldsymbol{\mu}}(\mathcal{A},T)}{\log(T)} \geq \sum_{i\in[K]\setminus i_\star} \frac{\mu_{i_\star,j_\star}-\mu_{i,j_\star}}{\mathrm{kl}(\mu_{i,j_\star},\mu_{i_\star,j_\star})} + \sum_{j\in[L]/j_\star} \frac{\mu_{i_\star,j_\star}-\mu_{i_\star,j}}{\mathrm{kl}(\mu_{i_\star,j},\mu_{i_\star,j_\star})}.$$

Lower bound on the regret for the KL-armed bandit problem, Lai and Robbins (1985):

$$\liminf_{T\to\infty} \frac{R_{\boldsymbol{\mu}}(\mathcal{A},T)}{\log(T)} \geq \sum_{(i,j)\in[K]\times[L]\setminus(i_\star,j_\star)} \frac{\mu_{i_\star,j_\star}-\mu_{i,j}}{\mathrm{kl}(\mu_{i,j},\mu_{i_\star,j_\star})}.$$

$\rightarrow$ A good rank-one algorithm should sample arms which are not in the best row/column only $o(\log T)$ times.

$\mathrm{kl}(x,y) = x\ln(x/y) + (1-x)\ln((1-x)/(1-y))$

Previous work for Bernoulli Rank-one bandits

Lower bound on the regret for Bernoulli Rank-one bandits, Katariya et al. (2017a).

$$\liminf_{T\to\infty} \frac{R_{\boldsymbol{\mu}}(\mathcal{A}, T)}{\log(T)} \geq \sum_{i\in[K]\setminus i_\star} \frac{\mu_{i_\star,j_\star} - \mu_{i,j_\star}}{\mathrm{kl}(\mu_{i,j_\star}, \mu_{i_\star,j_\star})} + \sum_{j\in[L]/j_\star} \frac{\mu_{i_\star,j_\star} - \mu_{i_\star,j}}{\mathrm{kl}(\mu_{i_\star,j}, \mu_{i_\star,j_\star})}.$$

Upper bound on the regret for `Rank1ElimKL`, Katariya et al. (2017a):

$$R(T) \leq \frac{160}{\mu\gamma} \left(\sum_{i=1}^{K} \frac{1}{\bar{\Delta}_i^U} + \sum_{j=1}^{L} \frac{1}{\bar{\Delta}_i^V} \right) \log T + (6e + 82)(K + L)$$

However, does not match the LB:
$\rightarrow$ Is the lower bound achievable? Yes, by seeing the rank-one bandits as a *Graphical Unimodal bandit problem*.

Unimodal structure of the Rank-one model

Definition (Unimodal structure)

Let $G = (V, E)$ an undirected graph.
A vector $\boldsymbol{\mu} = (\mu_k)_{k \in V}$ is *unimodal with respect to G* if

- there exists a unique $k_\star \in V$ such that $\mu_{k_\star} = \max_i \mu_i$
- from any $k \neq k_\star$, we can find an increasing path to the optimal arm. ($p = (k, k_2, \dots, k_\star)$ such that $\mu_k < \mu_{k_2} < \dots < \mu_{k_\star}$)

Graph for the Rank-one model:

- $V = \{1, \dots, K\} \times \{1, \dots, L\}$
- $((i,j),(k,\ell)) \in E$ iff $(i,j) \neq (k,\ell)$ and $(i = k$ or $j = \ell)$

$$\begin{bmatrix} (u_1 v_1) & (u_1 v_2) & (\boldsymbol{u_1 v_3}) & (u_1 v_4) \\ (u_2 v_1) & (u_2 v_2) & (\boldsymbol{u_2 v_3}) & (u_2 v_4) \\ (\boldsymbol{u_3 v_1}) & (\boldsymbol{u_3 v_2}) & \boxed{(u_3 v_3)} & (\boldsymbol{u_3 v_4}) \\ (u_4 v_1) & (u_4 v_2) & (\boldsymbol{u_4 v_3}) & (u_4 v_4) \end{bmatrix}$$

Neighbors of arm $(3,3)$ displayed in bold

Unimodal structure of the Rank-one model

Definition (Unimodal structure)

Let $G = (V, E)$ an undirected graph.
A vector $\boldsymbol{\mu} = (\mu_k)_{k \in V}$ is *unimodal with respect to G* if

- there exists a unique $k_\star \in V$ such that $\mu_{k_\star} = \max_i \mu_i$
- from any $k \neq k_\star$, we can find an increasing path to the optimal arm. $(p = (k, k_2, \ldots, k_\star)$ such that $\mu_k < \mu_{k_2} < \cdots < \mu_{k_\star})$

Increasing path from arm $(3, 3)$ to the optimal arm $(1, 1)$:

$$\rightarrow p = ((3, 3), (1, 3), (1, 1))$$

$$\begin{bmatrix} (\boldsymbol{u_\star v_\star}) & (u_1 v_2) & (\boldsymbol{u_1 v_3}) & (u_1 v_4) \\ (u_2 v_1) & (u_2 v_2) & (u_2 v_3) & (u_2 v_4) \\ (u_3 v_1) & (u_3 v_2) & (\boldsymbol{u_3 v_3}) & (u_3 v_4) \\ (u_4 v_1) & (u_4 v_2) & (u_4 v_3) & (u_4 v_4) \end{bmatrix}$$

LB for Unimodal and Rank-one bandits

By considering the previous graph, the lower bound for Unimodal bandits (Combes and Proutière (2014))...

$$\liminf_{T\to\infty} \frac{R_{\boldsymbol{\mu}}(\mathcal{A}, T)}{\ln(T)} \geq \sum_{k\in\mathcal{N}_G(k_\star)} \frac{\mu_{k_\star} - \mu_k}{\mathrm{kl}\left(\mu_k, \mu_{k_\star}\right)}$$

...matches the lower bound for Rank-one bandits:

$$= \sum_{i\in[K]\setminus i_\star} \frac{\mu_{i_\star,j_\star} - \mu_{i,j_\star}}{\mathrm{kl}(\mu_{i,j_\star}, \mu_{i_\star,j_\star})} + \sum_{j\in[L]/j_\star} \frac{\mu_{i_\star,j_\star} - \mu_{i_\star,j}}{\mathrm{kl}(\mu_{i_\star,j}, \mu_{i_\star,j_\star})}$$

$\rightarrow$ Asymptotically optimal algorithm for Unimodal bandits are also asymptotically optimal for Rank-one bandits
$\rightarrow$ There exists optimal algorithms for Graphical Unimodal bandits: OSUB Combes and Proutière (2014), UTS Paladino et al. (2017)

Algorithm 1 Unimodal Thompson Sampling for Rank-one bandits

Input: $\gamma \in \mathbb{N}, \gamma \geq 2$.

Warm-up phase: Draw each arm once

for $t = KL + 1, \dots, T$ **do**

  Compute the leader $L(t) = (I_L(t), J_L(t)) = \underset{(i,j)\in[K]\times[L]}{\operatorname{argmax}} \hat{\mu}_{i,j}(t)$

  Update the leader count $\ell_{L(t)} \leftarrow \ell_{L(t)} + 1$

  if $\ell_{L(t)} \equiv 0\,[\gamma]$ **then**

    Draw the leader $(I(t), J(t)) = L(t)$

  else

    Perform TS over the extended neighborhood of the leader

    for $k \in \{(I_L(t), j) : j \in [L]\} \bigcup \{(i, J_L(t)) : i \in [K]\}$ **do**

      $\theta_k \sim \mathrm{Beta}\,(S_k + 1, N_k - S_k + 1)$

    end for

    $(I(t), J(t)) = \underset{k}{\operatorname{argmax}}\, \theta_k$.

  end if

  Receive reward $R_t \sim \mathcal{B}(\mu_{(I_t,J_t)})$, Update statistics

end for

Upper bound on the regret for UTS

Let $\boldsymbol{\mu}$ be a unimodal bandit instance with respect to a graph G. For all $\gamma \geq 2$, *epsilon* > 0,

$$\mathcal{R}_{\boldsymbol{\mu}}(T, \mathtt{UTS}(\gamma)) \leq (1+\varepsilon) \sum_{k \in \mathcal{N}(k_\star)} \frac{(\mu_\star - \mu_k)}{\mathrm{kl}(\mu_k, \mu_\star)} \log(T) + C(\boldsymbol{\mu}, \gamma, \varepsilon),$$

where $C(\boldsymbol{\mu}, \gamma, \varepsilon)$ is some constant depending on the environment $\boldsymbol{\mu}$, on ε and on γ.

$\rightarrow$ Matches the lower bound for Unimodal bandits problem:

$$\limsup_{T \to \infty} \frac{\mathcal{R}_{\boldsymbol{\mu}}(T, \mathtt{UTS}(\gamma))}{\log(T)} \leq \sum_{k \in \mathcal{N}(k_\star)} \frac{(\mu_\star - \mu_k)}{\mathrm{kl}(\mu_k, \mu_\star)}$$

Sketch of proof for the Upper Bound of UTS

Outline of the proof:

$$\mathcal{R}(T) = \sum_{k \neq k_\star} \Delta_k \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}(K(t) = k)\right]$$

$$= \left.\sum_{k \in \mathcal{N}(k_\star)} \Delta_k \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}(K(t) = k, L(t) = k_\star)\right]\right\}\mathcal{R}_1(T)$$

$$+ \left.\sum_{k \neq k_\star} \Delta_k \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}(K(t) = k, L(t) \neq k_\star)\right]\right\}\mathcal{R}_2(T)$$

$\mathcal{R}_1(T)$: When the leader is the optimal arm $k_\star$. Similar proof to that of TS restricted to $\mathcal{N}(k_\star)$.
$\mathcal{R}_2(T)$: When the leader is a suboptimal arm

Sketch of proof for the Upper Bound of UTS

Denote by $\mathcal{B}_{\mathcal{N}(k)} = \text{argmax}_{\ell \in \mathcal{N}(k)} \mu_\ell$, the set of best neighbors of k.

$$\mathcal{R}_2(T) \leq \sum_{k \neq k_\star} \sum_{t=1}^{T} \mathbb{P}\left(L(t) = k\right)$$

$$= \sum_{k \neq k_\star} \underbrace{\sum_{t=1}^{T} \mathbb{P}\left(L(t) = k, \forall k_2 \in \mathcal{B}_{\mathcal{N}(k)}, N_{k_2}(t) \leq (\ell_k(t))^b\right)}_{\substack{\text{With TS, it is unlikely that the optimal} \\ \text{arm in the neighborhood of } k \text{ are not often drawn often}}}$$

$$+ \sum_{k \neq k_\star} \underbrace{\sum_{t=1}^{T} \mathbb{P}\left(L(t) = k, \exists k_2 \in \mathcal{B}_{\mathcal{N}(k)}, N_{k_2}(t) > (\ell_k(t))^b\right)}_{\substack{k \text{ is unlikely to remain leader because of leader exploration,} \\ \text{and } k_2 \text{ is drawn often}}}$$

$\square$

Leader exploration parameter γ

$\rightarrow$ UTS draws the leader every γ times it has been leader.

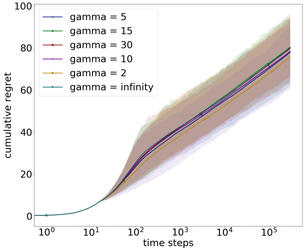


Figure: Cumulative regret of UTS for $\gamma \in \{2, 5, 10, 15, 30, +\infty\}$, $K = L = 4$.

- In our analysis, γ can be set to any arbitrary value in $\mathbb{N}$
- Empirically, forced exploration of the leader does not seem mandatory
- But $\gamma = 2$ yields good performance

Comparison with other algorithms

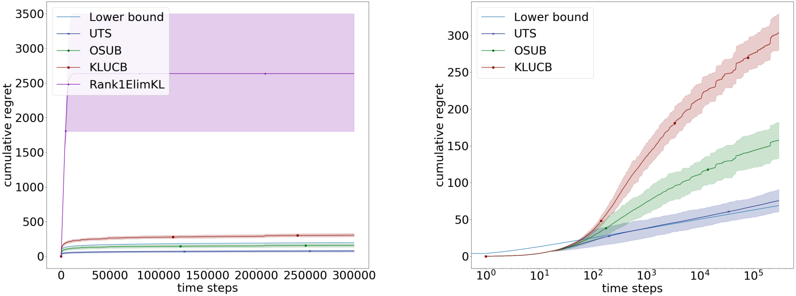


Figure: Cumulative regret of `Rank1ElimKL`, `OSUB`, `UTS`, `KL-UCB`, on 4×4 rank-one matrices (left). Regret in log-scale: the lower bound (in blue) shows the optimal asymptotic logarithmic growth of the regret. `UTS` and `OSUB` align with it, while `KL-UCB` has a larger slope (right).

Richard Combes and Alexandre Proutière. Unimodal bandits: Regret lower bounds and optimal algorithms. 2014.

Sumeet Katariya, Branislav Kveton, Csaba Szepesvári, Claire Vernade, and Zheng Wen. Bernoulli rank-1 bandits for click feedback. In *IJCAI*, 2017a.

Sumeet Katariya, Branislav Kveton, Csaba Szepesvári, Claire Vernade, and Zheng Wen. Stochastic rank-1 bandits. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, 2017b.

Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6 (1):4–22, 1985.

Stefano Paladino, Francesco Trovò, Marcello Restelli, and Nicola Gatti. Unimodal thompson sampling for graph-structured arms. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.