Contextual Combinatorial Cascading Bandits

Authors: Shuai Li , Baoxiang Wang , Shengyu Zhang , Wei Chen

(ICML'16: Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48, June 2016)

Instructor: Dr. Chicheng Zhang

Presenter: Rupal Jain

THE UNIVERSITY OF ARIZONA

Multi-Armed Bandits Problem

- System of base arms, ***K***
- Reward of each arm ***i*** are random samples from unknown distributions with unknown means.
    - best arm $\mu^* = max\, \mu_i$ (mean $= \mu_i$)

- In each round ***t*****,** the learning agent selects one arm i_t to play and observe the reward, $R_t(i_t)$
- Regret after playing T rounds: $\text{Regret} = T\mu^* - \mathbb{E}[\sum_{t=1}^{T} R_t(i_t)]$

Goal: Minimize the Cumulative regret

Problem of MAB: dealing with trade-off between exploitation and exploration

Feedbacks

At every time step, a learning agent chooses a subset of ground items (super arm) under certain combinatorial constraints.

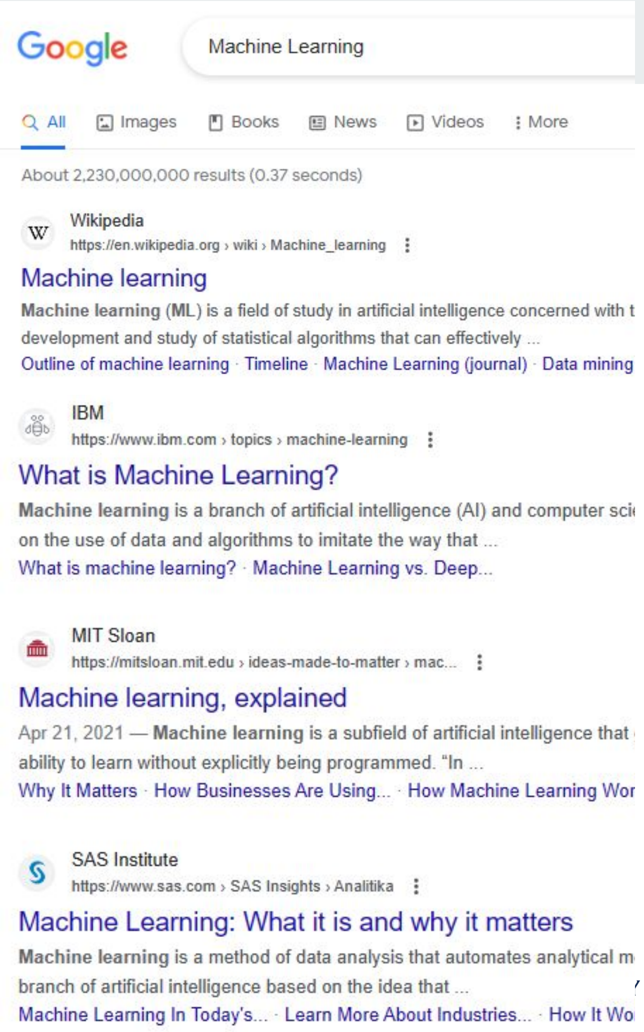
1. **Bandit feedback**: only reward of chosen super arm is obtained
2. **Semi-bandit feedback**: observe outcomes of the individual base arms in super arm
3. **Cascading feedback**: can obtain the reward of the super arm and the weights of some base arms in the chosen subset of arm, according to some stopping criterion.

THE UNIVERSITY OF ARIZONA

Related Work

Contextual Combinatorial Multi-Armed Bandit
• Involves Semi-bandit feedback and nonlinear reward • May have other combinatorial constraints Ref: Qin et al., 2014

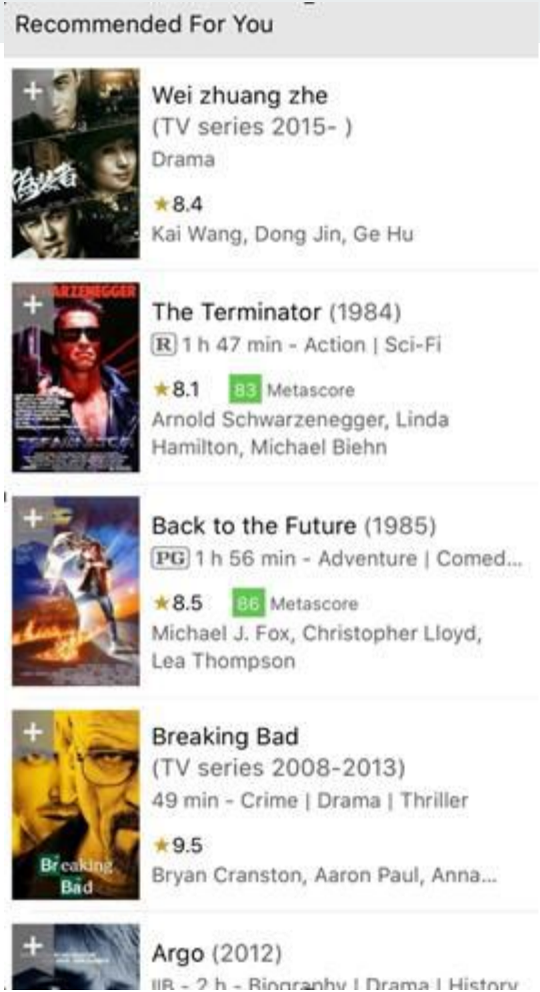
Cascading Bandit
• Feasible action is Sequential list • Stopping at first satisfactory item from the list • Feedback is recorded as Clicks • Disjunctive objective: reward of an action is 1 if there is at least one “good” item in the list Ref: Kveton et al. 2015a



Related Work - Cont.

Position Discount

- **Combes et al. 2015a** considered a **cascading model** with a particular case of **contextual information**,
- User is recommended with **K** items.
- The position discounts introduced makes the recommended set in the decreasing order of UCBs, which is more realistic
- Considered in the list order, so that the agent's reward is discounted depending on the position where the stopping criterion is met.

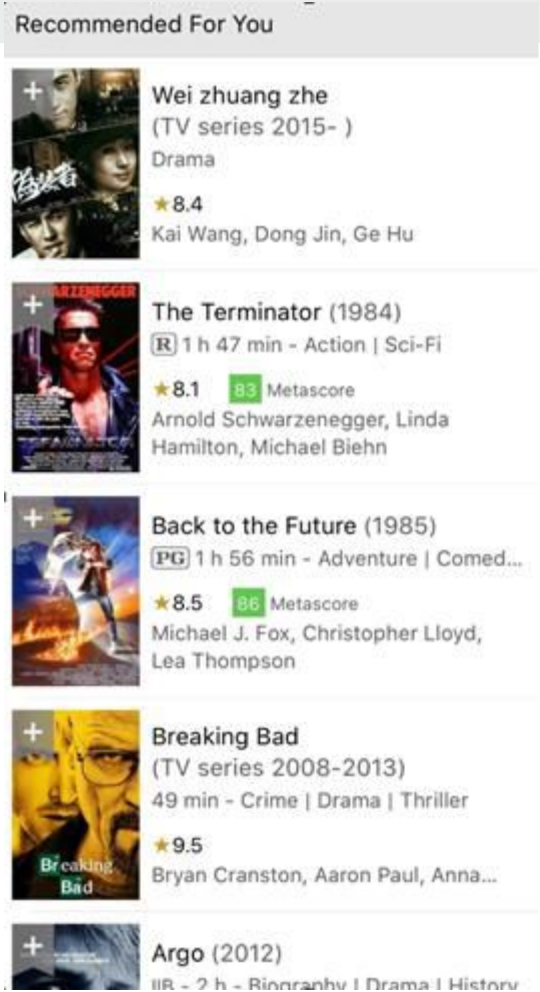


Related Work - Cont.

Combinatorial Cascading Multi-Armed Bandit

- Feasible action is subset of items under combinatorial constraints
- Can involve Semi-bandit feedback
- Conjunctive objective: reward of an action is 1 if all the items are “good” in the list
- Challenges:
    - Exponential number of actions
    - Offline optimization may already be hard

Ref: Kveton et al., 2015c



Author's Contribution

- Formulate the **Contextual Combinatorial Cascading Bandits** problem which handles
  - contextual information
  - cascading feedback
  - position discount
  - general reward function (non-linear)
- Proposed C^3 - Algorithm
- Satisfy Monotonicity and Lipschitz continuity conditions

	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	Yes	Yes	Yes	Yes

Ref:
1: Chen et al. 2016
2: Qin et al., 2014
3: Kveton et al., 2015c

THE UNIVERSITY OF ARIZONA

Learning Protocol

- $E=\{1,\ldots,L\}$: set of base arms
- **Action** $A = (a_1,\ldots,a_k)$: $a_1,\ldots,a_k \in$ **E**; a sequence of base arms (each with length of k)
  - There is a feasible action set $\mathcal{S}$ with length at most K
- At each time $t \geq 1$
  - set of contexts $\{x_{t,a}\}_{a\in E}$ are given (e.g. user/keyword features) and is ≤ 1
  - learning agent recommends a feasible action to user, $A_t = \left(a_1^t, \ldots, a_{|A_t|}^t\right) \in \mathcal{S}$
  - Cascading Feedback Model: The user checks from the first item and stops at O_t-th item.
  - Feedback: observe weights ("quality") of first O_t items in A_t at time t, $w_t(a^t_k)$, $k \leq O_t$.

$$\mathbb{E}[w_t(a)|H_t] = \underbrace{\theta_*^T}_{\text{Fixed but unknown}} x_{t,a} = w_{t,a}$$

Learning Protocol - Cont.

- **Assume the expected reward of an action A is a function $f(A,w_t)$ of expected weight, $w_t=\{w_{t,a}\}_{a\in E} = (\theta_*^{T} x_{t,a})_{a\in E}$ of each base arm.**
- **Satisfy the following assumptions:**
  - **Monotonicity:** expected reward function $\boldsymbol{f(A,w)}$ is a non-decreasing with respect to **w**: for any $w,w' \in [0, 1]^E$, if $w(a) \le w'(a)$, we have $f(A,w) \le f(A,w')$.
  - **Lipschitz continuity:** The expected reward function $\boldsymbol{f(A,w)}$ is B-Lipschitz continuous with respect to $\boldsymbol{w}$ together with position discount parameters $\gamma_k \in [0,1]$, $k \le K$. For any $w,w' \in [0, 1]^E$, we h

$$|f(A,w) - f(A,w')| \le B \sum_{k=1}^{|A|} \gamma_k \, |w(a_k) - w'(a_k)|$$

where $A = (a_1, \ldots, a_{|A|})$.

Learning Protocol - Cont.

- An Oracle $\mathcal{O}_{\mathcal{S}}(\boldsymbol{w})$ is called an α-approximation oracle for some $\alpha \leq 1$, if on given input $\boldsymbol{w}$, the oracle returns an action $\mathbf{A} = \mathcal{O}_{\mathcal{S}}(\boldsymbol{w}) \in \mathcal{S}$ satisfying $\boldsymbol{f(A,w)} \geq \boldsymbol{\alpha}\, \boldsymbol{f(A^*,w)}$ where $\boldsymbol{A} = \mathbf{argmax}_{\mathbf{A}\in\mathcal{S}}\, \boldsymbol{f(A,w)}$
- Regret in $\boldsymbol{T}$ rounds

$$\textbf{Regret} = \underbrace{\sum_{t=1}^{T} f_t^*}_{\text{Best cumulative reward}} - \mathbb{E}\left[\sum_{t=1}^{T} f(A_t, w_t)\right]$$

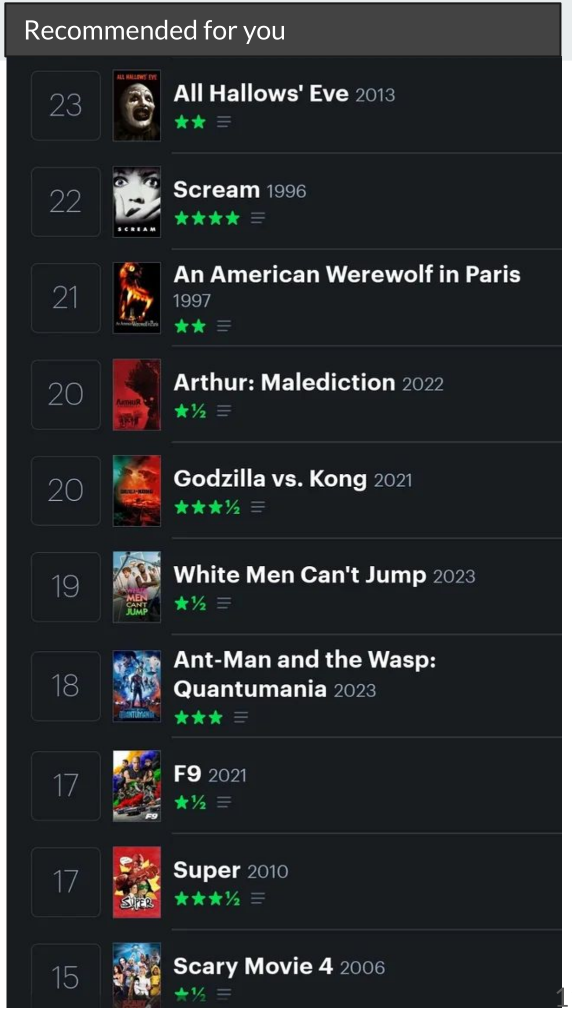
f_t^* :max expected reward in round t

THE UNIVERSITY OF ARIZONA

Example - Movie Recommendation

- Each Movie i has a feature vector m_i.
- At time t,
  - A random user comes with feature vector $\mathbf{u}_t$
  - Use $\boldsymbol{x}_{i,t} = \mathbf{g}(\mathbf{m}_i, \mathbf{u}_t)$, a function of $\mathbf{m}_i$ and $\mathbf{u}_t$, as context
  - The learning agent recommends a list of movies $\boldsymbol{A}_t$.
  - The user checks from the first movie and stops at the attractive one.
  - The learning agent received reward γ_k if the user stops at position k.

    $$1 = \gamma_k \geq \ldots \geq \gamma_k \geq 0$$



Algorithm

- To estimate the expected reward, we get an estimate $\boldsymbol{\theta}_*$ using ridge regression problem on context vector $\boldsymbol{x}$ and observed rewards $\boldsymbol{w}$.
- We get an l^2 **-regularized least-squares estimation** of $\boldsymbol{\theta}_*$ with **regularization parameter, λ > 0**

$$\hat{\theta}_t = \left(X_t^T X_t + \lambda I\right)^{-1} X_t^T Y_t$$

Where $X_t \in \mathbb{R}^{(\sum_{s=1}^{t} O_s).d}$: matrix with rows as $\gamma_k \, x_{s,a_k^s}^T$ and Y_t: column vector with elements as $\gamma_k \, w_s(a_k^s), k \in [O_s], s \in [t]$

- Let V_t is a symmetric positive definite matrix, $\in R^{d \times d}$

$$V_t = X_t^T X_t + \lambda I = \lambda I + \sum_{s=1}^{t} \sum_{k=1}^{O_s} \gamma_k^2 \, x_{s,a_k^s} \, x_{s,a_k^s}^T$$

THE UNIVERSITY OF ARIZONA

Algorithm

- **Lemma 1**: $\beta_t(\delta) = R\sqrt{ln(\frac{det(V_t)}{\lambda^d \delta^2})} + \sqrt{\lambda}$ **. For any δ>0, with probability at least 1-δ, for all t>0, we have $||\hat{\theta}_t - \theta_* ||_{V_t} \leq B_t(\delta)$**

Proof: By theorem 2 in (Abbasi-Yadkori et al., 2011), we can obtain a good estimate of difference between θ_* and $\hat{\theta}_t$.

Interpretation: With high probability, the estimate $\hat{\theta}$ lies in the ellipsoid centered at θ_* with confidence radius $\beta_t(\delta)$ under V_t norm.

- The upper confidence bound (UCB) of the expected weight for each base arms:

$$U_t(a) = min\left\{\hat{\theta}_{t-1}^T x_{t,a} + \beta_{t-1}(\delta)\left||x_{t,a}\right||_{V_{t-1}^{-1}}, 1\right\}$$

THE UNIVERSITY OF ARIZONA

- **Lemma 2**: When $||\boldsymbol{\hat{\theta}}_t - \boldsymbol{\theta}_* ||_{Vt} \leq \mathbf{B}_t\ (\boldsymbol{\delta})$ holds for time t-1, we have

$$0 \leq U_t(a) - w_{t,a} \leq 2\beta_{t-1}(\delta)||x_{t,a}||_{V_{t-1}^{-1}}$$

Proof: Note $w_{t,a} = (\theta_*^T x_{t,a})$

By Holder's inequality

$$\begin{aligned}
|\hat{\theta}_{t-1}^T x_{t,a} - \theta_*^T x_{t,a}| &= |[V_{t-1}^{1/2}(\hat{\theta}_{t-1}^T - \theta_*]^T (V_{t-1}^{-1/2} x_{t,a})| \\
&\leq ||V_{t-1}^{1/2}(\hat{\theta}_{t-1} - \theta_*||_2 ||V_{t-1}^{-1/2} x_{t,a}||_2 \\
&= ||\hat{\theta}_{t-1} - \theta_*||_{V_{t-1}} ||x_{t,a}||_{V_{t-1}^{-1}} \\
&\leq \beta_{t-1}(\delta)||x_{t,a}||_{V_{t-1}^{-1}}
\end{aligned}$$

Because $1 - \theta_*^T x_{t,a} \geq 0$ and

$$\begin{aligned}
0 &\leq (\hat{\theta}_{t-1}^T x_{t,a} + \beta_{t-1}(\delta)||x_{t,a}||_{V_{t-1}^{-1}}) - \theta_*^T x_{t,a} \\
&\leq 2\beta_{t-1}(\delta)||x_{t,a}||_{V_{t-1}^{-1}}
\end{aligned}$$

Claimed result is obtained…

THE UNIVERSITY OF ARIZONA

Algorithm

1. Learning agent computer (UCBs), $\boldsymbol{U_t} \in [0,1]^E$ for expected weights of all base arms in E.

2. Uses $\boldsymbol{U_t}$ to select an action $\boldsymbol{A}=(a^t_1,\ldots,a^t_{|A_t|})$

3. User selects from the first base arm in $\boldsymbol{A_t}$ and stops at $\boldsymbol{O_t}$-th base arm,

Agent observes $\boldsymbol{w_t}(a^t_k)$, $k \le O_t$

4. **Learning agent updates $\boldsymbol{V_t}$, X_t Y_t to get newer estimates $\hat{\boldsymbol{\theta}}_t$ and $\boldsymbol{\theta}_*$ and new confidence radius, $\boldsymbol{\beta_t}(\delta)$**

Algorithm 1 C^3-UCB

```
Parameters:
    {γ_k ∈ [0,1]}_{k≤K}; δ = 1/√n; λ ≥ C_γ = Σ_{k=1}^K γ_k^2
Initialization:
    θ̂_0 = 0, β_0(δ) = 1, V_0 = λI, X_0 = ∅, Y_0 = ∅
for all t = 1, 2, ..., n do
    Obtain context x_{t,a} for all a ∈ E                                    (1)
    ∀a ∈ E, compute
      U_t(a) = min{θ̂_{t-1}^T x_{t,a} + β_{t-1}(δ)||x_{t,a}||_{V_{t-1}^{-1}}, 1}
    //Choose action A_t using UCBs U_t                                      (2)
    A_t = (a_1^t, ..., a_{|A_t|}^t) ← O_S(U_t)
    Play A_t and observe O_t, w_t(a_k^t), k ∈ [O_t]                          (3)
    //Update statistics                                                     (4)
    V_t ← V_{t-1} + Σ_{k=1}^{O_t} γ_k^2 x_{t,a_k^t} x_{t,a_k^t}^T
    X_t ← [X_{t-1}; γ_1 x_{t,a_1^t}^T; ...; γ_{O_t} x_{t,a_{O_t}^t}^T]
    Y_t ← [Y_{t-1}; γ_1 w_t(a_1^t); ...; γ_{O_t} w_t(a_{O_t}^t)]
    θ̂_t ← (X_t^T X_t + λI)^{-1} X_t^T Y_t
    β_t(δ) ← R√(ln(det(V_t)/(λ^d δ^2))) + √λ
end for t
```

Result

Theorem: Suppose the expected reward function ***f(A,w)*** is a function of expected weights and satisfies the requirements of monotonicity and B-Lipschitz continuity. Then the α-regret of **C^3-UCB** algorithm, satisfies

$$Regret^{\alpha}(n) = O(\frac{dB}{p^*}\sqrt{TKln(T)})$$

Proof:

$$R^{\alpha}(n) \leq \frac{2\sqrt{2}B}{p^*}\sqrt{nKdln(1+C_{\gamma}n/(\lambda d))}$$
$$.(R\sqrt{ln[(1+C_{\gamma}n/(\lambda d))^d n]}+\sqrt{\lambda})+\alpha\sqrt{n}$$
$$= O(\frac{dBR}{p^*}\sqrt{nKln(C_{\gamma}n)})$$

$\mathbf{p_{t,A}}$ be the probability of full observation of A.

$\mathbf{p^* = \min_{1 \leq t \leq T} \min_{A \in S} p_{t,A}}$: Minimal probability that action has all base arms observed all time.

d**:** dimension of latent and feature vectors;

K**:** largest length of the sequence

n = T (number of rounds)

R is the sub-Gaussian constant

$C\gamma \leq K$ (sum of discounts $\leq$ number of positions)

$$C_{\gamma} = \sum_{k=1}^{K} \gamma_k^2 \leq K$$

THE UNIVERSITY OF ARIZONA

Experiment 1

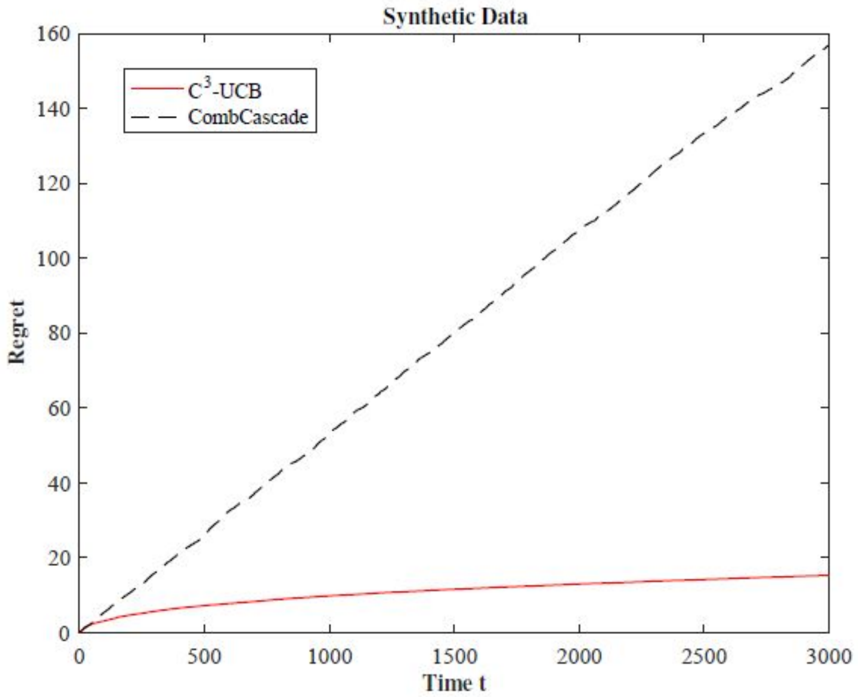
<u>Synthetic Data:</u> compare C^3-UCB to CombCascade (Combinatorial Cascading) on synthetic problems.

Setup:

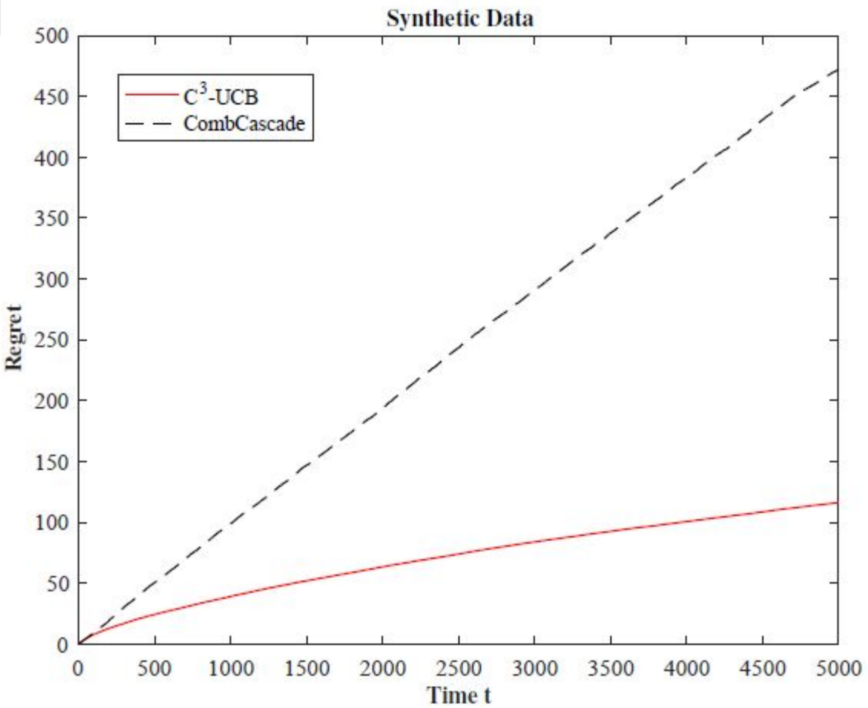
1. The problem is a contextual cascading bandit with L = 200 items and K = 4, d = 20 where at each time t.
2. The agent recommends K items to the user.
3. At first, we randomly choose a $\boldsymbol{\theta}$ in R^{d-1} with $\|\boldsymbol{\theta}\|_2 = \mathbf{1}$ and let $\boldsymbol{\theta}_* = (\boldsymbol{\theta}/2, 1/2)$.
4. Then at each time t, we randomly assign $\mathbf{x}'_{t,a}$ in $\mathbf{R}^{d-1}$ with $\|\mathbf{x}'_{t,a}\|_2 = \mathbf{1}$ to arm **a** and use $\mathbf{x}_{t,a} = (\mathbf{x}'_{t,a}, \mathbf{1})$ to be the contextual information for arm a.

This processing will guarantee the inner product $\theta_*^\top x_{t,a} = \frac{1}{2}(\theta^\top x'_{t,a} + 1) \in [0, 1]$

Next we generate the weight for arm a at time t by a random sample from the Bernoulli distribution with mean $\boldsymbol{\theta}_*^{\mathrm{T}} \boldsymbol{x}_{t,a}$

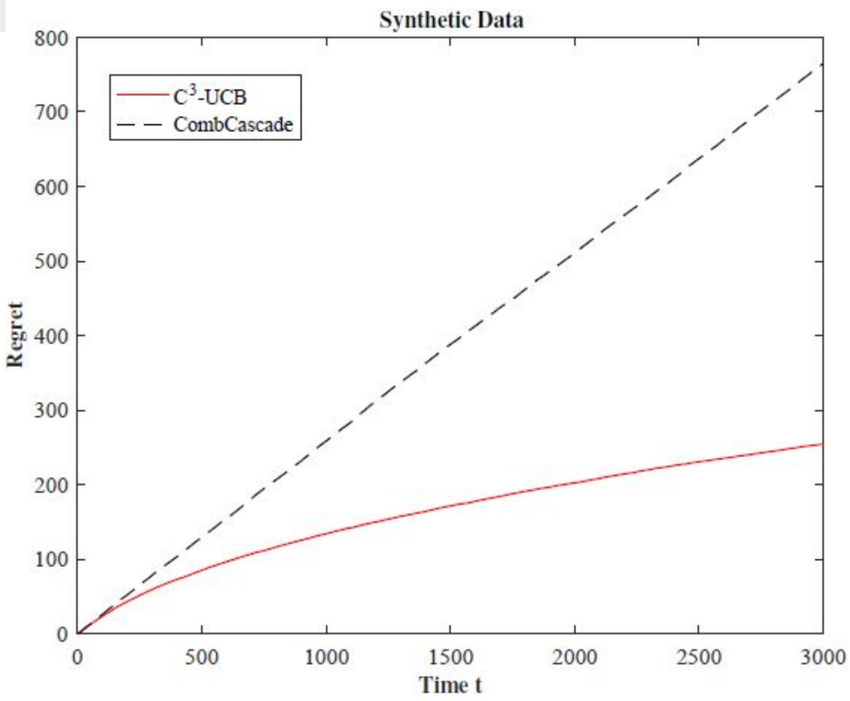


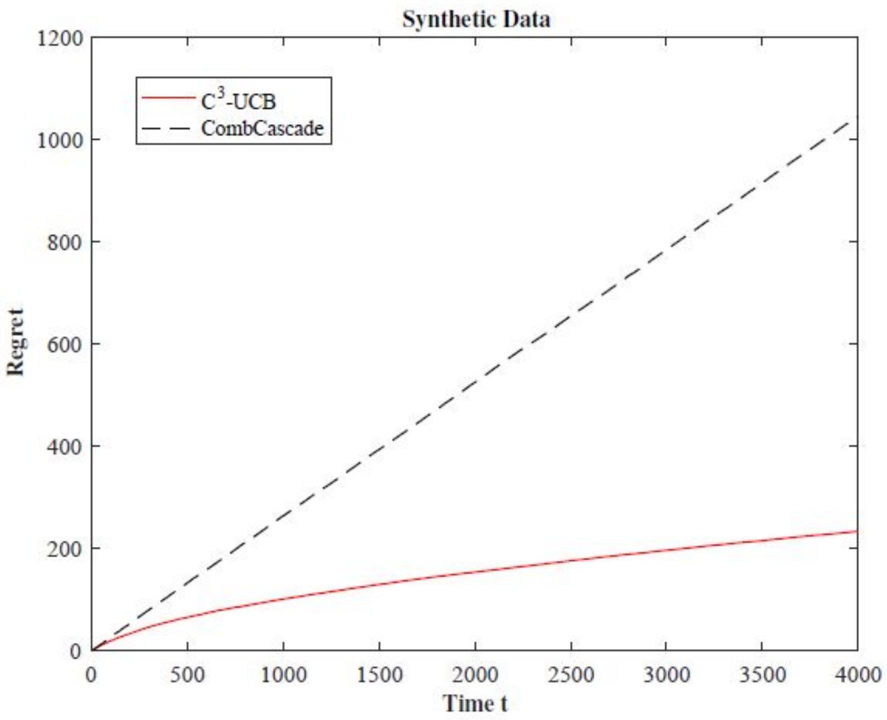
Disjunctive, $\gamma = 1$, 9.77%



Disjunctive, $\gamma = 0.9^{k-1}$, 24.6%

The above two settings are of disjunctive objective where the learning agent chooses a set of K items out of L ground items and observes a prefix of the chosen K items until the first one with weight 1;





Conjunctive, $\gamma = 1$, 3.33%

Conjunctive, $\gamma = 0.9^{k-1}$, 22.4%

The above two settings are of conjunctive objective where the learning agent chooses a set of K items out of L ground items and observes from the first item until the first one with weight 0.

THE UNIVERSITY OF ARIZONA

Experiment 2

Movie Recommendation: evaluate C3-UCB algorithm with dataset MovieLens (Lam & Herlocker, 2015) of 2015

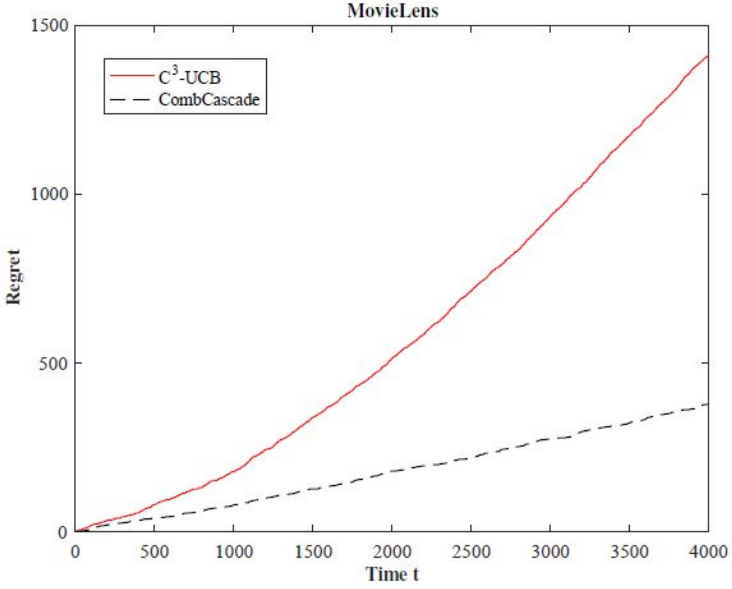
Problem:

1. There is a big sparse matrix $A \in \{0, 1\}^{N1 \times N2}$ where **A(i; j) = 1 denotes user i has watched movie j**.
2. Split A as H + F according to a Bernoulli distribution ~Ber(p) for some fixed p. (H : known history "what users have watched" and F as future criterion.)
3. Derive **feature vectors** of both users and movies via SVD, $\mathbf{U = (u_1, . . ., u_{N1}) \text{ and } M = (m_1, . . ., m_{N2})}$.
4. At every time t, we randomly choose a user $I_t \in [N_1]$. From Li et al., 2010, use $\mathbf{x_{t;j} = u_{I_t} m^T_j}$ as the contextual information for each movie **j**.
5. User provides cascading feedback, stopping at a movie they like and Agent receives discounted reward based on stop position.
6. The real weight of movie **j** at time t, $\mathbf{w_t(j)}$**, is** $\mathbf{F(I_t, m_j)}$.

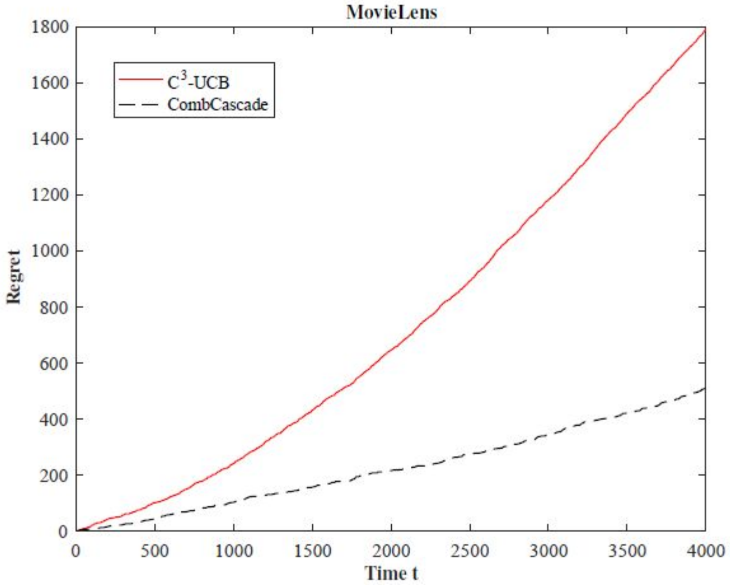
THE UNIVERSITY OF ARIZONA

Setup:

1. The problem is a contextual cascading bandit with L = 400 movies and K = 4, d = 400 where at each time t.
2. We experiment with both $\gamma = 1$ (no position discount) and $\gamma = 0{:}9$, and compare our algorithm with CombCascade.



γ=1



γ=0.9

Result: The rewards of our algorithms are 3.52 and 3.736 times of those of CombCascade (for $\gamma = 1$ and 0:9, respectively), which demonstrate the advantage to involve contextual information in real applications.

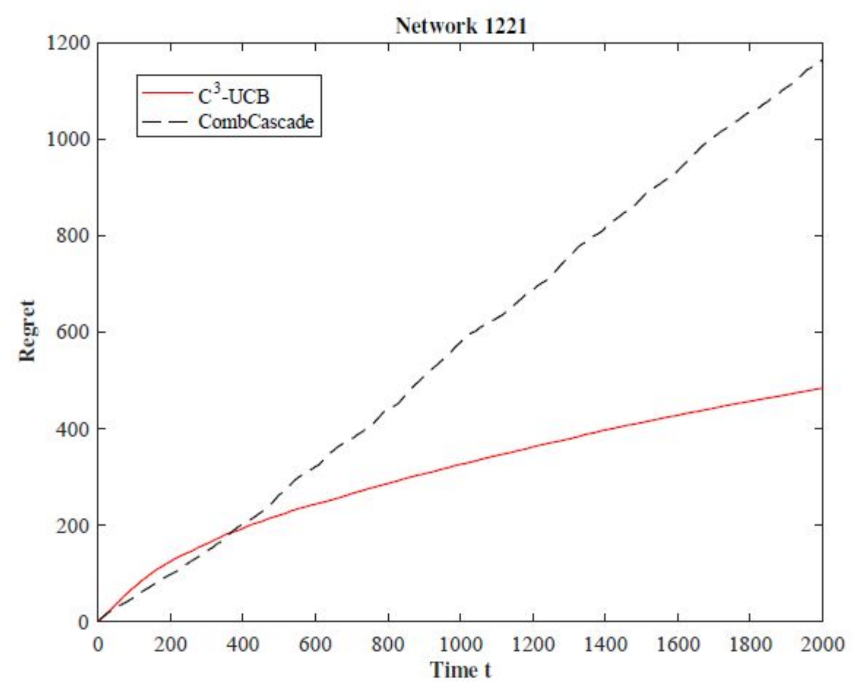
Experiment 3

Network Routing Problem: evaluate C^3-UCB with RocketFuel dataset (Spring et al., 2004).

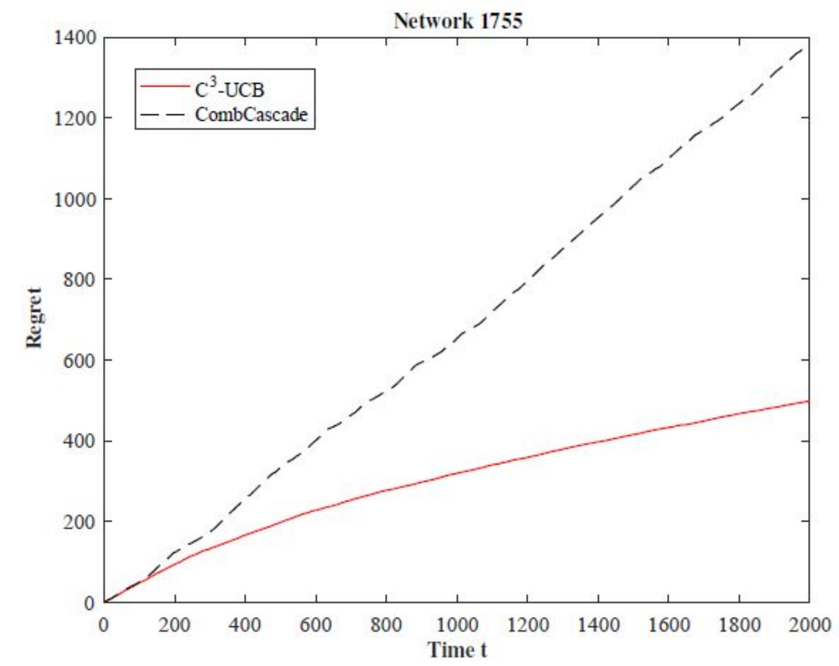
Setup:

1. The ground set E is the set of links in the network.
2. Before learning, the environment randomly chooses a d-dimensional vector $\boldsymbol{\theta} \in [0, 1]^d$ **(d = 5)**
3. At each time t, a pair of source and destination nodes are randomly chosen
4. The feasible action set $\mathbf{S_t}$ at time t contains all simple paths, paths without cycles, between the source and destination.
5. Any edge a in the set $\mathbf{S_t}$ is assigned with a random d-dimensional contextual information vector $\mathbf{x_{t,a}}$.

THE UNIVERSITY OF ARIZONA



Network 1221, γ=1



Network 1221, γ=0.9

Result: Both θ and x have been processed same as experiment 1, such that $\boldsymbol{\theta_*}^{\mathbf{T}} \boldsymbol{x} \in [0, 1]$. The weight for each edge a is a sample from Bernoulli distribution with mean $\boldsymbol{\theta_*}^{\mathbf{T}} \boldsymbol{x_{t,a}}$. Then the learning agent recommends a feasible path A to maximize the expected reward in the conjunctive objective. This experiment is on different position discounts.

Conclusion:

- Incorporating contextual information to cascading bandit with position discounts
- Each action is an ordered list and only a prefix of the action is observed each time.
- demonstrate the advantage to involve contextual information and position discounts.
- Application potential
    - Any sequential list recommendation (search, ads, mobile recommendations)
        - Need online (real time) feedback

Future Work:

- investigate on lower bounds of the regret and cascading on general graphs

THE UNIVERSITY OF ARIZONA

References:

- https://icml.cc/2016/reviews/581.txt
- Kveton, Branislav, Wen, Zheng, Ashkan, Azin, and Szepesvari, Csaba. Combinatorial cascading bandits. Advances in Neural Information Processing Systems, 2015c.
- Kveton, Branislav, Szepesv´ari, Csaba, Wen, Zheng, and Ashkan, Azin. Cascading bandits: learning to rank in the cascade model. In Proceedings of the 32th International Conference on Machine Learning, 2015a.
- Qin, Lijing, Chen, Shouyuan, and Zhu, Xiaoyan. Contextual combinatorial bandit and its application on diversified online recommendation. In Proceedings of the 2014 SIAM International Conference on Data Mining (SDM), 2014.
- Combes, Richard, Magureanu, Stefan, Proutiere, Alexandre, and Laroche, Cyrille. Learning to rank: Regret lower bounds and efficient algorithms. In Proceedings of the 2015 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems, pp. 231–244. ACM, 2015a.