Contextual Combinatorial Cascading Bandits

Shuai Li[1]; Baoxiang wang[1]; Shengyu Zhang[1]; Wei Chen[2]
[1]The Chinese University of Hong Kong, [2]Microsoft Research

Motivation

- Cascading feedback
  - Websites search results
  - Recommended movies
  - All are sequential lists
    - Users go through the list from top down
    - Stop at the first satisfactory item
    - Click it
    - This online feedback helps improving future list quality
- Contexts
  - User profiles, search keywords
  - Important for search, recommendations, etc.
- Combinatorial
  - Action is selection of a sequence of items.
  - May have other combinatorial constraints (e.g. paths in networks)

Setting

- A finite set $E = \{1, \ldots, L\}$ of L base arms.
- Let $\mathcal{S}$ be the set of feasible actions, which are tuples from E with length at most K.
- Position discounts $\gamma_k \epsilon [0,1], k \leq K$.
- α-approximation oracle $\mathcal{O}_{\mathcal{S}}$
- At time t,
  - For each $a \in E$, a feature vectors $x_{t,a} \in \mathbb{R}^{d\times 1}$ with $\|x_{t,a}\|_2 \leq 1$ is revealed to the learning agent.
  - Let $\mathcal{H}_t$ denote the history so far.
  - The learning agent recommends $\boldsymbol{A}_t = (\boldsymbol{a}_1^t, \ldots, \boldsymbol{a}_{|A_t|}^t) \in \mathcal{S}$ to the user.
  - The user checks from the first item of $\boldsymbol{A}_t$ and stops at $\boldsymbol{O}_t$-th item under some stopping criterion.
  - The learning agent observes the weights of first $\boldsymbol{O}_t$ base arms in $\boldsymbol{A}_t$, $\boldsymbol{w}_t(\boldsymbol{a}_k^t), k \leq \boldsymbol{O}_t$.
- Assume given $\mathcal{H}_t$,$\boldsymbol{w}_t(a)$'s are mutually independent R-sub Gaussian random variable with
$$\mathbb{E}[\boldsymbol{w}_t(a)|\mathcal{H}_t] = \theta_*^\top x_{t,a}$$
for some unknown $\theta_* \in \mathbb{R}^{d\times 1}$ with $\|\theta_*\|_2 \leq 1,\ 0 \leq \theta_*^\top x_{t,a} \leq 1$.
- Assume the expected reward of action A is a function $f(A,w)$ of expected weight w satisfying
  - monotonicity
  - B-Lipschitz continuity
- The α-regret of action A on time t is
$$R^\alpha(t,A) = \alpha f_t^* - f(A, w_t)$$
- Minimize α-regret of n rounds
$$R^\alpha(n) = \mathbb{E}\left[\sum_{t=1}^{n} R^\alpha(t, \boldsymbol{A}_t)\right].$$

Table 1. Comparisons of our setting with previous ones.

	context	cascading	Position discount	General reward
Combinatorial UCB[1]	No	Yes	No	Yes
Contextual Combinatorial UCB[2]	Yes	No	No	Yes
Comb-Cascade[3]	No	Yes	No	No
C[3]-UCB(ours)	Yes	Yes	Yes	Yes

Algorithm: C^3-UCB

1. Parameters:
$$\{\gamma_k \epsilon [0,1]\}_{k\leq K};\ \delta = \frac{1}{\sqrt{n}};\ \lambda \geq C_\gamma = \sum_{k=1}^{K} \gamma_k^2$$
2. Initialization:
$$\hat{\theta}_0 = 0, \beta_0(\delta) = 1, V_0 = \lambda I, X_0 = \emptyset, Y_0 = \emptyset$$
3. For all $t = 1,2,\ldots,n$ do
   1) Obtain context $x_{t,a}$ for all $a \in E$
   2) For any $a \in E$, compute
   $$\boldsymbol{U}_t(a) = \min\left\{\underbrace{\hat{\boldsymbol{\theta}}_{t-1}^\top x_{t,a}}_{\text{exploitation}} + \underbrace{\boldsymbol{\beta}_{t-1}(\delta)\|x_{t,a}\|_{\boldsymbol{V}_{t-1}^{-1}}}_{\text{exploration}}, 1\right\}$$
   3) Choose action $\boldsymbol{A}_t$ using UCBs $\boldsymbol{U}_t$
   $$\boldsymbol{A}_t = (\boldsymbol{a}_1^t, \ldots, \boldsymbol{a}_{|A_t|}^t) \leftarrow \mathcal{O}_{\mathcal{S}}(\boldsymbol{U}_t)$$
   4) Play $\boldsymbol{A}_t$ and observe $\boldsymbol{O}_t; \boldsymbol{w}_t(\boldsymbol{a}_k^t), k \leq \boldsymbol{O}_t$.
   5) Update statistics
   $$\boldsymbol{V}_t \leftarrow \boldsymbol{V}_{t-1} + \sum_{k=1}^{\boldsymbol{O}_t} \gamma_k^2 x_{t,a_k^t} x_{t,a_k^t}^\top$$
   $$\boldsymbol{X}_t \leftarrow \left[\boldsymbol{X}_{t-1};\ \gamma_1 x_{t,a_1^t}^\top; \ldots; \gamma_{O_t} x_{t,a_{O_t}^t}^\top\right]$$
   $$\boldsymbol{Y}_t \leftarrow \left[\boldsymbol{Y}_{t-1};\ \gamma_1 \boldsymbol{w}_t(\boldsymbol{a}_1^t); \ldots; \gamma_{O_t} \boldsymbol{w}_t(\boldsymbol{a}_{O_t}^t)\right]$$
   $$\underbrace{\hat{\theta}_t \leftarrow (\boldsymbol{X}_t^\top \boldsymbol{X}_t + \lambda I)^{-1} \boldsymbol{X}_t^\top \boldsymbol{Y}_t}_{\text{Linear regression}}$$
   $$\boldsymbol{\beta}_t(\delta) \leftarrow R\sqrt{\ln(\det(\boldsymbol{V}_t)/(\lambda^d \delta^2))} + \sqrt{\lambda}$$

   End for t

Results

Theorem 1. Suppose the expected reward function $f(A,w)$ is a function of expected weights and satisfies monotonicity and B-Lipschitz continuity. Then the α-regret of our algorithm, C^3-UCB, satisfies
$$R^\alpha(n) = O\left(\frac{dBR}{p^*}\sqrt{nK}\ln(C_\gamma n)\right),$$
where R is the sub-Gaussian constant and $C_\gamma = \sum_{k=1}^{K}\gamma_k^2 \leq K$.

Corollary 2. In the problem of cascading recommendation, the expected reward is disjunctive
$$f(A,w) = \sum_{k=1}^{|A|} \gamma_k \prod_{i=1}^{k-1}(1 - w(a_i)) w(a_k)$$
where $1 = \gamma_1 \geq \cdots \geq \gamma_K \geq 0$. Then the α-regret of C^3-UCB satisfies
$$R^\alpha(n) = O\left(\frac{d}{1 - f^*}\sqrt{nK}\ln(C_\gamma n)\right),$$
where $f^* = \max f_t^*$, the maximal expected reward in n rounds.

Theorem 3. Suppose $1 = \gamma_1 \geq \cdots \geq \gamma_K \geq 1 - \frac{\alpha}{4} f_*$, where $f_* = \min f_t^*$. Then the α-regret of C^3-UCB for the conjunctive objective
$$f(A,w) = \sum_{k=1}^{|A|}(1-\gamma_k)\prod_{i=1}^{k-1} w(a_i)(1 - w(a_k)) + \prod_{i=1}^{|A|} w(a_i)$$
satisfies
$$R^\alpha(n) = O\left(\frac{d}{\alpha f_*}\sqrt{nK}\ln(C_\gamma n)\right).$$

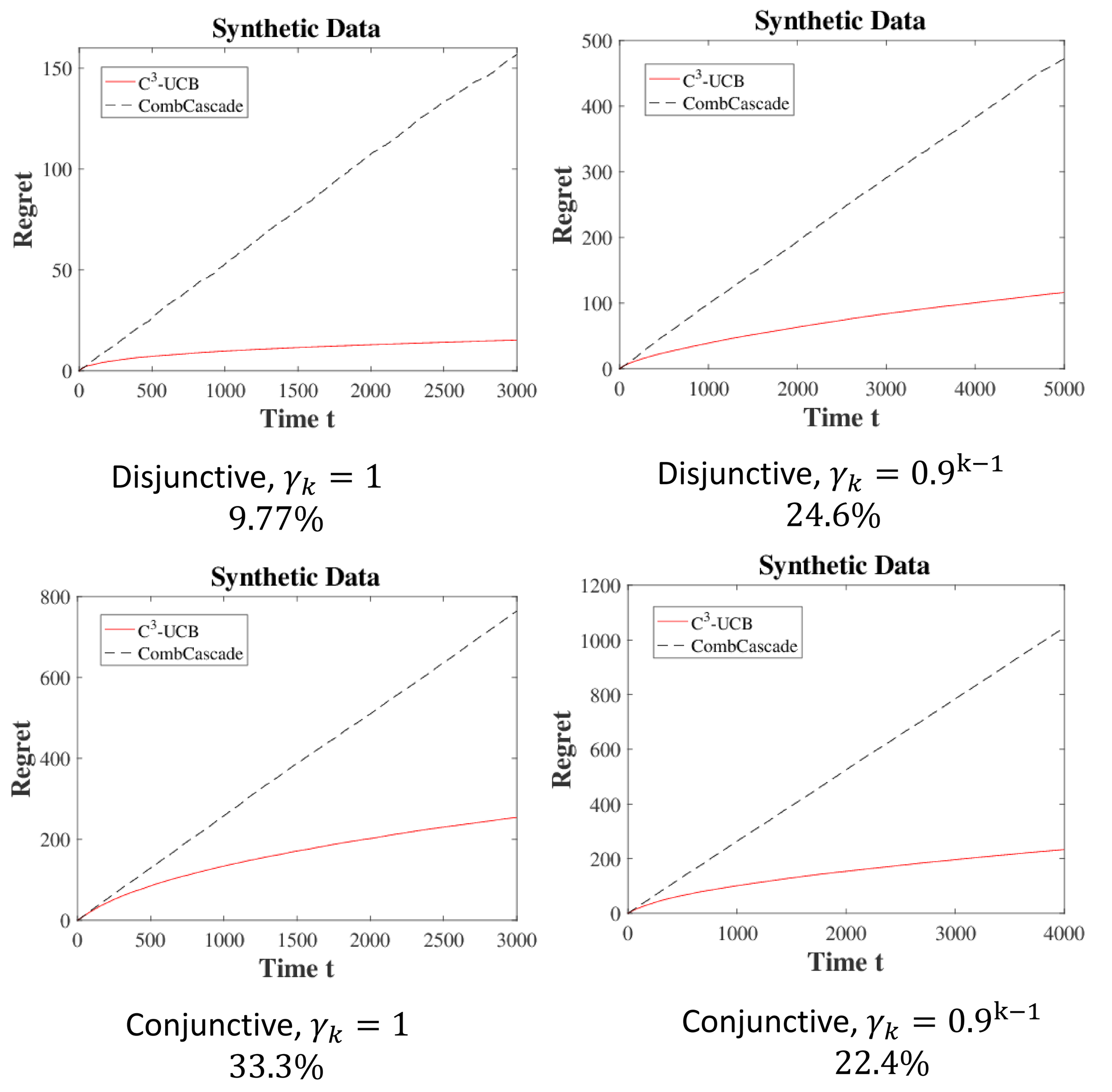

Figure 1. Experimental results for synthetic data. 100 items, select 4 items. Latent and feature vector dimension = 4.

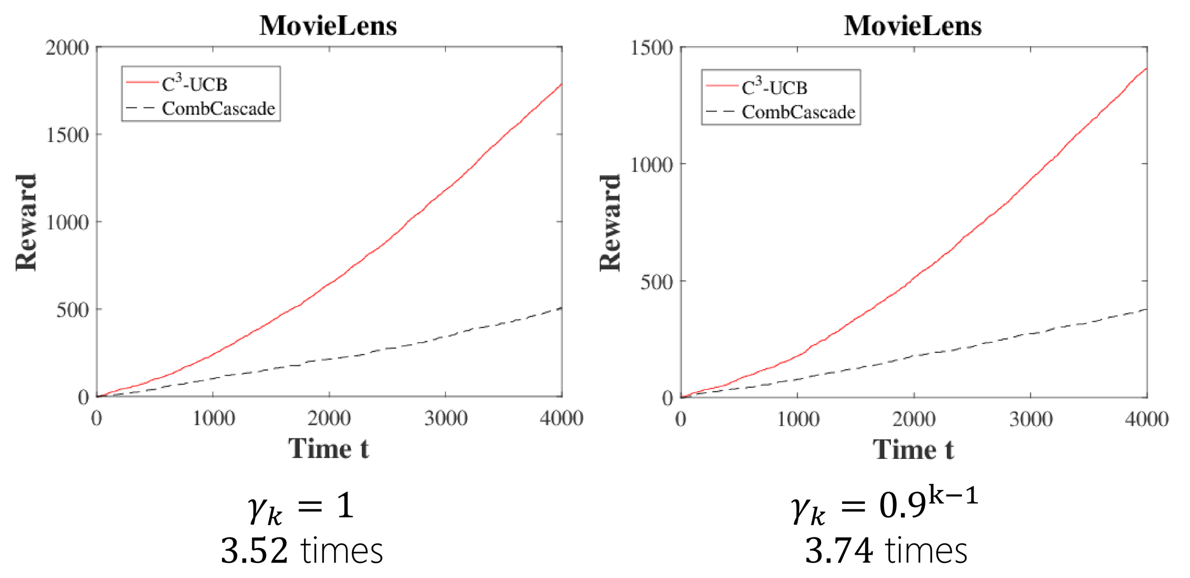

Figure 2. Experimental results on MovieLens dataset, 200 movies, select 4 items. d= 400 (By SVD decomposition)

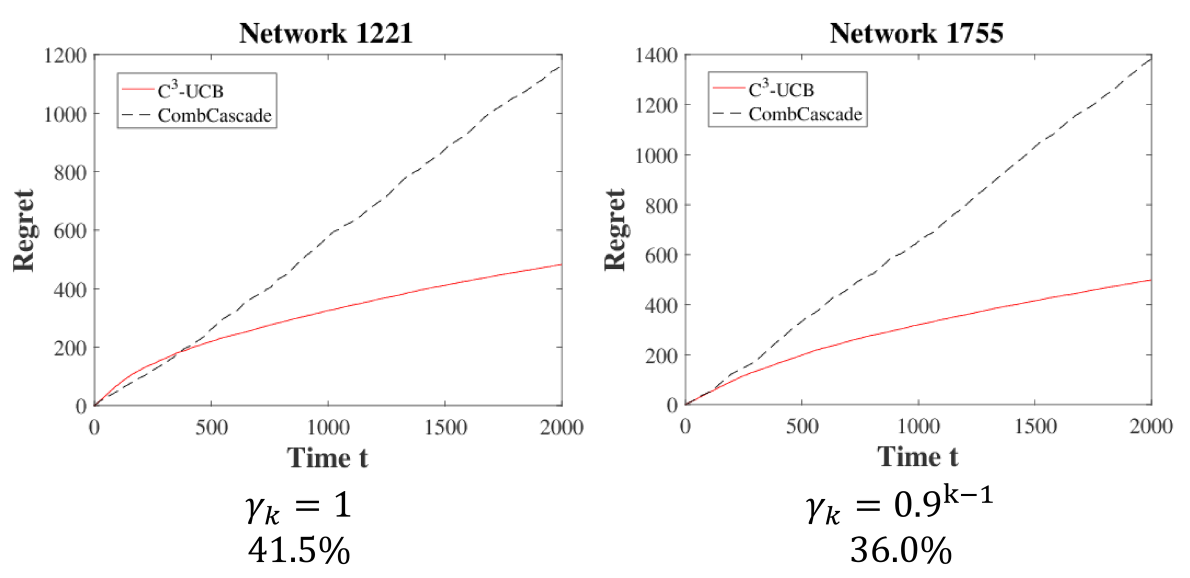

Figure 3. Experimental results on ISP dataset, $d = 5, K = 4$.

Conclusions

- Formulate Contextual Combinatorial Cascading Bandits problem
- Propose C^3-UCB algorithm that can handle
  - contextual information
  - cascading feedback
  - position discount
  - general reward function
- Theoretical analysis and empirical evaluation

Contact

Shuai Li
Email: shuaili@cse.cuhk.edu.hk

Shengyu Zhang
Email: syzhang@cse.cuhk.edu.hk

Baoxiang Wang
Email: bxwang@cse.cuhk.edu.hk

Wei Chen
Email: weic@Microsoft.com

References

1. Chen, Wei, Yajun Wang, and Yang Yuan. "Combinatorial multi-armed bandit: General framework and applications." Proceedings of the 30th International Conference on Machine Learning. 2013.
2. Qin, Lijing, Shouyuan Chen, and Xiaoyan Zhu. "Contextual Combinatorial Bandit and its Application on Diversified Online Recommendation." SDM. 2014.
3. Kveton, Branislav, Zheng Wen, Azin Ashkan, and Csaba Szepesvari. "Combinatorial Cascading Bandits." In Advances in Neural Information Processing Systems, pp. 1450-1458. 2015.