上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

上海交通大学
约翰·霍普克罗夫特计算机科学中心
John Hopcroft Center for Computer Science

Online Learning to Rank: An Overview

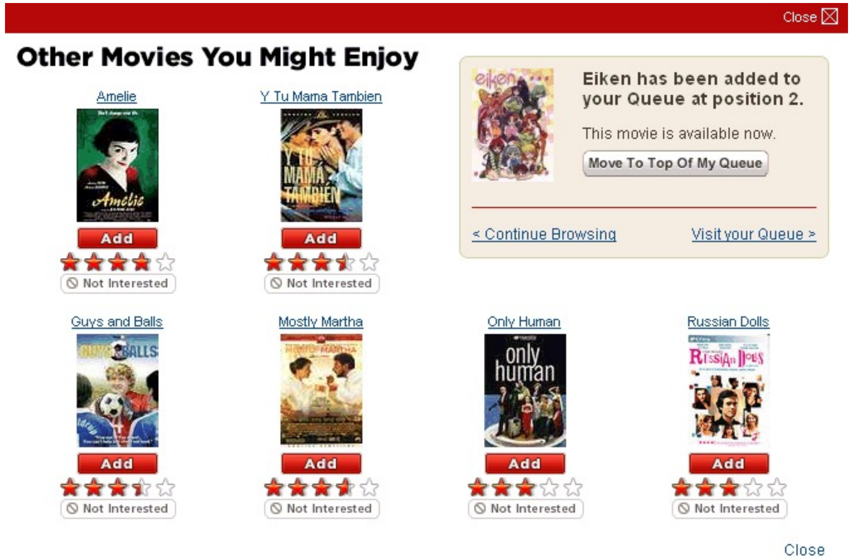
Shuai Li

http://shuaili8.github.io/

July 15, 2021

DRL4IR Workshop @ SIGIR 2021

Motivation - Learning to Rank



- Amazon, YouTube, Facebook, Netflix, Taobao

Multi-armed Bandits

Bandits

Time	1	2	3	4	5	6	7	8	9	10	11	12
Left arm	$1	$0			$1	$1	$0					
Right arm			$1	$0								

- Five rounds to go. Which arm would you choose next?

Multi-armed Bandit Problem

- A special case of reinforcement learning
- There are L arms items/products/movies/news/…
  - Each arm a has an unknown reward distribution on $[0,1]$ with unknown mean $\alpha(a)$ CTR/profit/…
  - The best arm is $a^* = \operatorname{argmax} \alpha(a)$

MAB Setting

- At each time t
  - The learning agent selects one arm a_t
  - Observe the reward $X_{a_t,t}$
- Objective:
  - Maximize the expected cumulative reward in T rounds $\mathbb{E}[\sum_{t=1}^{T} \alpha(a_t)]$
  - Minimize the regret in T rounds

$$R(T) = T \cdot \alpha(a^*) - \mathbb{E}\left[\sum_{t=1}^{T} \alpha(a_t)\right]$$

- Balance the trade-off between exploration and exploitation
  - Exploitation: Select arms that yield good results so far
  - Exploration: Select arms that have not been tried much before

UCB – Upper Confidence Bound

- Principle: optimism in face of uncertainty
- UCB policy: Select

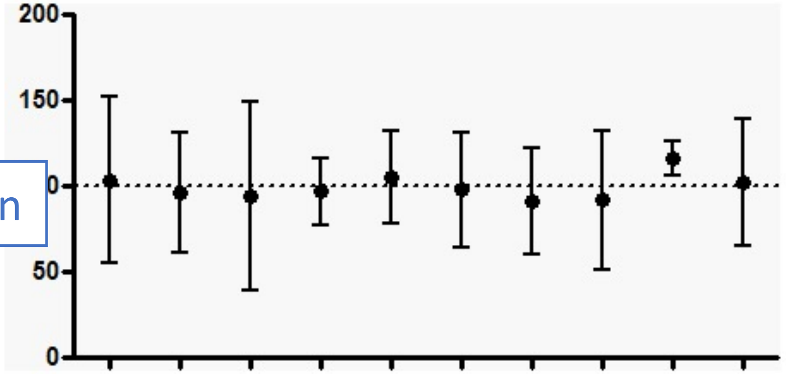
exploitation

exploration

$$a_t = \mathrm{argmax}_a \hat{\alpha}_a + \sqrt{\frac{2\log t}{T_a(t)}}$$

sample mean

selection times of arm a
till round t



- Gap-dependent regret bound $O(\frac{L}{\Delta}\log T)$ where $\Delta = \min_{\alpha_a<\alpha^*} \alpha^* - \alpha_a$ is the minimal gap
  match lower bound
- Gap-free bound $O(\sqrt{LT\log T})$ tight up to a factor of $\sqrt{\log T}$

- Auer, P., Cesa-bianchi, N., & Fischer, P. Finite-time Analysis of the Multiarmed Bandit Problem. Machine Learning, 2002.

Online Learning to Rank

Setting: Online Learning to Rank

- There are L items
  - Each item a with an unknown attractiveness $\alpha(a)$
- There are K positions
- At each time t
  - The learning agent recommends a list of items $A_t = (a_1^t, a_2^t, \dots, a_K^t)$
  - Receives the binary click feedback vector $C_t \in \{0,1\}^K$
- Objective: minimize the regret over T rounds

$$R(T) = T \cdot r(A^*) - \mathbb{E}\left[\sum_{t=1}^{T} r(A_t)\right]$$

where
  - $r(A)$ is the reward of list A
  - $A^* = (1,2, \dots, K)$ by assuming arms are ordered by $\alpha(1) \geq \alpha(2) \geq \cdots \geq \alpha(L)$

Click Models

- Describe how users interact with a list of items

- Cascade model (CM)
  - Assumes the user checks the list from position 1 to position K, clicks at the first satisfying item and stops
  - There is at most 1 click
  - $r(A) = 1 - \prod_{k=1}^{K}(1 - \alpha(a_k))$
  - The meaning of received feedback $(0,0,1,0,0)$

- Chuklin, A., Markov, I., & Rijke, M. D. Click models for web search. Synthesis Lectures on Information Concepts, Retrieval, and Services, 2015.

Key Point for Analysis

- The regret is defined on the whole list

$$R(T) = T \cdot r(A^*) - \mathbb{E}\left[\sum_{t=1}^{T} r(A_t)\right]$$

- But the received feedback is partial and random
- A key lemma

$$r(A^*; w_t) - r(A_t; w_t)$$
$$\leq \sum_{k=1}^{K} \prod_{i=1}^{k-1} (1 - w_t(a_{t,i}))\left[w_t(a^*_{t,k}) - w_t(a_{t,k})\right]$$

✗

✗

✓

?

?

Regret Bound

- For the cascade click model

$$R(T) = O\left(\frac{L}{\Delta}\log T\right)$$

- Contextual cascading bandits
  - The click rate of each item is a linear form

$$R(T) = O\left(d\sqrt{KT}\log T + e^{K}\right)$$

  - d is the feature dimension

- Kveton, B., Szepesvari, C., Wen, Z., & Ashkan, A. Cascading bandits: Learning to rank in the cascade model. ICML, 2015.
- Li, S., Wang, B., Zhang, S., & Chen, W. (2016). Contextual combinatorial cascading bandits. ICML, 2016.
- Zong, S., Ni, H., Sung, K., Ke, N. R., Wen, Z., & Kveton, B. Cascading bandits for large-scale recommendation problems. UAI, 2016.
- Li, S., & Zhang, S. Online clustering of contextual cascading bandits. AAAI, 2018.

Click Models – Dependent Click Model (DCM)

- Allow multiple clicks
- Assume there is a probability of satisfaction after each click
- $r(A) = 1 - \prod_{k=1}^{K}(1 - \alpha(a_k)\gamma_k)$
  - γ_k: satisfaction probability after click on position k
- The meaning of received feedback $(0,1,0,1,0)$

✗no click

✓click, not satisfied

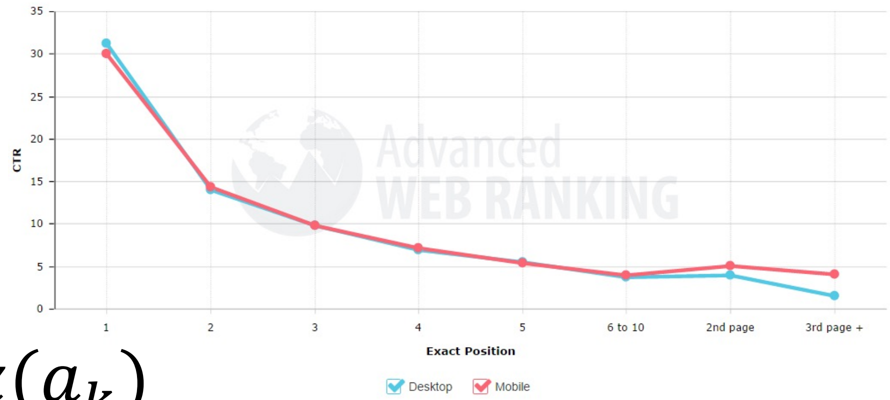
✗no click

✓click, satisfied?

?

Click Models – Position-based Model (PBM)

- Most popular model in industry
- Assumes the user click probability on an item a of position k can be factored into $\beta_k \cdot \alpha(a)$
- β_k is position bias. Usually $\beta_1 \geq \beta_2 \geq \cdots \geq \beta_K$



- $r(A) = \sum_{k=1}^{K} \beta_k \cdot \alpha(a_k)$
- The meaning of received feedback $(0,1,0,1,0)$

Bandit Works for Specific Click Models

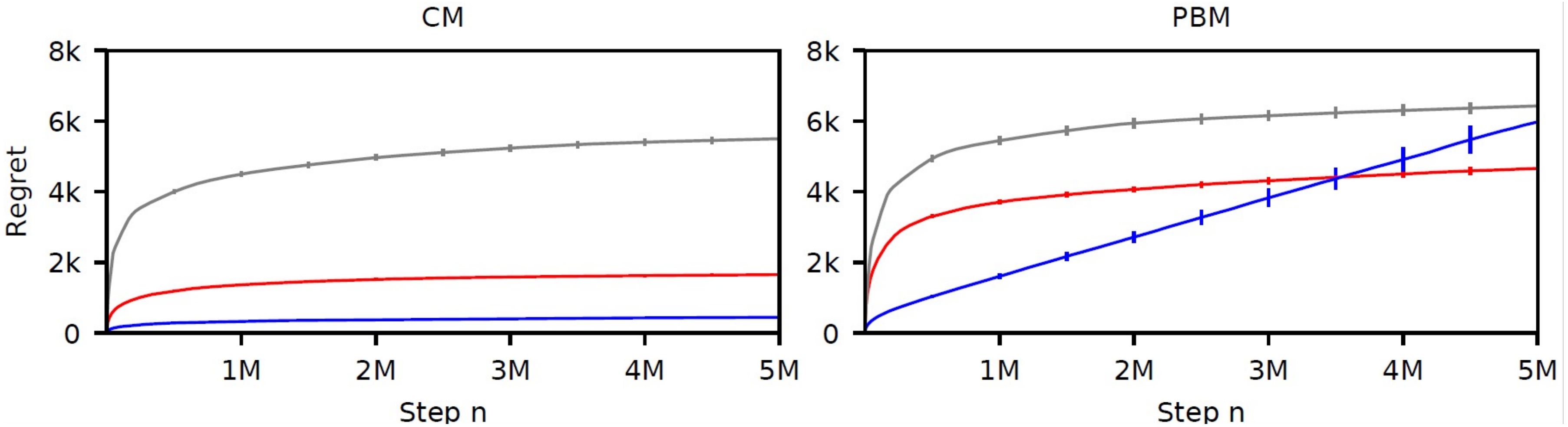
	Context	Click Model	Regret
[KSWA, 2015]	-	CM	$\Theta\left(\frac{L}{\Delta}\log T\right)$
[LWZC, ICML'2016] [ZNSKWK, UAI'2016] [LZ, AAAI'2018]	GL	CM	$O\left(d\sqrt{TK}\log T\right)$
[KKSW, 2016]	-	DCM	$O\left(\frac{L}{\Delta}\log T\right)$
[LLZ, COCOON'2018]	GL	DCM	$O\left(dK\sqrt{TK}\log T\right)$
[LVC, 2016]	-	PBM with known β	$O\left(\frac{L}{\Delta}\log T\right)$

- Katariya, S., Kveton, B., Szepesvari, C., & Wen, Z. DCM bandits: Learning to rank with multiple clicks. ICML, 2016.
- Lagrée, P., Vernade, C., & Cappe, O. Multiple-play bandits in the position-based model. NeurIPS, 2016.
- Komiyama, J., Honda, J., & Takeda, A. Position-based multiple-play bandit problem with unknown position bias. NeurIPS, 2017.
- Liu, W., Li, S., & Zhang, S. Contextual dependent click bandit algorithm for web recommendation. COCOON, 2018.

General Click Models

Modeling Bias

- CascadeKLUCB is the best algorithm under Cascade Model, but suffers linear regret in the environment of Position-based Model



- TopRank BatchRank are two algorithms designed for general click model

- Zoghi, M., Tunys, T., Ghavamzadeh, M., Kveton, B., Szepesvari, C., & Wen, Z. Online learning to rank in stochastic click models. ICML, 2017.
- Lattimore, T., Kveton, B., Li, S., & Szepesvari, C. (2018). TopRank: A practical algorithm for online stochastic ranking. NeurIPS, 2018.

General Click Models

- Common observations for click models
  - The click-through-rate (CTR) of list A on position k can be factored as
    $$v(A,k) = \mathrm{CTR}(A,k) = \chi(A,k)\alpha(a_k)$$
    where $\chi(A,k)$ is the examination probability of list A on position k
  - $\chi(A,k) = \prod_{i=1}^{k-1}(1-\alpha(a_i))$ in Cascade Model
  - $\chi(A,k) = \beta_k$ in Position-based Model
- Difficulties on General Click Models
  - χ depends on both click models and lists

Assumptions

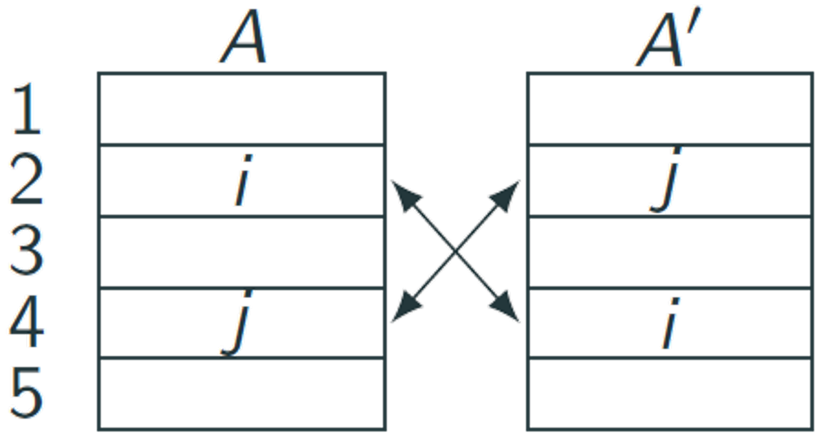
1. $v(A, k) = 0$ for all $k > K$

2. $A^* = (1,2, \dots, K)$ has the highest value $\sum_{k=1}^{K} v(A, k)$, where $\alpha(1) \geq \alpha(2) \geq \cdots \geq \alpha(L)$

3. Suppose $\alpha(i) \geq \alpha(j)$ and $\sigma: [L] \to [L]$ only exchanges i and j. Then for any list A

$$v\left(A, A^{-1}(i)\right) \geq \frac{\alpha(i)}{\alpha(j)} v\left(\sigma \circ A, A^{-1}(i)\right)$$

   - Illustration: $\alpha(i) \geq \alpha(j)$. Then $\chi(A, 2) \geq \chi(A', 2)$ and $\chi(A, 4) \leq \chi(A', 4)$

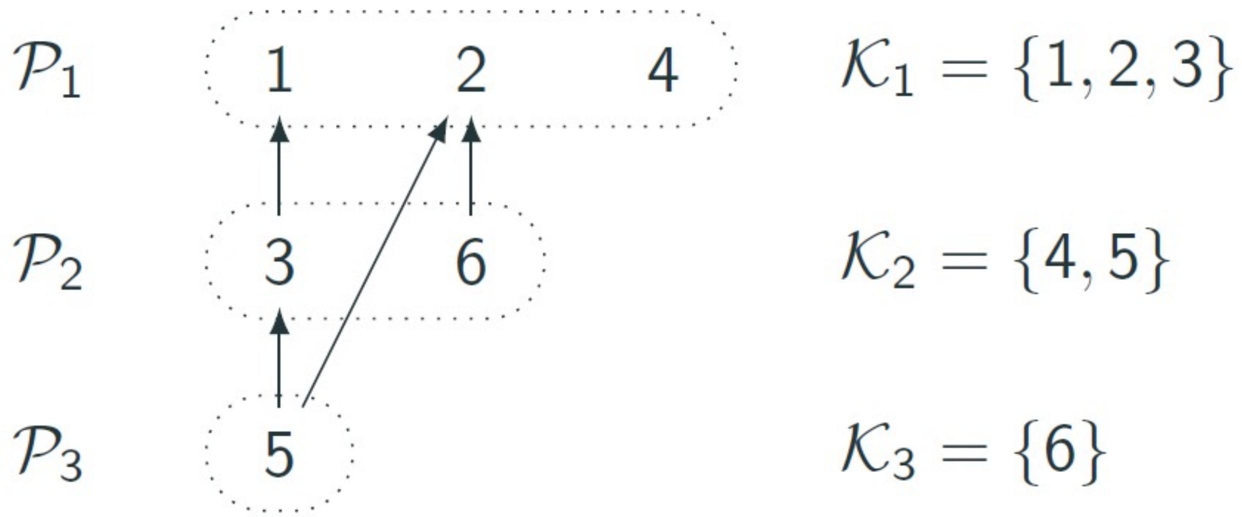


4. $\chi(A, k) \geq \chi(A^*, k)$

It can be checked that CM and PBM both satisfy these assumptions

TopRank [LKLS, NeurIPS’18]

- TopRank: Topological Ranking
- It maintains a set of order relationships between pairs of items: item b is worse than item a



e.g. (2,1,4,3,6,5), (4,1,2,6,3,5)

TopRank [LKLS, NeurIPS'18] 2

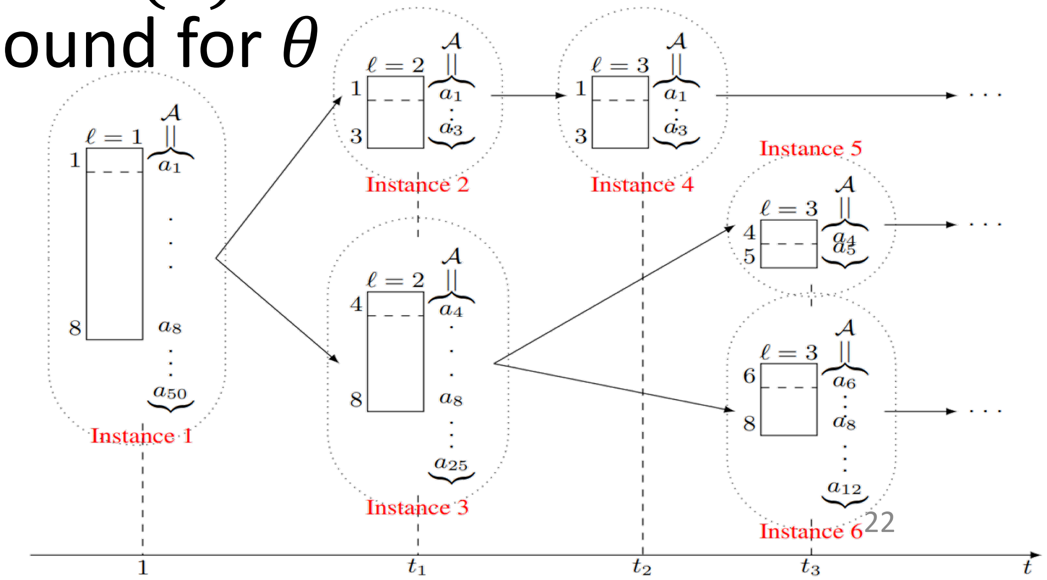
- TopRank ranks items randomly in each partition
- Based on the received click-or-not feedback, it is equivalent to draw a click difference X_{ab} on $\{-1,0,1\}$ for each pair of items (a, b) in the same partition
  - $X_{ab} = 1$ if a is clicked but b is not clicked
  - $\mathbb{E}[X_{ab} | X_{ab} \neq 0] \geq \frac{\alpha(a)-\alpha(b)}{\alpha(a)+\alpha(b)}$
- b is worse than a if $S_{ab} \geq \sqrt{2N_{ab} \log\left(\frac{c}{\delta}\sqrt{N_{ab}}\right)}$ and $N_{ab} > 0$
  - $S_{ab} = \sum_t X_{ab,t}$ is the sum of click difference in the same partition
  - $N_{ab} = \sum_t |X_{ab,t}|$ is the sum of times there is a click difference in same partition
  - This concentration bound is better to use N_{ab} instead of the number of samples

Online Learning to Rank with Features (LLS, ICML'19)

- Each item a is represented by a feature vector $x_a \in \mathbb{R}^d$
- The attractiveness of item a is $\alpha(a) = \theta^\top x_a$
- The concentration bound

$$\mathbb{E}[X_{ab} | X_{ab} \neq 0] \geq \frac{\alpha(a) - \alpha(b)}{\alpha(a) + \alpha(b)}$$

  can't be transferred to a concentration bound for θ
- RecurRank (Recursive Ranking)



- Li, S., Lattimore, T., & Szepesvari, C. Online Learning to Rank with Features. ICML, 2019.

Bandit Works for OLTR with Click Models

	Context	Click Model	Regret
[KSWA, ICML'2015]	-	CM	$\Theta\left(\frac{L}{\Delta}\log T\right)$
[LWZC, ICML'2016] [ZNSKWK, UAI'2016] [LZ, AAAI'2018]	GL	CM	$O\left(d\sqrt{TK}\log T\right)$
[KKSW, ICML'2016]	-	DCM	$O\left(\frac{L}{\Delta}\log T\right)$
[LLZ, COCOON'2018]	GL	DCM	$O\left(dK\sqrt{TK}\log T\right)$
[LVC, NeurIPS'2016]	-	PBM with known β	$O\left(\frac{L}{\Delta}\log T\right)$
[ZTGKSW, ICML'2017] [LKLS, NeurIPS'2018]		General	$O\left(\frac{LK}{\Delta}\log T\right)$ $O\left(\sqrt{K^3LT\log T}\right)$ $\Omega\left(\sqrt{LKT}\right)$
[LLS,ICML'2019]	Linear	General	$O\left(K\sqrt{dT\log(nT)}\right)$

Best-of-both-worlds

Adversarial MAB

- There are n arms
  - An adversary secretly preselects all loss vectors $\{l_{t,a}\}_{t,a}$ from $[0,1]$
  - The best arm is $a^* = \text{argmin} \sum_{t=1}^{T} l_{t,a}$

Setting of Adversarial MAB

- At each time t
  - The learning agent selects one arm a_t
  - Observe the loss l_{t,a_t}
- Objective:
  - Minimize the expected cumulative loss in T rounds $\mathbb{E}\left[\sum_{t=1}^{T} l_{t,a_t}\right]$
  - Minimize the regret in T rounds

$$R(T) = \mathbb{E}\left[\sum_{t=1}^{T} l_{t,a_t}\right] - \min_a \sum_{t=1}^{T} l_{t,a}$$

- Balance the trade-off between exploration and exploitation
  - Exploitation: Select arms that yield good results so far
  - Exploration: Select arms that have not been tried much before

Exp3: Exponential Weight Algorithm for Exploration and Exploitation

- Importance-weight estimator

$$\hat{l}_{t,i} = \frac{\mathbb{I}\{a_t = i\} \cdot l_{t,a_t}}{\mathbb{P}(a_t = i)}$$

- For each time t
  - Calculate the sampling distribution

Learning rate

Exponential weighting

$$\mathbb{P}(a_t = i) = \frac{\exp(-\eta \hat{L}_{t-1,i})}{\sum_{j=1}^{n} \exp(-\eta \hat{L}_{t-1,j})}$$

  - Sample $a_t \sim \mathbb{P}(a_t = i)$ and observe l_{t,a_t}
  - Calculate $\hat{L}_{t,i} = \sum_{s=1}^{t} \hat{l}_{t,i}$
- Regret bound $O(\sqrt{LT \log L})$

- Auer, P., Cesa-Bianchi, N., Freund, Y., & Schapire, R. E. The nonstochastic multiarmed bandit problem. SIAM journal on computing, 2002.

Comparison between Stochastic and Adversarial Environments

- Stochastic
- Reward fixed distribution on $[0,1]$ with fixed mean
- Best arm $a^* = \mathrm{argmax}\, \alpha(a)$
- Regret bound $O(\log T)$
- Runs in adversarial setting
  - Regret may not even converge

- Adversarial
- Loss arbitrary on $[0,1]$
- Best arm $a^* = \mathrm{argmin}\, \sum_{t=1}^{T} l_{t,a}$
- Regret bound $O(\sqrt{T})$
- Runs in stochastic setting
  - Regret bound $O(\sqrt{T})$

Can we design algorithms that achieve $O(\log T)$ regret if run in stochastic setting and $O(\sqrt{T})$ if run in adversarial setting?

Best of Both Worlds

Best-of-both-worlds in OLTR

Adversarial setting under PBM

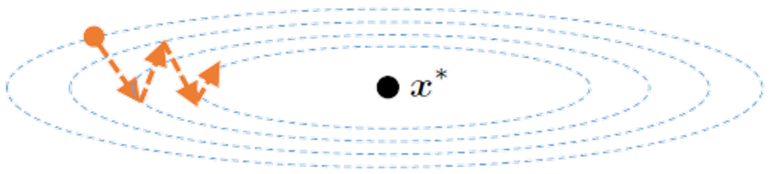
- The adversary secretly preselects the loss vectors $l_{t,i,j} \in \{0,1\}$ for any round t, item i at position j
- The action set can be rewritten as

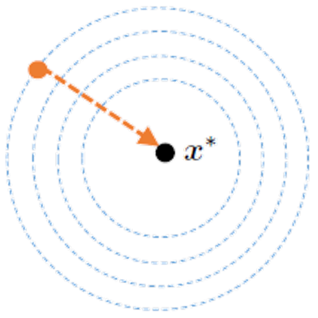
$$\chi = \left\{X \in \{0,1\}^{L\times K}: \sum_{i=1}^{L} X_{i,j} = 1, \forall j \in [K]; \sum_{j=1}^{K} X_{i,j} \leq 1, \forall i \in [L]\right\}$$

- Objective: Minimize the regret over T rounds

$$R(T) = \mathbb{E}\left[\sum_{t=1}^{T} \langle X_t, l_t\rangle - \min_{X\in\chi} \sum_{t=1}^{T} \langle X, l_t\rangle\right]$$

Algorithm: Follow the Regularized Leader (FTRL)





- Input χ
- $\hat{L}_0 = 0_{L \times K}, \eta_t = 1/(2\sqrt{t})$
- For $t = 1,2, \ldots$
  - Compute $x_t = \arg\min_{x \in \mathrm{Conv}(\chi)} \langle x, \hat{L}_{t-1} \rangle + \eta_t^{-1} \Psi(x)$
  - Sample $X_t \sim P(x_t)$
  - Compute the loss estimator $\hat{l}_{t,i,j} = \frac{\mathbb{I}\{X_{t,i,j}=1\} \cdot l_{t,i,j}}{x_{t,i,j}}$
  - Compute $\hat{L}_t = \hat{L}_{t-1} + \hat{l}_t$

Potential Function

Proof idea in the adversarial setting

- $\Psi(x) = \sum_i -\sqrt{x_i}$ for $x \in [0,1]^L$ ½-Tsallis entropy
- Let $\Phi_t(\cdot) = \max_{x \in \mathrm{Conv}(\chi)} \langle x, \cdot \rangle - \eta_t^{-1} \Psi(x)$ Fenchel conjugate

$$R(T) = \underbrace{\mathbb{E}\left[\sum_{t=1}^{T} \langle X_t, l_t \rangle + \Phi_t(-\hat{L}_t) - \Phi_t(-\hat{L}_{t-1})\right]}_{\text{Stability Term}}$$

$$+ \underbrace{\mathbb{E}\left[-\Phi_t(-\hat{L}_t) + \Phi_t(-\hat{L}_{t-1}) - \min_{X \in \chi} \sum_{t=1}^{T} \langle X, l_t \rangle\right]}_{\text{Regularization Penalty Term}}$$

Proof idea in the adversarial setting 2

- $R_{stab} \leq \sum_{t=4}^{T} \left[2\eta_t \sum_{j=1}^{K} \sum_{i \neq I_j^*} \left(\sqrt{\mathbb{E}[x_{t,i,j}]} + \mathbb{E}[x_{t,i,j}] \right) \right] + O(K \log T)$
- $R_{pen} \leq \sum_{t=1}^{T} \sum_{j=1}^{K} \sum_{i \neq I_j^*} \frac{1}{\sqrt{t}} \left(2\sqrt{\mathbb{E}[x_{t,i,j}]} - \mathbb{E}[x_{t,i,j}] \right)$
- By Cauchy-Schwartz Theorem

$$R(T) = O(K\sqrt{LT})$$

Proof idea in the stochastic setting

- In the stochastic case, $l_{t,i,j} \sim \mathrm{Ber}(1 - \alpha_i \beta_j)$
- Define gap for PBM as

$$\Delta_{i,j} = \begin{cases} (\beta_j - \beta_{j+1})(\alpha_j - \alpha_i), & j < i \\ 0, & j = i \\ (\beta_{j-1} - \beta_j)(\alpha_i - \alpha_j), & j > i \end{cases}$$

  the minimal regret incurred as putting item i at position j
- $R(T) \geq \sum_{t=1}^{T} \sum_{i=1}^{L} \sum_{j=1}^{K} \frac{1}{2} \Delta_{i,j}\, \mathbb{E}\left[x_{t,i,j}\right]$ self-bounding constraint
- $R(T) = O\left(\sum_{i \neq j} \frac{\log T}{\Delta_{i,j}}\right)$
- Also provide a lower bound $R(T) = \Omega(K\sqrt{LT})$

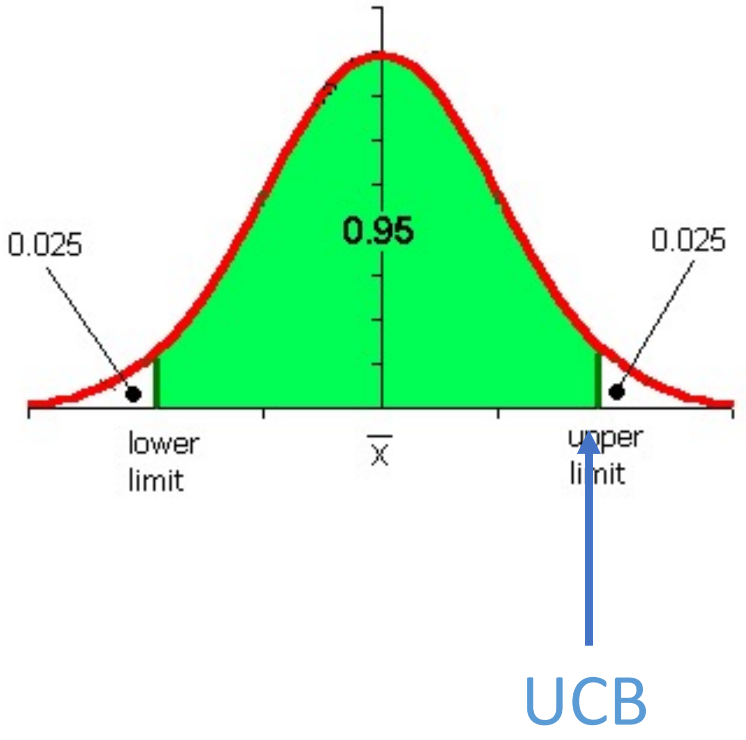
Contributions to Existing Works

	Context	Click Model	Regret	Adversarial
[KSWA, ICML'2015]	-	CM	$\Theta\left(\frac{L}{\Delta}\log T\right)$	
[LWZC, ICML'2016] [ZNSKWK, UAI'2016] [LZ, AAAI'2018]	GL	CM	$O\left(d\sqrt{TK}\log T\right)$	
[KKSW, ICML'2016]	-	DCM	$O\left(\frac{L}{\Delta}\log T\right)$	
[LLZ, COCOON'2018]	GL	DCM	$O\left(dK\sqrt{TK}\log T\right)$	
[LVC, NeurIPS'2016]	-	PBM with known β	$O\left(\frac{L}{\Delta}\log T\right)$	
[ZTGKSW, ICML'2017] [LKLS, NeurIPS'2018]		General	$O\left(\frac{LK}{\Delta}\log T\right)$ $O\left(\sqrt{K^3LT\log T}\right)$ $\Omega\left(\sqrt{LKT}\right)$	
[LLS,ICML'2019]		General	$O\left(K\sqrt{dT\log(nT)}\right)$	
In submission		PBM	$O\left(\frac{KL}{\delta_\beta\Delta}\log T\right)$ $\Omega\left(K\sqrt{LT}\right)$	$O\left(K\sqrt{LT}\right)$

Other Related Works

Thompson Sampling Algorithms

- Cascade Model:
  - Cheung, W. C., Tan, V., & Zhong, Z. A Thompson sampling algorithm for cascading bandits. AISTATS, 2019.
- Position-based Model:
  - Gauthier, C. S., Gaudel, R., & Fromont, E. Position-Based Multiple-Play Bandits with Thompson Sampling. arXiv preprint arXiv:2009.13181.
- Dependent Click Model:
  - In submission



Other metrics

- Safety
  - Li, C., Kveton, B., Lattimore, T., Markov, I., de Rijke, M., Szepesvári, C., & Zoghi, M. BubbleRank: Safe online learning to re-rank via implicit click feedback. UAI, 2020.
- Diversity
  - Hiranandani, G., Singh, H., Gupta, P., Burhanuddin, I. A., Wen, Z., & Kveton, B. Cascading linear submodular bandits: Accounting for position bias and diversity in online learning to rank. UAI, 2020.
- Differential privacy
  - Wang, K., Dong, J., Wang, B., Li, S., & Shao, S. (2021). Cascading Bandit under Differential Privacy. arXiv preprint arXiv:2105.11126.

Click Models

- Imitation learning
  - Dai, X., Lin, J., Zhang, W., Li, S., Liu, W., Tang, R., ... & Yu, Y. (2021, April). An Adversarial Imitation Click Model for Information Retrieval. WWW, 2021.
- Graph NN
  - Lin, J., Liu, W., Dai, X., Zhang, W., Li, S., Tang, R., ... &Yu, Y. (2021). A Graph-Enhanced Click Model for Web Search. SIGIR 2021.

Summary: OLTR

- Stochastic
  - UCB: CM, DCM, PBM, General (matched upper/lower bounds)
  - TS: CM, PBM
- Adversarial
  - PBM
- Best-of-both-worlds
  - PBM (not tight in the stochastic case)

Future Directions

- Stochastic: Thompson sampling algorithm for generalized click model
- Adversarial: Cascade click model
- Adversarial: General click model
- Best-of-both-worlds: General click model
- Corruption & attack

You are welcome to contact me if you are interested in any of these topics.

Other Research Projects

- Online influence maximization
  - Analysis for Thompson sampling
- Online matching markets
  - Many-to-one matching
  - Thompson sampling algorithms
- Best-of-both-worlds
  - Online learning with graph feedback
  - Multi-agent with communication graph
- Online clustering of bandits
- Conversation aided recommendations
  - The application of bandit/RL algorithms

Thanks!
&
Questions?

Shuai Li

- Assistant Professor at John Hopcroft Center Shanghai Jiao Tong University
- Research interests: Bandit/RL algorithms
- Personal website: http://shuaili8.github.io/