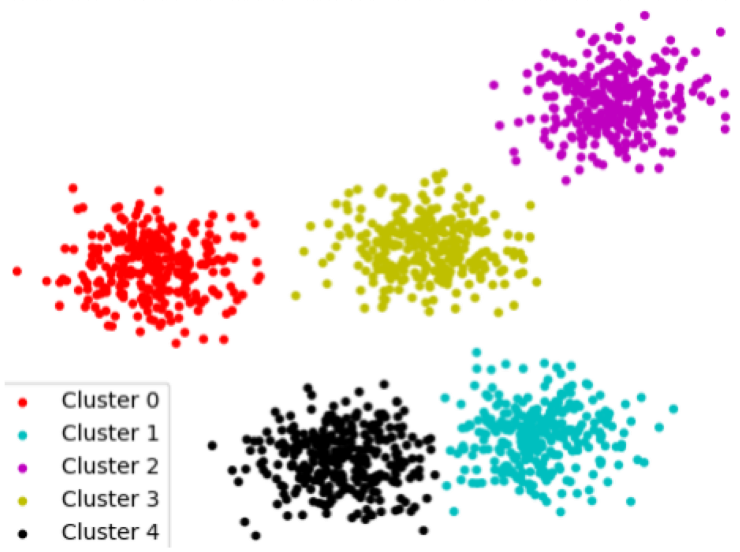
Online Clustering of Contextual Cascading Bandits

Shuai Li*[1]   Shengyu Zhang*[1,2]

1. The Chinese University of Hong Kong.   2. Tencent

Motivation

- Cascading feedback
  - Scenarios: web search results, online recommendation systems, …
  - Model: On an ordered list
    - A user goes through the list from top down,
    - stops at the first satisfactory item, and
    - clicks it.
    - Task: Use this online feedback to help improve future list recommendation
- Contexts: Features for an item or and item-user pair
  - Important for recommendations
- Combinatorial:
  - An action is an ordered sequence of items
- Clustering
  - Users have a clustering structure
  - We only see the user indices
  - Need to learn the user similarities online



Algorithm: CLUB-cascade

1. Parameters: $\lambda, \alpha, \beta > 0$

2. Initialization:
   - G is a complete graph over users;
   - $S_i = 0_{d\times d}, b_i = 0_{d\times 1}, T_i = 0$ for any user i.

3. For all $t = 1,2,\dots,n$ do
   1) Obtain user index I_t and item set $D_t \subset \mathbb{R}^{d\times 1}$
   2) Find the connected component V_t of user I_t and compute
   $$M = \lambda I + \sum_{i\in V_t} S_i, b = \sum_{i\in V_t} b_i, T = \sum_{i\in V_t} T_i$$
   3) Compute
   $$\hat{\theta} = M^{-1} b$$
   Linear regression
   4) For any $x \in D_t$, compute
   $$U_t(x) = \min\{\hat{\theta}^\top x + \beta \|x\|_{M^{-1}}, 1\}$$
   exploitation
   exploration
   5) Recommend the K items with largest U_t values and observe $\boldsymbol{K}_t; \boldsymbol{w}_t(\boldsymbol{x}_k^t), k \le \boldsymbol{K}_t$.
   6) Update statistics
   $$S_i = S_i + \sum_{k=1}^{\boldsymbol{K}_t} x_{t,k}\, x_{t,k}^\top ,$$
   $$b_i = b_i + x_{t,K_t} w_t(x_{t,K_t}),$$
   $$T_i = T_i + K_t$$
   $$\hat{\theta}_i = (\lambda I + S_i)^{-1} b_i$$
   7) Delete edge (i, l) if
   $$\|\hat{\theta}_i - \hat{\theta}_l\| \ge \alpha \left(\sqrt{\frac{\ln T_i}{T_i}} + \sqrt{\frac{\ln T_l}{T_l}} \right)$$

   End for t

Setting

- n_u of users.
- Each action is an ordered list of K items.
- At time step t,
  - User I_t comes to be served with items $D_t \subset \mathbb{R}^d$.
  - Let $\boldsymbol{\mathcal{H}}_t$ denote the history so far.
  - The learning agent recommends $\boldsymbol{A}_t = (\boldsymbol{x}_1^t, \dots, \boldsymbol{x}_K^t)$ to the user.
  - The user checks from the first item of $\boldsymbol{A}_t$ and stops at $\boldsymbol{K}_t$-th item.
  - The learning agent observes the weights of first $\boldsymbol{K}_t$ base arms in $\boldsymbol{A}_t$, $\boldsymbol{w}_t(\boldsymbol{a}_k^t), k \le \boldsymbol{K}_t$.
- Assume that given $\boldsymbol{\mathcal{H}}_t$,$\boldsymbol{w}_t(a)$'s are mutually independent Bernoulli random variables with
$$\mathbb{E}[\boldsymbol{w}_t(a) | \boldsymbol{\mathcal{H}}_t] = \theta_{I_t}^\top x_{t,a}$$
for some unknown $\theta_{I_t} \in \mathbb{R}^{d\times 1}$ with $\|\theta_{I_t}\|_2 \le 1,\ 0 \le \theta_{I_t}^\top x_{t,a} \le 1$.
- Cluster regularity: All users in the same cluster have the same θ. Users in different clusters have noticeably different θ's:
$$\|\theta - \theta'\| \ge \gamma > 0.$$
- User uniformness: At each time, the user is drawn uniformly from the set of all users, independently over the past.
- Item regularity. At each time step, the items are drawn independently from a fixed distribution where $\mathbb{E}[xx^\top]$ has minimal eigenvalue $\lambda_x > 0$.
- The regret of action A on time t is
$$R(t, A) = f_t^* - f(A, w_t)$$
where $f_t^* = \max_{A^*} f(A^*, w_t)$
- Task: Minimize the cumulative regret of n rounds
$$R(n) = \mathbb{E}\left[\sum_{t=1}^{n} R(t, \boldsymbol{A}_t)\right].$$

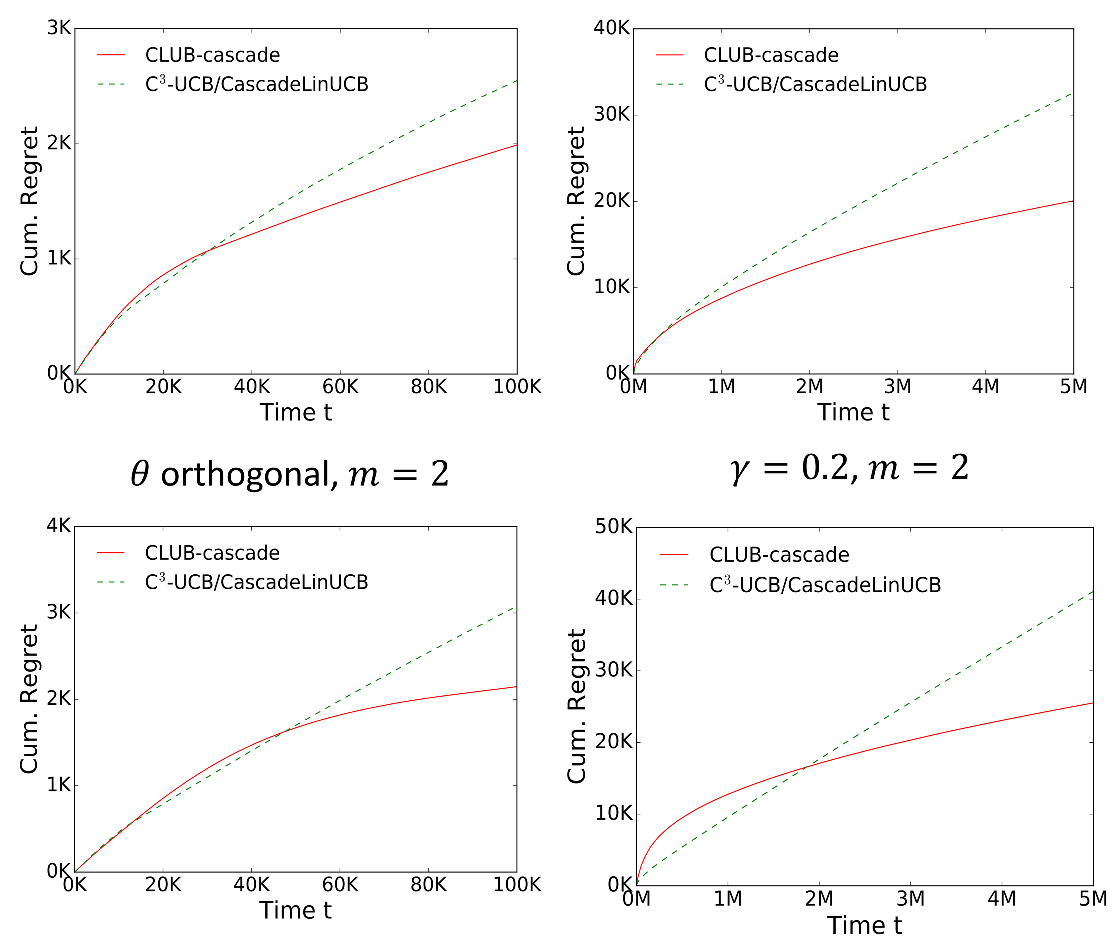
Theoretical analysis

Theorem 1. .Let $\beta = \sqrt{d \ln(1 + n/\lambda d) + 2 \ln 4mn} + \sqrt{\lambda}$ and $\alpha = 4\sqrt{d}/\lambda_x$. Then the regret of our algorithm, CLUB-cascade, satisfies
$$R(n) = O\left(d\sqrt{mnK} \ln n\right).$$

Corollary 2. When the number of clusters $m = 1$, the regret satisfies
$$R(n) = O\left(d\sqrt{nK} \ln n\right)$$
which is better than the results in [2][3].

Theorem 3. Consider a generalized linear reward function
$$\mu(\theta_{I_t}^\top x_{t,a}),$$
where μ is strictly increasing, continuously differentiable, and Lipschitz with constant κ. Let $c = \inf_{a\in[-2,2]} \mu'(a)$. Then the regret satisfies
$$R(n) = O\left(\frac{\kappa d}{c} \sqrt{mnK} \ln n\right)$$

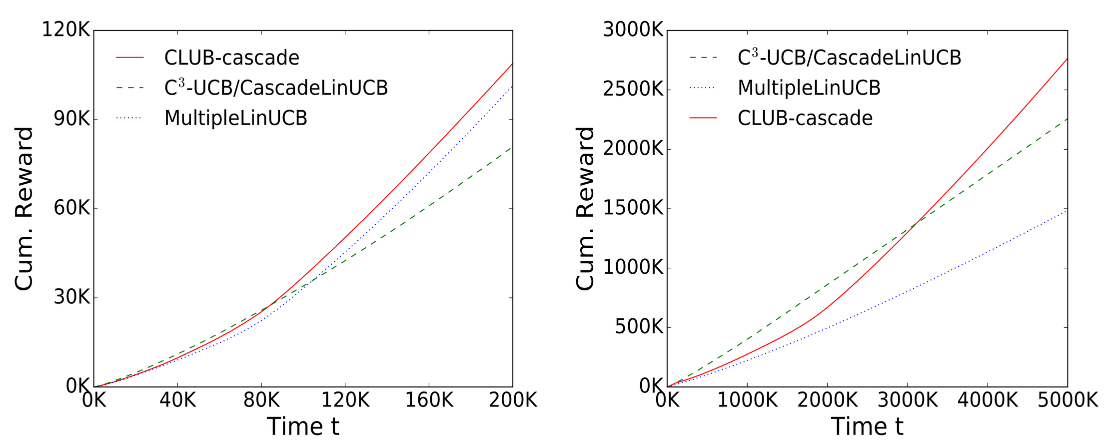
Experiments



θ orthogonal, $m = 2$ $\gamma = 0.2, m = 2$

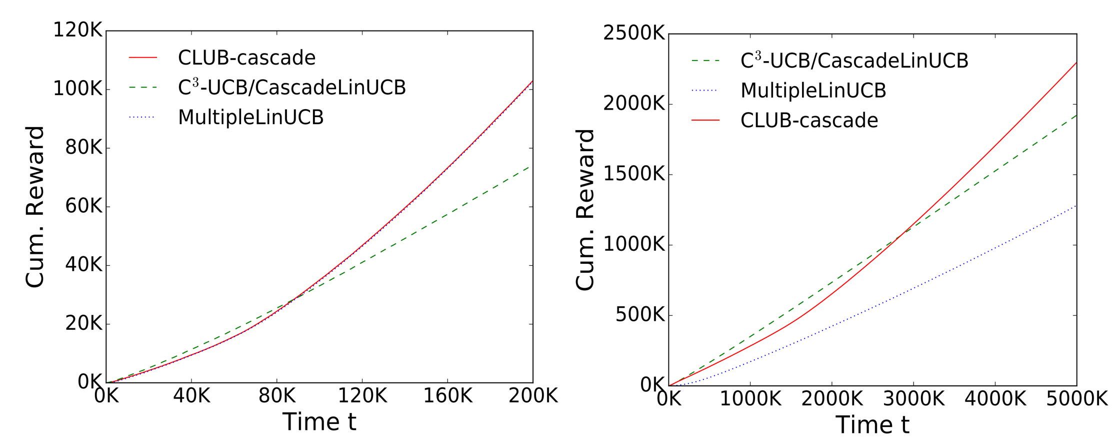
θ orthogonal, $m = 5$ $\gamma = 0.2, m = 5$

Figure 1. Experimental results for synthetic data. 40 users, 200 items, $K = 4, d = 20$.



Yelp, $u = 40$ Yelp, $u = 900$

Figure 2. Cumulative rewards on Yelp dataset. $K = 4, d = 20, L = 1k$.



MovieLens, $u = 40$ MovieLens, $u = 900$

Figure 3. Cumulative rewards on MovieLens dataset. $K = 4, d = 20, L = 1k$.

Conclusions

- Formulate Online Clustering of Contextual Cascading Bandits problem.
- Propose a CLUB-cascade algorithm that can learn clustering over users and, at the same time, effectively handle
  - contextual information
  - cascading feedback
- Theoretical analysis
- Empirical evaluation

Contact

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

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

References

1. Gentile, Li, and Zappella. "Online clustering of bandits." Proceedings of the 31st International Conference on Machine Learning. 2014.
2. Li, Wang, Zhang, Chen. "Contextual combinatorial cascading bandits." International Conference on Machine Learning. 2016.
3. Zong, Ni, Sung, Ke, Wen, Kveton. "Cascading bandits for large-scale recommendation problems." Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence. AUAI Press, 2016.