Thompson Sampling for Cascading Bandits

(Arxiv 1810.01187)

Wang Chi Cheung
NUS ISEM

Vincent Y. F. Tan
NUS ECE and Math

Zixin Zhong
NUS Math

National University of Singapore

ITA 2019

12 Feb 2019

What is the Cascading Bandits Problem?

Ground set

A set of all available items $[L] := \{1, \dots, L\}$.

What is the Cascading Bandits Problem?

Ground set

A set of all available items $[L] := \{1, \dots, L\}$.

Click probability/weight of item $i \in [L]$

A user clicks item i with prob. $w(i) \in [0, 1]$, independently of other items.

What is the Cascading Bandits Problem?

Ground set

A set of all available items $[L] := \{1, \dots, L\}$.

Click probability/weight of item $i \in [L]$

A user clicks item i with prob. $w(i) \in [0, 1]$, independently of other items.

Whether item i is clicked at time t

This is revealed by a random variable $W_t(i) \sim \mathrm{Bern}\,(w(i))$.

- $W_t(i) = 1$ iff the user examines and clicks on i at time t.
- $W_t(i) = 0$ iff the user examines but does not click on i at time t.

What is the Cascading Bandits Problem?

Ground set

A set of all available items $[L] := \{1, \ldots, L\}$.

Click probability/weight of item $i \in [L]$

A user clicks item i with prob. $w(i) \in [0, 1]$, independently of other items.

Whether item i is clicked at time t

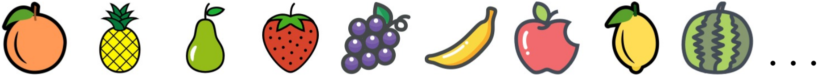
This is revealed by a random variable $W_t(i) \sim \mathrm{Bern}\,(w(i))$.

- $W_t(i) = 1$ iff the user examines and clicks on i at time t.
- $W_t(i) = 0$ iff the user examines but does not click on i at time t.

Applications

- Online recommender systems
- Movie suggestions by Netflix; Restaurant recommendations by Yelp

Cascading Bandits Setting (Kveton et al., 2015)



For time step $t = 1, 2, \ldots, T$:

1. The agent selects a list of K items $S_t := (i_1^t, \ldots, i_K^t) \in \pi_K(L)$ to the user, where $\pi_K(L) = \{$all K-permutations of $[L]\}$;

Cascading Bandits Setting (Kveton et al., 2015)

Recommendation 

For time step $t = 1, 2, \ldots, T$:

1. The agent selects a list of K items $S_t := (i_1^t, \ldots, i_K^t) \in \pi_K(L)$ to the user, where $\pi_K(L) = \{$all K-permutations of $[L]\}$;

Cascading Bandits Setting (Kveton et al., 2015)

Recommendation
Attractiveness

For time step $t = 1, 2, \ldots, T$:

1. The agent selects a list of K items $S_t := (i_1^t, \ldots, i_K^t) \in \pi_K(L)$ to the user, where $\pi_K(L) = \{$all K-permutations of $[L]\}$;
2. The user examines the items from i_1^t to i_K^t:
   - If she is **attracted** by an item, **clicks** on it;
   - If not, she skips to the next item and checks if it is attractive;
   - Process stops when she clicks on one item or when she comes to the end of the list.

Cascading Bandits Setting (Kveton et al., 2015)

Recommendation  

Attractiveness ×

For time step $t = 1, 2, \ldots, T$:

1. The agent selects a list of K items $S_t := (i_1^t, \ldots, i_K^t) \in \pi_K(L)$ to the user, where $\pi_K(L) = \{$all K-permutations of $[L]\}$;
2. The user examines the items from i_1^t to i_K^t:
   - If she is **attracted** by an item, **clicks** on it;
   - If not, she skips to the next item and checks if it is attractive;
   - Process stops when she clicks on one item or when she comes to the end of the list.

Cascading Bandits Setting (Kveton et al., 2015)

Recommendation 

Attractiveness $\times$ $\times$

For time step $t = 1, 2, \ldots, T$:

1. The agent selects a list of K items $S_t := (i_1^t, \ldots, i_K^t) \in \pi_K(L)$ to the user, where $\pi_K(L) = \{$all K-permutations of $[L]\}$;
2. The user examines the items from i_1^t to i_K^t:
   - If she is **attracted** by an item, **clicks** on it;
   - If not, she skips to the next item and checks if it is attractive;
   - Process stops when she clicks on one item or when she comes to the end of the list.

Cascading Bandits Setting (Kveton et al., 2015)

Recommendation 

Attractiveness × × ×

For time step $t = 1, 2, \ldots, T$:

1. The agent selects a list of K items $S_t := (i_1^t, \ldots, i_K^t) \in \pi_K(L)$ to the user, where $\pi_K(L) = \{$all K-permutations of $[L]\}$;
2. The user examines the items from i_1^t to i_K^t:
   - If she is **attracted** by an item, **clicks** on it;
   - If not, she skips to the next item and checks if it is attractive;
   - Process stops when she clicks on one item or when she comes to the end of the list.

Cascading Bandits Setting (Kveton et al., 2015)

Recommendation 

Attractiveness $\times$ $\times$ $\times$ $\surd$

For time step $t = 1, 2, \ldots, T$:

1. The agent selects a list of K items $S_t := (i_1^t, \ldots, i_K^t) \in \pi_K(L)$ to the user, where $\pi_K(L) = \{$all K-permutations of $[L]\}$;
2. The user examines the items from i_1^t to i_K^t:
   - If she is **attracted** by an item, **clicks** on it;
   - If not, she skips to the next item and checks if it is attractive;
   - Process stops when she clicks on one item or when she comes to the end of the list.

Cascading Bandits Setting (Kveton et al., 2015)

The agent **maximize his overall reward** over a fixed time horizon.

Instantaneous reward of the agent at time t

$$R(S_t|\boldsymbol{w}) := 1 - \prod_{k=1}^{K} \left(1 - W_t(i_k^t)\right) \in \{0, 1\}.$$

The agent gets a reward of

$R(S_t|\boldsymbol{w}) = 1$ if some i_k^t is clicked, (some $W_t(i_k^t) = 1$)

$R(S_t|\boldsymbol{w}) = 0$ if none of i_k^t is clicked. (all $W_t(i_k^t) = 0$)

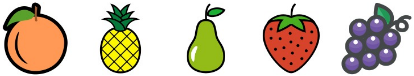
Cascading Bandits Setting (Kveton et al., 2015)

Feedback of the agent at time t

$$k_t := \min \left\{1 \leq k \leq K : W_t(i_k^t) = 1\right\},$$

where $\min \emptyset = \infty$. If $k_t < \infty$.

Cascading Bandits Setting (Kveton et al., 2015)



Recommendation					
Attractiveness	$\times$	$\times$	$\times$	$\surd$	?

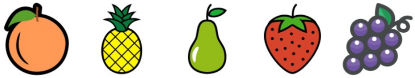
Feedback of the agent at time t

$$k_t := \min\left\{1 \le k \le K : W_t(i_k^t) = 1\right\},$$

where $\min \emptyset = \infty$. If $k_t < \infty$.

- $k_t < \infty$: the agent observes $W_t(i_k^t) = 0$ for $1 \le k < k_t$, and $W_t(i_k^t) = 1$, but does not observe $W_t(i_k^t)$ for $k > k_t$;

Cascading Bandits Setting (Kveton et al., 2015)



Recommendation

Attractiveness $\times$ $\times$ $\times$ $\times$ $\times$

Feedback of the agent at time t

$$k_t := \min\left\{1 \le k \le K : W_t(i_k^t) = 1\right\},$$

where $\min\emptyset = \infty$. If $k_t < \infty$.

- $k_t < \infty$: the agent observes $W_t(i_k^t) = 0$ for $1 \le k < k_t$, and $W_t(i_k^t) = 1$, but does not observe $W_t(i_k^t)$ for $k > k_t$;
- $k_t = \infty$: the agent observes $W_t(i_k^t) = 0$ for $1 \le k \le K$.

Cascading Bandits Setting (Kveton et al., 2015)

Expected instantaneous reward

$$r(S|\boldsymbol{w}) = \mathbb{E}[R(S|\boldsymbol{w})] = 1 - \mathbb{E}\left[\prod_{i_k \in S} (1 - W(i_k))\right] = 1 - \prod_{i_k \in S} (1 - w(i_k)).$$

Cascading Bandits Setting (Kveton et al., 2015)

Expected instantaneous reward

$$r(S|\boldsymbol{w}) = \mathbb{E}[R(S|\boldsymbol{w})] = 1 - \mathbb{E}\left[\prod_{i_k \in S}(1 - W(i_k))\right] = 1 - \prod_{i_k \in S}(1 - w(i_k)).$$

Optimal K-subset S^*

Assume that $w(1) \geq w(2) \geq \ldots \geq w(L)$, then any permutation of $\{1, \ldots, K\}$ maximizes the mean reward. We let

$$S^* = (1, \ldots, K).$$

Cascading Bandits Setting (Kveton et al., 2015)

In T steps, we aim to minimize...

Expected cumulative regret (Criterion of algorithm)

$$\mathrm{Reg}(T) := T \cdot r(S^*|\boldsymbol{w}) - \sum_{t=1}^{T} r(S_t|\boldsymbol{w}),$$

- $\boldsymbol{w} \in [0,1]^L$, the vector of click probabilities, is not known to the agent;
- S_t is chosen online, i.e., dependent on previous choices and the previous rewards.

Difficulties of Analyzing TS for Cascading Bandits

Cascading Bandits for Large-Scale Recommendation Problems

Shi Zong
Dept of Electrical and Computer Engineering
Carnegie Mellon University
szong@andrew.cmu.edu

Hao Ni
Dept of Electrical and Computer Engineering
Carnegie Mellon University
haon@cmu.edu

Kenny Sung
Dept of Electrical and Computer Engineering
Carnegie Mellon University
tsung@andrew.cmu.edu

Nan Rosemary Ke
Dépt d'informatique et de recherche opérationnelle
Université de Montréal
nke001@gmail.com

Zheng Wen
Adobe Research
San Jose, CA
zwen@adobe.com

Branislav Kveton
Adobe Research
San Jose, CA
kveton@adobe.com

UCB-based algorithms proposed in Zong *et al.* (UAI 2016)

Difficulties of Analyzing TS for Cascading Bandits

Recent work [21, 24] demonstrated close relationships between UCB-like algorithms and Thompson sampling algorithms in related bandit problems. Therefore, we believe that a similar regret bound to that in Theorem 1 also holds for CascadeLinTS. However, it is highly non-trivial to derive a regret bound for CascadeLinTS. Unlike in [24], CascadeLinTS cannot be analyzed from the Bayesian perspective because the Gaussian posterior is inconsistent with the fact that $\bar{w}(e)$ is bounded in $[0, 1]$. Moreover, a subtle statistical dependence between partial monitoring and Thompson sampling prevents a frequentist analysis similar to that in [4]. Therefore, we leave the formal analysis

Mentioned difficulties of analysis of Thompson sampling

TS-CASCADE algorithm

Algorithm 1: TS-CASCADE, TS for Cascading Bandits with Gaussians

1: Initialize $\hat{\mu}_1(i) = 0$, $N_1(i) = 0$ for all $i \in [L]$.
for $t = 1, 2, \ldots$ **do**
2: Sample a 1-dim r.v. $Z_t \sim \mathcal{N}(0, 1)$.
3: Construct Thompson sample $\theta_t(i)$ for all $i \in [L]$ with Alg 2.
for $i \in [L]$ **do**
4: Extract $i_k^t \in \text{argmax}_{i \in [L] \setminus \{i_1^t, \ldots i_{k-1}^t\}} \theta_t(i)$.
end
5: Pull arm $S_t = (i_1^t, i_2^t, \ldots, i_K^t)$.
6: Update $\hat{\mu}_{t+1}(i)$, $N_{t+1}(i)$ for all $i \in [L]$ with Bayes rule, i.e., Alg 3.
end

TS-CASCADE Algorithm

Algorithm 2: Construct Thompson sample

1: Calculate the empirical variance $\hat{\nu}_t(i) = \hat{\mu}_t(i)(1 - \hat{\mu}_t(i))$.

2: Calculate std. dev. of the Thompson sample

$$\sigma_t(i) = \max\left\{\sqrt{\frac{\hat{\nu}_t(i)\log(t+1)}{N_t(i)+1}}, \frac{\log(t+1)}{N_t(i)+1}\right\}.$$

3: Construct the Thompson sample $\theta_t(i) = \hat{\mu}_t(i) + Z_t\sigma_t(i)$.

TS-CASCADE Algorithm

Algorithm 2: Construct Thompson sample

1: Calculate the empirical variance $\hat{\nu}_t(i) = \hat{\mu}_t(i)(1 - \hat{\mu}_t(i))$.

2: Calculate std. dev. of the Thompson sample

$$\sigma_t(i) = \max\left\{\sqrt{\frac{\hat{\nu}_t(i)\log(t+1)}{N_t(i)+1}}, \frac{\log(t+1)}{N_t(i)+1}\right\}.$$

3: Construct the Thompson sample $\theta_t(i) = \hat{\mu}_t(i) + Z_t\sigma_t(i)$.

Algorithm 3: Update parameters

1: If $W_t(i)$ is observed for arm i, update parameters as follows:

$$\hat{\mu}_{t+1}(i) = \frac{N_t(i)\hat{\mu}_t(i) + W_t(i)}{N_t(i)+1}, \quad N_{t+1}(i) = N_t(i) + 1.$$

2: For $j \neq i$ params. unchanged: $\hat{\mu}_{t+1}(j) = \hat{\mu}_t(j)$, $N_{t+1}(j) = N_t(j)$.

TS-CASCADE Algorithm

Use Gaussians to estimate the probability $w(i)$

- More natural to use a Beta-Bernoulli update to maintain a Bayesian estimate on the probability $w(i)$ (Russo et al., 2018).

TS-CASCADE Algorithm

Use Gaussians to estimate the probability $w(i)$

- More natural to use a Beta-Bernoulli update to maintain a Bayesian estimate on the probability $w(i)$ (Russo et al., 2018).
- Gaussian is useful: can be readily generalized the algorithm and analyses to the contextual setting (Li et al., 2010), the online setting (Li et al., 2016), and the linear bandits setting (Zong et al., 2016) for handling a large L.

TS-CASCADE Algorithm

Use Gaussians to estimate the probability $w(i)$

- More natural to use a Beta-Bernoulli update to maintain a Bayesian estimate on the probability $w(i)$ (Russo et al., 2018).
- Gaussian is useful: can be readily generalized the algorithm and analyses to the contextual setting (Li et al., 2010), the online setting (Li et al., 2016), and the linear bandits setting (Zong et al., 2016) for handling a large L.
- Difficulties of analysis comes from that $\theta_t(i)$ is not in $[0, 1]$ with probability one. Our proof shows that this replacement of the Beta by the Gaussian does not incur any significant loss in terms of the regret.

Upper Bound of Regret of TS-Cascade

Theorem

For $T \geq L$, TS-Cascade *incurs an expected regret at most*

$$\mathrm{Reg}(T) = O\big(\sqrt{KLT}\log T\big).$$

Upper Bound of Regret of TS-CASCADE

Theorem

For $T \geq L$, TS-CASCADE incurs an expected regret at most

$$\mathrm{Reg}(T) = O\big(\sqrt{KLT}\log T\big).$$

- For the vanilla MAB problem, problem-independent regrets (e.g., KL-UCB or Thompson sampling) scale like

$$\tilde{O}(\sqrt{LT}).$$

Upper Bound of Regret of TS-CASCADE

Theorem

For $T \geq L$, TS-CASCADE incurs an expected regret at most

$$\mathrm{Reg}(T) = O\big(\sqrt{KLT}\log T\big).$$

- For the vanilla MAB problem, problem-independent regrets (e.g., KL-UCB or Thompson sampling) scale like

$$\tilde{O}(\sqrt{LT}).$$

- Proof Ideas:
  1. Appropriate definitions of nice events ($\hat{\mu}_t$ and θ_t concentrate);
  2. Typicality of θ_t w.r.t. cascading objective;
  3. Anti-concentration to ensure exploration of unsaturated super-arms;
  4. Martingale-style analysis of empirical variance (Audibert et al., 2009).

Proof Sketch I: Nice Events

Concentration of Nice Events (Audibert et al., 2009)

Define

$$\mathcal{E}_{\hat{\mu},t} := \{\forall\, i \in [L] : |\hat{\mu}_t - w(i)| \le g_t(i)\}$$
$$\mathcal{E}_{\theta,t} := \{\forall\, i \in [L] : |\theta_t(i) - \hat{\mu}_t| \le h_t(i)\}$$

Proof Sketch I: Nice Events

Concentration of Nice Events (Audibert et al., 2009)

Define

$$\mathcal{E}_{\hat{\mu},t} := \{\forall\, i \in [L] : |\hat{\mu}_t - w(i)| \le g_t(i)\}$$
$$\mathcal{E}_{\theta,t} := \{\forall\, i \in [L] : |\theta_t(i) - \hat{\mu}_t| \le h_t(i)\}$$

where (cf. UCB-V by Audibert et al. (2009))

$$g_t(i) := \sqrt{\frac{16\hat{\nu}_t(i)\log(t+1)}{N_t(i)+1}} + \frac{24\log(t+1)}{N_t(i)+1}$$
$$h_t(i) := \sqrt{\log(t+1)}\,g_t(i).$$

Proof Sketch I: Nice Events

Concentration of Nice Events (Audibert et al., 2009)

Define

$$\mathcal{E}_{\hat{\mu},t} := \{\forall\, i \in [L] : |\hat{\mu}_t - w(i)| \leq g_t(i)\}$$
$$\mathcal{E}_{\theta,t} := \{\forall\, i \in [L] : |\theta_t(i) - \hat{\mu}_t| \leq h_t(i)\}$$

where (cf. UCB-V by Audibert et al. (2009))

$$g_t(i) := \sqrt{\frac{16\hat{\nu}_t(i)\log(t+1)}{N_t(i)+1}} + \frac{24\log(t+1)}{N_t(i)+1}$$
$$h_t(i) := \sqrt{\log(t+1)}\,g_t(i).$$

Then,

$$\Pr[\mathcal{E}_{\hat{\mu},t}] \geq 1 - \frac{3L}{(t+1)^3}, \quad \text{and} \quad \Pr[\mathcal{E}_{\theta,t}|\mathcal{E}_{\hat{\mu},t}] \geq 1 - \frac{1}{2(t+1)^2}$$

Proof Sketch II: Unsaturated Super-Arms

Unsaturated Super-Arms

For $S = (i_1, i_2, \ldots, i_K)$, define the weighted statistical gap

$$F(S, t) := \sum_{k=1}^{K} \left[\prod_{j=1}^{k-1} (1 - w(i_j)) \right] \big(g_t(i_k) + h_t(i_k) \big)$$

The unsaturated superarms (Agrawal and Goyal, 2013) are in the set

$$\mathcal{S}_t := \left\{ S = (i_1, \ldots, i_K) \in \pi_K(L) : F(S, t) \geq r(S^* | \boldsymbol{w}) - r(S | \boldsymbol{w}) \right\}$$

Proof Sketch II: Unsaturated Super-Arms

Unsaturated Super-Arms

For $S = (i_1, i_2, \ldots, i_K)$, define the weighted statistical gap

$$F(S,t) := \sum_{k=1}^{K} \left[\prod_{j=1}^{k-1} (1 - w(i_j)) \right] (g_t(i_k) + h_t(i_k))$$

The unsaturated superarms (Agrawal and Goyal, 2013) are in the set

$$\mathcal{S}_t := \left\{ S = (i_1, \ldots, i_K) \in \pi_K(L) : F(S,t) \geq r(S^*|\boldsymbol{w}) - r(S|\boldsymbol{w}) \right\}$$

- Arms in $\mathcal{S}_t$ (S^* is a prime e.g.) can lack observations, while arms in $\mathcal{S}_t^c$ are observed enough, and are believed to be suboptimal.

Proof Sketch II: Unsaturated Super-Arms

Unsaturated Super-Arms

For $S = (i_1, i_2, \ldots, i_K)$, define the weighted statistical gap

$$F(S,t) := \sum_{k=1}^{K} \left[\prod_{j=1}^{k-1} (1 - w(i_j)) \right] \big(g_t(i_k) + h_t(i_k)\big)$$

The unsaturated superarms (Agrawal and Goyal, 2013) are in the set

$$\mathcal{S}_t := \left\{ S = (i_1, \ldots, i_K) \in \pi_K(L) : F(S,t) \geq r(S^*|\boldsymbol{w}) - r(S|\boldsymbol{w}) \right\}$$

- Arms in $\mathcal{S}_t$ (S^* is a prime e.g.) can lack observations, while arms in $\mathcal{S}_t^c$ are observed enough, and are believed to be suboptimal.
- For any suboptimal $i \in [L] \setminus [K]$ and optimal $k \in [K]$, we hope that

$$g_t(i) + h_t(i) \geq w(k) - w(i)$$

but this is too optimistic. Hope that $S_t \in \mathcal{S}_t$.

Proof Sketch II: Unsaturated Super-Arms

Exploration of Unsaturated Super-Arms

Define typical event

$$\mathcal{T} := \left\{ \sum_{k=1}^{K} \left[\prod_{j=1}^{k-1} (1 - w(j)) \right] \theta_t(k) \geq \sum_{k=1}^{K} \left[\prod_{j=1}^{k-1} (1 - w(j)) \right] w(k) \right\}$$

Then,

$$\mathcal{E}_{\hat{\mu},t} \cap \mathcal{E}_{\theta,t} \cap \mathcal{T} \subset \{S_t \in \mathcal{S}_t\}.$$

Proof Sketch II: Unsaturated Super-Arms

Exploration of Unsaturated Super-Arms

Define typical event

$$\mathcal{T} := \left\{ \sum_{k=1}^{K} \left[\prod_{j=1}^{k-1} (1 - w(j)) \right] \theta_t(k) \geq \sum_{k=1}^{K} \left[\prod_{j=1}^{k-1} (1 - w(j)) \right] w(k) \right\}$$

Then,

$$\mathcal{E}_{\hat{\mu},t} \cap \mathcal{E}_{\theta,t} \cap \mathcal{T} \subset \{S_t \in \mathcal{S}_t\}.$$

Anticoncentration of Exploration of Unsaturated Super-Arms

For any history H_t, there exists $c > 0$ such that

$$\Pr_{\boldsymbol{\theta}_t} \left[\mathcal{E}_{\theta,t} \cap \mathcal{T} \,\middle|\, H_t \right] \geq c.$$

Proof Sketch III: Bounding Regret

Bounding Regret

Assuming H_t is typical,

$$\mathbb{E}_{\boldsymbol{\theta}_t}\left[r(S^*|\boldsymbol{w}) - r(S_t|\boldsymbol{w}) \,\middle|\, H_t\right]$$
$$\leq \left(1+\frac{4}{c}\right)\mathbb{E}_{\boldsymbol{\theta}_t}\left[\underbrace{F(S_t,t)}_{\text{weighted statistical gap}} \,\middle|\, H_t\right] + \frac{L}{2(t+1)^2}.$$

Proof Sketch III: Bounding Regret

Bounding Regret

Assuming H_t is typical,

$$\mathbb{E}_{\boldsymbol{\theta}_t}\left[r(S^*|\boldsymbol{w}) - r(S_t|\boldsymbol{w}) \,\middle|\, H_t\right]$$
$$\leq \left(1 + \frac{4}{c}\right)\mathbb{E}_{\boldsymbol{\theta}_t}\left[\underbrace{F(S_t, t)}_{\text{weighted statistical gap}} \,\middle|\, H_t\right] + \frac{L}{2(t+1)^2}.$$

- Relies on truncating the Thompson sample $\boldsymbol{\theta}_t \in \mathbb{R}^L$ to $\widetilde{\boldsymbol{\theta}}_t \in [0,1]^L$.

Proof Sketch III: Bounding Regret

Bounding Regret

Assuming H_t is typical,

$$\mathbb{E}_{\boldsymbol{\theta}_t}\left[r(S^*|\boldsymbol{w}) - r(S_t|\boldsymbol{w}) \,\middle|\, H_t\right] \leq \left(1 + \frac{4}{c}\right)\mathbb{E}_{\boldsymbol{\theta}_t}\left[\underbrace{F(S_t, t)}_{\text{weighted statistical gap}} \,\middle|\, H_t\right] + \frac{L}{2(t+1)^2}.$$

- Relies on truncating the Thompson sample $\boldsymbol{\theta}_t \in \mathbb{R}^L$ to $\widetilde{\boldsymbol{\theta}}_t \in [0,1]^L$.
- Finish the proof by summing the per time-step regret and using a standard telescoping property of the summation.

Upper bound of TS-Cascade

Comparison To State-Of-The-Art

Algorithm	Bounds	Indep.
TS-Cascade	$O\left(\sqrt{KLT}\log T\right)$	$\surd$
CUCB (Wang and Chen, 2017)	$O\left(\sqrt{KLT\log T}\right)$	$\surd$
CUCB1 (Kveton et al., 2015)	$O\left((L-K)(\log T)/\Delta\right)$	$\times$
CKL-UCB (Kveton et al., 2015)	$O\left((L-K)\log(T/\Delta)/\Delta\right)$	$\times$
Lower Bd (Kveton et al., 2015)	$\Omega\left((L-K)(\log T)/\Delta\right)$	$\times$

- Upper bounds on the T-regret of TS-Cascade, CUCB, CascadeUCB1 and CascadeKL-UCB
- Lower bound of all cascading bandits algorithms

Upper bound of TS-Cascade

Comparison To State-Of-The-Art

Algorithm	Bounds	Indep.
TS-Cascade	$O\left(\sqrt{KLT}\log T\right)$	$\surd$
CUCB (Wang and Chen, 2017)	$O\left(\sqrt{KLT\log T}\right)$	$\surd$
CUCB1 (Kveton et al., 2015)	$O\left((L-K)(\log T)/\Delta\right)$	$\times$
CKL-UCB (Kveton et al., 2015)	$O\left((L-K)\log(T/\Delta)/\Delta\right)$	$\times$
Lower Bd (Kveton et al., 2015)	$\Omega\left((L-K)(\log T)/\Delta\right)$	$\times$

- Optimal items $i \in S^*$ have the same click probability w_1
- Suboptimal items $i \notin S^*$ have the same click probability w_2
- Gap $\Delta := w_1 - w_2$: measure the difficulty of the problem

Comparison To State-Of-The-Art

Algorithm	Bounds	Indep.
TS-CASCADE	$O\big(\sqrt{KLT}\log T\big)$	$\surd$
CUCB (Wang and Chen, 2017)	$O\big(\sqrt{KLT\log T}\big)$	$\surd$
CUCB1 (Kveton et al., 2015)	$O\big((L-K)(\log T)/\Delta\big)$	$\times$
CKL-UCB (Kveton et al., 2015)	$O\big((L-K)\log(T/\Delta)/\Delta\big)$	$\times$
Lower Bd (Kveton et al., 2015)	$\Omega\big((L-K)(\log T)/\Delta\big)$	$\times$

- Our upper bound grows like $\sqrt{T}$ just like the others.
  - Matches the state-of-the-art UCB bound (up to log factors) by Wang and Chen (2017).

Upper bound of TS-Cascade

Comparison To State-Of-The-Art

Algorithm	Bounds	Indep.
TS-Cascade	$O\big(\sqrt{KLT}\log T\big)$	$\surd$
CUCB (Wang and Chen, 2017)	$O\big(\sqrt{KLT\log T}\big)$	$\surd$
CUCB1 (Kveton et al., 2015)	$O\big((L-K)(\log T)/\Delta\big)$	$\times$
CKL-UCB (Kveton et al., 2015)	$O\big((L-K)\log(T/\Delta)/\Delta\big)$	$\times$
Lower Bd (Kveton et al., 2015)	$\Omega\big((L-K)(\log T)/\Delta\big)$	$\times$

- Our upper bound grows like $\sqrt{T}$ just like the others.
  - Matches the state-of-the-art UCB bound (up to log factors) by Wang and Chen (2017).
- When $T \geq L$, our bound is $\sqrt{\log T}$ factor worse than the problem independent bound in Wang and Chen (2017).
  - First TS analysis for stochastic combinatorial bandits with partial feedback

Experiments

Evaluate TS-Cascade against CascadeKL-UCB and CascadeUCB1 in Kveton et al. (2015).

Experiments

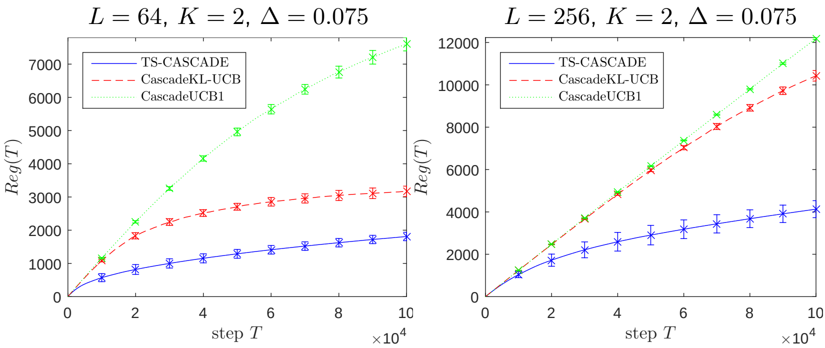
Evaluate TS-Cascade against CascadeKL-UCB and CascadeUCB1 in Kveton et al. (2015).

Experiment setting

- Optimal items $i \in S^*$ have the same click probability w_1
- Suboptimal items $i \notin S^*$ have the same click probability w_2
- Gap $\Delta := w_1 - w_2 > 0$

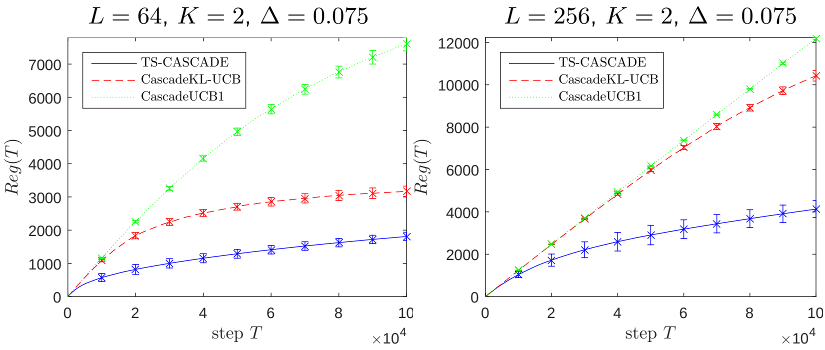
Setting $w_1 = 0.2$, $T = 10^5$, we conduct 20 independent simulations with each algorithm under each setting of L, K, and Δ.

Numerical results



$\mathrm{Reg}(T)$ of TS-CASCADE, CASCADEKL-UCB and CASCADEUCB1 with each line indicates the average $\mathrm{Reg}(T)$ (over 20 runs) and the length of each errorbar above and below each data point is the standard deviation.

Numerical results



- TS-Cascade outperforms the two UCB algorithms.
- When $L = 256$ is large, the UCB-based algorithms do not demonstrate the $\sqrt{T}$ behavior even after $T = 10^5$ iterations.
- $\mathrm{Reg}(T)$ for TS-Cascade $\sim O(\sqrt{T})$, implying that the empirical performance corroborates the theoretical result.

Other Results

- **Problem-independent lower bound**
  Judicious construction of an adversarial bandit example +
  information-theoretic technique of Auer et al. (2002).

  Lower Bound on Regret:

$$\mathrm{Reg}(T) = \tilde{\Omega}\big(\sqrt{LT}\big).$$

Other Results

Other Results

- **Problem-independent lower bound**
  Judicious construction of an adversarial bandit example + information-theoretic technique of Auer et al. (2002).

  Lower Bound on Regret:

  $$\mathrm{Reg}(T) = \tilde{\Omega}\big(\sqrt{LT}\big).$$

- **Generalization to the contextual and linear settings** (Agrawal and Goyal, 2013; Li et al., 2010, 2016; Qin et al., 2014)
  The click probs. $w(i) = x(i)^T \beta$ for some unknown $\beta \in \mathbb{R}^d$,

  Regret Under Linear Generalization:

  $$\mathrm{Reg}_{\mathrm{lin}}(T) = \tilde{O}\big(dK\sqrt{T}\big)$$

References I

S. Agrawal and N. Goyal. Thompson sampling for contextual bandits with linear payoffs. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28, pages 127–135, 2013.

J.-Y. Audibert, R. Munos, and C. Szepesvri. Exploration-exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410(19):1876 – 1902, 2009. Algorithmic Learning Theory.

P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal of Computing*, 32(1):48–77, 2002.

B. Kveton, C. Szepesvari, Z. Wen, and A. Ashkan. Cascading bandits: Learning to rank in the cascade model. In *International Conference on Machine Learning*, pages 767–776, 2015.

L. Li, W. Chu, J. Langford, and R. E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, pages 661–670, 2010.

S. Li, B. Wang, S. Zhang, and W. Chen. Contextual combinatorial cascading bandits. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48, pages 1245–1253, 2016.

L. Qin, S. Chen, and X. Zhu. Contextual combinatorial bandit and its application on diversified online recommendation. In *SDM*, pages 461–469. SIAM, 2014.

D. Russo, B. V. Roy, A. Kazerouni, I. Osband, and Z. Wen. A tutorial on Thompson sampling. *Foundations and Trends in Machine Learning*, 11(1):1–96, 2018.

References II

Q. Wang and W. Chen. Improving regret bounds for combinatorial semi-bandits with probabilistically triggered arms and its applications. In *Advances in Neural Information Processing Systems*, pages 1161–1171, 2017.

S. Zong, H. Ni, K. Sung, N. R. Ke, Z. Wen, and B. Kveton. Cascading bandits for large-scale recommendation problems. In *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence*, UAI'16, pages 835–844, 2016.