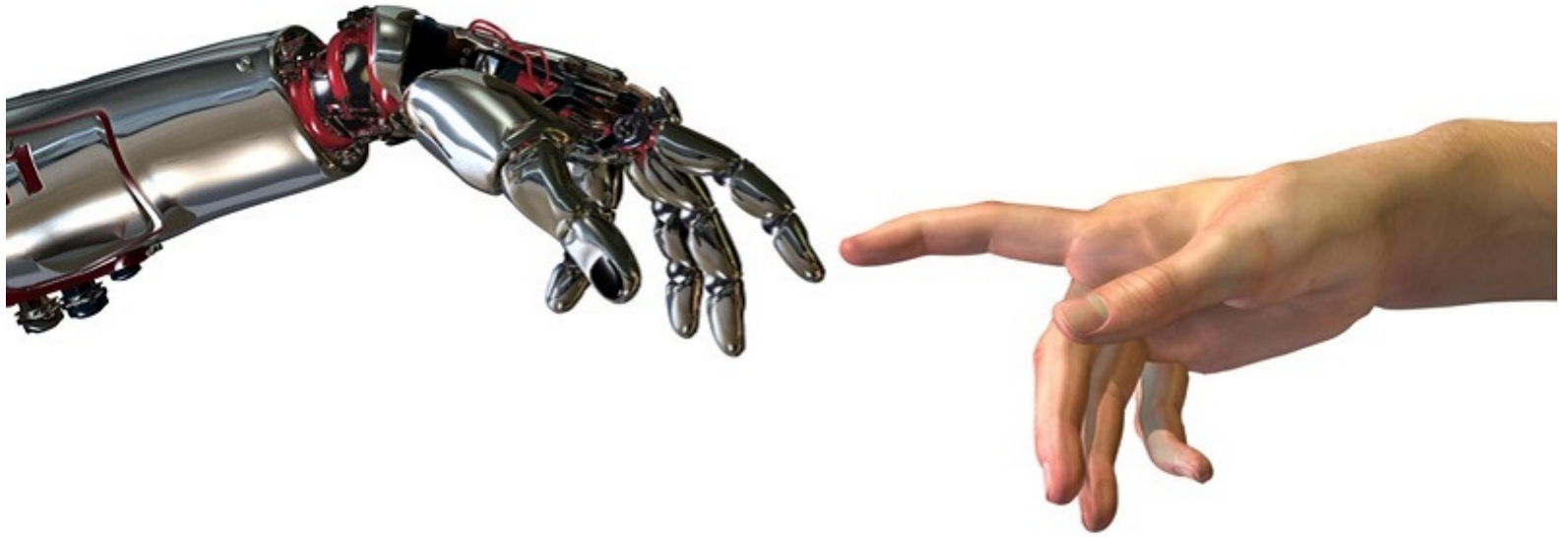


Interactive Information Retrieval with Bandit Feedback

Huazheng Wang, Yiling Jia, Hongning Wang

Department of Computer Science

University of Virginia

Outline of this tutorial

- Motivation
- Background: bandit algorithms
- Bandit learning for recommender systems
- Bandit learning for retrieval systems
- Ethical considerations in IR with bandit learning
- Conclusion & future directions

Information retrieval is everywhere

- A predominant interface between users and massive amount of information indexed in modern online systems

Check media feeds

Listen to music and podcasts

Attend to emails

Find a restaurant

Online shopping

Search online

Apply for jobs

Take courses

Goal of information retrieval

- Satisfy users' information need

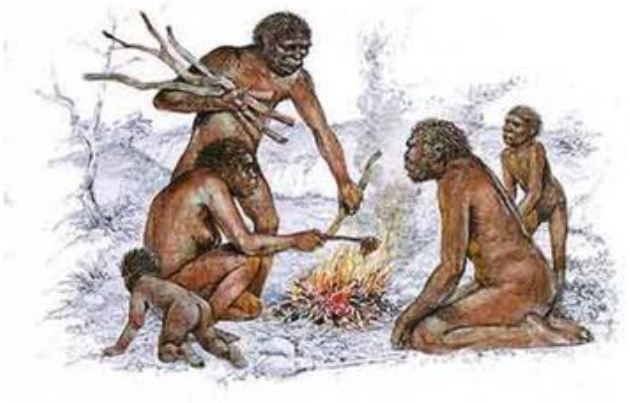
Result ranking is essential

- Probability ranking principle [Rob77] – a theoretical justification
  - The overall utility of the system to its users is maximized when the results are ranked in a descending order of usefulness to the users

- Sequential examination
- Independent relevance evaluation

How do we estimate utility?

- Classical IR methods
  - Document retrieval
    - BM25, language models, page rank
  - Recommendation
    - Content-based recommendation, collaborative filtering
  - Unsupervised, and rely on empirical parameter tuning



How do we estimate utility?

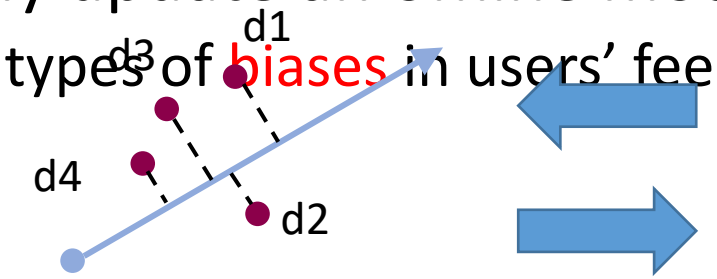
- Learning-based methods
  - Document retrieval
    - Learning to rank
  - Recommendation
    - Latent factor models, neural network models
  - Supervised, hungry for labeled training data!

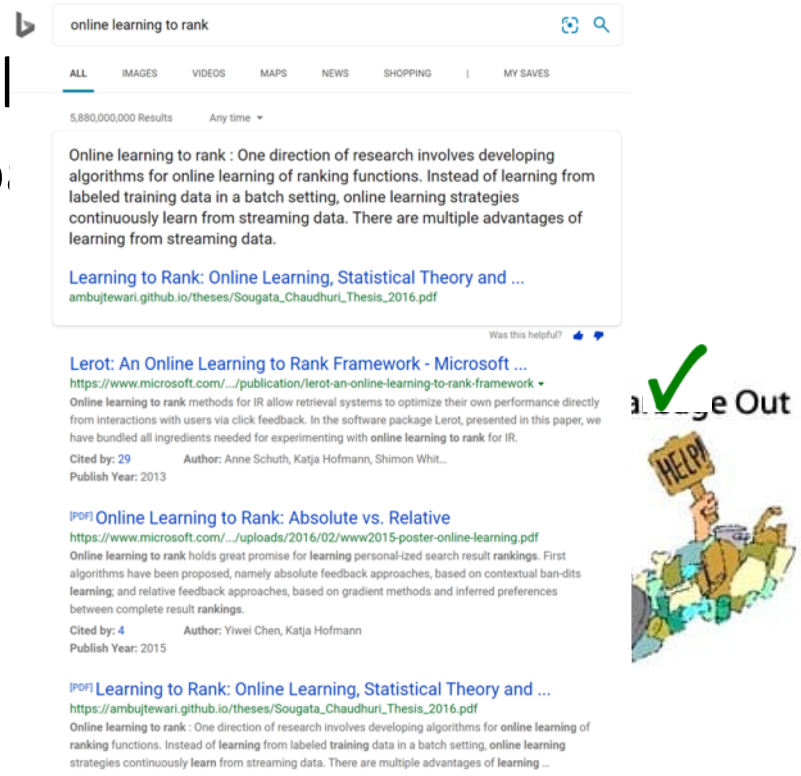
Learning from users' implicit feedback

- User behavior oriented result feedback
  - Low cost
  - Large scale
  - Natural usage context and utility

Interactive information retrieval

- Learning by interacting with users
  - Eliminates offline methods' heavy dependency on manual relevance annotations
  - **NOT** simply update an offline model
    - Various types of biases in users' feedb





Limitations in users' implicit feedback

- In a recommender system
  - Presentation bias

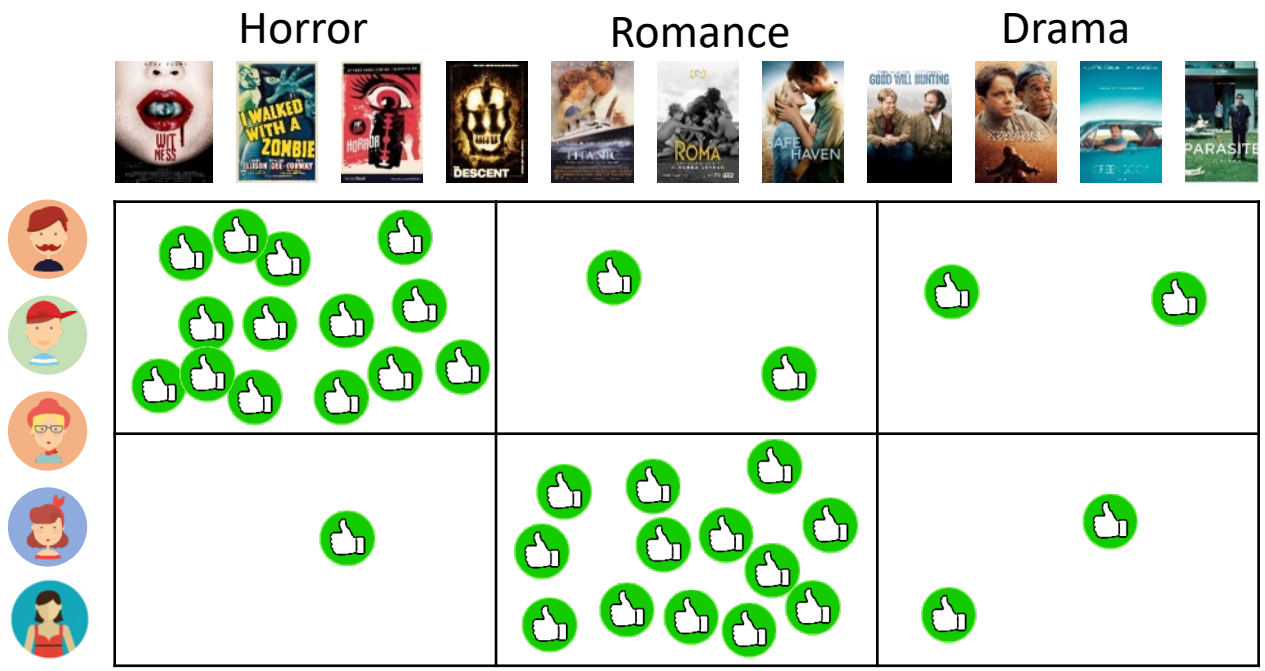


Figure credit: Schnabel et al. 2016 [SSSCJ16]

Matthew effect: we still don't know what we don't know!

Limitations in users' implicit feedback

- In a retrieval system
  - Presentation bias
  - Position bias
  - Trust bias

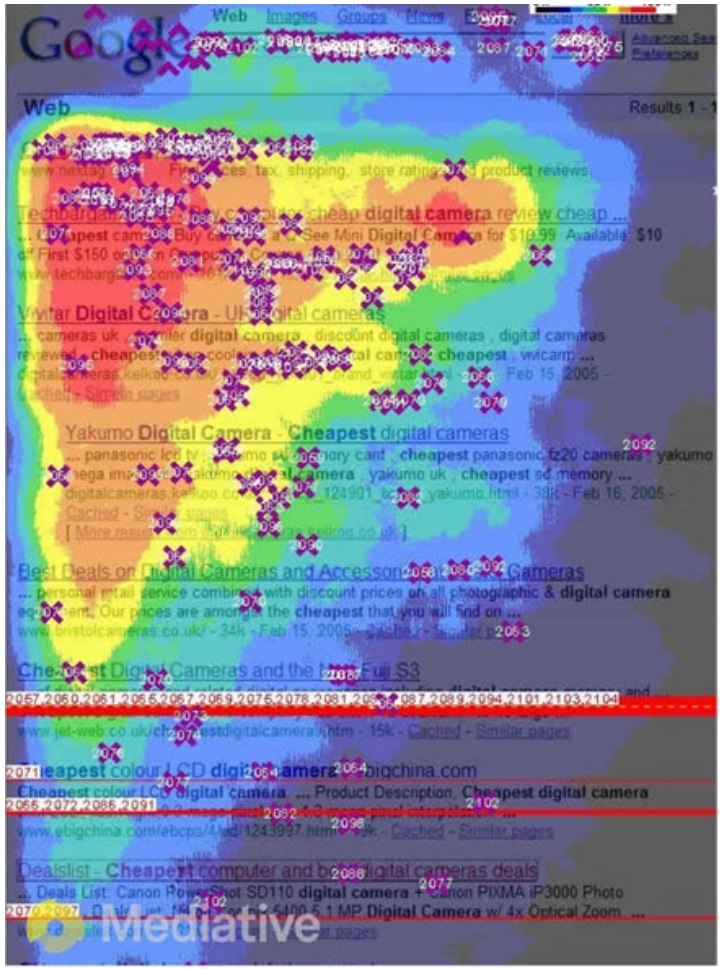


Figure credit: Baeza-Yates 2018 [Bae18]

Interactive information retrieval

- Learning by interacting with users
  - Eliminates offline methods' heavy dependency on manual relevance annotations
  - **NOT** simply update an offline model online
    - Various types of biases in users' feedback
    - Learning while serving the users

Exploitation

Present the best results estimated so far to satisfy users

Interactive IR System

Exploration

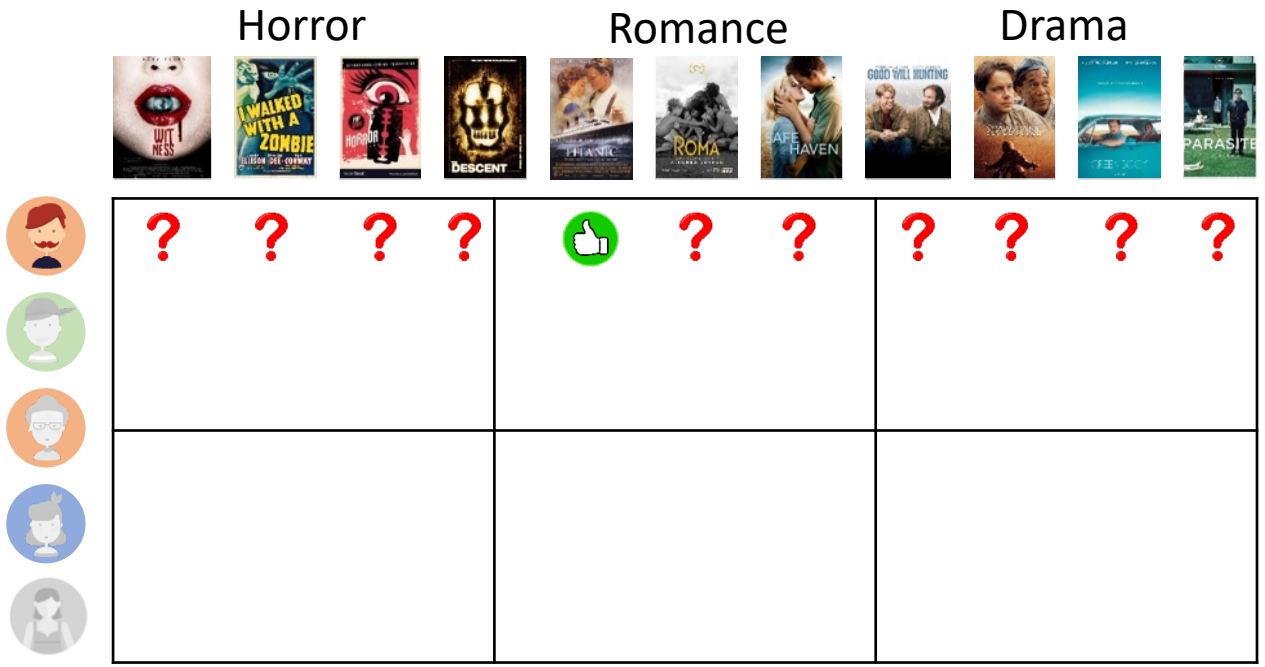
Present currently underestimated results to best improve the ranker

Efficient exploration is crucial

- Huge search space
  - Exploration is costly

Efficient exploration is crucial

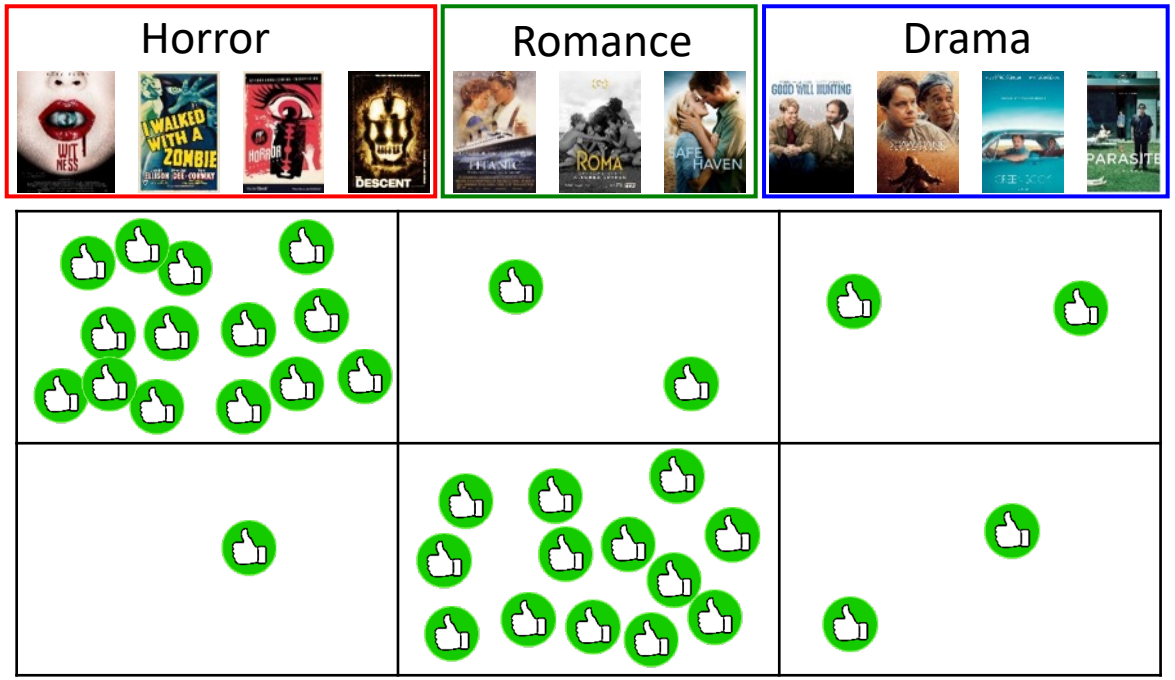
- One only gets answers to the questions she asked
  - Bandit feedback



Efficient exploration is crucial

- The problem space can be structured
  - The observations might not be independent

Clustered, correlated responses?



Social influence?

Figure credit: Schnabel et al. 2016 [SSSCJ16]

Efficient exploration is crucial

- The problem space is evolving over time
  - Any real-world environment is non-stationary
  - Recognize outdated model is important

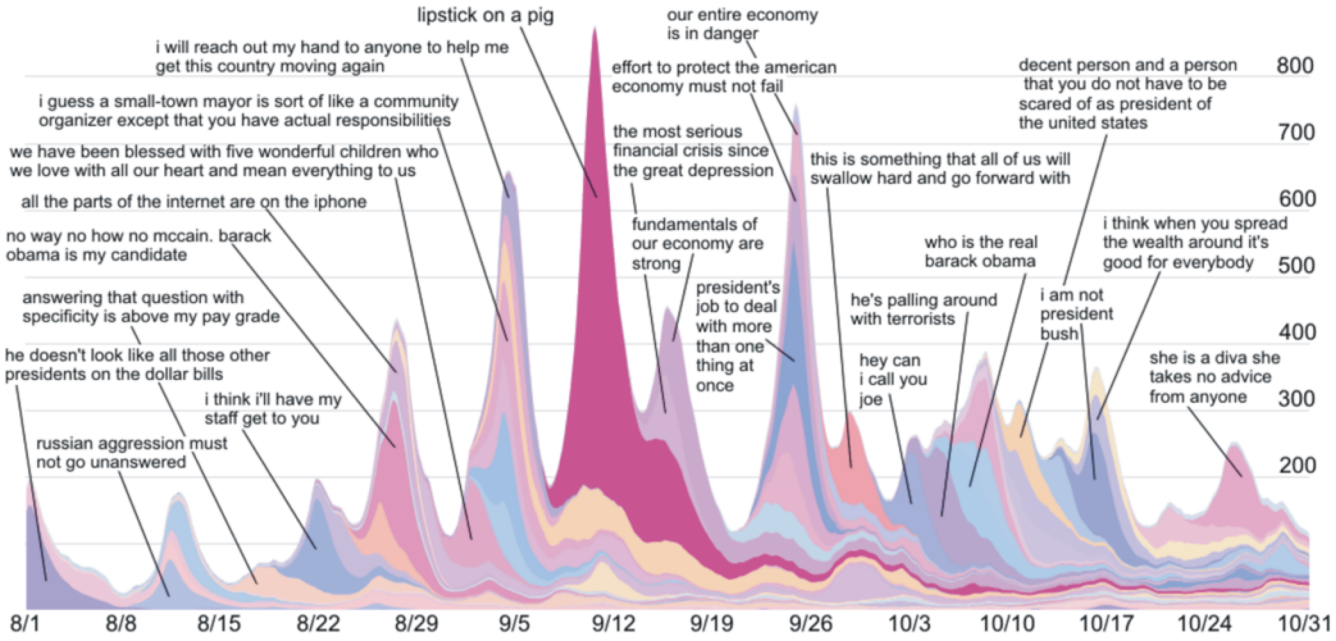
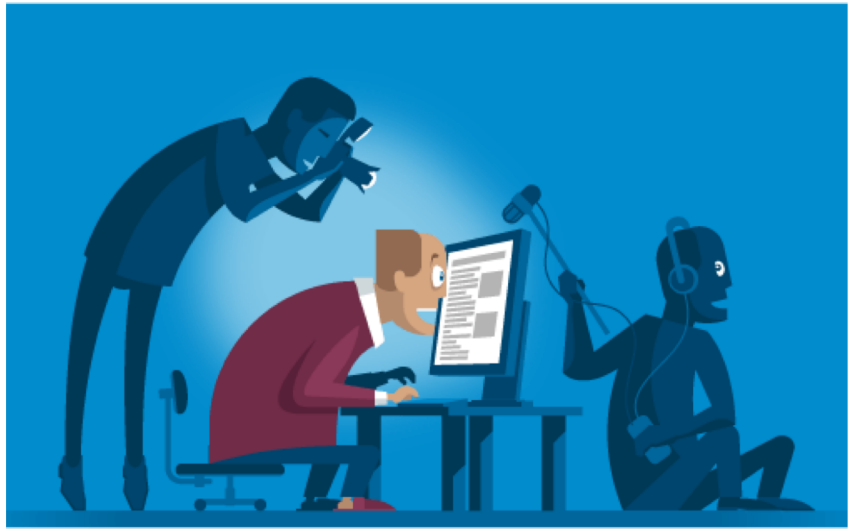
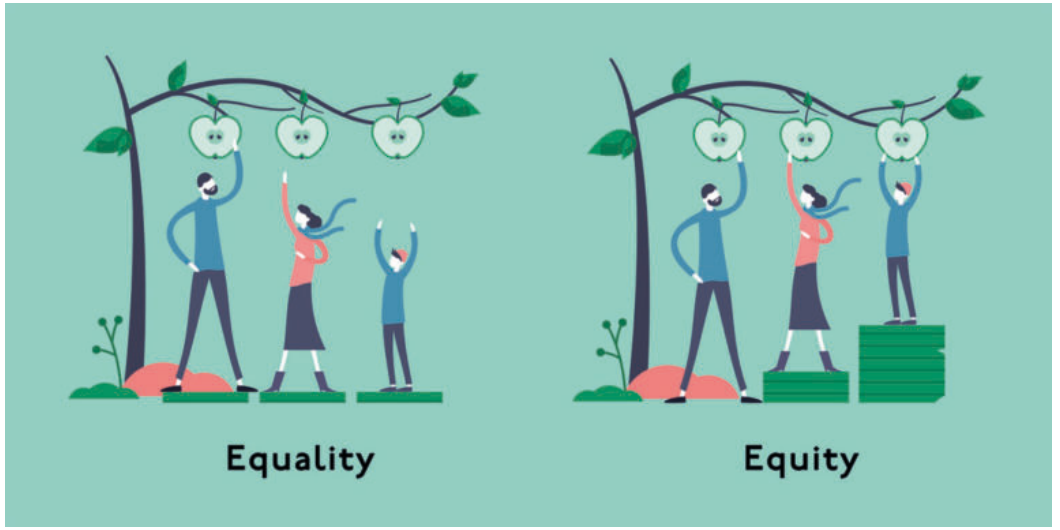


Figure credit: Leskovec et al. 2016 [LBK09]

Exploration also induces risks

- There are also privacy, fairness, and ethic concerns in exploration





Outline of this tutorial

- Motivation
- Background: bandit algorithms
- Bandit learning for recommender systems
- Bandit learning for retrieval systems
- Ethical considerations in IR with bandit learning
- Conclusion & future directions

Problem formulation

- (stochastic) Multi-arm bandit

Reward distribution

$$\mathbb{E}[r_{a,t}] = r_a^*$$

Goal: maximize the accumulated reward over T rounds

Actions

$$\mathcal{A} = \{a_1, a_2, \ldots, a_K\}$$

Agent/learner/ policy

$$\pi_\theta(a|\mathcal{H}_t)$$

Reward

$$r_{a,t}$$

History $\mathcal{H}_t = \{(a_i, r_i)\}_{i=1}^{t-1}$

Map it to an IR problem (e.g., recommender systems):

- Environment: a user or users
- Agent: recommendation algorithm
- Actions/arms: recommendation candidates
- Reward: click, purchase

Problem formulation

- (stochastic) Multi-arm bandit

Reward distribution

$\mathbb{E}[r_{a,t}] = r_a^*$

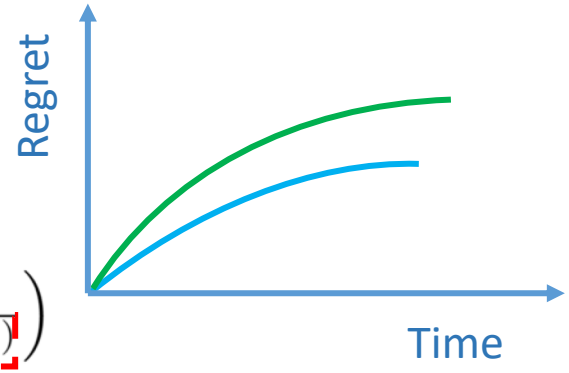
Actions

$\mathcal{A} = \{a_1, a_2, \ldots, a_K\}$

Goal: maximize the accumulated reward over T rounds

(Pseudo) Regret: expected loss due to not playing the best arm

$\mathbf{R}(T) = \sum_{t=1}^{T} (\mathbb{E}[r_{a^*}] - \mathbb{E}[r_{a,t}])$



Agent/learner/ policy

$\pi_\theta(a|\mathcal{H}_t)$

Reward

$r_{a,t}$

History $\mathcal{H}_t = \{(a_i, r_i)\}_{i=1}^{t-1}$

Lower regret bound [ARK85]:

$\mathbf{R}(T) = \Omega\left(\gamma\sqrt{TK}\right)$ $\frac{\log T}{KL(p_{a^*}, p_{a_i})}$

Finite variance is required!

Difficulty of the problem

Sub-linear regret is preferred!

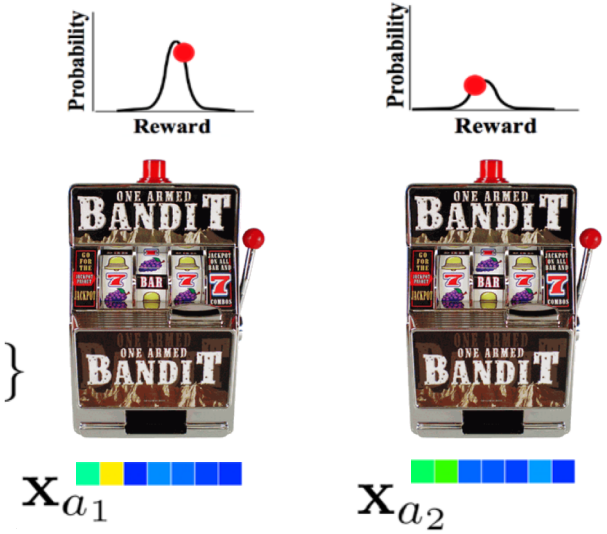
Problem formulation

In linear contextual bandit [LCLS10],
$r_{a,t} \sim N(\mathbf{x}_a^\mathsf{T}\boldsymbol{\theta}, \sigma^2),\ \theta \in \mathbb{R}^d$

- Contextual/structured bandit

Reward distribution
$\mathbb{E}[r_{a,t}] = f_{\boldsymbol{\theta}}(\mathbf{x}_a)$

Actions
$\mathcal{A}_t = \{\mathbf{x}_{a_i,t} \in \mathbb{R}^d\}_{i=1}^K\}$



Agent/learner/ policy
$\pi_\theta(a|\mathcal{A}_t, \mathcal{H}_t)$

Reward
$r_{a,t}$

History $\mathcal{H}_t = \{(\mathcal{A}_i, a_i, r_i)\}_{i=1}^{t-1}$

Goal: maximize the accumulated reward over T rounds

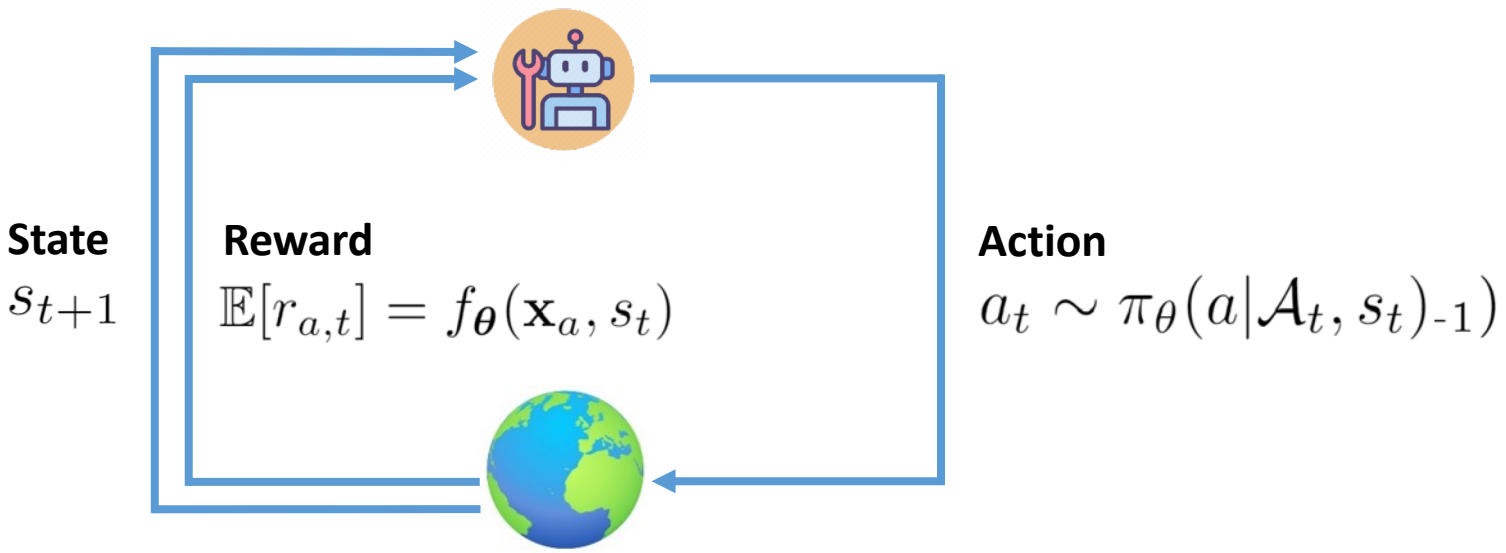
(Pseudo) Regret: expected loss due to not playing the best arm
$\mathbf{R}(T) = \sum_{t=1}^T (\mathbb{E}[r_{a^*}] - \mathbb{E}[r_{a,t}])$

Lower regret bound in linear contextual bandit [CLRS11]:
$$\mathbf{R}(T) = \Omega\left(\gamma\sqrt{Td}\right), \text{when } d \leq \sqrt{T}$$

Problem formulation

⚠ ***Out of scope of this tutorial***

- Reinforcement learning [SB18]



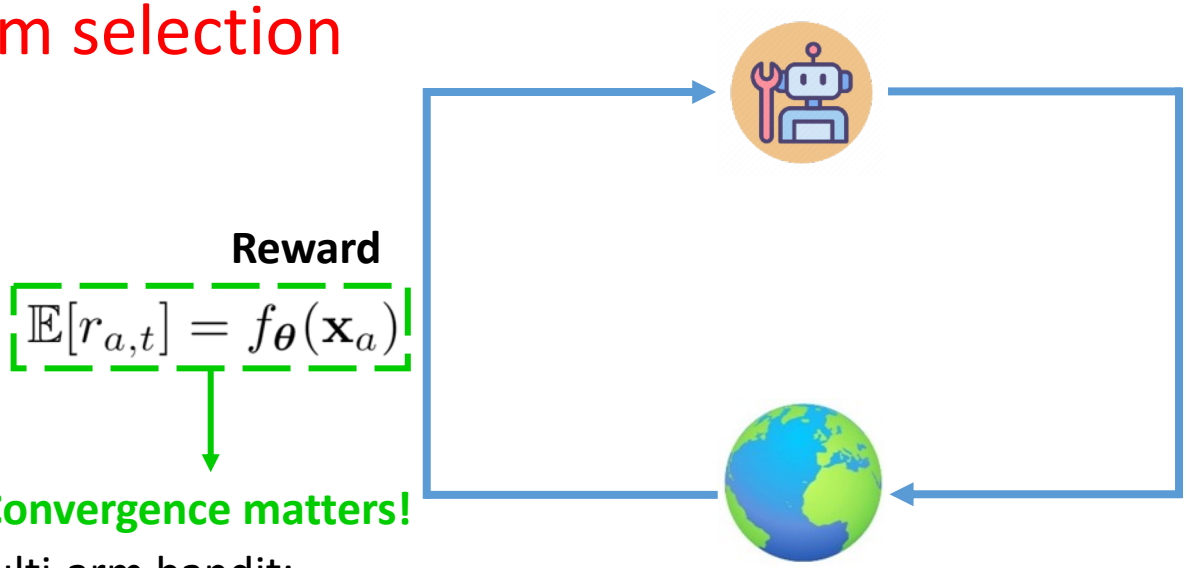
Problem formulation

In linear contextual bandit [LCLS10],

$\hat{\theta}_t = \arg\min_{\theta \in \mathbf{\Theta}} \sum_{(a_i, r_i) \in \mathcal{H}_t} (\mathbf{x}_{a_i}^\mathsf{T} \theta - r_i)^2 + \lambda \theta^\mathsf{T} \theta$

Ridge regression, closed form solution exists!

- Key problems
  - Reward estimation
  - Arm selection



Reward

$\mathbb{E}[r_{a,t}] = f_{\boldsymbol{\theta}}(\mathbf{x}_a)$

Convergence matters!

Action

$a_t \sim \pi_\theta(a | \mathcal{A}_t, \mathcal{H}_{t-1})$

Exploration matters!

1. Multi-arm bandit:
   $\hat{r}_{k,t} = \frac{\sum_{(a_i, r_i) \in \mathcal{H}_t} \mathbb{1}\{a_i = k\} r_i}{n_t^k}$
2. Contextual bandit:
   $\hat{\theta}_t = \arg\min_{\theta \in \mathbf{\Theta}} \mathcal{L}(\mathcal{H}_t, \theta) + \mathcal{R}(\theta)$

Loss function Regularization

1. Adaptive v.s., non-adaptive
2. Independent v.s., collaborative
3. Unconstrained v.s., constrained

Problem formulation

- Map to IR problems
  - Reward estimation
    - Document retrieval: document relevance under a given query
    - Recommendation: item utility for a given user
  - Arm selection
    - Document retrieval: ranked list of documents (i.e., top-k ranking)
    - Recommendation: the item of choice (i.e., top-1 ranking)

Classical bandit learning algorithms

- Random exploration
- Optimism in the face of uncertainty
- Posterior sampling

Random exploration

Reward estimation in MAB:

$$\hat{r}_{k,t} = \frac{\sum_{(a_i,r_i)\in\mathcal{H}_t} \mathbb{1}\{a_i=k\} r_i}{n_t^k} \frac{}{p_{a_i})}$$

- ϵ-greedy [Aue02]

Constant ϵ_t leads to linear regret!

By setting $\epsilon_t = \min\left(1, \frac{cK}{t\Delta_{\min}^2}\right)$, with $c > 0$, we have $\mathbf{R}(T) \leq C \sum_{i=1}^{K} \left(\Delta_i + \frac{\Delta_i}{\Delta_{\min}^2} \log \max\left\{e, \frac{T\Delta_{\min}^2}{K}\right\}\right)$

$c_t = \text{Bernoulli}(\epsilon_t)$

$c_t = 1$ $\quad$ $c_t = 0$

Arm selection

$a_t = \arg\max_{a\in\mathcal{A}} \hat{r}_{a,t}$ $\quad$ Randomly choose a_t from $\mathcal{A}$

Observe reward r_t and update $\mathcal{H}_t$ by (a_t, r_t)

Non-adaptive; often leads to sub-optimal performance in practice.

Optimism in the face of uncertainty

- UCB1 [Aue02]
  - Upper confidence bound based

Chernoff-Hoeffding bound

Lower regret bound of MAB:

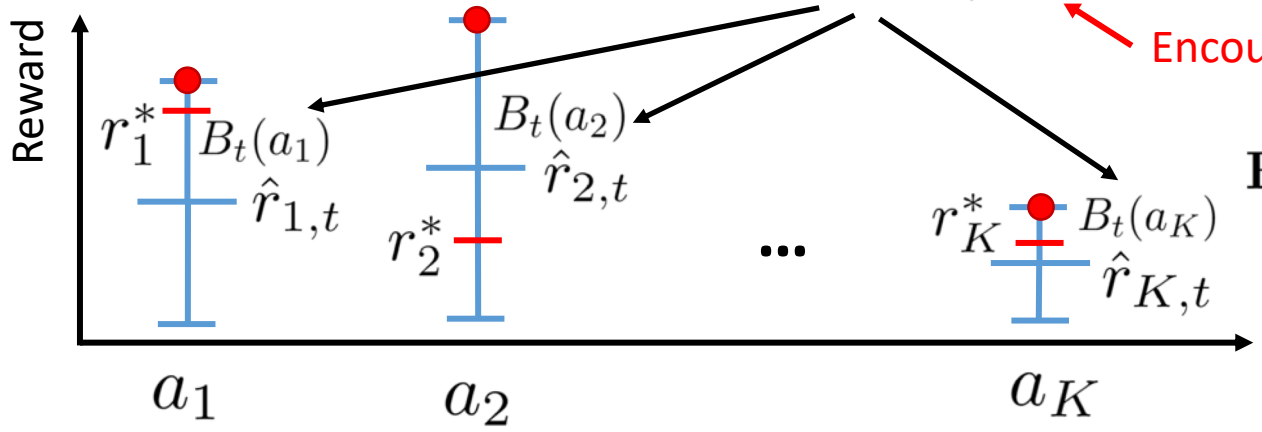
$$\mathbf{R}(T) = \Omega\left(\gamma\sqrt{TK}\right)\frac{1}{p_{a^*}, p_{a_i}} a^2/n)$$

where $\mathfrak{s}_n = \sum_{i=1} x_i$

$$\mathbb{E}[x_n | x_1, \ldots, x_{n-1}] = \mu$$

$$B_t(a_k) = \sqrt{\frac{2\log t}{n_t^k}}$$

Encourage underestimated arms



$$\mathbf{R}(T) = O(\sqrt{KT\log T})(1+\tfrac{\pi^2}{3})\sum_{i=1}^{K}\Delta_i$$

$$a_t = \arg\max_{a\in\mathcal{A}} \hat{r}_{a,t} + B_t(a)$$

Optimism in the face of uncertainty

- LinUCB [LCLS10]
  - Upper confidence bound based

Lower regret bound in linear contextual bandit [CLRS11]: $\mathbf{R}(T) = \Omega\left(\gamma\sqrt{Td}\right)$, when $d \leq \sqrt{T}$

$$\hat{r}_{a,t} = \mathbf{x}_a^\top \hat{\boldsymbol{\theta}}_t$$

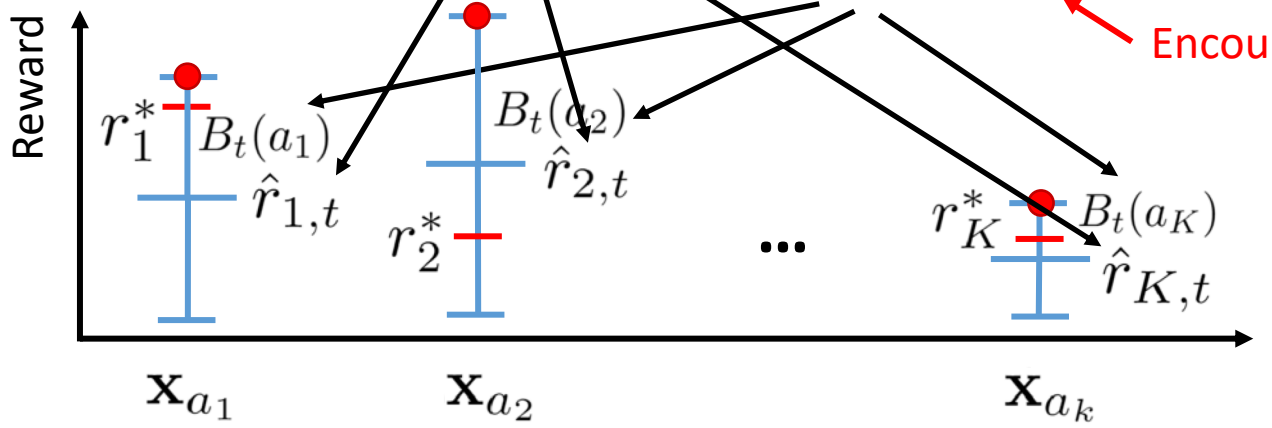
$$B_t(a_k) = \alpha_t ||\mathbf{x}_{a_k}||_{\mathbf{A}_t^{-1}}$$

Confidence ellipsoid

$$\alpha_t = ||\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^*||_{A_t} = O\left(\sqrt{d\log(t/\lambda)}\right)$$

$$\mathbf{R}(T) = O\left(\sqrt{Td}\log(\lambda + T/d)\right)$$

Encourage underestimated arms



$$a_t = \arg\max_{a \in \mathcal{A}} \hat{r}_{a,t} + B_t(a)$$

Closed form estimator

$$\mathbf{A}_t = \lambda\mathbf{I} + \sum_{(a_i,r_i)\in\mathcal{H}_t} \mathbf{x}_{a_i}\mathbf{x}_{a_i}^\top$$

$$\mathbf{b}_t = \sum_{(a_i,r_i)\in\mathcal{H}_t} r_{a_i}\mathbf{x}_{a_i}$$

$$\hat{\boldsymbol{\theta}}_t = \mathbf{A}_t^{-1}\mathbf{b}_t$$

Optimism in the face of uncertainty

- GLM-UCB [FCGS10]
  - A generalized linear model

In generalized linear bandit [LCLS10],

$$\mathbb{E}[r_{a,t}] = \mu(\mathbf{x}_a^\top \boldsymbol{\theta}), \boldsymbol{\theta} \in \mathbb{R}^d$$

Link function, e.g., $\mu(y) = 1/(1+e^{-y})$

Confidence ellipsoid

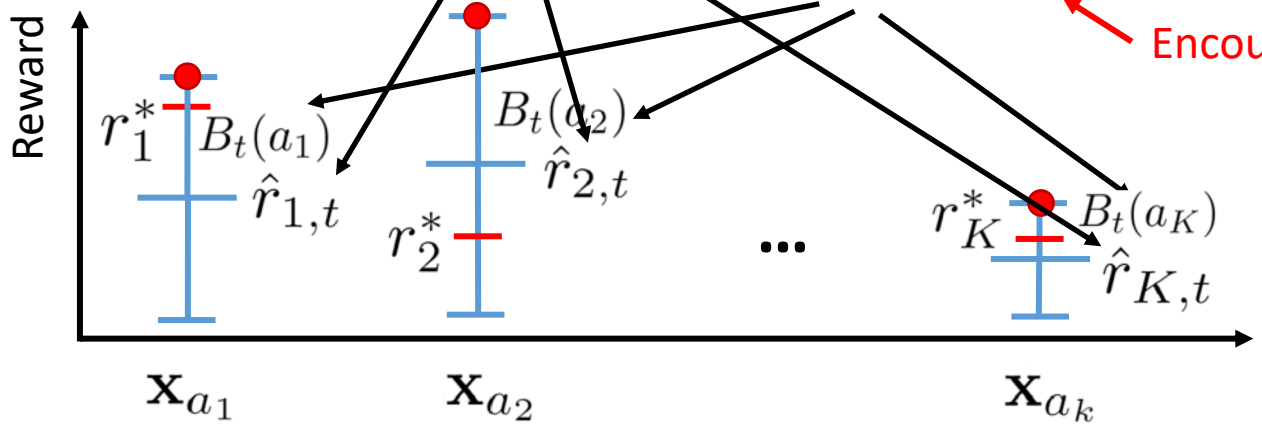
$$\alpha_t = O\left(\sqrt{d\log(t)}\right)$$

$$\mathbf{R}(T) = O\left(\sqrt{T\log(2dT)}\log(T)\right)$$

$$\hat{r}_{a,t} = \mu(\mathbf{x}_a^\top \hat{\boldsymbol{\theta}}_t)$$

$$B_t(a_k) = \alpha_t ||\mathbf{x}_{a_k}||_{\mathbf{A}_t^{-1}}$$

Encourage underestimated arms



$$a_t = \arg\max_{a\in\mathcal{A}} \hat{r}_{a,t} + B_t(a)$$

No closed form solution for $\hat{\boldsymbol{\theta}}_t$

Using gradient descent to solve

$$\arg\min_{\boldsymbol{\theta}\in\Theta} \left\| \sum_{(a_i,r_i)\in\mathcal{H}_t} \mathbf{x}_{a_i}\left(r_i - \mu(\mathbf{x}_{a_i}^\top\boldsymbol{\theta})\right)\right\|_{A_t^{-1}}$$

where $\mathbf{A}_t = \lambda\mathbf{I} + \sum_{(a_i,r_i)\in\mathcal{H}_t} \mathbf{x}_{a_i}\mathbf{x}_{a_i}^\top$

Posterior sampling

In linear contextual bandit [LCLS10], $r_{a,t} \sim N(\mathbf{x}_a^\mathsf{T}\boldsymbol{\theta}, \sigma^2),\ \theta \in \mathbb{R}^d$

Predictive distribution:

$$P(r_{a,t}|\mathbf{x}_a, \mathcal{H}_t, \lambda) = N\left(\mathbf{x}_a^\mathsf{T}\mathbf{A}_t^{-1}\mathbf{b}_t, \tfrac{1}{\sigma^2} + \mathbf{x}_a^\mathsf{T}\mathbf{A}_t^{-1}\mathbf{x}_a\right)$$

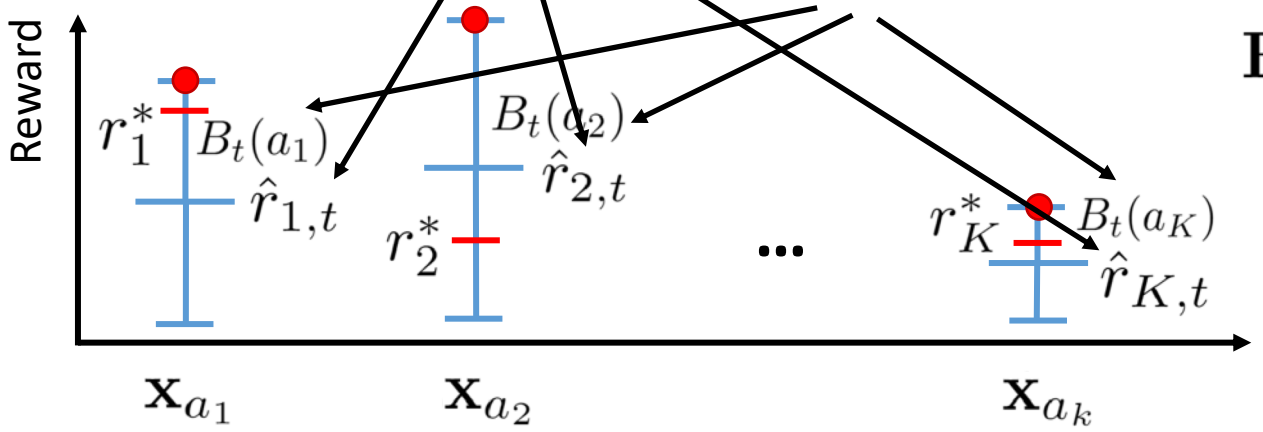
- Thompson sampling [AG13]

Source of uncertainty

A Bayesian perspective of reward estimation: $\hat{\boldsymbol{\theta}}_t \sim P(\boldsymbol{\theta}|\mathcal{H}_t, \lambda)$ prior

$$\hat{r}_{a,t} = \mathbf{x}_a^\mathsf{T}\hat{\boldsymbol{\theta}}_t$$

$$B_t(a_k) = \alpha_t ||\mathbf{x}_{a_k}||_{\mathbf{A}_t^{-1}}$$



$$a_t = \arg\max_{a \in \mathcal{A}} \hat{r}_{a,t}$$

related to prior

$$\mathbf{R}(T) = O\left(d^2\sqrt{T^{1+\epsilon}}\log(Td)\right)$$

Analytic form of posterior

Prior: $P(\boldsymbol{\theta}|\lambda) = N(0, \lambda\mathbf{I})$

$$\mathbf{A}_t = \lambda\mathbf{I} + \tfrac{1}{\sigma^2}\textstyle\sum_{(a_i,r_i)\in\mathcal{H}_t} \mathbf{x}_{a_i}\mathbf{x}_{a_i}^\mathsf{T}$$

$$\mathbf{b}_t = \tfrac{1}{\sigma^2}\textstyle\sum_{(a_i,r_i)\in\mathcal{H}_t} r_{a_i}\mathbf{x}_{a_i}$$

$$\hat{\boldsymbol{\theta}}_t \sim N(\mathbf{A}_t^{-1}\mathbf{b}_t, \mathbf{A}_t)$$

Posterior sampling

⚠ *No formal regret known yet; and some analysis from Bayesian regret perspective*

- Thompson sampling [RVKOW18]
  - No analytic posterior?
  - Approximate posterior inference!
    - Gibbs sampling
    - Particle sampling
    - Laplace approximation
    - Bootstrapping

References I

[Rob77] Robertson, Stephen E. "The probability ranking principle in IR." *Journal of documentation* (1977).

[SSSCJ16] Schnabel, T., Swaminathan, A., Singh, A., Chandak, N., & Joachims, T. (2016, June). Recommendations as treatments: debiasing learning and evaluation. In Proceedings of the 33rd International Conference on International Conference on Machine Learning-Volume 48 (pp. 1670-1679).

[Bae18] Baeza-Yates, R. (2018). Bias on the web. Communications of the ACM, 61(6), 54-61.

[LBK09] Leskovec, J., Backstrom, L., & Kleinberg, J. (2009, June). Meme-tracking and the dynamics of the news cycle. In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 497-506).

[LR85] Lai, T. L., & Robbins, H. (1985). Asymptotically efficient adaptive allocation rules. Advances in applied mathematics, 6(1), 4-22.

[Aue95] Auer, P., Cesa-Bianchi, N., Freund, Y., & Schapire, R. E. (1995, October). Gambling in a rigged casino: The adversarial multi-armed bandit problem. In Proceedings of IEEE 36th Annual Foundations of Computer Science (pp. 322-331). IEEE.

[LCLS10] Li, L., Chu, W., Langford, J., & Schapire, R. E. (2010, April). A contextual-bandit approach to personalized news article recommendation. In Proceedings of the 19th international conference on World wide web (pp. 661-670).

[CLRS11] Chu, W., Li, L., Reyzin, L., & Schapire, R. (2011, June). Contextual bandits with linear payoff functions. In Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics (pp. 208-214).

[SB18] Sutton, R. S., & Barto, A. G. (2018). Reinforcement learning: An introduction. MIT press.

[Aue02] Auer, P. (2002). Using confidence bounds for exploitation-exploration trade-offs. Journal of Machine Learning Research, 3(Nov), 397-422.

References II

[LZ08] Langford, J., & Zhang, T. (2008). The epoch-greedy algorithm for multi-armed bandits with side information. In Advances in neural information processing systems (pp. 817-824).

[FCGS10] Filippi, S., Cappe, O., Garivier, A., & Szepesvári, C. (2010). Parametric bandits: The generalized linear case. In Advances in Neural Information Processing Systems (pp. 586-594).

[GC11] Garivier, A., & Cappé, O. (2011, December). The KL-UCB algorithm for bounded stochastic bandits and beyond. In Proceedings of the 24th annual conference on learning theory (pp. 359-376).

[AG13] Agrawal, S., & Goyal, N. (2013, February). Thompson sampling for contextual bandits with linear payoffs. In International Conference on Machine Learning (pp. 127-135).

[RVKOW18] Russo, D. J., Van Roy, B., Kazerouni, A., Osband, I., & Wen, Z. (2018). A Tutorial on Thompson Sampling. Foundations and Trends® in Machine Learning, 11(1), 1-96.

[KSGB19a] Kveton, B., Szepesvári, C., Ghavamzadeh, M., & Boutilier, C. (2019, August). Perturbed-history exploration in stochastic multi-armed bandits. In Proceedings of the 28th International Joint Conference on Artificial Intelligence (pp. 2786-2793). AAAI Press.

[KSGB19b] Kveton, B., Szepesvari, C., Ghavamzadeh, M., & Boutilier, C. (2019). Perturbed-history exploration in stochastic linear bandits. arXiv preprint arXiv:1903.09132.

[KMRWW18] Kannan, S., Morgenstern, J. H., Roth, A., Waggoner, B., & Wu, Z. S. (2018). A smoothed analysis of the greedy algorithm for the linear contextual bandit problem. In Advances in Neural Information Processing Systems (pp. 2227-2236).

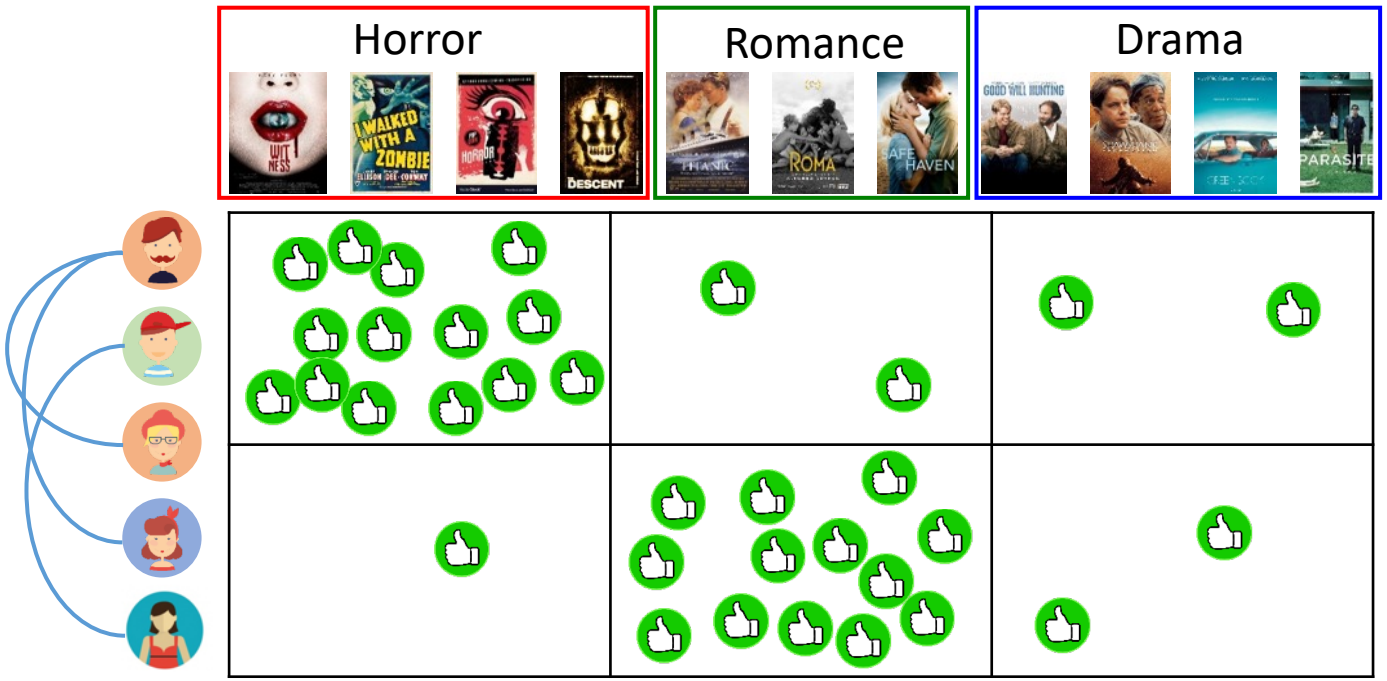
Outline of this tutorial

- Motivation
- Background: bandit algorithms
- Bandit learning for recommender systems
- Bandit learning for retrieval systems
- Ethical considerations in IR with bandit learning
- Conclusion & future directions

Real-world challenges: recap

- Huge exploration space, but the problem space has structure
- User's preference can be non-stationary

Clustered, correlated responses?



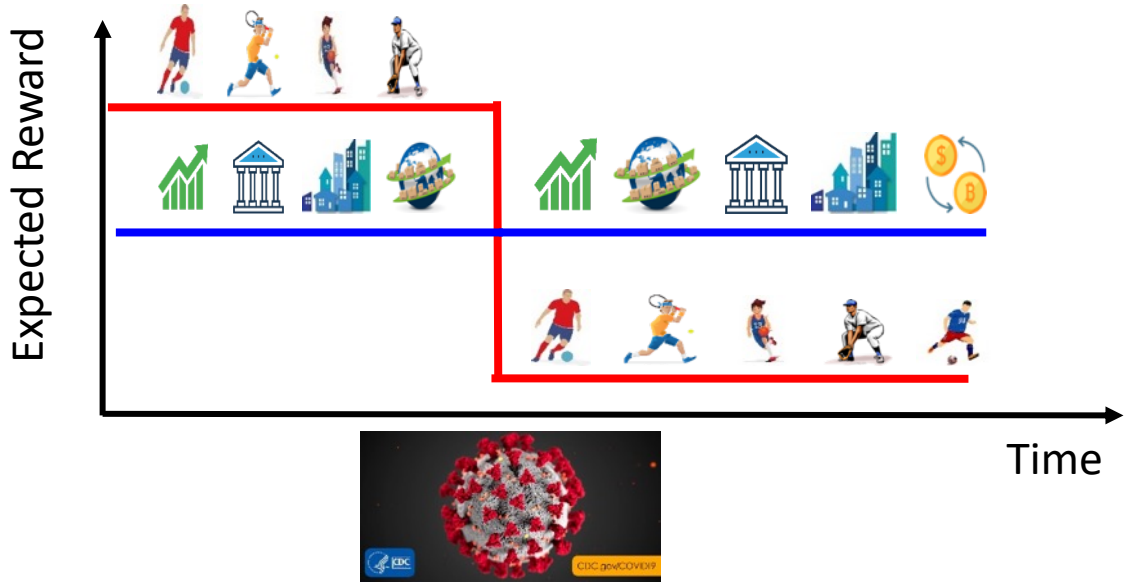


Figure credit: Schnabel et al. 2016 [SSSCJ16]

Social influence?

Bandit learning for recommender systems

- Contextual bandits for recommendation
- Collaborative bandit learning
- Learning in a non-stationary environment

Contextual bandits for recommendation

- Linear bandit formulation:
  - User has an unknown preference $\boldsymbol{\theta}$
  - Each item a is represented by vector x_a
  - Linear reward assumption:

$$r_{a,t} \sim N(\mathbf{x}_a^\mathsf{T}\boldsymbol{\theta}, \sigma^2),\ \theta \in \mathbb{R}^d$$

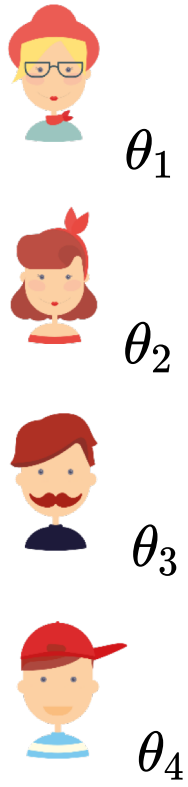
- Classical method: LinUCB [LCLS10]
  - Recap: optimism in the face of uncertainty
- Non-linear reward function
  - Logistic reward [FCGS10]
  - Neural network [ZLG20]

Bandit learning for recommender systems

- Contextual bandits for recommendation
- Collaborative bandit learning
  - With user-dependency structure
  - Online user / item clustering
  - Matrix factorization for low-rank structure
  - Warm-start exploration
- Learning in a non-stationary environment

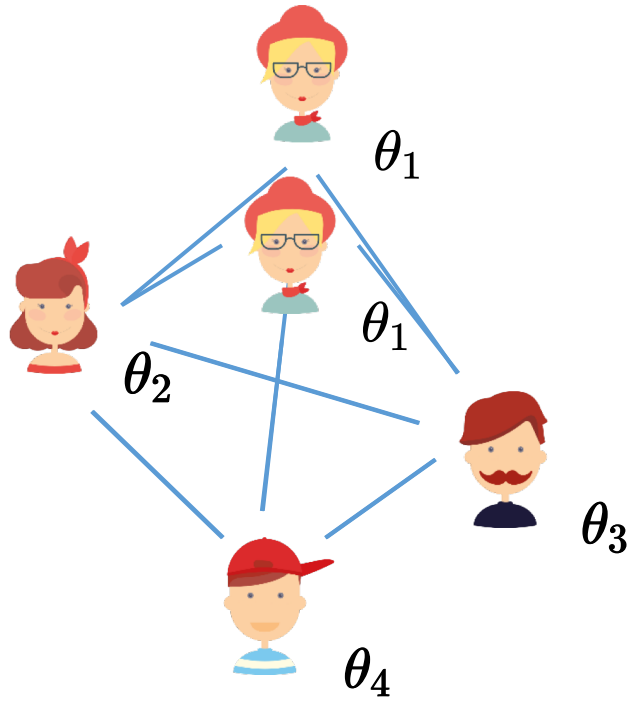
Collaborative bandit learning

- Multi-agent linear bandits: N users, each user has his/her own θ
- Build independent LinUCB for each user?
  - Cold start challenge
  - Users are not independent
- Leverage user dependency for efficient exploration
  - Use existing user dependency information
  - Discover dependency online (via clustering)



Collaborative bandit learning

- GOBLin [CGZ13]
  - Connected users are assumed to share similar model parameters
  - Graph Laplacian based regularization upon ridge regression to model dependency



Graph E is input.
Regularization term:

$$\sum_i \|\boldsymbol{\theta}_i\|_2 + \sum_{(i,j)\in E} \|\boldsymbol{\theta}_i - \boldsymbol{\theta}_j\|_2$$

Collaborative bandit learning

- GOBLin [CGZ13]
  - Graph Laplacian based regularization upon ridge regression to model dependency
  - Encode graph Laplacian in context, formulate as a *dN-dimensional* LinUCB

Graph Laplacian

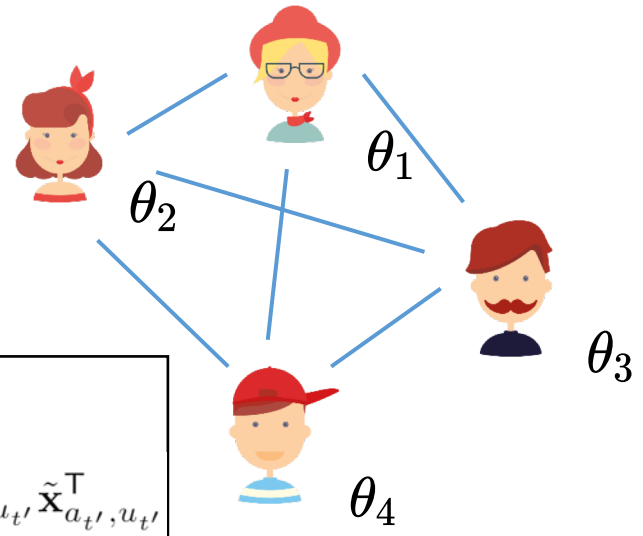
$$\tilde{\mathbf{x}}_{a_t,u_t} = ((\mathbf{L} + \mathbf{I}_N) \otimes \mathbf{I}_d)^{-1/2} \mathrm{Vec}(\mathring{\mathbf{X}}_{a_t,u_t})$$

$$\mathring{\mathbf{X}}_{a,u} = \{\mathbf{0}, \cdots, \underbrace{\mathbf{x}_a}_{u\text{-th column}}, \cdots, \mathbf{0}\}$$

$$\boldsymbol{\theta}^* = ((\mathbf{L} + \mathbf{I}_N) \otimes \mathbf{I}_d)^{1/2} (\boldsymbol{\theta}_1^*, \cdots, \boldsymbol{\theta}_N^*)$$

Regret for empty graph: $O(N\sqrt{T} \ln \frac{T}{N})$

Regret for complete graph: $O(N\sqrt{T} \ln \frac{T}{N^2})$



Closed form estimator

$$\mathbf{A}_t = \lambda \mathbf{I}_{dN} + \sum_{t'-1}^{t-1} \tilde{\mathbf{x}}_{a_{t'},u_{t'}} \tilde{\mathbf{x}}_{a_{t'},u_{t'}}^\mathsf{T}$$

$$\mathbf{b}_t = \sum_{t'=1}^{t-1} \tilde{\mathbf{x}}_{a_{t'},u_{t'}} r_{a_{t'},u_{t'}}$$

$$\hat{\boldsymbol{\theta}}_t = \mathbf{A}_t^{-1} \mathbf{b}_t$$

Collaborative bandit learning

- CoLin [WWGW16]
  - Social influence among users: content and opinion sharing in social network W
  - Reward: weighted average of expected reward among friends
  - A *dN-dimensional* LinUCB

$$\tilde{\mathbf{x}}_{a_t,u_t} = \mathrm{Vec}(\mathring{\mathbf{X}}_{a_t,u_t}\mathbf{W}^\top)$$

Closed form estimator

$$\mathbf{A}_t = \lambda \mathbf{I}_{dN} + \sum_{t'=1}^{t-1} \tilde{\mathbf{x}}_{a_{t'},u_{t'}} \tilde{\mathbf{x}}_{a_{t'},u_{t'}}^\top$$

$$\mathbf{b}_t = \sum_{t'=1}^{t-1} \tilde{\mathbf{x}}_{a_{t'},u_{t'}} r_{a_{t'},u_{t'}}$$

$$\hat{\theta}_t = \mathbf{A}_t^{-1}\mathbf{b}_t$$

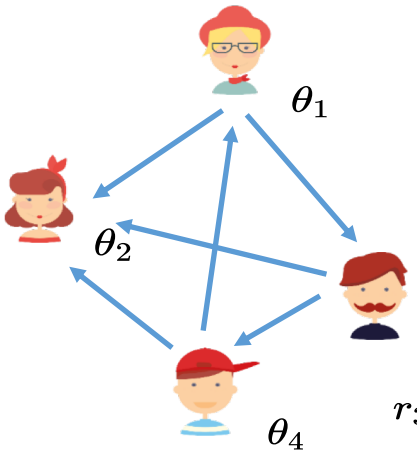
$$\mathbf{R}(T) \leq 2\alpha_T \sqrt{2dN \ln(1 + \frac{1 + \sum_{t=1}^{T}\sum_{j=1}^{N} w_{u_t j}^2}{\lambda dN})}$$

Regret for empty graph ($W = I$) $O(N\sqrt{T}\ln\frac{T}{N})$

Regret for $W = U$: $O(N\sqrt{T}\ln\frac{T}{N^2})$

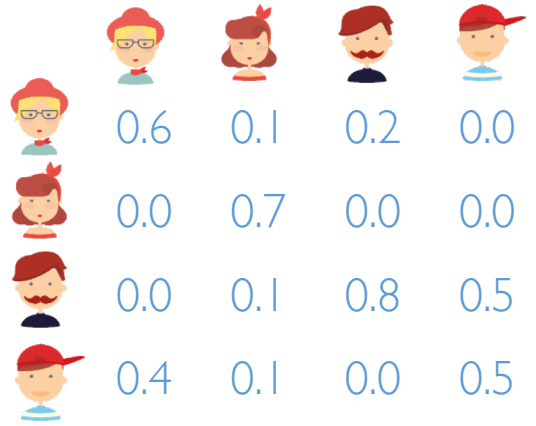
When W is uniform, i.e, all users are uniformly connected to share:

$$\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}^2 = 1$$



$$r_3 = \sum_{j=1}^{N} w_{3j} r_j$$

$\mathbf{W}$ in which $\sum_{j=1}^{N} w_{ij} = 1$



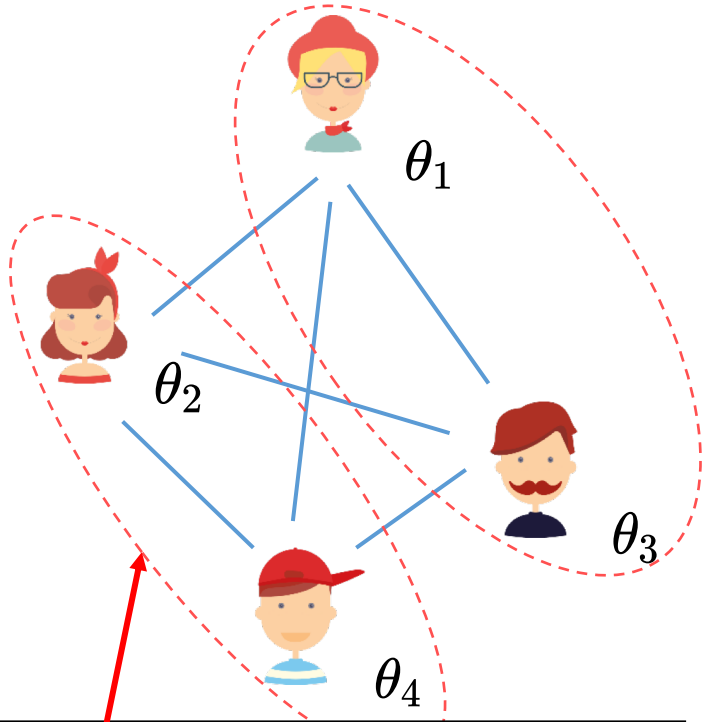
Online Clustering

- Discover user dependency structure on the fly
- CLUB[GLZ14]
  - Adaptively cluster users into groups by keep removing edges
  - Threshold to remove edges is based on closeness of the users' models
  - Build LinUCB on each cluster
- Regret: $O(\sqrt{mT}\log T)$. Reduce regret from n (users) to m (clusters)

Remove edge if

$$\|\hat{\theta}_{i,t} - \hat{\theta}_{j,t}\|_2 \geq \tilde{B}_{i,t} + \tilde{B}_{j,t}$$

$$\tilde{B}_{i,t} = \alpha\sqrt{\frac{1+\log(1+T_{i,t})}{1+T_{i,t}}}$$



Closed form estimator

$$\mathbf{A}_t = \lambda\mathbf{I} + \sum_{(a_i,r_i)\in\mathcal{H}_t} \mathbf{x}_{a_i}\mathbf{x}_{a_i}^\top$$

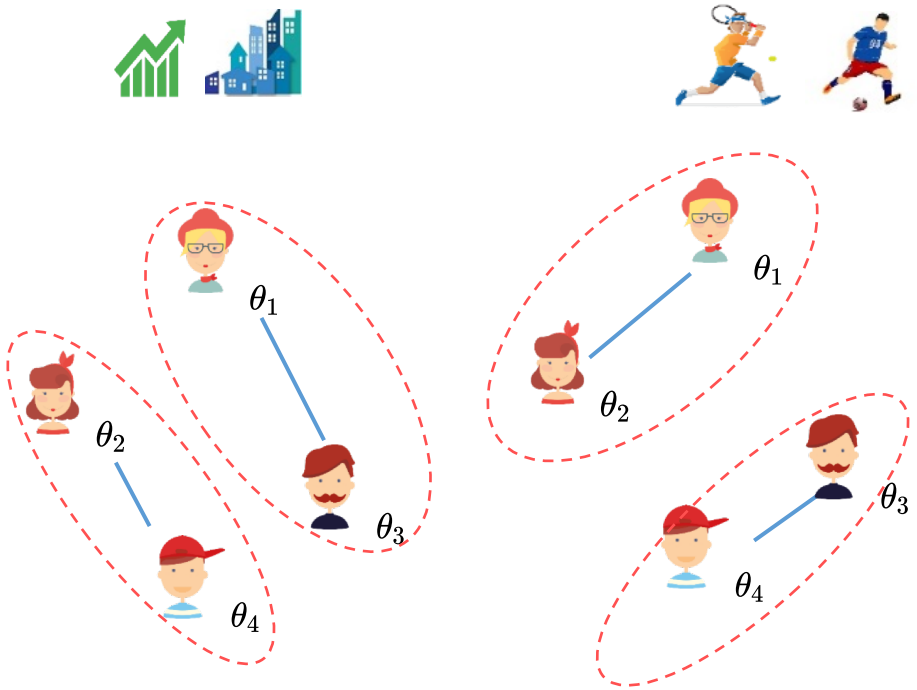
$$\mathbf{b}_t = \sum_{(a_i,r_i)\in\mathcal{H}_t} r_{a_i}\mathbf{x}_{a_i}$$

$$\hat{\boldsymbol{\theta}}_t = \mathbf{A}_t^{-1}\mathbf{b}_t$$

Online Clustering

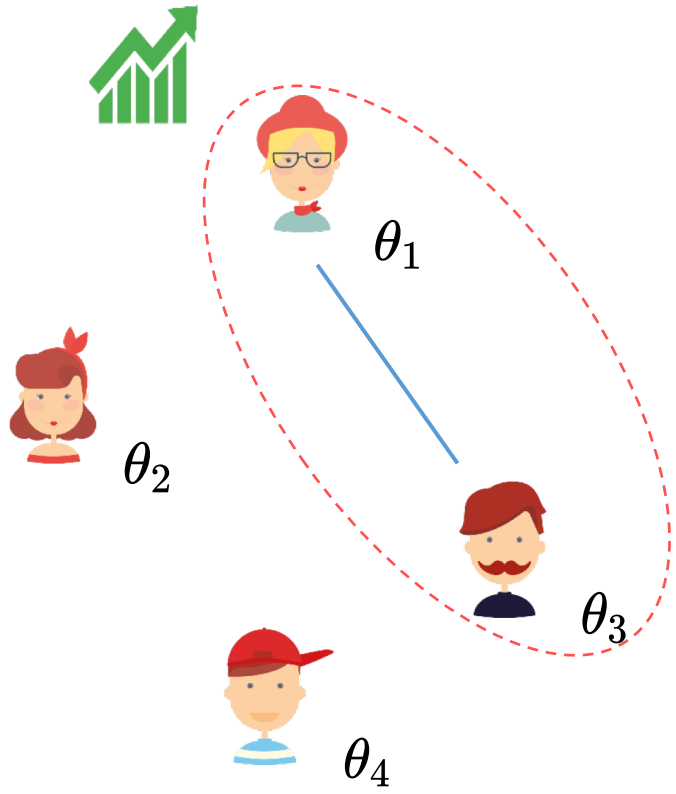
- COFIBA [LKG16]:
  - Collaborative filtering via user clustering & item clustering
  - Each item cluster is associated with its own user clustering
  - To remove an edge in user cluster: same as CLUB
  - To remove an edge in item cluster: for the two items, user i forms different neighboring user set $\{j\}$ based on

$$\|\mathbf{x}^T\hat{\theta}_{i,t} - \mathbf{x}^T\hat{\theta}_{j,t}\|_2 \leq CB_{i,t} + CB_{j,t}$$



Online Clustering

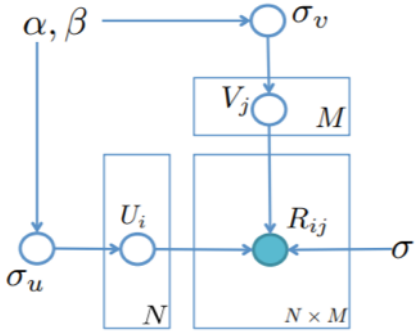
- CAB [GLKKZE17]:
  - Context-dependent clustering
  - For current user i , find neighboring user set $\{j\}$ for every candidate item x_a
  - Then aggregate the history rewards / predictions within the user cluster.



Low rank structures

- Particle Thompson Sampling (PTS) [KBKTC15]
  - Probabilistic Matrix Factorization framework
  - Particle filtering for online Bayesian parameter estimation
  - Thompson Sampling for exploration

$$
\begin{aligned}
U_i \text{ i.i.d.} \sim \quad & \mathcal{N}(0, \sigma_u^2 I_K) \\
V_j \text{ i.i.d.} \sim \quad & \mathcal{N}(0, \sigma_v^2 I_K) \\
r_{ij} | U, V \text{ i.i.d.} \sim \quad & \mathcal{N}(U_i^\top V_j, \sigma^2)
\end{aligned}
$$



Generative Model

Low rank structures

- Hidden LinUCB [WWW16]
  - Matrix Factorization framework: user & item factors
  - Alternating Least Squares for optimization
  - Exploration considers uncertainty from two factors

Source of uncertainty in confidence bound estimation

$$r_{a_t,u} = (\mathbf{x}_{a_t}, \mathbf{v}_{a_t})^\intercal (\boldsymbol{\theta}_u^x, \boldsymbol{\theta}_u^v) + \epsilon_t$$

Hidden feature (of an item): known to the environment, but unknown to the learner

$$a_t = \arg\max_{a \in \mathcal{A}} \Big((\mathbf{x}_a, \hat{\mathbf{v}}_{a,t})^\intercal \hat{\boldsymbol{\theta}}_{u,t} + \alpha_t^u \sqrt{(\mathbf{x}_a, \hat{\mathbf{v}}_{a,t}) \mathbf{A}_{u,t}^{-1} (\mathbf{x}_a, \hat{\mathbf{v}}_{a,t})^\intercal} + \alpha_t^a \sqrt{\hat{\boldsymbol{\theta}}_{u,t}^{\mathbf{v}} \mathbf{C}_{a,t}^{-1} \hat{\boldsymbol{\theta}}_{u,t}^{\mathbf{v}\intercal}} \Big)$$

Uncertainty of user preference θ_u estimation

Uncertainty of hidden feature v_a estimation

Low rank structures

- Projected Stochastic Linear Bandit [LAAH19]
- Assume item features $\{x_a \in R^d\}$ is rank-k ($k \ll d$)
- Idea: run PCA on all item features
  - Construct projection matrix P with first k eigenvectors.
- Reward estimation: $(Px_a)^T \hat{\theta}$

Warm-start exploration

- Have some offline data $\{(x, r)\}$ before the bandits start.
  - E.g., from human annotations
- Leverage historical data to warm start model, reduce the need of exploration
- Key challenge: historical data could come from different distribution
  - Historical data generated by θ' while environment follows θ^*

Warm-start exploration

- Leverage historical data to warm start model, reduce the need of exploration
- Adaptive Reweighting (ARROW-CB) [ZADLN19]
  - Based on ϵ-greedy algorithm
  - Reweight historical data based on bandits' observation
    - λ on historical data, $1-\lambda$ on bandits' observation
  - Online model selection to pick the weight λ
    - Pre-defined $|\Lambda|$ (hyperparameter, set to 8 in the paper) candidates
- Regret reduction when historical data and environment have similar distribution

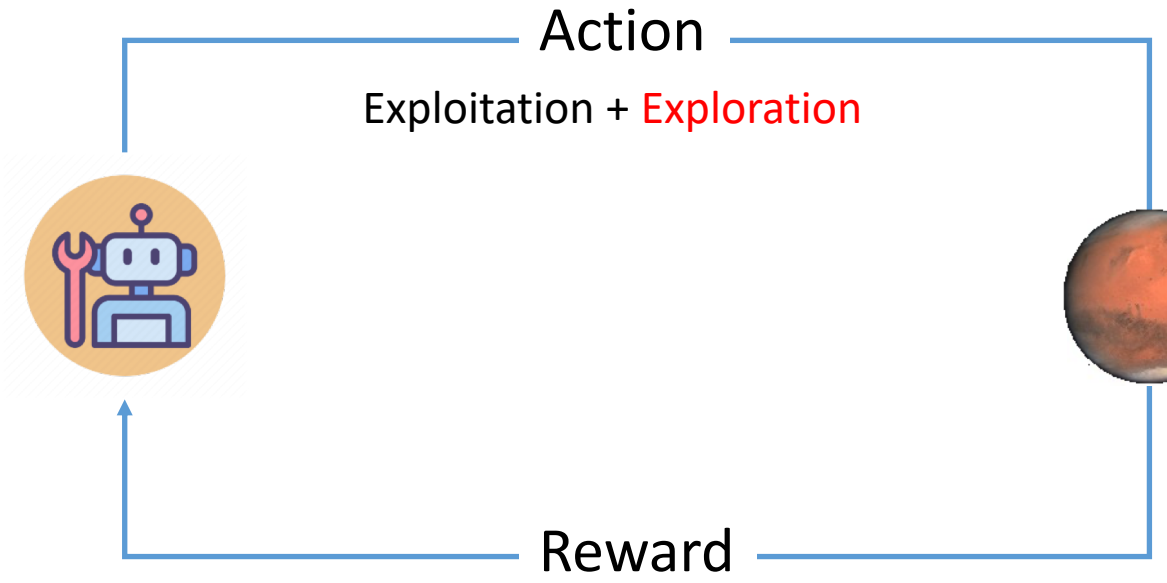
Open questions

- What is the problem-related (structure-related) regret lower bound
  - E.g., user dependency structure, low rank, offline data
  - Did current algorithms fully utilize the information in problem structure?
- Efficient exploration for other structures in real-world problem
  - E.g., sparse structure, ranking structure, etc.

Bandit learning for recommender systems

- Contextual bandits for recommendation
- Collaborative bandit learning
- Learning in a non-stationary environment
  - Passively adaptive approaches
  - Actively adaptive approaches
  - Unifying clustering and non-stationarity detection

Exploration in non-stationary environments



- Stationary assumption is commonly used
- The learner is usually facing a non-stationary environment ☹

Problem formulation

- (stochastic) Multi-arm bandit

Reward distribution

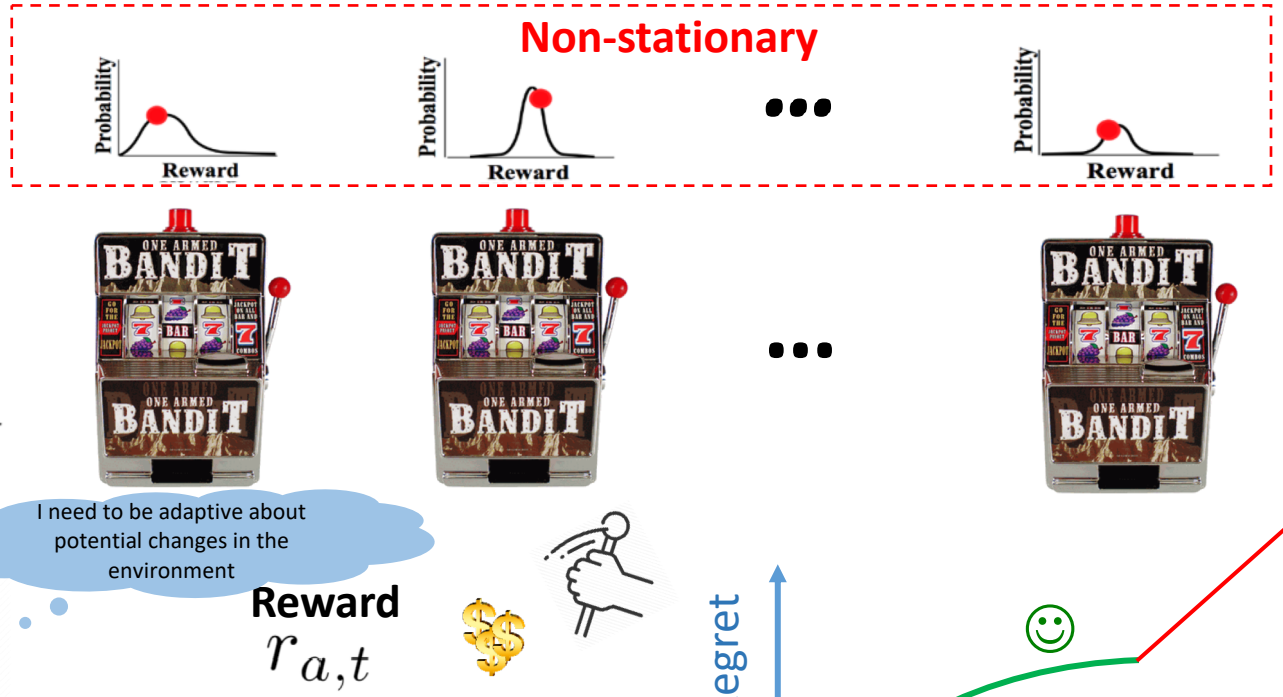
$\mathbb{E}[r_{a,t}] = \cancel{r_a^*}\ r_{a,t}^*$

Actions

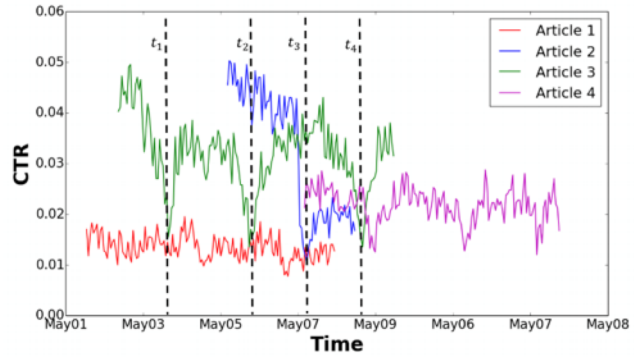
$\mathcal{A} = \{a_1, a_2, \ldots, a_k\}$

Agent/learner/ policy

$\pi_\theta(a|\mathcal{H}_t)$



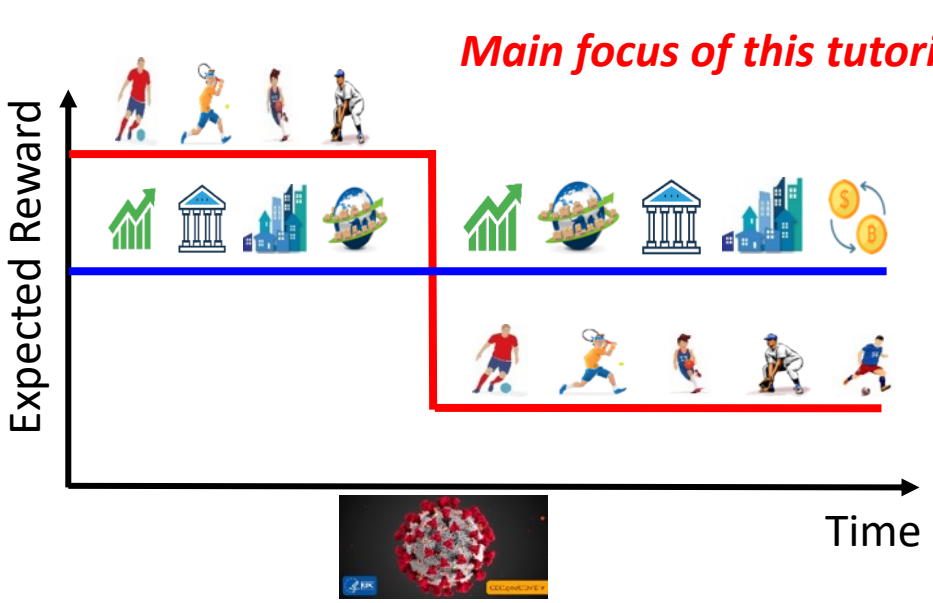
History is no longer helpful and may even be misleading in the player's decision-making

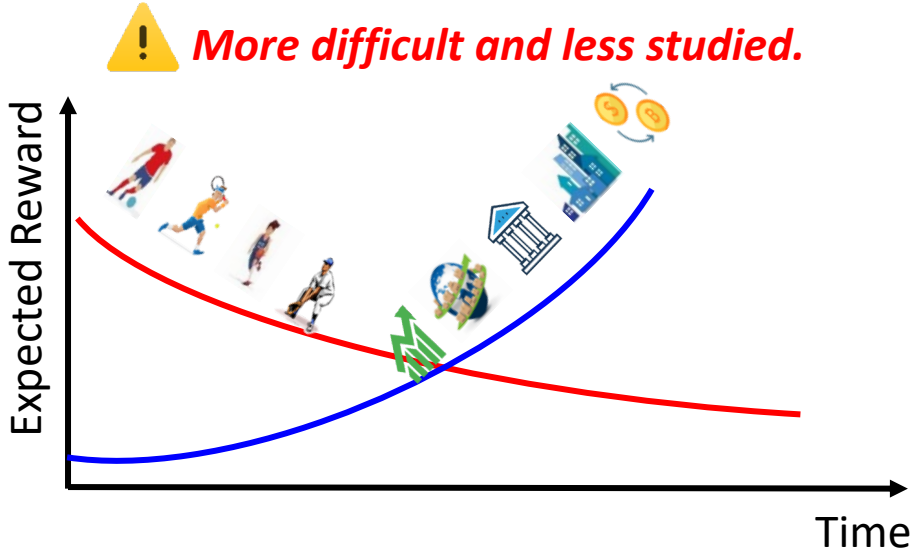


Real-time click-through-rate of 4 news articles collected from Yahoo user click log dataset

Problem formulation

- Different types of non-stationary environments
  - Piecewise stationary environments
  - Gradually changing environments





Problem to solve

- Changes (when and how) are **unknown** to the learner

(otherwise we can just restart the learner)

- **Online** learning setting and **bandit** feedback: incomplete knowledge

(change detection in the offline batch setting has been extensively studied in statistics and control theory [BN93])

Approaches

- Passively adaptive approaches
  - Key idea: design a proper mechanism to forget old observations
  - Assumption: old observations are less relevant
- Actively adaptive approaches
  - Key idea: actively detect potential changes in the environment during online decision making
  - Assumption: abrupt changes, i.e., piece-wise stationary environment

Passively adaptive approaches

- Discounted UCB and Sliding window UCB [GM08] $\gamma \in (0,1]$

$$\hat{r}_{a,t}(\gamma) = \frac{1}{N_{a,t}(\gamma)} \sum_{s=1}^{t} \gamma^{t-s} r_{a,t} \mathbb{1}_{\{a_t = i\}} \qquad N_{a,t}(\gamma) = \sum_{s=1}^{t} \gamma^{t-s} \mathbb{1}_{\{a_t = a\}}$$

Discount the old observations

$$a_t = \arg\max_{a \in \mathcal{A}} \hat{r}_{a,t}(\gamma) + B_{a,t}(\gamma) \rightarrow B_t(a,\gamma) = \alpha \sqrt{\frac{\epsilon \log n_t(\gamma)}{N_{a,t}(\gamma)}}$$

$$a_t = \arg\max_{a \in \mathcal{A}} \hat{r}_{a,t}(\tau) + B_{a,t}(\tau) \rightarrow B_t(a,\tau) = \alpha \sqrt{\frac{\epsilon \log n_t(\tau)}{N_{a,t}(\tau)}}$$

$$\hat{r}_{a,t}(\tau) = \frac{1}{N_{a,t}(\tau)} \sum_{s=t-\tau+1}^{t} r_{a,t} \mathbb{1}_{(a_t = i)} \qquad N_{a,t}(\tau) = \sum_{s=t-\tau+1}^{t} \mathbb{1}_{\{a_t = a\}}$$

Only utilize the most recent τ observations

- Weighted linear bandit [RVC19] $\tau > 1$
  - Discounted-UCB in the linear contextual bandit setting

Passively adaptive approaches

- Pros
  - Simple: easy to implement and have almost no computation overhead
  - Have provable theoretical guarantee in piece-wise non-stationary environment and environments with slow changes
- Cons
  - Passive: always assuming old observations are less relevant
  - Not practical: very sensitive to hyper-parameters discount factor, sliding-window size

Actively adaptive approaches

- General framework

Inputs: Hyper-parameters.
Initialization: Set initial values for statistics about arm selection π_0 and change detection Π_0 of the bandit model

1: **for** t = 1, ..., T **do**
2: Observe arm set $\mathcal{A}_t$ (and potentially also context features)
3: Select arm $a_t = \pi_t(\mathcal{A}_t)$
4: Observe reward r_{a_t}
5: Update Π_{t+1}
6: CHANGED = NonstationaryTest(Π_{t+1})
7: **if** CHANGED **then**
8: Restart or partially reset π_{t+1} and Π_{t+1}
9: **else**
10: Update π_{t+1}
11: **end if**
12: **end for**

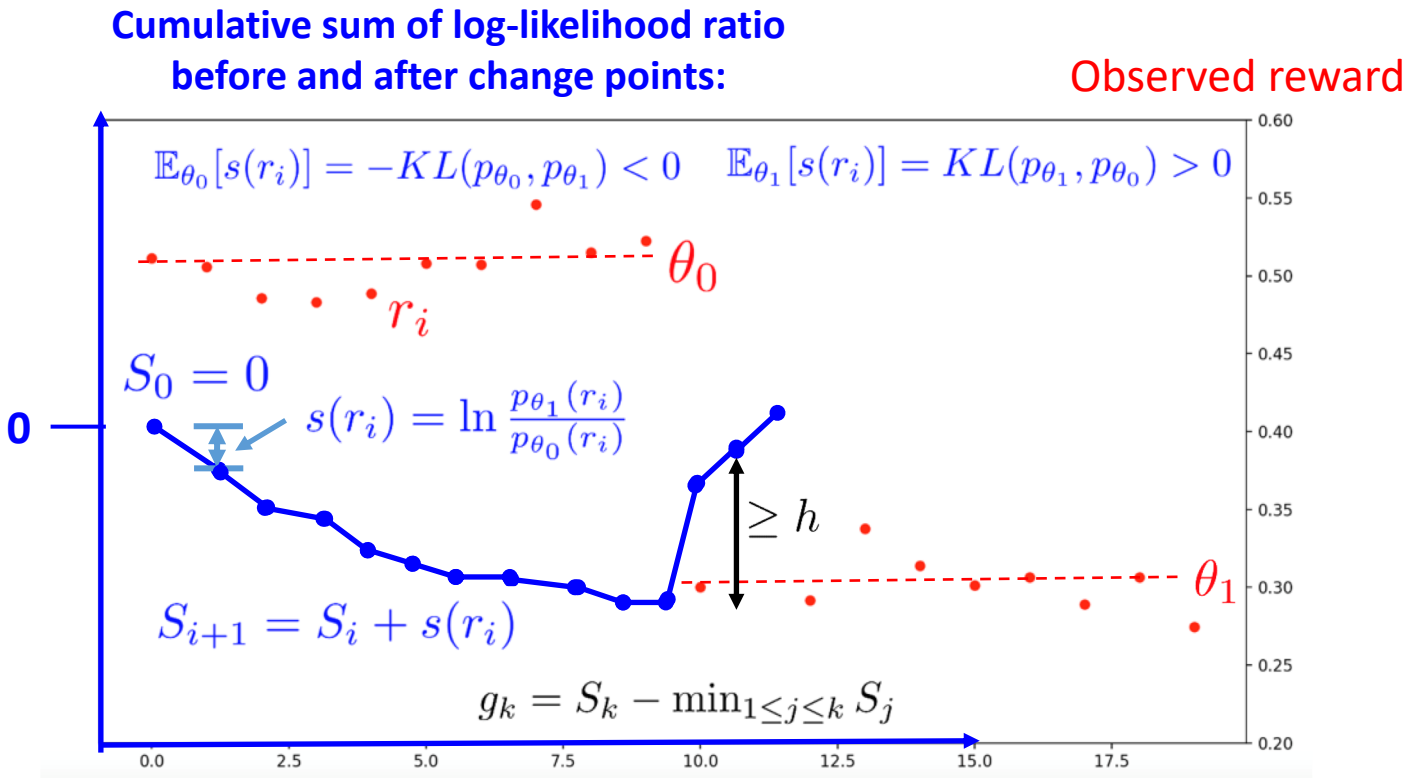
Statistics about the stationarity of the environment

Non-stationarity detection with bandit feedback

- Cumulative sum control chart (CUSUM) based detection
  - AdTS [HMB15], CD-UCB [LLS18]
- Generalized likelihood ratio test
  - GLR-kl-UCB [BK19]
- Online reward mean-shift detection
  - WMD-UCB1 [YM09], M-UCB [CWKX19]
- Confidence bound based detection
  - dLinUCB [WIW18], DenBand [WWLW19]
- Other
  - Ada-ILTCB, Ada-greedy [LWAL18], Ada-ILTCB$^{[+]}$ [CLLW19] …

CUSUM-based online change detection

- CUSUM in the offline setting



CUSUM-based online change detection

- Tailored CUSUM in the Bernoulli bandit setting
  - Intuition: $\hat{r}_a(M)$

$$\left| r_{i,a} - \mathbb{E}[r_{i,a}] \right| - \epsilon$$

Requirement on the minimum magnitude of changes on the reward

has negative mean drift before the change point and positive after the change point

  - Assumptions:
    - Piecewise stationary with detectability assumption
    - Bernoulli bandit

- **CUSUM in CD-UCB [LLS18]**

Each arm needs to have at least M observations

Inputs: Hyperparameters M, h

$s^+_{i,a} = (r_{i,a} - \hat{r}_a(M) - \epsilon)\mathbb{1}_{\{i>M\}}$

$s^-_{i,a} = (\hat{r}_a(M) - r_{i,a} - \epsilon)\mathbb{1}_{\{i>M\}}$

$g^+_{i,a} = \max\{0, g^+_{i-1,a} + s^+_{i,a}\}$

$g^-_{i,a} = \max\{0, g^-_{i-1,a} + s^-_{i,a}\}$

if $g^+_{i,a} > h$ or $g^-_{i,a} > h$ **then**

  Return **True**

end if

Detection threshold

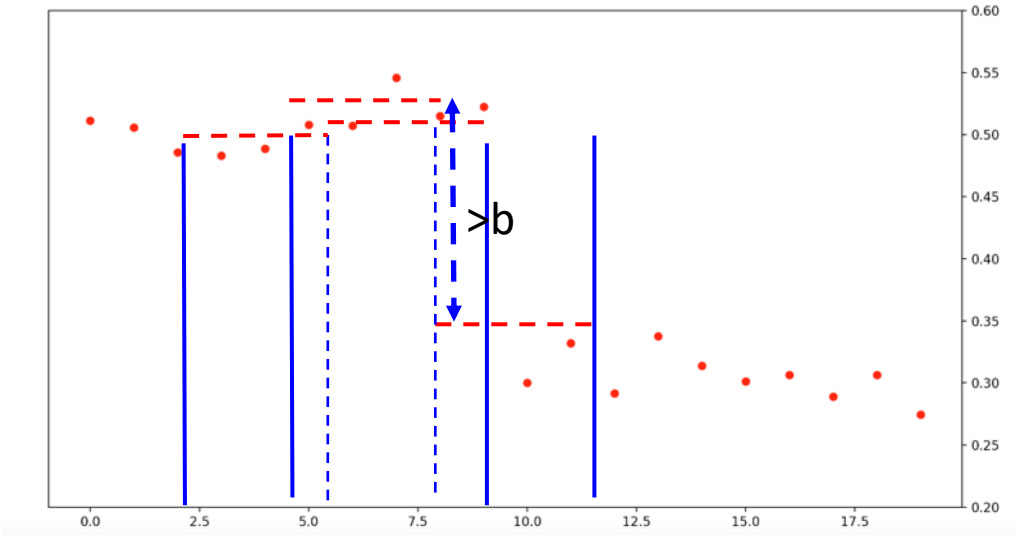
  - Local restart: restart the related statistics for the changed arms

Generalized likelihood ratio test based online change detection

- CUSUM requires the pre-change and post-change environment parameters to be known to get the log-likelihood.
- Unknown pre-change and/or post-change parameters -> Generalized Likelihood Ratio Test (GLRT)
  - GLRT with Bernoulli reward -> GLR-klUCB [BK19]
  - GLRT can achieve asymptotically optimal detection delay with sub-Gaussian reward assumption [Mai19]

Online mean-shift detection

- Monitored-UCB [CWKX19]
  - Compare running sample means over a sliding window



Inputs: Hyperparameter w (an even number)

1: **if** $\sum_{i=w/2+1}^{w} r_{i,a} - \sum_{i=1}^{w/2} r_{i,a} > b$ **then**

2: Return **True**

3: **end if**

- Global restart: once a change is detected, restart the related statistics for all the arms
- Need to periodically perform uniform arm selection (uniform exploration)
  to ensure sufficient data can be gathered for all arms to perform CD
- Arm selection between detected change points (except for the uniform exploration iterations): UCB1

CB based online change detection

- Utilizing the reward estimation confidence bound [WNW18, WWLW19]
  - If the environment is stationary, the reward prediction residual should be within a confidence bound with high probability,

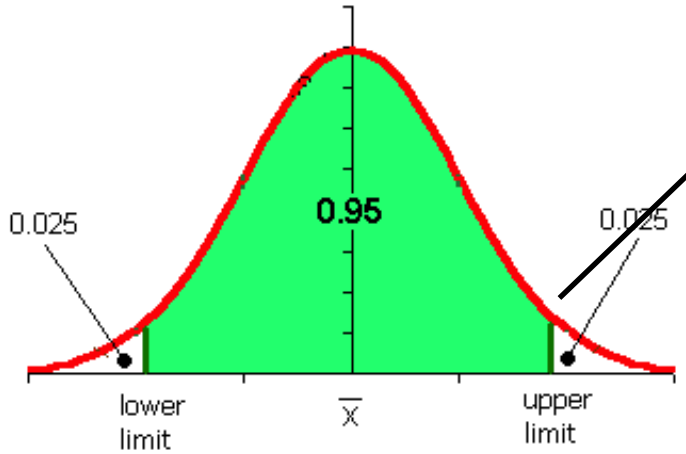
Predicted reward of bandit model m

Observed reward

Confidence bound of reward estimation

$$|\hat{r}_{a,t}(m) - r_{a,t}| \leq B_{a,t}(m) + \delta$$

We can bound this to further exclude this possibility!



Unexpected outcome is caused by,

1) Something wrong in the observed reward: large noise
2) Something wrong in the predicted reward: the model is incorrect

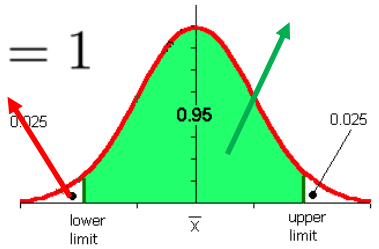
CB based online change detection

- Prediction badness at interaction i:

$$e_i(m) = \mathbb{1}\{|\hat{r}_{a,t}(m) - r_{a,t}| \leq B_{a,t}(m) + \delta\}$$

$e_i(m) = 0$

$e_i(m) = 1$



Similar to the **goodness of fit** concept in chi-squared test

- Badness of model *m* over a sliding time window:

$$\hat{e}_t(m) = \frac{\sum_{i=t-\tau}^{t} e_i(m)}{\hat{\tau}(m)}$$

a sliding time window to collect badness observations

- Detection threshold:
  - Bad enough:

$$\hat{e}_t(m) > \delta_1 + \sqrt{\frac{\ln(1/\delta_2)}{2\tilde{\tau}(m)}}$$

Change detected. (abandon learner *m*)

  - Good enough:

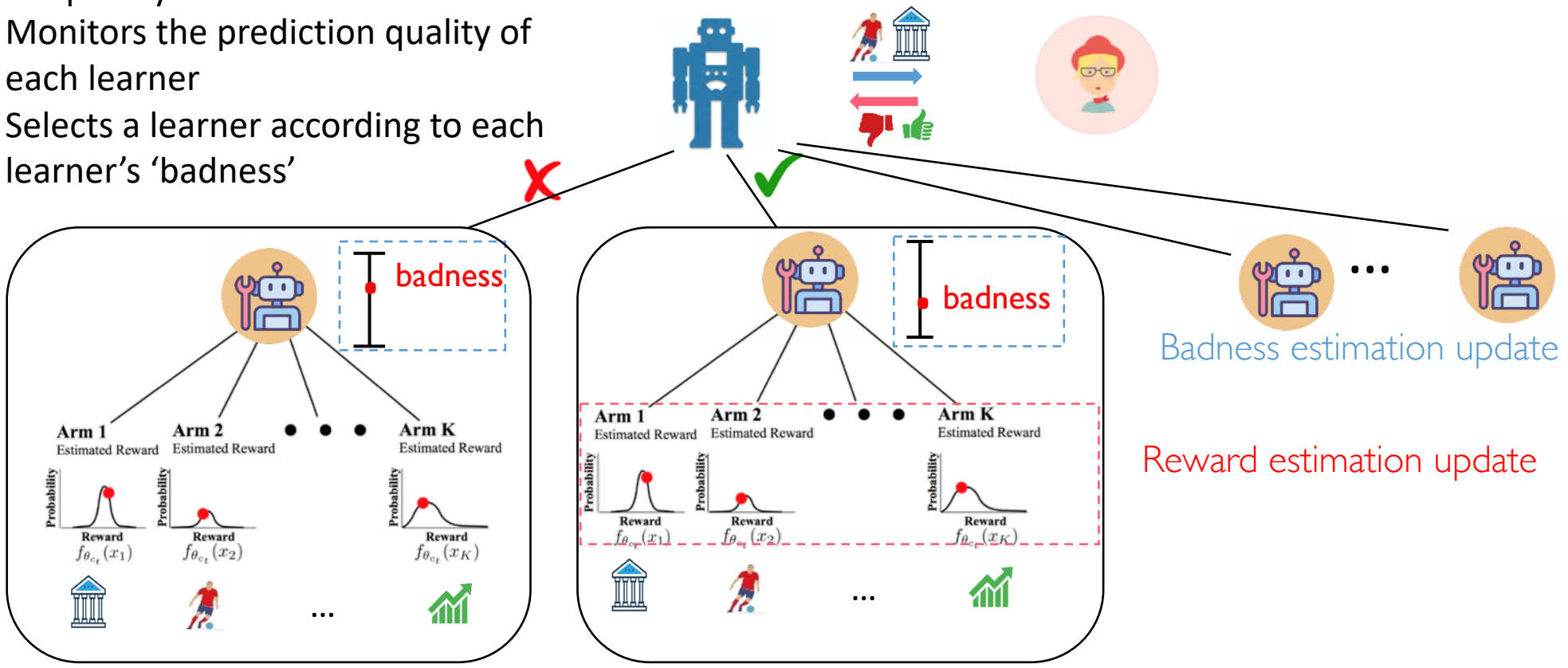
$$\hat{e}_t(m) \leq \tilde{\delta}_1 + \sqrt{\frac{\ln(1/\delta_2)}{2\tilde{\tau}(m)}}$$

$$\tilde{\delta}_1 < \delta_1$$

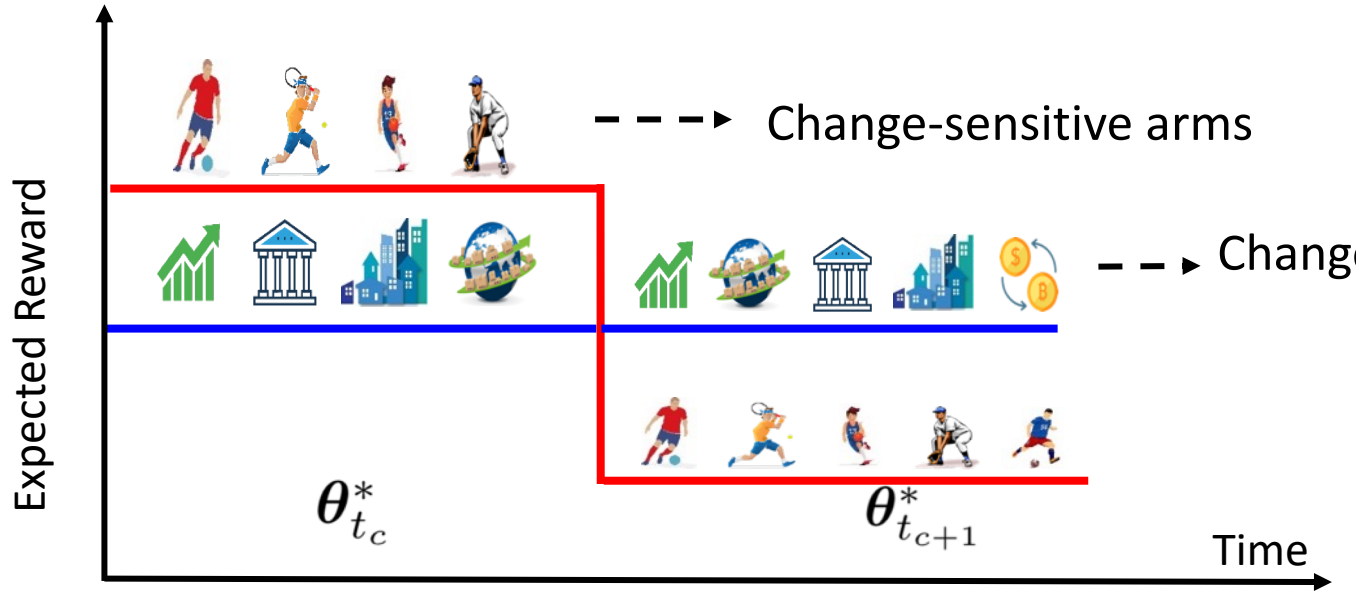
When no learner is good enough, will also report a change. (but do not abandon learner m)

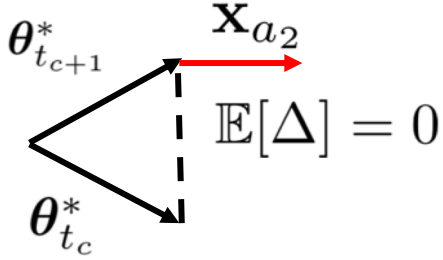
CB based online change detection

- dLinUCB [WNW18]
  - Adaptively creates new learners
  - Monitors the prediction quality of each learner
  - Selects a learner according to each learner's 'badness'

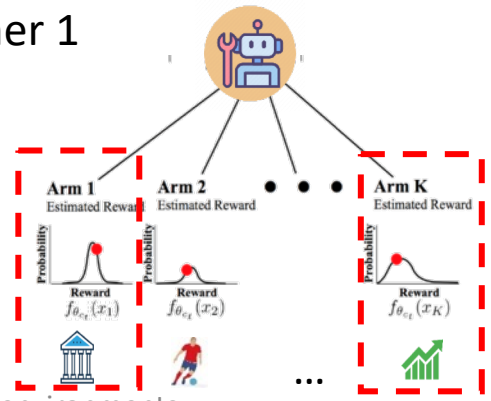


Detection of context-dependent changes

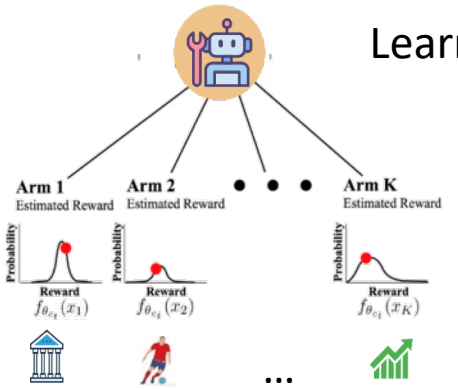




Learner 1



Learner 2



Reward estimation on change-invariant arms

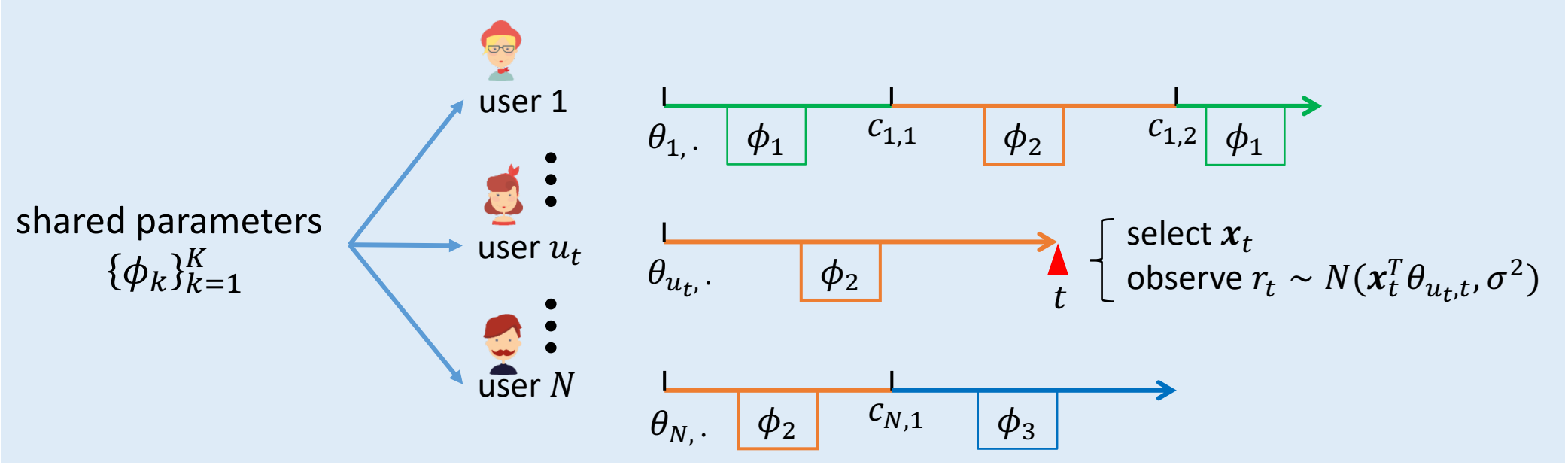
DenBand [WWLW19]

Reuse the experiences learnt in old models for change-invariant arms

Save unnecessarily exploration

Unifying clustering and change detection

- A clustered and non-stationary environment

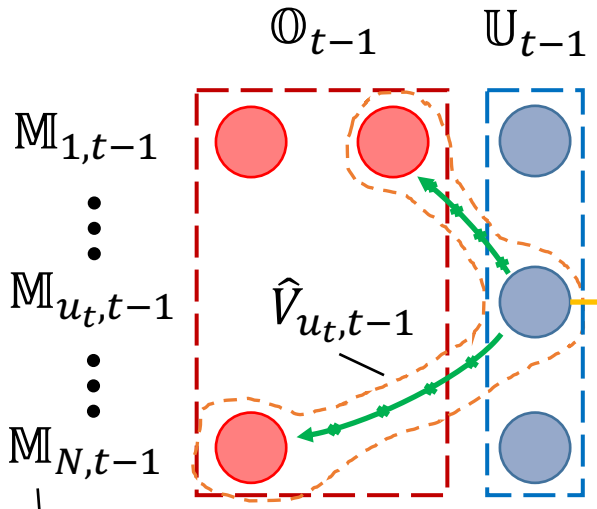


Unifying clustering and change detection

- Connection: both are testing homogeneity between data sequences
- More formally, given two data sequences:
  - $\mathcal{H}_1 = \{(x_i, r_i)\}_{i=1}^{t_1}$ and $\mathcal{H}_2 = \{(x_j, r_j)\}_{j=1}^{t_2}$
  - $\forall (x_i, r_i) \in \mathcal{H}_1, r_i \sim \mathcal{N}(x_i^\top \theta_1, \sigma^2)$ and $\forall (x_j, r_j) \in \mathcal{H}_2, r_j \sim \mathcal{N}(x_j^\top \theta_2, \sigma^2)$
  - Decide whether $\theta_1 = \theta_2$
- When data are from two user: clustering
- When data are from different periods of same user: change detection

Dynamic Clustering (DyClu) [LWW21a]

Lower level:
maintain local user models



- observation history $\mathcal{H}_{N,t-1}$
- sufficient statistics $\boldsymbol{A}_{N,t-1}, \boldsymbol{b}_{N,t-1}$

Upper level:
aggregate observations in $\hat{V}_{u_t,t-1}$

- select $\boldsymbol{x}_t$ using aggregated model

replace outdated user models

- outdated model set $\mathbb{O}_{t-1}$
- up-to-date model set $\mathbb{U}_{t-1}$

$\{(\boldsymbol{x}_t, y_t)\}$

cluster identification:
$s(\mathcal{H}_{u_t,t-1}, \mathcal{H}_{j,t-1})$

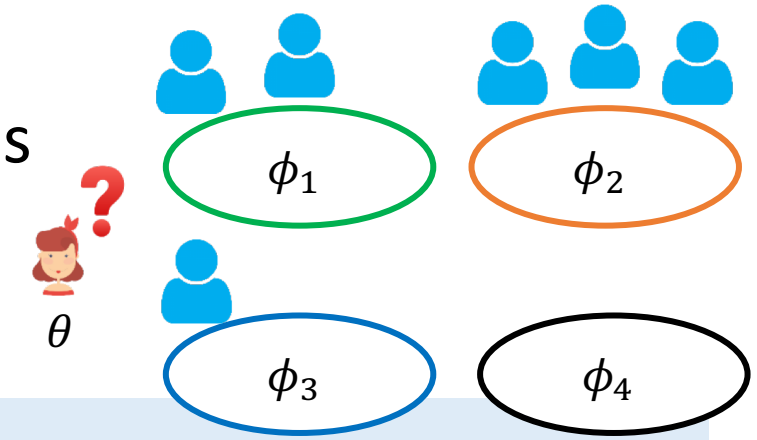
change detection:
$s(\mathcal{H}_{u_t,t-1}, \{(x_t, y_t)\})$

$s()$ is the test statics of homogeneity

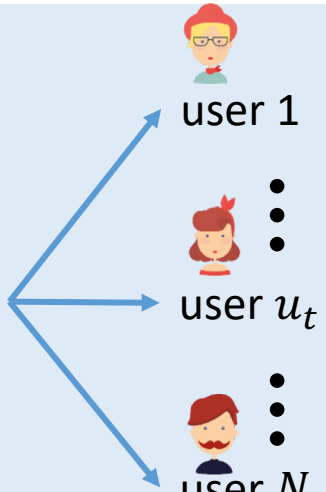
DyClu: one model per stationary period
Detect change; cluster individual bandit models; select arm

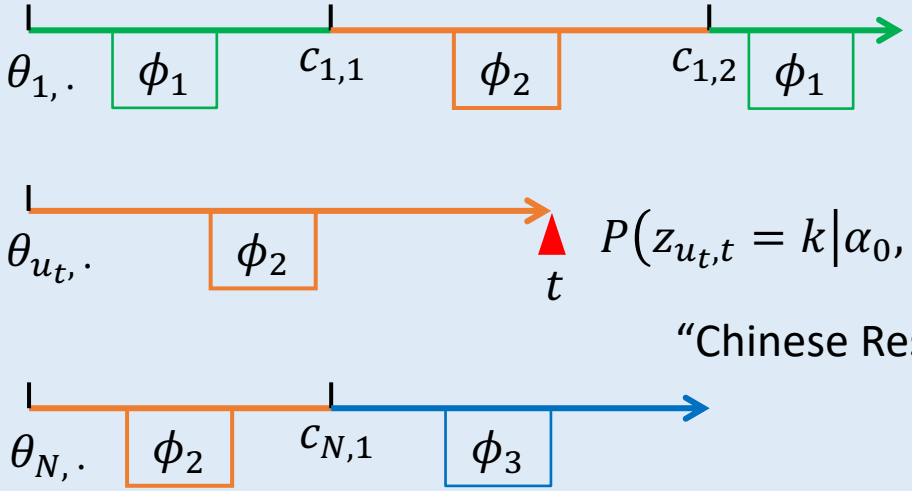
CoDBand [LWW21b]

- Dirichlet Process (DP) mixture of regression models
  - θs are samples from infinite mixture model G with prior $\mathrm{DP}(\alpha_0, G_0)$
  - The set of unique bandit parameters: $\{\phi_k\}_{k=1}^{K_{t-1}}$
  - Indicator variable $z_{u,c}$: $\theta_{u,c} = \phi_{z_{u,c}}$



shared parameters $\{\phi_k\}_{k=1}^{K}$



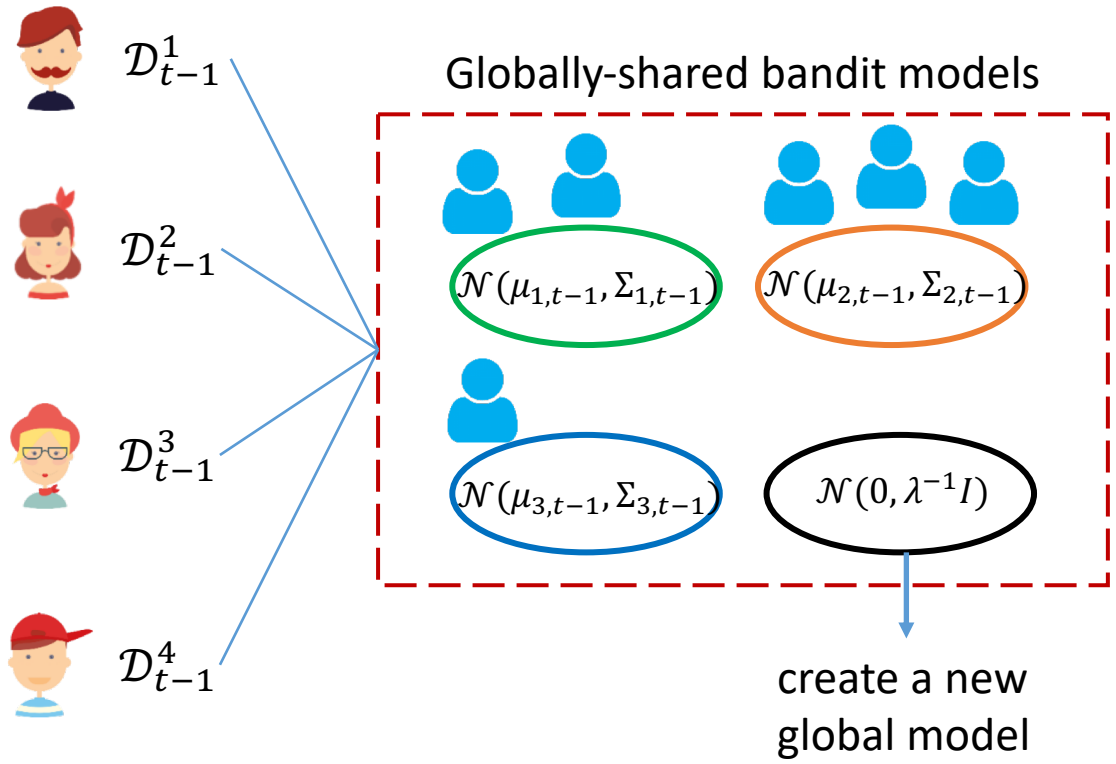


$$P(z_{u_t,t} = k | \alpha_0, \mathcal{Z}_{t-1}) \propto \begin{cases} n_{k,t-1} & k \in [K_{t-1}] \\ \alpha_0 & k \notin [K_{t-1}] \end{cases}$$

"Chinese Restaurant Process"

CoDBand [LWW21b]

- CoDBand: two-level Thompson sampling
- Arm selection: posterior sampling
  - Sample $\tilde{z}_{u_t,t} \sim P(z_{u_t,t}|\alpha_0, \mathcal{Z}_{t-1}, \mathcal{D}_{t-1}^{u_t})$
  - Sample $\tilde{\theta}_{u_t,t} \sim \mathcal{N}(\mu_{\tilde{z}_{u_t,t},t-1}, \Sigma_{\tilde{z}_{u_t,t},t-1})$
  - Select arm $x_t = \underset{x \in \mathcal{A}_t}{\operatorname{argmax}}\, x^T \tilde{\theta}_{u_t,t}$, observe r_t
- Model update
  - Update global model $\mathcal{N}(\mu_{\tilde{z}_{u_t,t},t-1}, \Sigma_{\tilde{z}_{u_t,t},t-1})$ with (x_t, r_t)
  - Update dataset $\mathcal{D}_t^{u_t} = \mathcal{D}_{t-1}^{u_t} \cup \{(x_t, r_t)\}$
  - If change is detected: Reset $\mathcal{D}_t^i = \emptyset$



CoDBand: one model per unique parameter
Detect change; select global bandit model; select arm

Open questions

- Can we re-use related historical observations (e.g., recurring environment, or context-dependent changes) and what's the benefit of it?

- How to handle gradually changing environment?

References III

[FCGS10] Filippi, S., Cappe, O., Garivier, A., & Szepesvári, C. (2010, December). Parametric Bandits: The Generalized Linear Case. In NIPS (Vol. 23, pp. 586-594).

[ZLG20] Zhou, D., Li, L., & Gu, Q. (2020, November). Neural contextual bandits with UCB-based exploration. In International Conference on Machine Learning (pp. 11492-11502). PMLR.

[CGZ13] Cesa-Bianchi, N., Gentile, C., & Zappella, G. (2013). A gang of bandits. In Advances in Neural Information Processing Systems (pp. 737-745).

[WWGW16] Wu, Q., Wang, H., Gu, Q., & Wang, H. (2016, July). Contextual bandits in a collaborative environment. In Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval (pp. 529-538).

[GLZ14] Gentile, C., Li, S., & Zappella, G. (2014, January). Online clustering of bandits. In International Conference on Machine Learning (pp. 757-765).

[GLKKZE17] Gentile, C., Li, S., Kar, P., Karatzoglou, A., Zappella, G., & Etrue, E. (2017, July). On context-dependent clustering of bandits. In International Conference on Machine Learning (pp. 1253-1262).

[LKG16] Li, S., Karatzoglou, A., & Gentile, C. (2016, July). Collaborative filtering bandits. In Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval (pp. 539-548).

[KBKTC15] Kawale, J., Bui, H. H., Kveton, B., Tran-Thanh, L., & Chawla, S. (2015). Efficient Thompson Sampling for Online Matrix-Factorization Recommendation. In Advances in neural information processing systems (pp. 1297-1305).

[KKSVW17] Katariya, S., Kveton, B., Szepesvari, C., Vernade, C., & Wen, Z. (2017, April). Stochastic rank-1 bandits. In Artificial Intelligence and Statistics (pp. 392-401).

[WWW16] Wang, H., Wu, Q., & Wang, H. (2016, October). Learning hidden features for contextual bandits. In Proceedings of the 25th ACM International on Conference on Information and Knowledge Management (pp. 1633-1642).

References IV

[ZADLN19] Zhang, C., Agarwal, A., Daumé III, H., Langford, J., & Negahban, S. N. (2019). Warm-starting contextual bandits: Robustly combining supervised and bandit feedback. arXiv preprint arXiv:1901.00301.

[LAAH19] Lale, S., Azizzadenesheli, K., Anandkumar, A., & Hassibi, B. (2019). Stochastic linear bandits with hidden low rank structure. arXiv preprint arXiv:1901.09490.

[BN93] Basseville, M., & Nikiforov, I. V. (1993). *Detection of abrupt changes: theory and application* (Vol. 104). Englewood Cliffs: prentice Hall.

[GM08] Garivier, A., & Moulines, E. (2008). On upper-confidence bound policies for non-stationary bandit problems. *arXiv preprint arXiv:0805.3415*.

[RVC19] Russac, Y., Vernade, C., & Cappé, O. (2019). Weighted linear bandits for non-stationary environments. In *Advances in Neural Information Processing Systems* (pp. 12040-12049).

[YM09] Yu, J. Y., & Mannor, S. (2009, June). Piecewise-stationary bandit problems with side observations. In *Proceedings of the 26th annual international conference on machine learning* (pp. 1177-1184).

[HMB15] Hariri, N., Mobasher, B., & Burke, R. (2015, June). Adapting to user preference changes in interactive recommendation. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*.

[WNW18] Wu, Q., Iyer, N., & Wang, H. (2018, June). Learning contextual bandits in a non-stationary environment. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval* (pp. 495-504).

References V

[LWAL18] Luo, H., Wei, C. Y., Agarwal, A., & Langford, J. (2018, July). Efficient contextual bandits in non-stationary worlds. In *Conference On Learning Theory* (pp. 1739-1776).

[CWKX19] Cao, Y., Wen, Z., Kveton, B., & Xie, Y. (2019, April). Nearly optimal adaptive procedure with change detection for piecewise-stationary bandit. In *The 22nd International Conference on Artificial Intelligence and Statistics* (pp. 418-427).

[LLS18] Liu, Fang, Joohyun Lee, and Ness Shroff. "A change-detection based framework for piecewise-stationary multi-armed bandit problem." Thirty-Second AAAI Conference on Artificial Intelligence. 2018.

[WWLW19] Wu, Q., Wang, H., Li, Y., & Wang, H. (2019, May). Dynamic Ensemble of Contextual Bandits to Satisfy Users' Changing Interests. In *The World Wide Web Conference* (pp. 2080-2090).

[CLLW19] Chen, Y., Lee, C. W., Luo, H., & Wei, C. Y. (2019). A new algorithm for non-stationary contextual bandits: Efficient, optimal, and parameter-free. *arXiv preprint arXiv:1902.00980*.

[BK19] Besson, L., & Kaufmann, E. (2019). The generalized likelihood ratio test meets klucb: an improved algorithm for piece-wise non-stationary bandits. *arXiv preprint arXiv:1902.01575*.

[Mai19] Maillard, O. A. (2019). Sequential change-point detection: Laplace concentration of scan statistics and non-asymptotic delay bounds.

[LWW21a] Li, C., Wu, Q., & Wang, H. (2021, March). Unifying Clustered and Non-stationary Bandits. In International Conference on Artificial Intelligence and Statistics (pp. 1063-1071). PMLR.

[LWW21b] Li, C., Wu, Q., & Wang, H. (2021). When and Whom to Collaborate with in a Changing Environment: A Collaborative Dynamic Bandit Solution. In SIGIR 2021.

Outline of this tutorial

- Motivation
- Background: bandit algorithms
- Bandit learning for recommender systems
- Bandit learning for retrieval systems
- Ethical considerations in IR with bandit learning
- Conclusion & future directions

Bandit learning for retrieval systems

- Why the bandit algorithms for recommendation systems do not apply?

Recommender systems

- Top-1 ranking
  - Treat each item as an arm
  - Linear exploration space
- Presentation bias
- No position bias, as the recommended item will always be examined

Retrieval systems

- Top-K ranking
  - Treat each ranking as an arm?
  - Exponential exploration space
- Presentation bias
- Position bias due to users' examination behavior on a ranked list

Online learning to rank

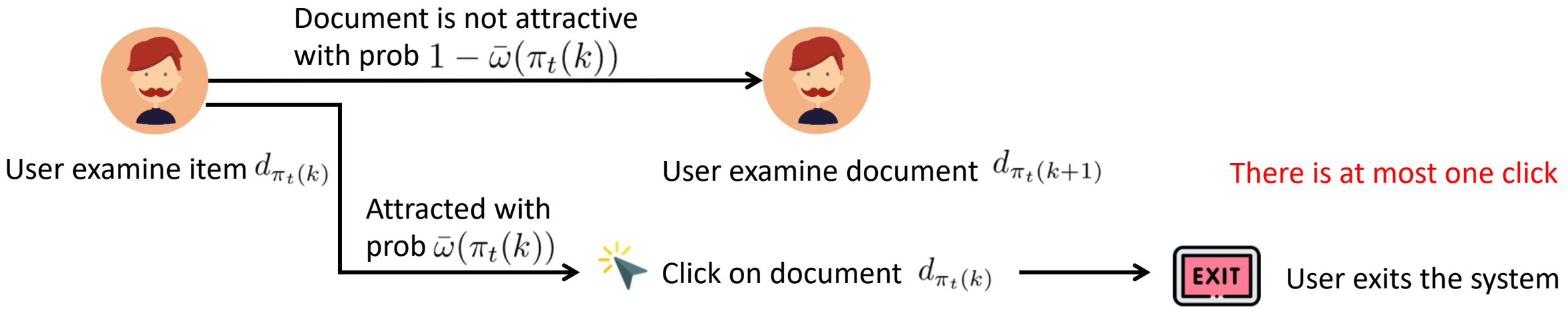
- OL2R under specific click models
- Dueling bandit gradient descent
- Online pairwise methods
- OL2R vs. offline unbiased learning to rank

OL2R under specific click models

- Assume users' behavior follows some specific click models, e.g., cascade model, position-based model
- Deal with the exponential ranking space $\longrightarrow$ document space

Cascading bandits [KSWA15]

- Cascade model is a popular model of user behavior in web search
  - A set of L documents $S = \{1, ..., L\}$
  - Attraction probabilities $\bar{\omega} \in [0, 1]^S$
  - User sequentially scan a list of K documents $\pi = (d_1, ..., d_K) \in \Pi_K(S)$



Cascading bandits

- Compute UCB on the attraction probability of each document.
- Rank the documents according to UCB and return the top-K documents.

- One model for each query
- Interaction at time t:
  - Environment draws attraction weights w_t for L candidates
  - The model chooses an ordered list of K documents, $\pi_t = (d_1^t, ..., d_K^t) \in \Pi_K(S)$
  - User clicks first attractive item in π_t, C_t
  - Update the weights of all observed items according to the feedback

$$\omega_t(\pi_t(k)) = \mathbf{1}\{C_t = k\}, k = 1, ..., \min(C_t, K)$$

  - Learning agent receives reward $f(\pi_t, w_t)$

$$f(\pi_t, \omega_t) = 1 - \prod_{k=1}^{K}(1 - \omega(\pi_t(k)))$$

At least one document is attractive.

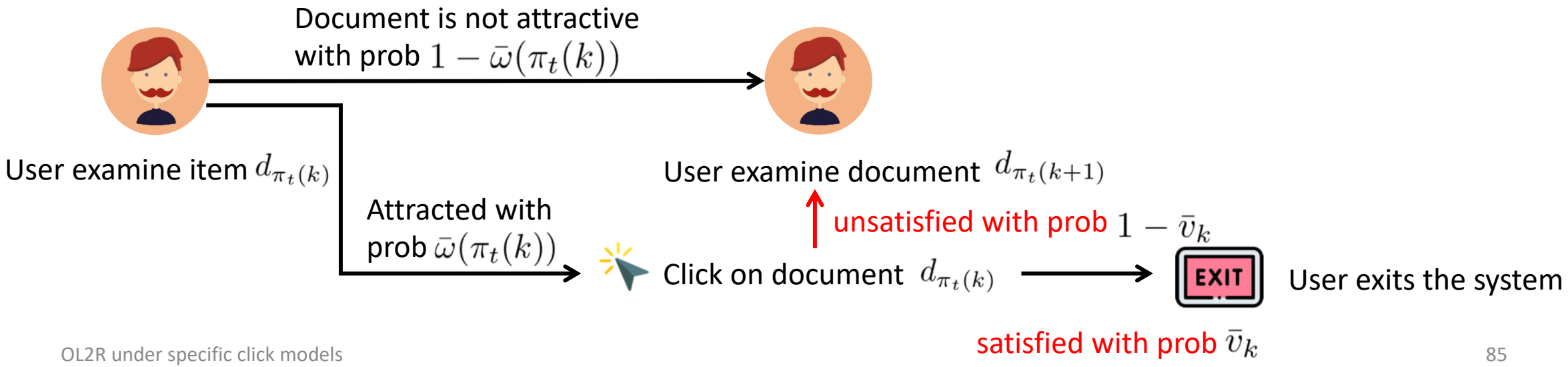
- Goal: minimize the expected cumulative regret over T steps

$$\mathbf{R}_T = \mathbb{E}[\sum_{t=1}^{T} f(\pi^*, \omega_t) - f(\pi_t, \omega_t)]$$

ne probability that none of the K
ocuments is attractive.

DCM bandits [KKSW16]

- Dependent click model (DCM)
  - Extend from cascade model where user may click on **multiple documents**
    - Attraction probability: $\bar{\omega} \in [0,1]^S$
    - termination probability: $\bar{v} \in [0,1]^S$ Position-dependent



DCM bandits [KKSW16]

- Dependent click model (DCM)
  - Extend from cascade model where user may click on **multiple documents**
    - Attraction probability: $\bar{\omega} \in [0,1]^S$
    - termination probability: $\bar{v} \in [0,1]^S$ Position-dependent
- The probability that at least one document in π is satisfactory:

$$f(\pi_t, \omega_t, v) = 1 - \prod_{k=1}^{K} (1 - v(k) w_t(\pi_t(k)))$$

- At each step, select the list that maximizes $f(\pi_t, \omega_t, v)$

TopRank: OL2R by topological sort [LKLS18]

- Observation: no single existing click model captures the behaviors of an entire population of users
- Motivation: to eliminate the dependency on click models, TopRank assumes,

$$P(C_{t,\pi_t(k)}|\pi_t) = \nu(\pi_t, k)$$ an unknown function

  - The click probability **does not** factor into the examination probability of the position and the attractiveness of the documents at that position

TopRank: OL2R by topological sort

Algorithm 1 TopRank

$G_1 \leftarrow \emptyset$ and $c \leftarrow \frac{4\sqrt{2/\pi}}{\mathrm{erf}(\sqrt{2})}$
for $t = 1, \ldots, n$ **do**
  $d \leftarrow 0$
  while $[L] \setminus \bigcup_{c=1}^{d} \mathcal{P}_{tc} \neq \emptyset$ **do**
    $d \leftarrow d + 1$
    $\mathcal{P}_{td} \leftarrow \min_{G_t} \left([L] \setminus \bigcup_{c=1}^{d-1} \mathcal{P}_{tc} \right)$
  Choose A_t uniformly at random from $\mathcal{A}(\mathcal{P}_{t1}, \ldots, \mathcal{P}_{td})$
  Observe click indicators $C_{ti} \in \{0, 1\}$ for all $i \in [L]$
  for all $(i, j) \in [L]^2$ **do**
    $U_{tij} \leftarrow \begin{cases} C_{ti} - C_{tj} & \text{if } i, j \in \mathcal{P}_{td} \text{ for some } d \\ 0 & \text{otherwise} \end{cases}$
    $S_{tij} \leftarrow \sum_{s=1}^{t} U_{sij}$ and $N_{tij} \leftarrow \sum_{s=1}^{t} |U_{sij}|$
  $G_{t+1} \leftarrow G_t \cup \left\{ (j, i) : S_{tij} \geq \sqrt{2 N_{tij} \log\left(\frac{c}{\delta} \sqrt{N_{tij}} \right)} \text{ and } N_{tij} > 0 \right\}$

The difference of the feedback received by document i and j at round t

The cumulative difference of the feedback received by document i and j until t

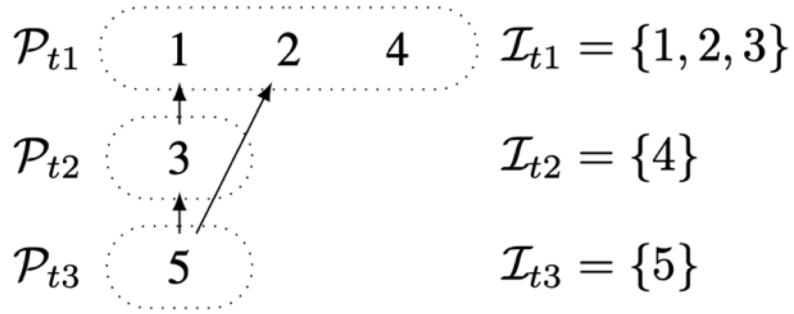
Document pair (j, i) is added to G_t when item i receives sufficiently more clicks than j

TopRank: OL2R by topological sort

Ranking example

$$G_t = \{(3,1),(5,2),(5,3)\}$$

i.e., Sufficient observations are received that document 1 is better than document 3



One possible ranking: 2, 1, 3, 4, 5

Algorithm 1 TopRank

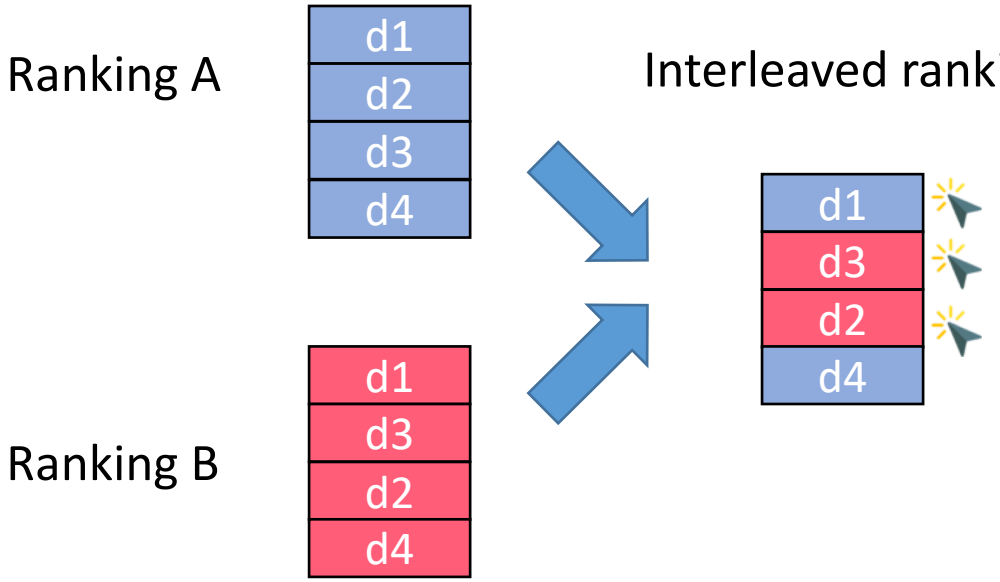
$G_1 \leftarrow \emptyset$ and $c \leftarrow \frac{4\sqrt{2/\pi}}{\mathrm{erf}(\sqrt{2})}$
for $t = 1, \ldots, n$ **do**
  $d \leftarrow 0$
  while $[L] \setminus \bigcup_{c=1}^{d} \mathcal{P}_{tc} \neq \emptyset$ **do**
    $d \leftarrow d + 1$
    $\mathcal{P}_{td} \leftarrow \min_{G_t} \left([L] \setminus \bigcup_{c=1}^{d-1} \mathcal{P}_{tc} \right)$
  Choose A_t uniformly at random from $\mathcal{A}(\mathcal{P}_{t1}, \ldots, \mathcal{P}_{td})$
  Observe click indicators $C_{ti} \in \{0,1\}$ for all $i \in [L]$
  for all $(i,j) \in [L]^2$ **do**
    $U_{tij} \leftarrow \begin{cases} C_{ti} - C_{tj} & \text{if } i,j \in \mathcal{P}_{td} \text{ for some } d \\ 0 & \text{otherwise} \end{cases}$
    $S_{tij} \leftarrow \sum_{s=1}^{t} U_{sij}$ and $N_{tij} \leftarrow \sum_{s=1}^{t} |U_{sij}|$
  $G_{t+1} \leftarrow G_t \cup \left\{ (j,i) : S_{tij} \geq \sqrt{2 N_{tij} \log\left(\frac{c}{\delta}\sqrt{N_{tij}}\right)} \text{ and } N_{tij} > 0 \right\}$

Mini summary

- Cascading bandits
  - Assume user behavior follows cascade model: only one click
  - Maximize the probability that at least one document is attractive
- DCM bandits
  - Assume user behavior follows DCM model to allow multiple clicks
  - Maximize the probability that at least one document is attractive
- TopRank
  - No specific click model assumption
  - Rank documents with topological sort with respect to the confidence of the preference between documents

Dueling bandit gradient descent [YJ09]

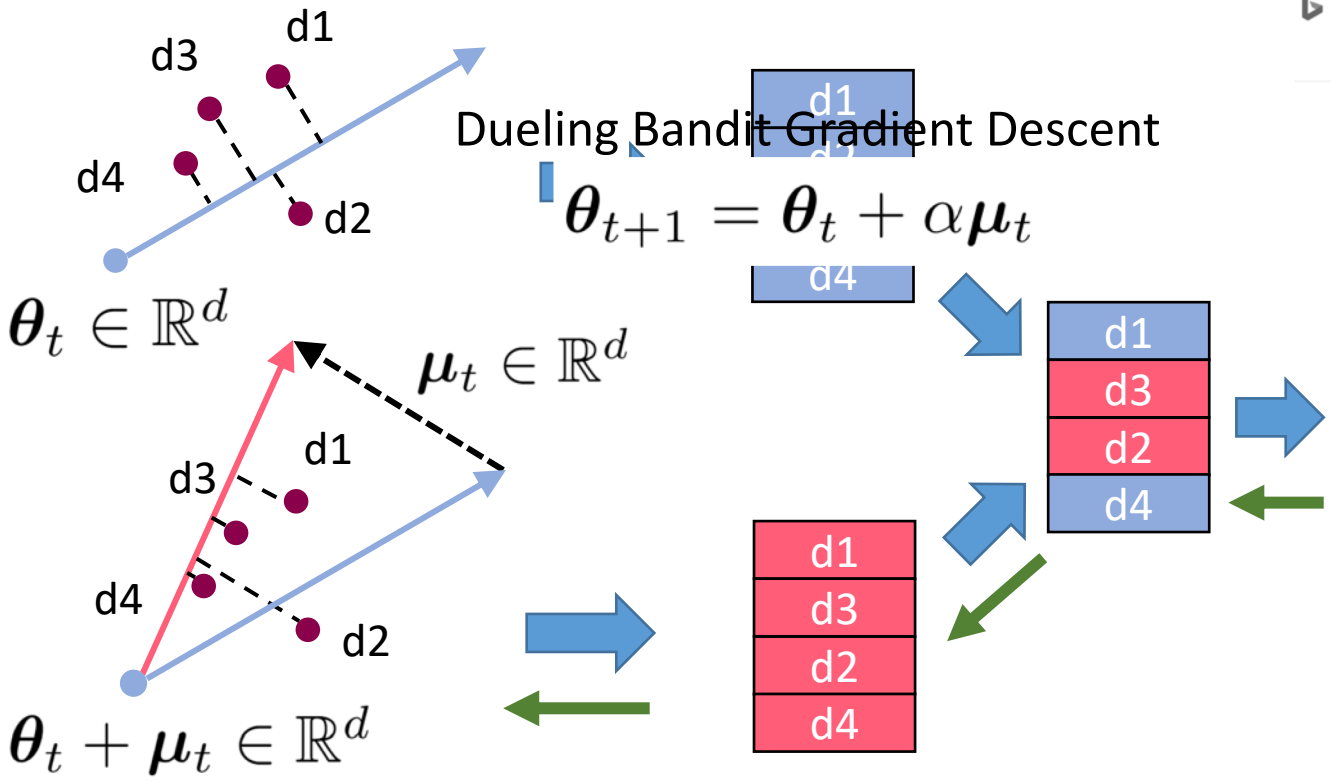
- DBGD is built on **interleaving**, an online evaluation method for rankings
  - Infer preference between two ranking lists based on clicks on the interleaved ranking
- DBGD: online gradient descent based on the inferred preference between models ⟶ explore the model space

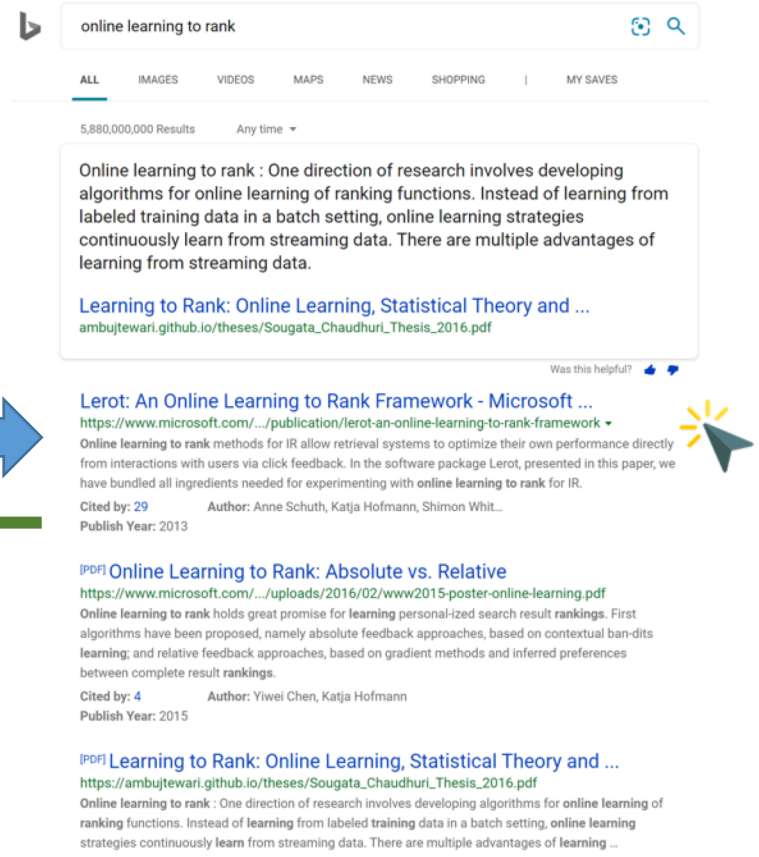


Ranking A receives one click.
Ranking B receives two clicks.

Dueling bandit gradient descent [YJ09]

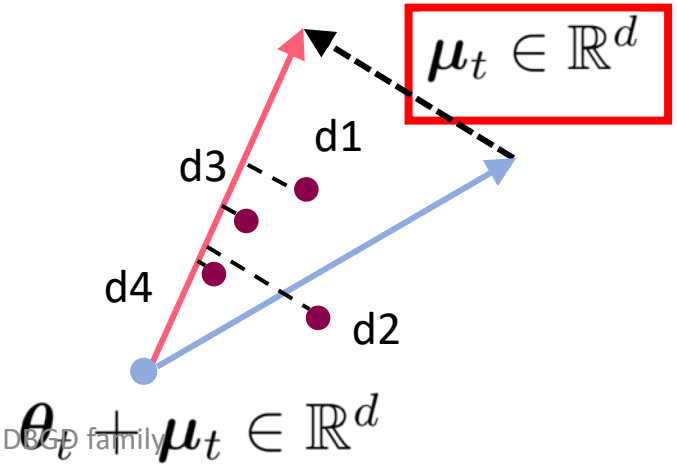
- Interleaving method offers a reliable mechanism for deriving **relative preferences** between retrieval functions





Dueling bandit gradient descent [YJ09]

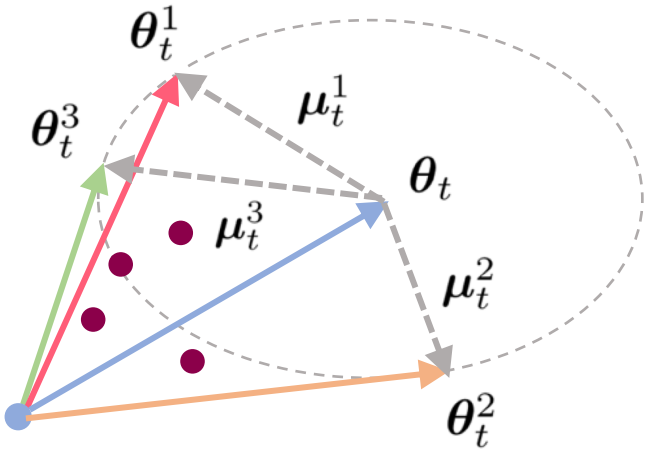
- Unform exploration by random sampling: unbiased gradient estimation
- Perform online gradient descent in expectation
- DBGD has a sublinear regret upper bound:
  - Regret: $\mathbf{R}_T = \sum_{t=1}^{T} \epsilon(\theta^*, \theta_t) + \epsilon(\theta^*, \theta_t') \leq O(\sqrt{d}T^{3/4})$
    - d is the number of feature dimensions, T is the number of interaction rounds
    - ϵ quantify the difference between two models

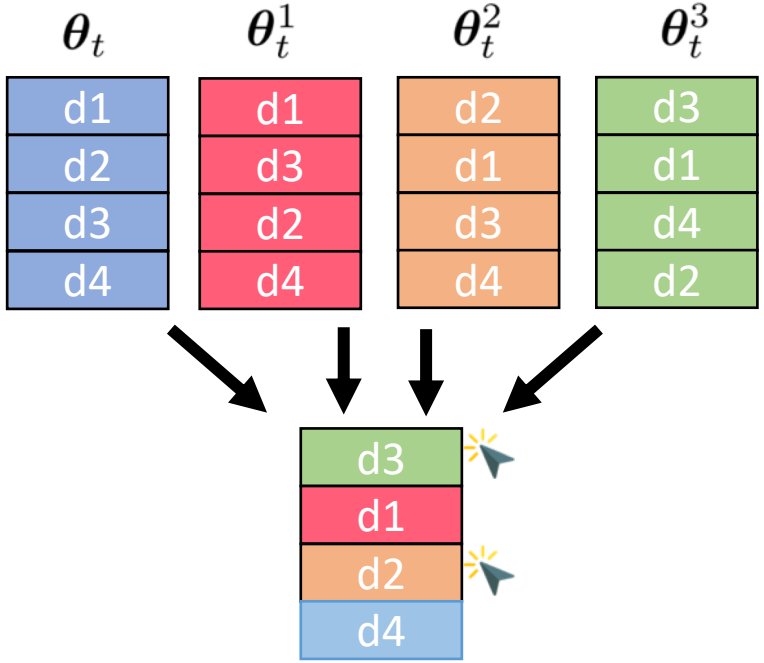


- Problems:
  - One single direction is explored at a time
  - μ_t is uniformly sampled in the d-dimension feature space: high variance and slow convergence

Multileave gradient descent [SOWR16]

- Compared to DBGD, MGD explores multiple directions uniformly from parameter space simultaneously
  - Reduce the updates
- Multileaved comparison among candidate rankers





Winners are θ_t^2 θ_t^3

- Winner takes all: randomly choose one winner

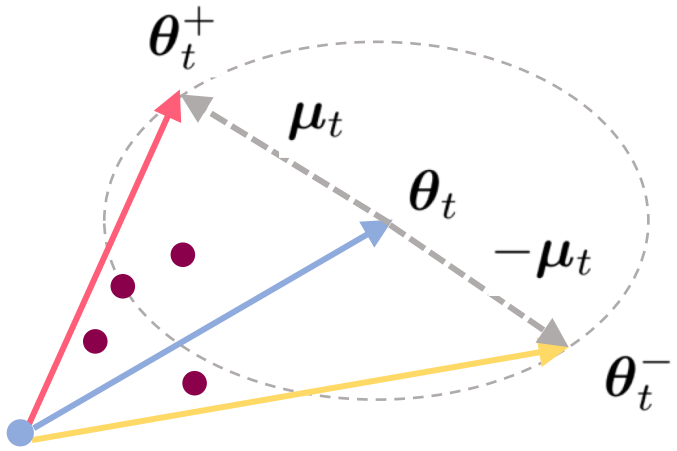
$$\theta_{t+1} = \theta_t + \alpha \mu_t^2$$

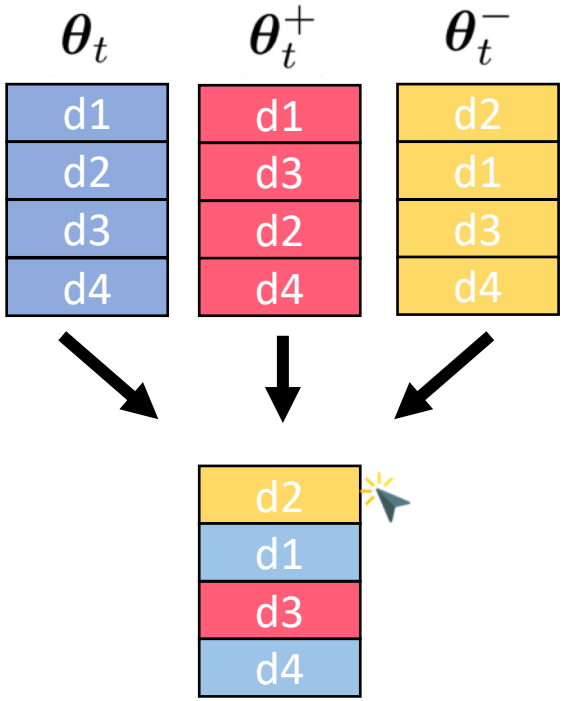
- Mean winner: compute the mean of the winner

$$\theta_{t+1} = \theta_t + \alpha(\mu_t^2 + \mu_t^3)/2$$

Dual-Point DBGD [ZK16]

- Dual-Point DBGD: explore two opposite directions simultaneously
  - Explore more efficiently than DBGD
  - Reduce uncertainty in the exploration with MGD



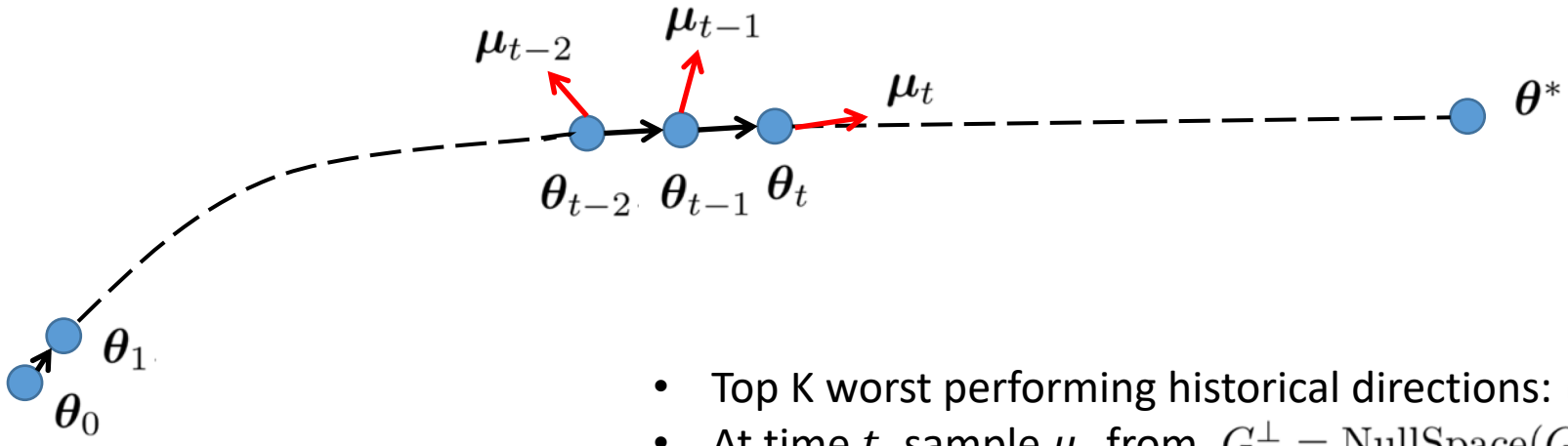


Update the model

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \alpha \boldsymbol{\mu}_t$$

Null space gradient descent [WLKMW18]

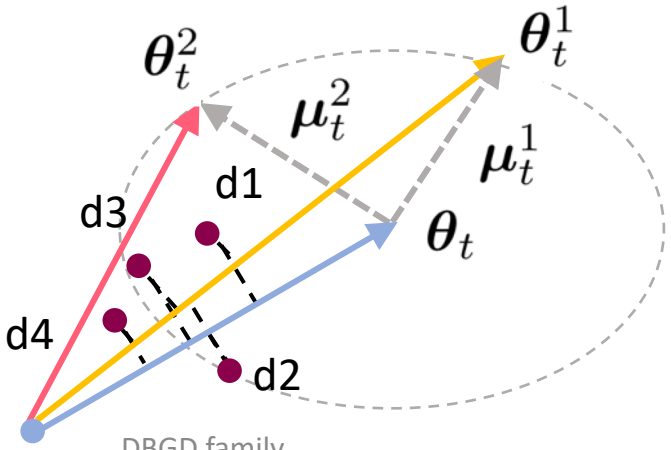
- Intuition: avoid making similar mistakes again
- Null space exploration:
  - Maintain a collection of recently explored gradients that performed poorly
  - Sample new directions from the null space of these gradients
  - Avoid repeatedly exploring poorly performing directions

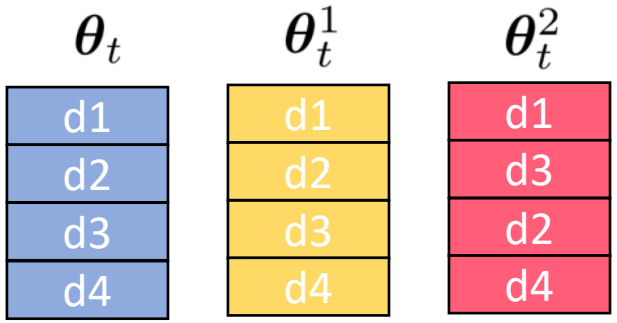


- Top K worst performing historical directions: $G = [\mu^1, \mu^2, ..., \mu^K]$
- At time t, sample μ_t from $G^{\perp} = \text{NullSpace}(G)$

Null space gradient descent [WLKMW18]

- Intuition: avoid making similar mistakes again
- Context-dependent ranker preselection
  - Construct the candidate ranker to maximize the chance that thy can be differentiated from the current ranker
    - At time t, sample $\{\mu_t^i\}_{i=1}^n$ from $G^\perp = \text{NullSpace}(G)$
    - Select top m directions that maximize $|\bar{x}^T \mu_t^i|$ from $\{\mu_t^i\}_{i=1}^n$





- θ_t and θ_t^1 rank the candidate documents in the same order
- No interleaved test can differentiate the ranking quality for current query
- NSGD favors θ_t^2 as it ranks the documents in a different order than θ_t

DBGD-document space projection [WKMWW19]

- Observation: users only examine m documents ($m \ll d$)
- Intuition: only consider the gradient belongs to the document space

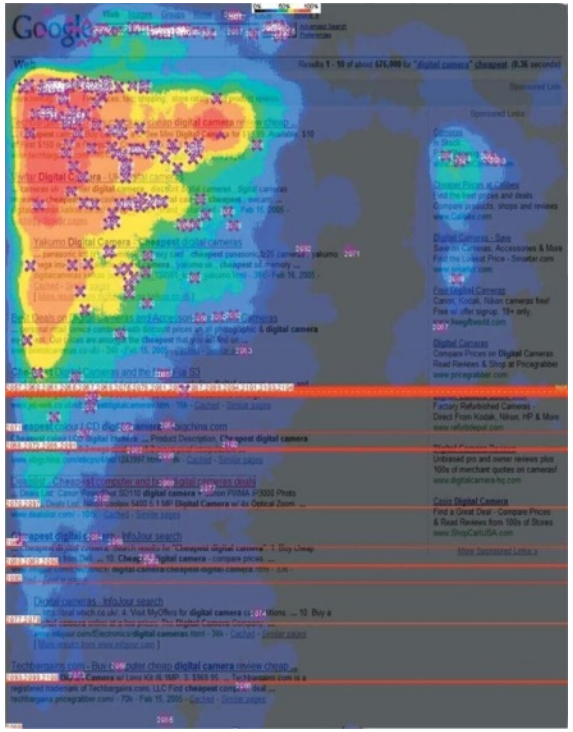


Figure credit: Baeza-Yates, 2018 [3]

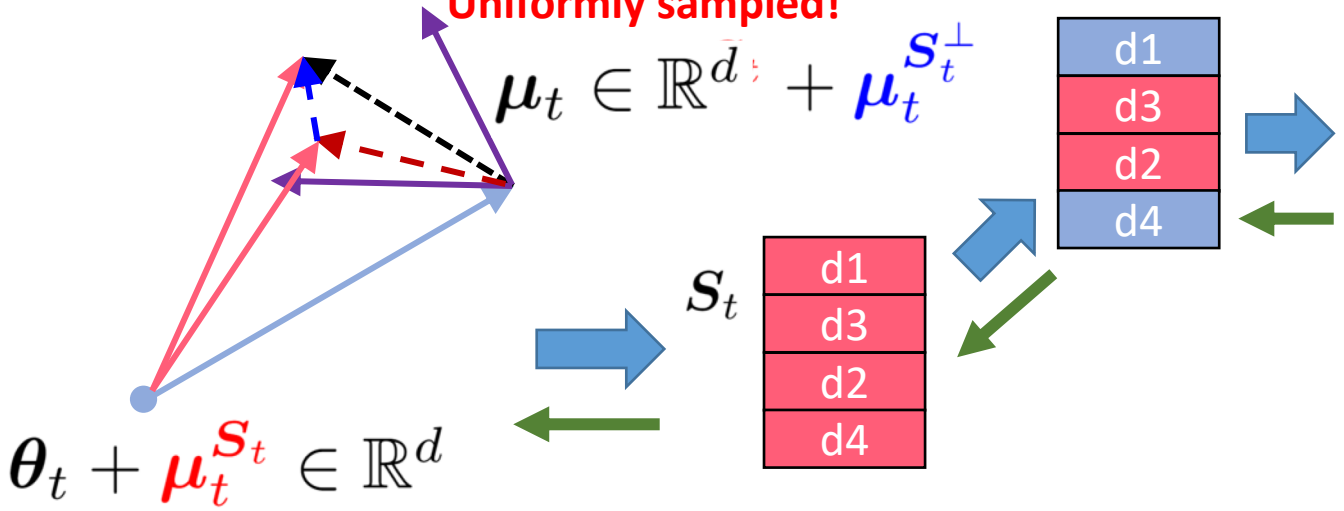
DBGD-document space projection [WKMWW19]

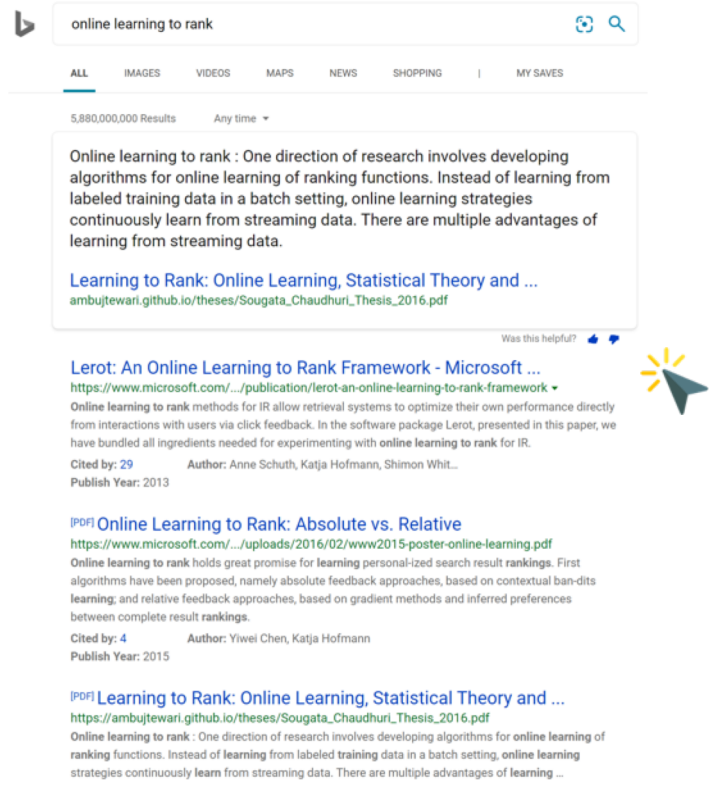
- Observation: users only examine m documents ($m \ll d$)
- Intuition: only consider the gradient belongs to the document space

Considerable regret reduction:

$$\mathcal{O}(\sqrt{d}T^{3/4}) \Rightarrow \mathcal{O}(\sqrt{m}T^{3/4})$$

Uniformly sampled!





Mini summary

- Explore in the entire model space
  - DBGD: randomly sample one single direction
  - MGD: randomly sample multiple directions
  - DP-BGD: randomly sample two opposite directions
  - NSGD: maintain a set of “bad” directions, and sample from the null space of them
  - DBGD-DSP: project the gradient into document space and use it to update model
- DBGD and its extensions
  - Explore in the entire model space
  - High variance and slow convergence
  - The assumption for theoretical analysis does not hold for all models
    - There is a single optimal model
    - The utility space is smooth

PDGD [OR18]

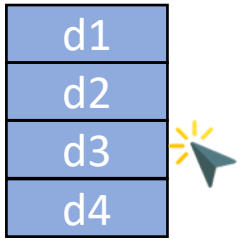
- Pairwise differentiable gradient descent
  - Optimize a Plackett Luce ranking model, which models a probabilistic distribution over documents
  - Infer the preference between document pairs
  - Intuition: provide an unbiased pairwise gradient to update the ranking model

PDGD [OR18]

- At each time t
  - Observe user query
  - Sample a ranking from the document distribution:

$$P(d|D,\theta) = \frac{\mathbf{e}^{f_{\theta_t}(d)}}{\sum_{d' \in D} \mathbf{e}^{f_{\theta_t}(d')}}$$

  - Present the ranking to the user
  - Infer pairwise preference from user clicks



$$d_3 > d_1, d_3 > d_2, d_3 > d_4$$

$$\nabla f_{\theta_t}(\cdot) = \sum_{d_i >_c d_j} \nabla P(d_i \succ d_j | D, \theta_t)$$

d_i receives click and d_j does not receive click.

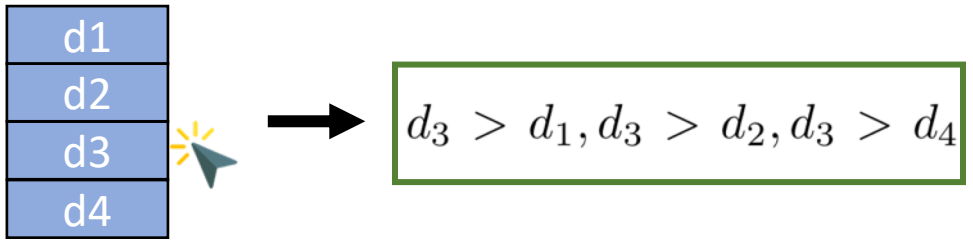
Biased!!

Some preferences are more likely to be observed due to position/selection bias.

PDGD [OR18]

- At each time t
  - Observe user query
  - Sample a ranking from the document distribution:

$$P(d|D,\theta) = \frac{\mathbf{e}^{f_{\theta_t}(d)}}{\sum_{d' \in D} \mathbf{e}^{f_{\theta_t}(d')}}$$



  - Present the ranking to the user
  - Infer pairwise preference from user clicks
  - Update model according to the estimated unbiased pairwise gradient:

Weighting function to deal with bias: the ratio between the probability of the ranking and the reversed pair ranking

$$\nabla f_{\theta_t}(\cdot) = \sum_{d_i >_c d_j} \rho(d_i, d_j, \pi_t) \nabla P(d_i \succ d_j | D, \theta_t)$$

$$\rho(d_i, d_j, \pi_t) = \frac{P(\pi^*(d_i,d_j,\pi_t)|f,D)}{P(\pi_t|f,D)+P(\pi^*(d_i,d_j,\pi_t)|f,D)}$$

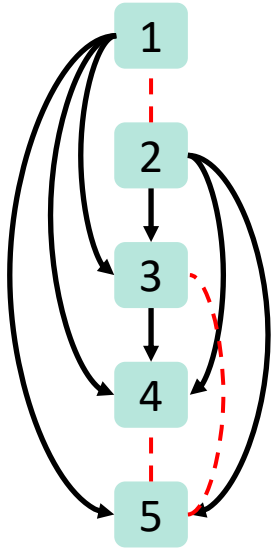
Compared to π_t, d_i and d_j are swapped.

The presented ranking

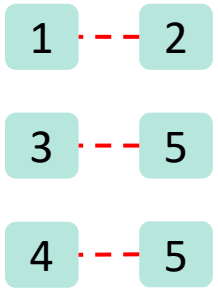
PairRank [JWGW21]

- Online learning to rank by divide-and-conquer
- Key insights:
  - A complete ranking can be decomposed into a series of pairwise comparisons
  - Only explore the pairs the ranker is currently still uncertain about the order
    - Divide-and-Conquer
  - Reduce exponentially sized action space to quadratic

Document graph



Uncertain rank orders



Explore the uncertain rank orders
Exploit the certain rank orders

Pairwise learning to rank

- Learn a pairwise ranking model online
- Ranking model: a single layer RankNet[10] model with sigmoid activation function
  - Loss function at time t:

Regularization term

Training data observed so far

$$L_t = \sum_{s=1}^{t} \sum_{(i,j)\in\Omega_t} -y_{ij}^s \log(\sigma({x_{ij}^s}^\top \theta)) - (1 - y_{ij}^s) \log(1 - \sigma({x_{ij}^s}^\top \theta)) + \frac{\lambda}{2}\|\theta\|^2$$

y_{ij}^s: whether the document i is preferred over document j in the click feedback

$x_{ij}^s = x_i^s - x_j^s$: feature difference between document i and document j

Sigmoid function to model the pairwise preference probability: $\sigma(s) = \frac{1}{1+\exp(-s)}$

Pairwise estimation uncertainty

- Pairwise feedback is noisy:
  - Given the documents are examined, the click feedback is independent from each other

$$y_{ij}^s = \sigma({x_{ij}^s}^\top \theta^*) + \epsilon_{ij}^s$$

- Confidence interval of pairwise preference estimation
  - At round t, with probability at least $1 - \delta_1$,

$$\left|\sigma({x_{ij}^t}^\top \theta_t) - \sigma({x_{ij}^t}^\top \theta^*)\right| \leq \alpha_t \|x_{ij}^t\|_{\mathbf{M}_t^{-1}}$$

Estimated pairwise preference between i and j

Ground truth pairwise preference between i and j

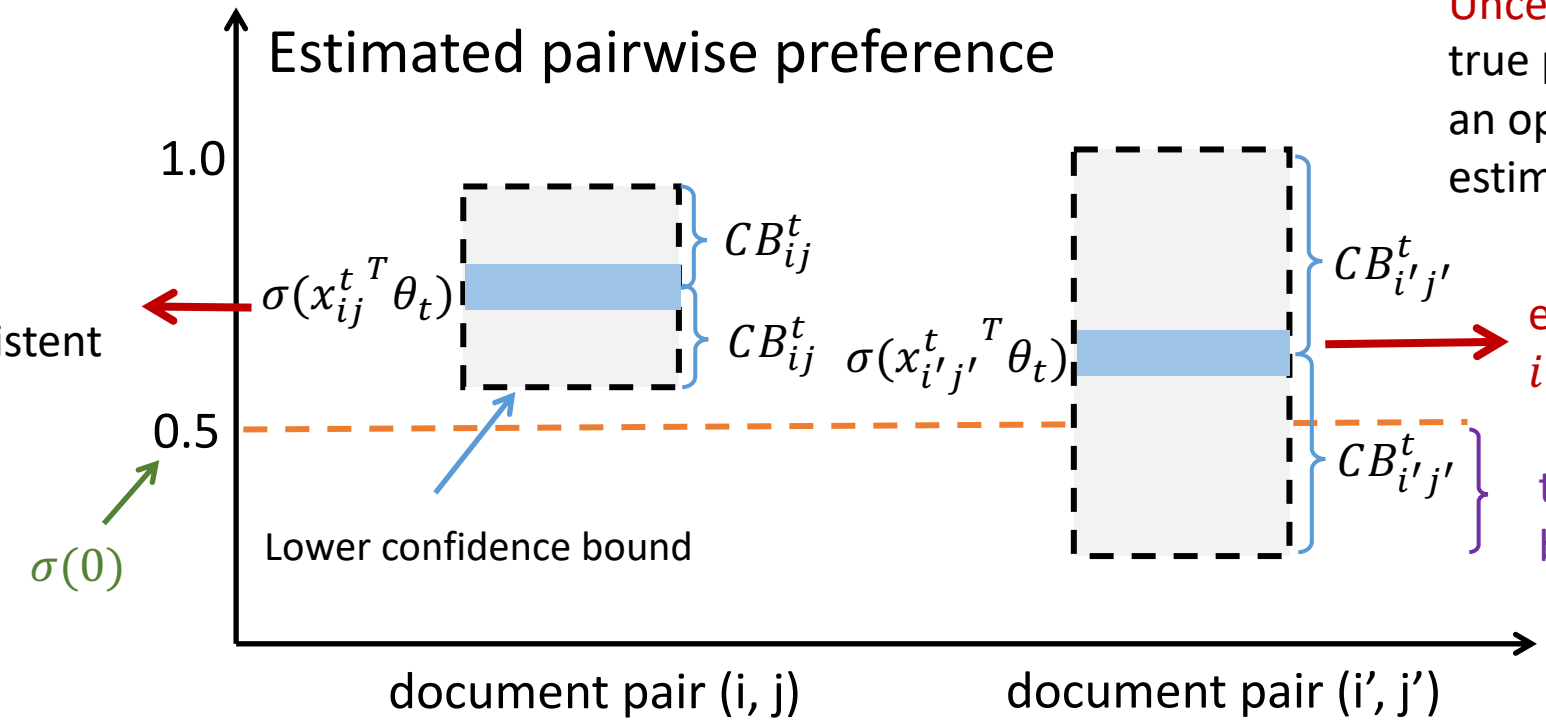
Covariance matrix constructed with previous observations

$$\mathbf{M}_t = \sum_{s=1}^{t-1} \sum_{(i,j)\in\Omega_s} x_{ij}^s {x_{ij}^s}^\top + \lambda \mathbf{I}$$

Certain rank order & uncertain rank order

- With high probability, the optimal value belongs to

Pairwise preference estimation $\left[\sigma(x_{ij}^{t\top}\theta_t) - CB_{ij}^t, \sigma(x_{ij}^{t\top}\theta_t) + CB_{ij}^t\right]$ Estimation uncertainty



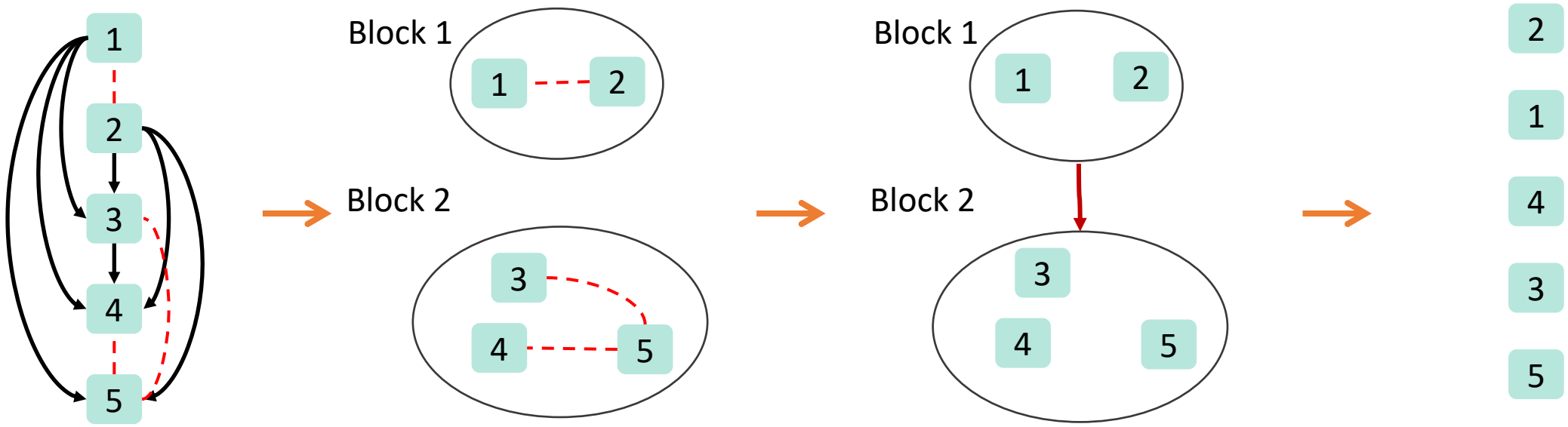
Uncertain rank order:
true preference may present an opposite order from the estimated order.

Certain rank order:
true preference is consistent with the estimated preference: $i \succ j$.

estimated preference: $i' \succ j'$

true preference might be: $j' \succ i'$

PairRank: explore via divide and conquer



Document graph

Uncertain order graph

Block graph

Ranked list
(possible one)

Constructed with the certain and uncertain rank orders between document pairs.

Explore the uncertain rank orders: construct blocks with the connected components.

Topological sort between blocks: exploit the certain rank orders.

Randomly shuffle the order of documents within each block: explore the uncertain rank orders.

Pairwise regret

- Pairwise regret for OL2R
  - The cumulative number of mis-ordered pairs from the presented ranking to the ideal one, i.e., the Kendall tau rank distance

$$\mathbf{R}_T = \mathbb{E}[\sum_{t=1}^{T} r_t] = \mathbb{E}[\sum_{t=1}^{T} K(\pi_t, \pi_t^*)]$$

  - Most ranking metrics deployed in real-world retrieval systems, such as ARP and NDCG, can be decomposed into pairwise comparisons
- Sublinear upper regret bound of PairRank: $\mathbf{R}_T \leq O(d \log^4(T))$

OL2R vs. offline unbiased L2R

- Goal: to find the best models that rank documents based on their utility
- Learn from user interactions, implicit feedback, e.g., clicks

OL2R vs. offline unbiased L2R

- **OL2R**
  - Interactively optimize and update a ranking model after every interaction
  - Combat bias by interventions, i.e., exploration
  - User experience might be hurt due to exploration
- **Offline unbiased L2R**
  - Learn a ranking model from a historical interaction log
  - Remove data bias by re-weighting strategies
  - There is no presentation bias
  - Do not affect user experience but cannot explore and limited to rankings in the historical log

OL2R vs. offline unbiased L2R

- To model or to intervene [JOR19]
  - Compare the counterfactual L2R and OL2R methods under different experimental conditions
  - Performance of OL2R and counterfactual L2R depend on the presence of selection bias, the degree of position bias and interaction noise
    - Counterfactual method performs best when there is litter bias or noise in the feedback
    - OL2R methods are more robust to bias and noise, but they may hurt user experience
- Unbiased learning to rank: online or offline [AYWM21]
  - Are counterfactual L2R and OL2R are two sides of the same coin for unbiased L2R?
  - Almost all unbiased L2R algorithms in offline learning can be directly applied to online learning

Unifying online and counterfactual learning to rank [OR21]

- Intervention-aware estimator
  - Bridge the online and counterfactual L2R divisions
- Key insights
  - Use the offline policy-aware estimator to correct position bias, presentation bias and user trust bias
  - Online intervention: take the entire collection of logging policies into consideration

Unifying online and counterfactual learning to rank [OR21]

- Intervention-oblivious estimator
  - Clicks follow an affine model, for item d displayed at rank k:

$$P(C = 1|d, k) = \alpha_k P(R = 1|d) + \beta_k$$

  - Conditioned on logging policy π, the click probability is:

$$P(C = 1|d, \pi) = \sum_{k=1}^{K} \pi(k|d)(\alpha_k P(R = 1|d) + \beta_k)$$

  - The estimator is based on the inverse:

$$P(R = 1|d) = \frac{P(C=1|d,\pi) - \mathbb{E}_k[\beta_k|d,\pi]}{\mathbb{E}_k[\alpha_k|d,\pi]}$$

Unifying online and counterfactual learning to rank [OR21]

- Intervention-aware estimator
  - Clicks follow an affine model, for item d displayed at rank k:

$$P(C = 1|d, k) = \alpha_k P(R = 1|d) + \beta_k$$

  - Conditioned on **the set of logging policies $\boldsymbol{\Pi}$**, the click probability is:

$$P(C = 1|d, \pi) = \frac{1}{|\Pi|} \sum_{\pi_t \in \Pi} \sum_{k=1}^{K} \pi_t(k|d)(\alpha_k P(R = 1|d) + \beta_k)$$

  - The estimator is based on the inverse:

$$P(R = 1|d) = \frac{P(C=1|d,\Pi) - \mathbb{E}_k[\beta_k|d,\Pi]}{\mathbb{E}_k[\alpha_k|d,\Pi]}$$

Open questions

- How to balance the efficiency and effectiveness of OL2R?
  - Online stochastic gradient descent
  - Perturbation-based/randomization-based exploration

References VI

- [KSWA15] Kveton, Branislav, Csaba Szepesvari, Zheng Wen, and Azin Ashkan. “Cascading Bandits: Learning to Rank in the Cascade Model.” In *International Conference on Machine Learning*, 767–76. PMLR, 2015.
- [KKSW16] Katariya, Sumeet, Branislav Kveton, Csaba Szepesvari, and Zheng Wen. “DCM Bandits: Learning to Rank with Multiple Clicks.” In *International Conference on Machine Learning*, 1215–24. PMLR, 2016.
- [LKLS18] Lattimore, Tor, Branislav Kveton, Shuai Li, and Csaba Szepesvári. "TopRank: A practical algorithm for online stochastic ranking." In *NeurIPS*. 2018.
- [YJ09] Yue, Yisong, and Thorsten Joachims. "Interactively optimizing information retrieval systems as a dueling bandits problem." In *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 1201-1208. 2009.
- [SOWR16] Schuth, Anne, Harrie Oosterhuis, Shimon Whiteson, and Maarten de Rijke. "Multileave gradient descent for fast online learning to rank." In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, pp. 457-466. 2016.
- [ZK16] Zhao, Tong, and Irwin King. "Constructing reliable gradient exploration for online learning to rank." In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pp. 1643-1652. 2016.
- [WLKMW18] Wang, Huazheng, Ramsey Langley, Sonwoo Kim, Eric McCord-Snook, and Hongning Wang. "Efficient exploration of gradient space for online learning to rank." In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pp. 145-154. 2018.
- [WKMWW19] Wang, Huazheng, Sonwoo Kim, Eric McCord-Snook, Qingyun Wu, and Hongning Wang. "Variance reduction in gradient exploration for online learning to rank." In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*
- [OR18] Oosterhuis, Harrie, and Maarten de Rijke. "Differentiable unbiased online learning to rank." In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pp. 1293-1302. 2018.
- [JWGW21] Jia, Yiling, Huazheng Wang, Stephen Guo, and Hongning Wang. "PairRank: Online Pairwise Learning to Rank by Divide-and-Conquer." In *Proceedings of the Web Conference 2021*, pp. 146-157. 2021.

References VII

- [JOR19] Jagerman, Rolf, Harrie Oosterhuis, and Maarten de Rijke. "To model or to intervene: A comparison of counterfactual and online learning to rank from user interactions." In *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval*
- [AYWM21] Ai, Qingyao, Tao Yang, Huazheng Wang, and Jiaxin Mao. "Unbiased Learning to Rank: Online or Offline?." *ACM Transactions on Information Systems (TOIS)* 39, no. 2 (2021): 1-29.
- [OR21] Oosterhuis, Harrie, and Maarten de Rijke. "Unifying Online and Counterfactual Learning to Rank: A Novel Counterfactual Estimator that Effectively Utilizes Online Interventions." In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pp. 463-471. 2021.

Outline of this tutorial

- Motivation
- Background: bandit algorithms
- Bandit learning for recommender systems
- Bandit learning for retrieval systems
- Ethical considerations in IR with bandit learning
- Conclusion & future directions

Ethical considerations

- Privacy concerns
  - Background: differential privacy (for continual release)
  - Global and local differentially private bandit learning
- Fairness concerns
  - Meritocratic fairness
  - Merit-based fair exposure
- Safety and security concerns

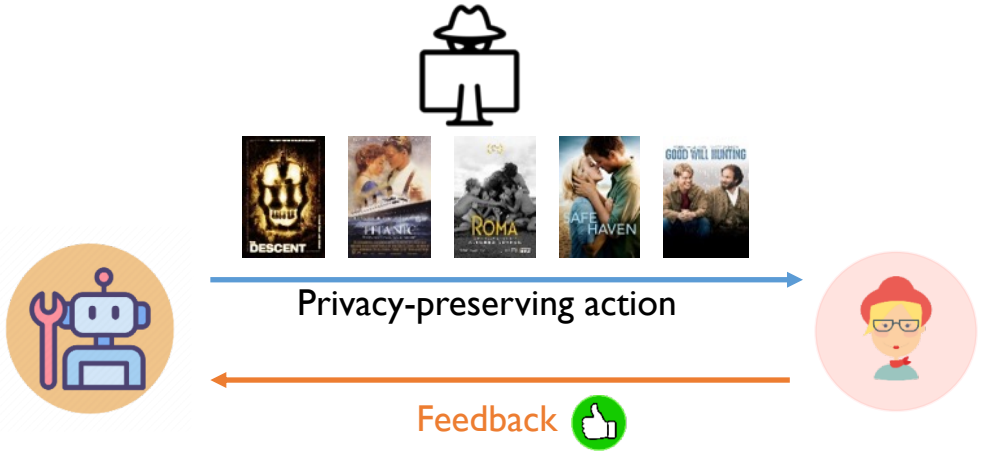
Ethical considerations

- Privacy concerns
  - Background: differential privacy (for continual release)
  - Global and local differentially private bandit learning
- Fairness concerns
  - Meritocratic fairness
  - Merit-based fair exposure
- Safety and security concerns: exploration with constraints
  - Regret constraint: conservative exploration [WSLS16, KGYR17]
  - Side constraint: [AAT19, KB20]
  - Will not cover the details due to time limit

Privacy concerns

- "A Face Is Exposed for AOL Searcher No. 4417749" [BZH06]
- "Robust De-anonymization of Large Datasets (How to Break Anonymity of the Netflix Prize Dataset)" [NS08]
- Tutorial "Differential Privacy for Information Retrieval" @ WSDM 2018
- In bandit learning: privacy for reward (and context) under extraction attacks
- Idea: exploration reconciles the need for learning and the need for privacy protection

[CKNFS11, Kor10]



Adversary tries to extract user's reward from recommendation sequence

Differential privacy [Dwo06]

- A randomized mechanism $A: X \to Y$ is (ϵ, δ)-differentially private if for all neighboring inputs X' and for all sets of outputs $\mathrm{O} \subseteq Y$,

$$P(A(X) \in O) \leq e^{\epsilon} P(A(X') \in O) + \delta$$

  - $\delta = 0$: ϵ-DP
- Intuition: cannot differentiate whether a data point is presented

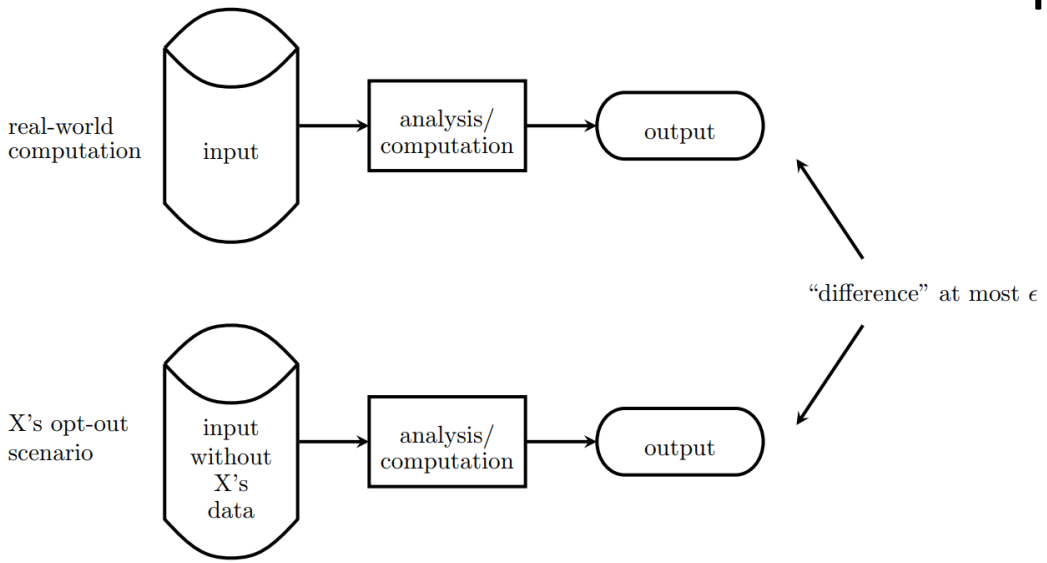
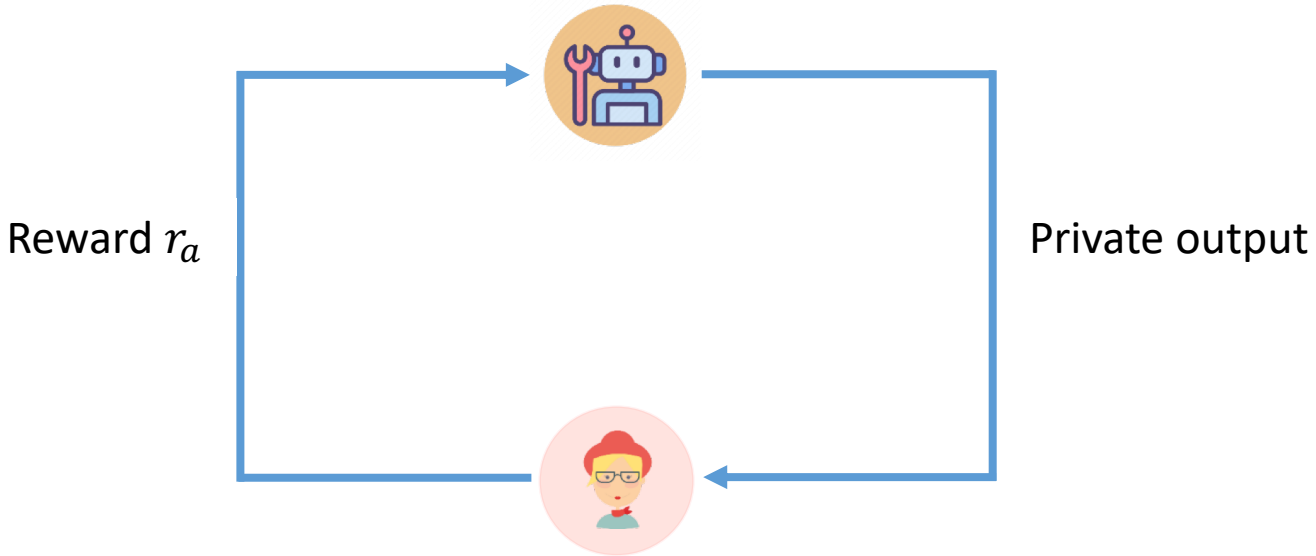


Figure from [Ngu19]

Differential privacy for continual observations

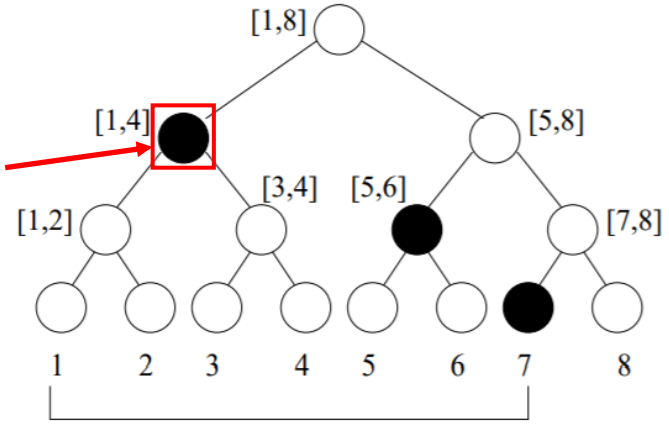
- Data stream has T examples, $r_i \in [0, 1]$
- Goal: Privately output the sum of first k data α_k, $k \in [1..T]$
- Aggregate k data by tree representation $a_i..b_i$
- Output private statistics of $\sum \alpha_{a_i..b_i}$



Tree-based aggregation [CSS10, DNPR10]

- Represent sum of r_1 to r_7 as $\alpha_{1..4} + \alpha_{5..6} + \alpha_7$
  - Separate into $\log(T)$ partial sums

Sum of r_1 to r_4 with $Lap\left(\frac{\log T}{\epsilon}\right)$



- Every node is a private partial sum with noise $Lap\left(\frac{\Delta \log T}{\epsilon}\right)$: $\frac{\epsilon}{\log T}$-DP
  - By composition theorem the total sum is ϵ-DP
- Noise (error) in private sum is bounded by $O\left(\frac{\log^{1.5} T}{\epsilon}\right)$

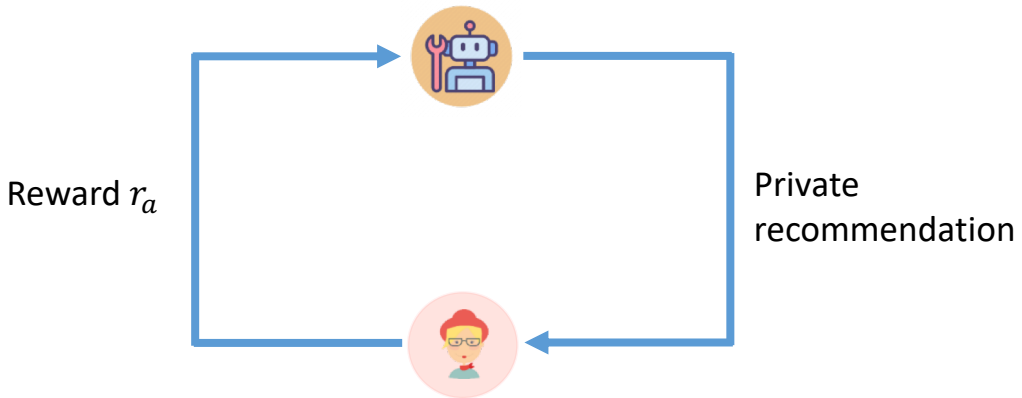
Differentially private UCB1 [MT15]

- Idea: keep the averaged reward $\hat{r}_{a,t} = \frac{1}{N_a}\sum_{\tau=1}^{N_a} r_{a,\tau}$ private
  - Post-processing invariant
- Use tree-based aggregation, add noise to $\sum_{\tau=1}^{N_a} r_{a,\tau}$ for private $\hat{r}^p_{a,t}$
- Arm selection strategy:

$$\arg\max_a \hat{r}^p_{a,t} + \mathrm{CB}(a,t) + O(\frac{\log^{1.5} N_{a,t}}{\epsilon})$$

  - Additional confidence term for exploration in private setting

- Regret: $O(\frac{k\log^2 T}{\epsilon\Delta})$



Differentially private LinUCB

- Privacy for reward [NR18]
- Use tree-based aggregation, add noise to b_t
- Arm selection strategy:

$$\arg\max_a \hat{r}^p_{a,t} + \mathrm{CB}(a,t) + O\left(\frac{\log^{1.5} N_{a,t} \log \frac{K}{\delta}}{\epsilon}\right)$$

- Regret: $O\left(K\sqrt{T}\log T + \frac{\sqrt{T}\log^{2.5} T}{\epsilon}\right)$

Closed form estimator of LinUCB

$$\mathbf{A}_t = \lambda \mathbf{I} + \sum_{(a_i,r_i)\in\mathcal{H}_t} \mathbf{x}_{a_i}\mathbf{x}_{a_i}^\mathsf{T}$$
$$\mathbf{b}_t = \sum_{(a_i,r_i)\in\mathcal{H}_t} r_{a_i}\mathbf{x}_{a_i}$$
$$\hat{\boldsymbol{\theta}}_t = \mathbf{A}_t^{-1}\mathbf{b}_t$$

Red terms are due to privacy mechanism

Differentially private LinUCB

- Privacy for both context and reward [SS18]
- Use tree-based aggregation, add noise to A_t and b_t
- Arm selection strategy: $\arg\max_a \hat{r}^p_{a,t} + \mathrm{CB}^p(a,t)$
- Regret: $O(\frac{\log T}{\Delta} + \frac{\log T \sqrt{d \log \frac{1}{\delta}/\epsilon}}{\Delta})$
- Also showed a matched gap-dependent lower bound
  - Constructing the lower bound based on definition of privacy for context

Closed form estimator of LinUCB

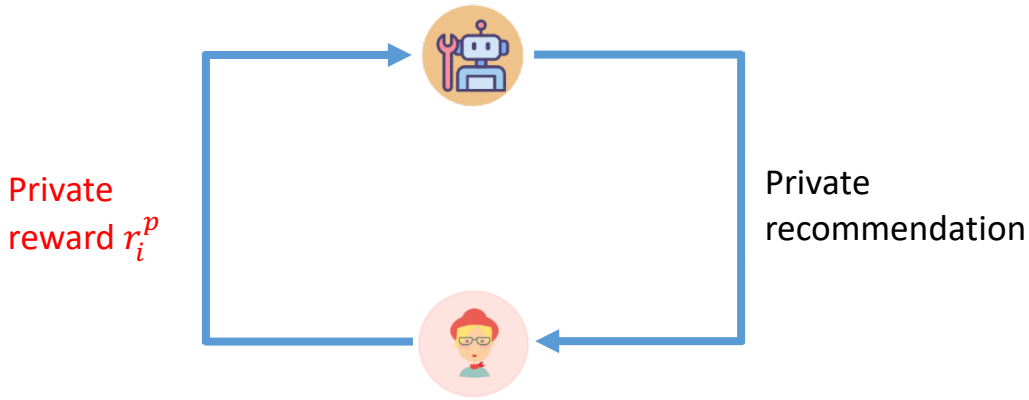
$$\mathbf{A}_t = \lambda \mathbf{I} + \sum_{(a_i, r_i) \in \mathcal{H}_t} \mathbf{x}_{a_i} \mathbf{x}_{a_i}^\mathsf{T}$$

$$\mathbf{b}_t = \sum_{(a_i, r_i) \in \mathcal{H}_t} r_{a_i} \mathbf{x}_{a_i}$$

$$\hat{\boldsymbol{\theta}}_t = \mathbf{A}_t^{-1} \mathbf{b}_t$$

Local differential privacy

- DP: user sent data to central server, and server adds noise to the aggregated result
  - Concerns: Data communication or even the center can be compromised
- LDP: Data is randomized on the user side before sent to aggregator
- LDP is a stronger privacy definition
  - Larger cost / regret is expected



UCB with local differential privacy

- LDP-UCB-Laplace [RZLS20]
  - User sends noisy feedback $r_{a,t} + Lap\left(\frac{1}{\epsilon}\right)$ to the server
- Arm selection strategy:

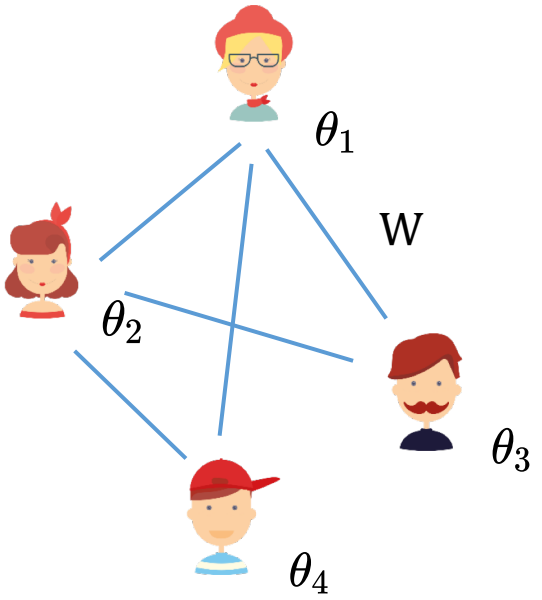
$$\arg\max_a \hat{r}^p_{a,t} + \mathrm{CB}(a,t) + O\left(\sqrt{\frac{32\log T}{\epsilon^2 N_{a,t}}}\right)$$

  - Added confidence term for exploration in private setting

- Regret: $O\left(\frac{k\log T}{\epsilon^2 \Delta}\right)$

Private Collaborative Bandits [WZWCKW20]

- Main idea: add Laplace noise $Lap\left(\frac{\Delta \log T}{\epsilon}\right)$ to the reward during model estimation;
  - Scale the noise based on sensitivity Δ and privacy budget ϵ
- Collaborative learning: calibrate sensitivity with respect to the user dependency structure W
  - $\Delta = \max_i 2\|W_i\|_2$
  - Vanilla (independent users) setting assumes rewards in [-1, 1], so $\Delta = 2$

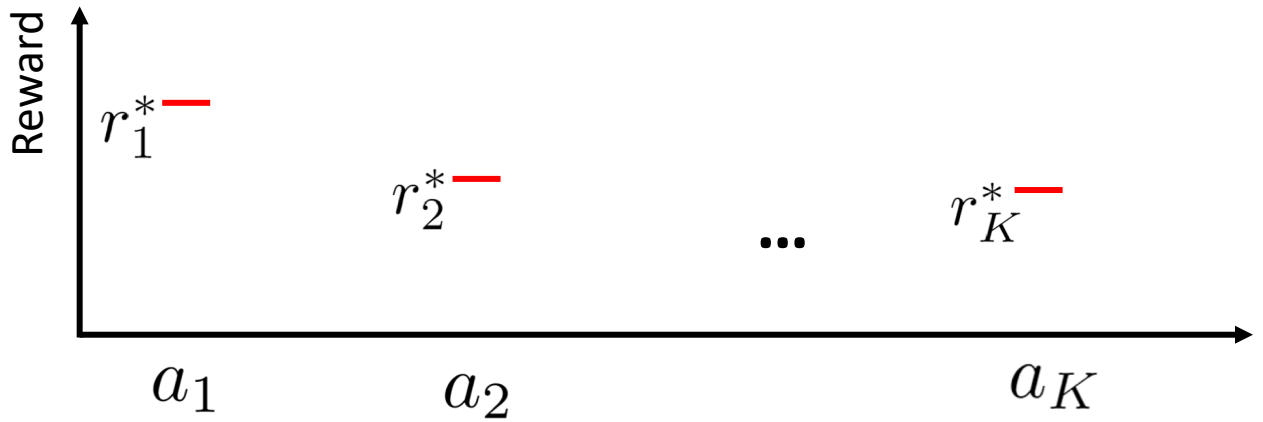


Fairness concerns

- Algorithmic bias – an important topic
  - Research of bias in data, model, algorithm etc.
    - E.g., discriminatory treatment of subpopulations
  - The need to explore
- Fairness guarantee during online decision making
  - Fair recommendation / fair LTR
- Many literatures in offline learning setting
  - Check "Tutorial on Fairness of Machine Learning in Recommender Systems" @ SIGIR 2021

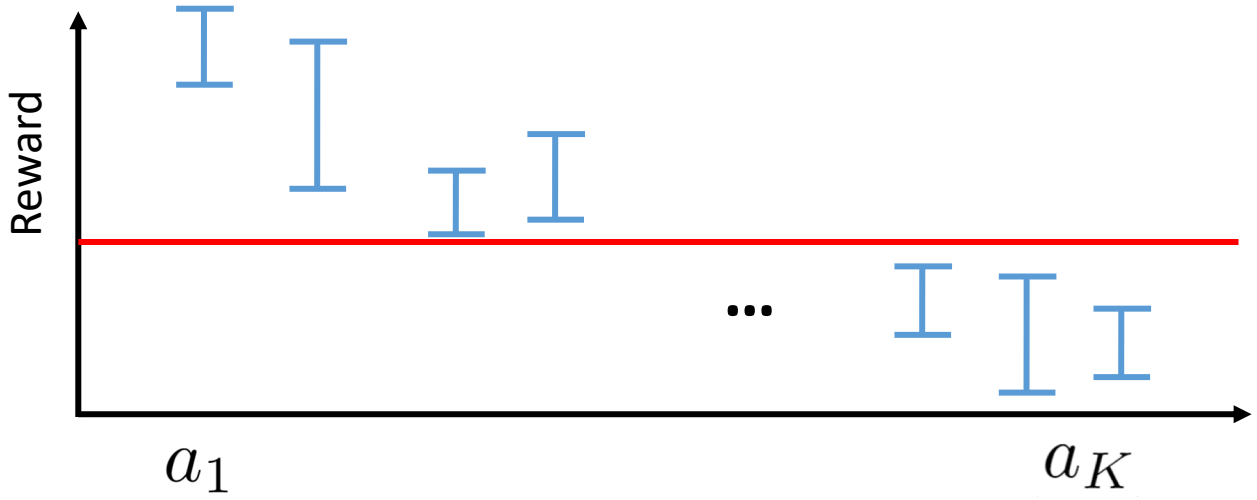
Weakly meritocratic fairness [JKMNR16]

- Fairness definition: if reward $r_a^* \geq r_b^*$ then $P(a) \geq P(b)$
  - A fair bandits should never favor a worse arm at any round
- Prefect strategy is fair: $P(a^*) = 1$ -- but we don't which arm is perfect at the beginning
- Uniformly random is fair: $P(a) = \frac{1}{K}, \forall a$ -- linear regret
- Somewhere in between?



FairUCB [JKMR16]

- Idea: uniformly pull arm within the first confidence interval chain
  - Start from the largest UCB, find overlapped confidence intervals
- Guaranteed fairness at every step with high probability
- Regret: $R(T) = O\left(\sqrt{k^3 T \ln \frac{Tk}{\delta}}\right)$



Merit-based fair exposure

- Fairness definition: given merit function f ,

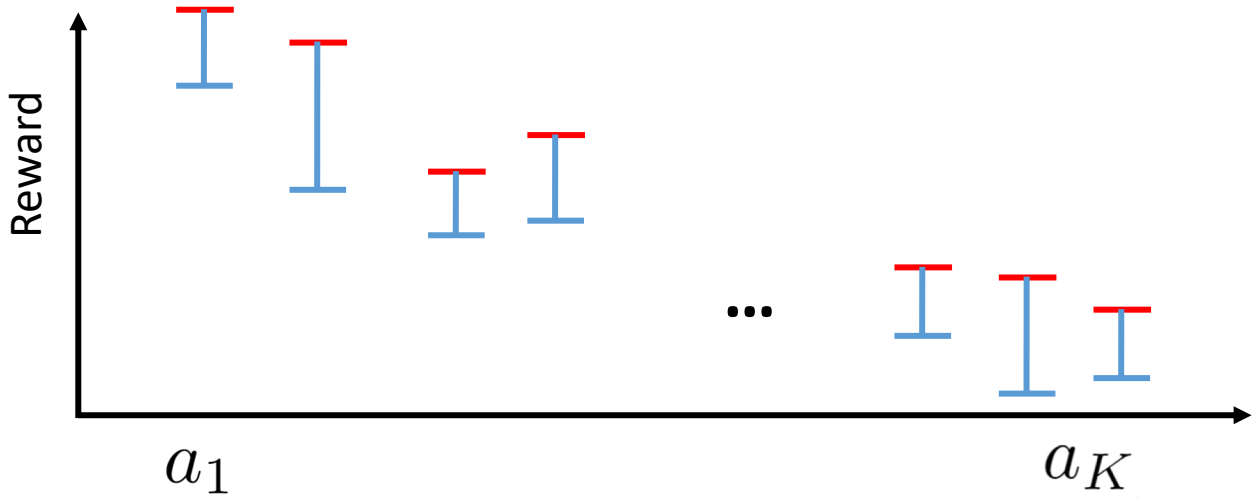
$$\frac{P(a)}{P(b)} = \frac{f(r_a^*)}{f(r_b^*)}$$

- Intuition: exposure should be proportional to the merit
- Compare with previous fairness definition: prefect strategy with $P(a^*) = 1$ is no longer fair

FairX-UCB [WBSJ21]

- Idea: pull arm proportional to the merit $f(\tilde{r}_a)$
- $\tilde{r}$ is an optimism reward prediction satisfying fairness constraint

$$\tilde{r} = \arg\max_{r \in C} \frac{\sum_a f(r_a) r_a}{\sum_a f(r_a)}, C = \{r : r_a \in [\hat{r}_a - \mathrm{CB}(a), \hat{r}_a + \mathrm{CB}(a)]\}$$



FairCo [MSHJ20]

- Controlling Fairness and Bias in Dynamic Learning-to-Rank
- Divide documents into groups $\{G_i\}$
  - Group fairness / individual fairness
- Exposure:

$$E_t(G_i) = \frac{\sum_{d \in G_i} p_t(d)}{|G_i|}$$

  - Averaged examination probability
- Merit function:

$$f(G_i) = \frac{\sum_{d \in G_i} r(d)}{|G_i|}$$

  - Averaged relevance
- Fairness: exposure should be proportional to the merit

$$\frac{E_t(G_i)}{E_t(G_j)} = \frac{f(G_i)}{f(G_j)}$$

FairCo [MSHJ20]

- Fairness: exposure should be proportional to the merit
- Idea: fairness constraints as an added error term
  - Ranking list $\sigma = \arg\operatorname*{sort}_d(\hat{r}(d|query) + \lambda err_T(d))$
- Error term is the exposure-based fairness disparity between two groups

$$\frac{\frac{1}{T}\sum_t E_t(G_i)}{f(G_i)} - \frac{\frac{1}{T}\sum_t E_t(G_j)}{f(G_j)}$$

Marginal fairness [YA21]

- Fairness in top-k settings
  - Proportional is not enough
- FairExposure@k
  - $E_t^k(G_i) = \frac{\sum_{d \in G_i \cap d \in \sigma^k} p_t(d)}{|G_i|}$
  - $\text{Unfairness@}k = \sum_i \sum_j \frac{\frac{1}{T}\sum_t E_t^k(G_i)}{f(G_i)} - \frac{\frac{1}{T}\sum_t E_t^k(G_j)}{f(G_j)}$
- Idea: minimize the marginal unfairness between

$$\text{Unfairness@}(k-1) - \text{Unfairness@}k$$

Open questions

- Regret Lower Bound for bandits with privacy guarantee
  - What is the minimum noise and regret to achieve ϵ-DP/LDP?
- Calibrate privacy and fairness with problem-dependent structure
  - Collaborative Bandits
  - Low-rank structure
  - Non-stationary environment
- Other fairness definition

References VIII

[BZH06] Barbaro, M., Zeller, T., & Hansell, S. (2006). A face is exposed for AOL searcher no. 4417749. New York Times, 9(2008), 8.

[NS08] Narayanan, A., & Shmatikov, V. (2008, May). Robust de-anonymization of large sparse datasets. In 2008 IEEE Symposium on Security and Privacy (sp 2008) (pp. 111-125). IEEE.

[Kor'10] Korolova, A. (2010, December). Privacy violations using microtargeted ads: A case study. In 2010 IEEE International Conference on Data Mining Workshops (pp. 474-482). IEEE.

[CKNFS'11] Calandrino, J. A., Kilzer, A., Narayanan, A., Felten, E. W., & Shmatikov, V. (2011, May). " You might also like:" Privacy risks of collaborative filtering. In 2011 IEEE symposium on security and privacy (pp. 231-246). IEEE.

[Dwo06] Dwork, C. (2006, July). Differential privacy. In Proceedings of the 33rd international conference on Automata, Languages and Programming-Volume Part II (pp. 1-12). Springer-Verlag.

[Ngu19] Nguyen, A (2019). Understanding Differential Privacy.Towards Data Science. towardsdatascience.com/understanding-differential-privacy-85ce191e198a.

[CSS10] Chan, T. H., Shi, E., & Song, D. (2010, July). Private and continual release of statistics. In International Colloquium on Automata, Languages, and Programming (pp. 405-417). Springer, Berlin, Heidelberg.

[DNPR10] Dwork, C., Naor, M., Pitassi, T., & Rothblum, G. N. (2010, June). Differential privacy under continual observation. In Proceedings of the forty-second ACM symposium on Theory of computing (pp. 715-724).

[MT15] Mishra, N., & Thakurta, A. (2015, July). (Nearly) optimal differentially private stochastic multi-arm bandits. In Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence (pp. 592-601).

[TD16] Tossou, A. C., & Dimitrakakis, C. (2016, March). Algorithms for differentially private multi-armed bandits. In Thirtieth AAAI Conference on Artificial Intelligence.

[SS18] Shariff, R., & Sheffet, O. (2018). Differentially private contextual linear bandits. In Advances in Neural Information Processing Systems (pp. 4296-4306).

[NR18] Neel, S., & Roth, A. (2018, July). Mitigating Bias in Adaptive Data Gathering via Differential Privacy. In International Conference on Machine Learning (pp. 3720-3729).

References IX

[RZLS20] Ren, W., Zhou, X., Liu, J., & Shroff, N. B. (2020). Multi-Armed Bandits with Local Differential Privacy. arXiv preprint arXiv:2007.03121.

[WZWCKW20] Wang, H., Zhao, Q., Wu, Q., Chopra, S., Khaitan, A., & Wang, H. (2020, September). Global and Local Differential Privacy for Collaborative Bandits. In Fourteenth ACM Conference on Recommender Systems (pp. 150-159).

[JKMR16] Joseph, M., Kearns, M., Morgenstern, J. H., & Roth, A. (2016). Fairness in learning: Classic and contextual bandits. In Advances in Neural Information Processing Systems (pp. 325-333).

[JKMNR16] Joseph, M., Kearns, M., Morgenstern, J., Neel, S., & Roth, A. (2016). Fair algorithms for infinite and contextual bandits. arXiv preprint arXiv:1610.09559.

[WSLS16] Wu, Y., Shariff, R., Lattimore, T., & Szepesvári, C. (2016, June). Conservative bandits. In International Conference on Machine Learning (pp. 1254-1262).

[KGYR17] Kazerouni, A., Ghavamzadeh, M., Yadkori, Y. A., & Van Roy, B. (2017). Conservative contextual linear bandits. In Advances in Neural Information Processing Systems (pp. 3910-3919).

[AAT19] Amani, S., Alizadeh, M., & Thrampoulidis, C. (2019). Linear stochastic bandits under safety constraints. In Advances in Neural Information Processing Systems (pp. 9256-9266).

[KB20] Khezeli, K., & Bitar, E. (2020). Safe Linear Stochastic Bandits. In AAAI (pp. 10202-10209).

[MSHJ20] Morik, M., Singh, A., Hong, J., & Joachims, T. (2020, July). Controlling fairness and bias in dynamic learning-to-rank.In SIGIR 2020 (pp. 429-438).

[WBSJ21] Wang, L., Bai, Y., Sun, W., & Joachims, T. (2021). Fairness of Exposure in Stochastic Bandits. arXiv preprint arXiv:2103.02735.

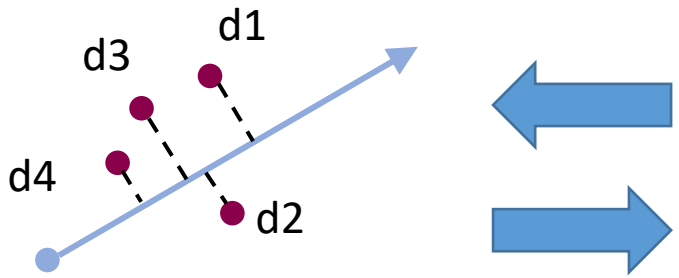
[YA21] Yang, T., & Ai, Q. (2021, April). Maximizing Marginal Fairness for Dynamic Learning to Rank. In Proceedings of the Web Conference 2021 (pp. 137-145).

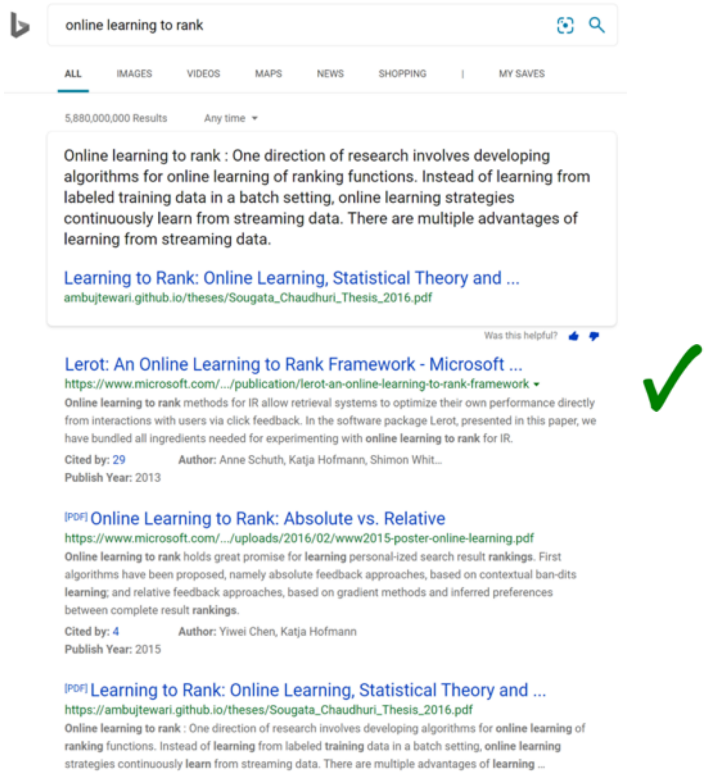
Outline of this tutorial

- Motivation
- Classical exploration strategies
- Efficient exploration in complicated real-world environments
- Exploration in non-stationary environments
- Ethical considerations of exploration
- Conclusion & future directions

Conclusions

- Interactive information retrieval with bandit feedback





Conclusions

- Interactive information retrieval with bandit feedback

Exploitation

Present the best results estimated so far to satisfy users

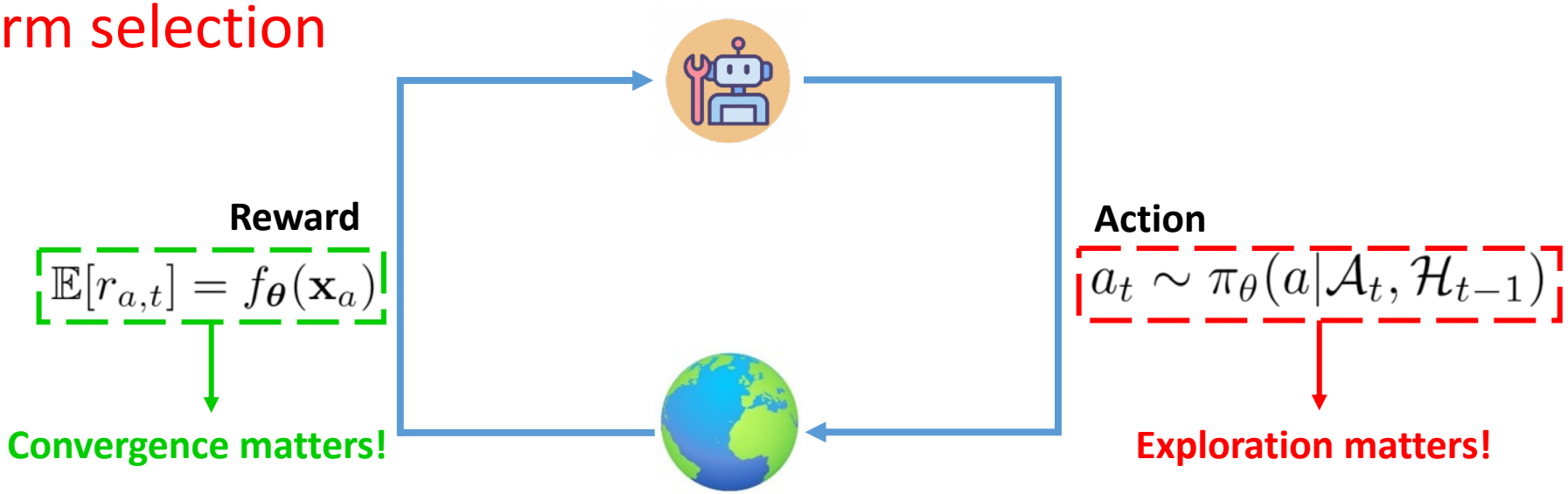
Interactive IR System

Exploration

Present currently underestimated results to best improve the ranker

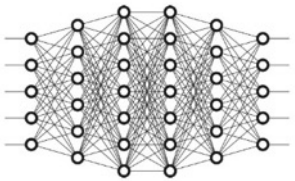
Conclusions

- Key problems
  - Reward estimation
  - Arm selection



Future research directions

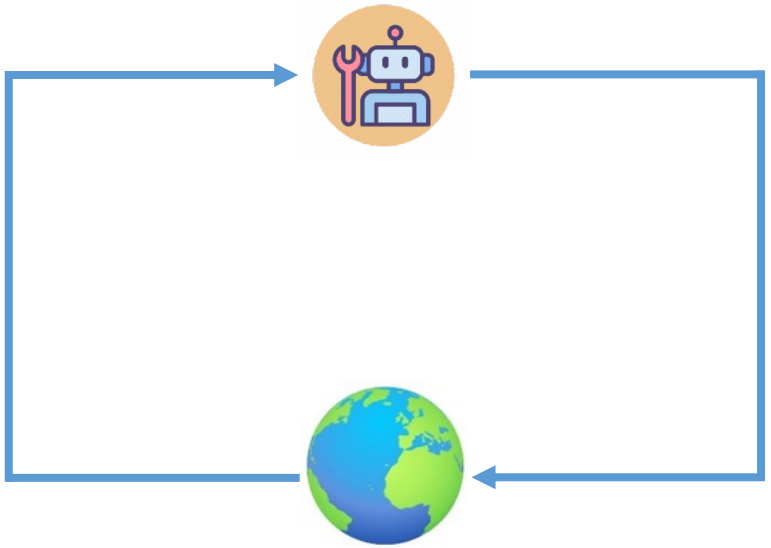
- Going beyond linear models
  - How about deep models?



Reward

$$\mathbb{E}[r_{a,t}] = f_{\boldsymbol{\theta}}(\mathbf{x}_a)$$

Convergence matters!



Action

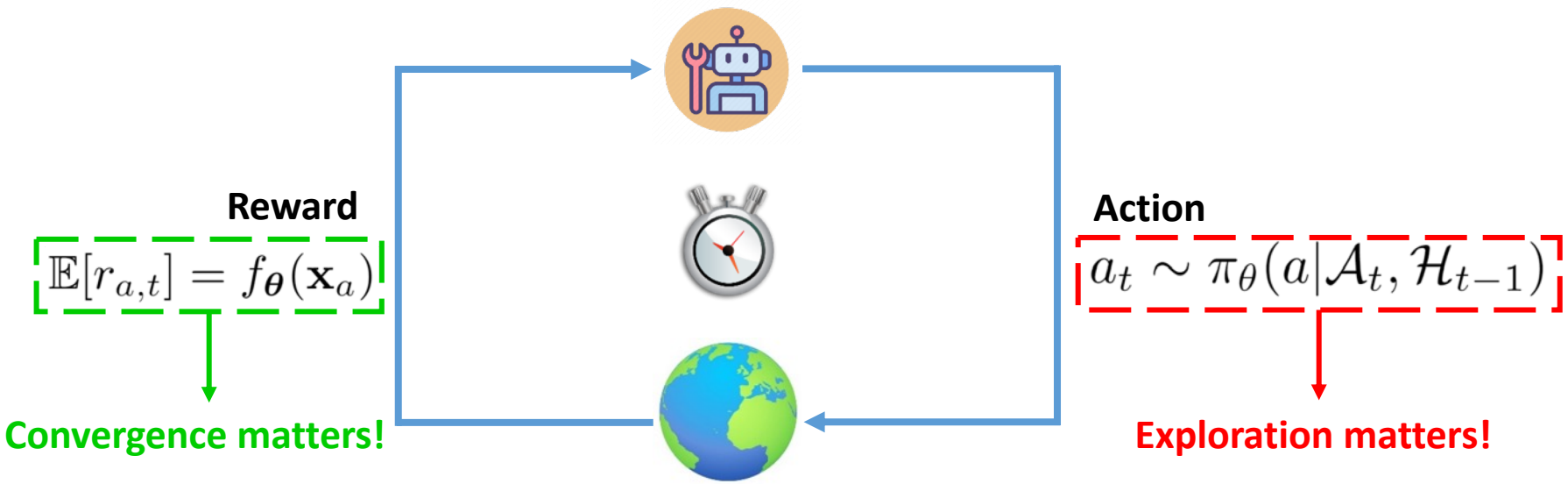
$$a_t \sim \pi_\theta(a|\mathcal{A}_t, \mathcal{H}_{t-1})$$

Exploration matters!

Some preliminary studies exist:
[ZLG19, AFB14]

Future research directions

- Online deployment
  - Policy update driven by every interaction in real-time



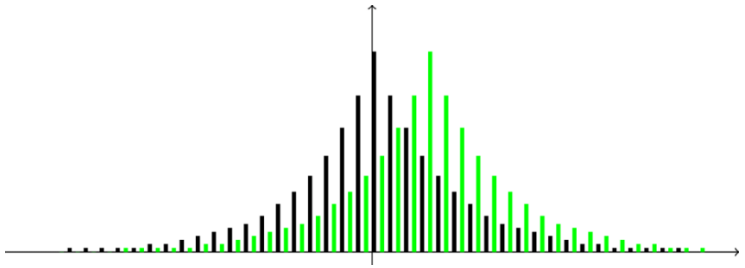
Some recent work: [WHCW19, KSL16, MWLS20]

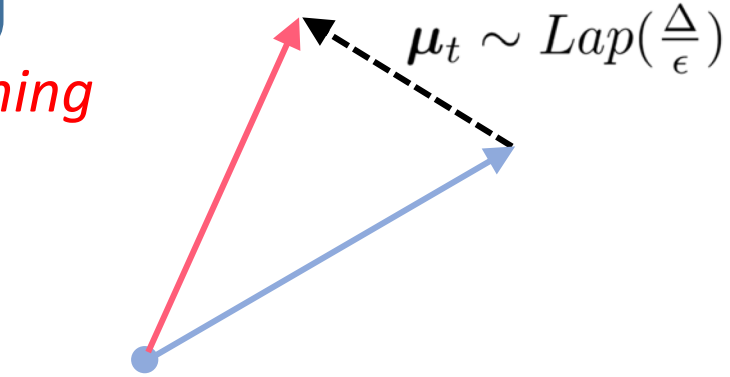
Future research directions

- Learning under adversarial contexts
  - Privacy breach under extraction attacks

[CKNFS11, Kor10]

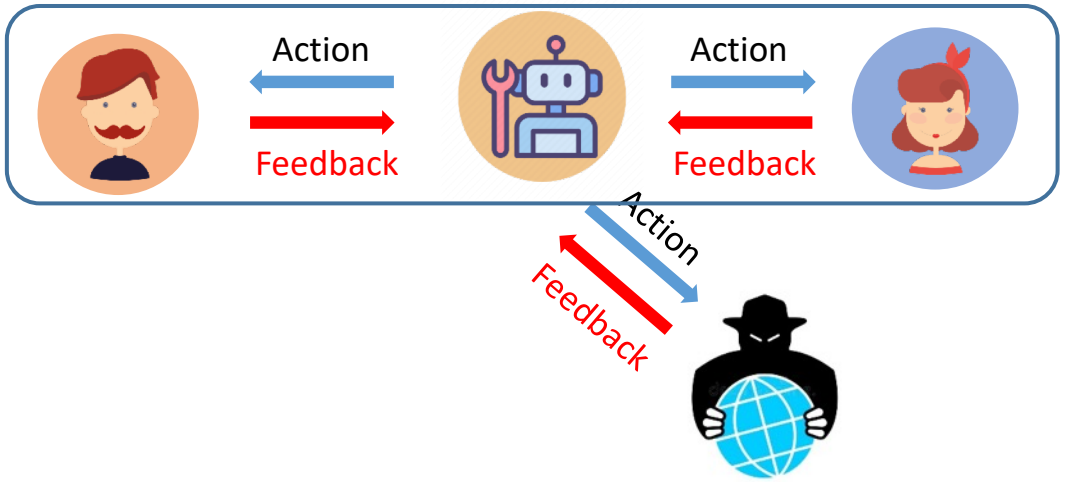
Exploration reconciles the need for learning and the need for privacy protection





Future research directions

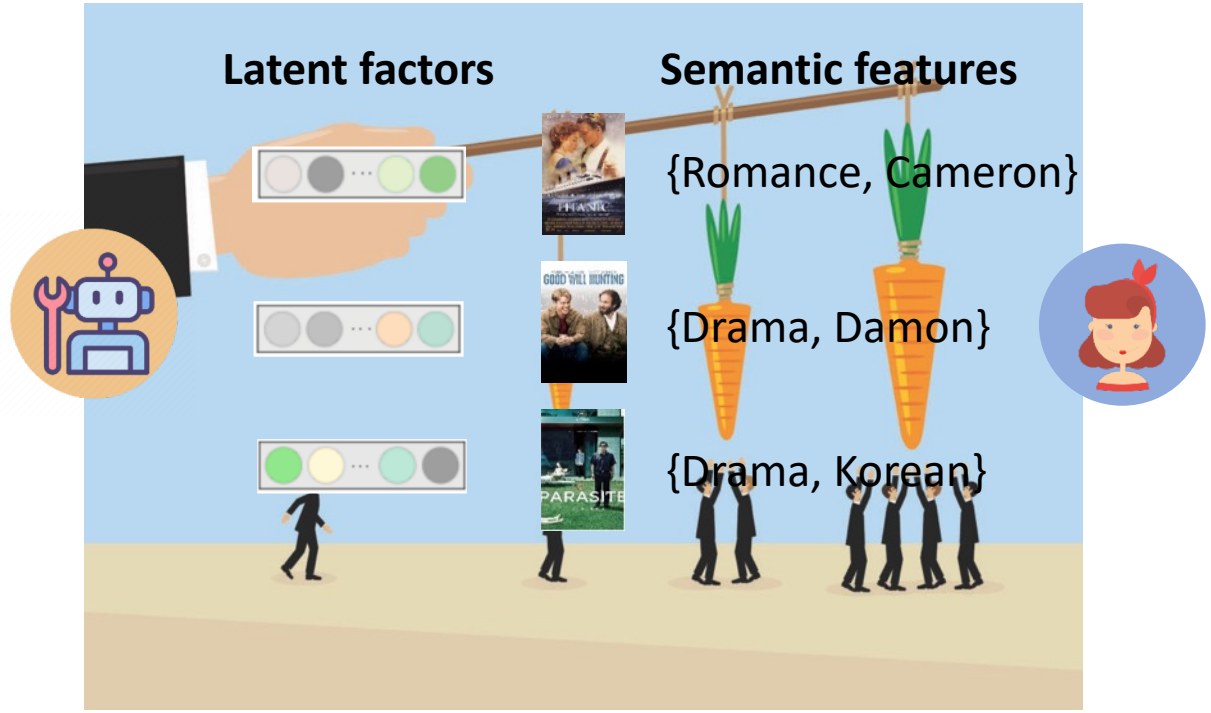
- Learning under adversarial contexts
  - Robustness under poisoning attacks



Randomness in exploration hardens the model against adversary; dependence structure among users improves privacy utility trade-off.

Future research directions

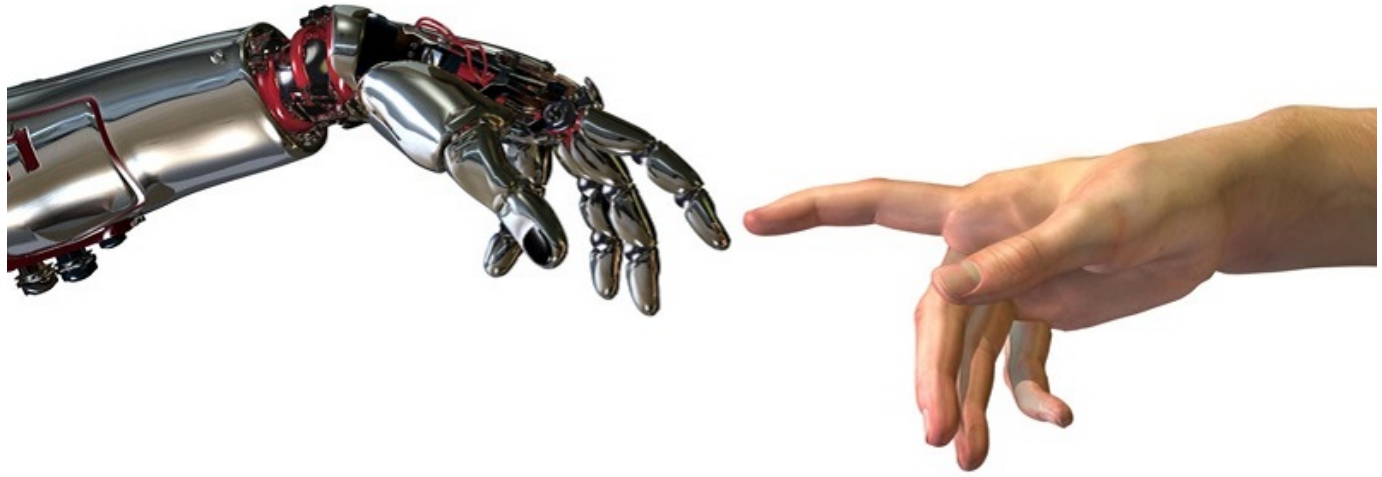
- Incentivize the exploration
  - No regret under information gap



Explanations?

Future research directions

- System learning in accordance with user learning
  - User is not omniscient, but also learns from interactions with the system



References X

[ZLG19] Zhou, D., Li, L., & Gu, Q. (2019). Neural Contextual Bandits with Upper Confidence Bound-Based Exploration. arXiv preprint arXiv:1911.04462.

[AFB14] Allesiardo, R., Féraud, R., & Bouneffouf, D. (2014, November). A neural networks committee for the contextual bandit problem. In International Conference on Neural Information Processing (pp. 374-381). Springer, Cham.

[WHCW19] Wang, Y., Hu, J., Chen, X., & Wang, L. (2019). Distributed bandit learning: Near-optimal regret with efficient communication. arXiv preprint arXiv:1904.06309.

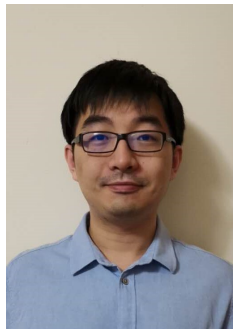
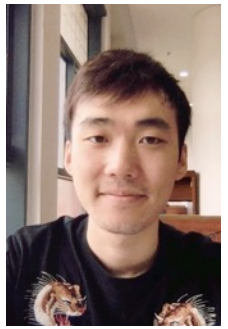
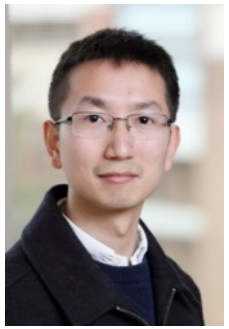
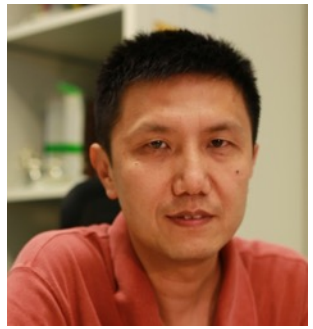
[KSL16] Korda, N., Szorenyi, B., & Li, S. (2016, June). Distributed clustering of linear bandits in peer to peer networks. In International Conference on Machine Learning (pp. 1301-1309).

[MWLS20] Mahadik, K., Wu, Q., Li, S., & Sabne, A. (2020, June). Fast distributed bandits for online recommendation systems. In Proceedings of the 34th ACM International Conference on Supercomputing (pp. 1-13).

[CKNFS11] Calandrino, J. A., Kilzer, A., Narayanan, A., Felten, E. W., & Shmatikov, V. (2011, May). " You might also like:" Privacy risks of collaborative filtering. In 2011 IEEE symposium on security and privacy (pp. 231-246). IEEE.

[Kor10] Korolova, A. (2010, December). Privacy violations using microtargeted ads: A case study. In 2010 IEEE International Conference on Data Mining Workshops (pp. 474-482). IEEE.

Acknowledgement


NSF IIS-1553568 IIS-1618948

Bloomberg

Walmart Labs

YAHOO! RESEARCH

UNIVERSITY of VIRGINIA ENGINEERING

Back-up slides

- Simulation-based evaluation
  - Simulate the non-stationary environment
- Semi-simulation-based evaluation
  - Real-world datasets (that do not have non-stationarity) + simulated changes
- Evaluation on real-world datasets
  - On real-world datasets that have non-stationarity

Empirical evaluations

- Simulation-based evaluation
  - Simulate the non-stationary environment
- Semi-simulation-based evaluation
  - Real-world datasets (that do not have non-stationarity) + simulated changes
- Evaluation on real-world datasets
  - On real-world datasets that have non-stationarity

Empirical evaluations

- Yahoo! Today Module

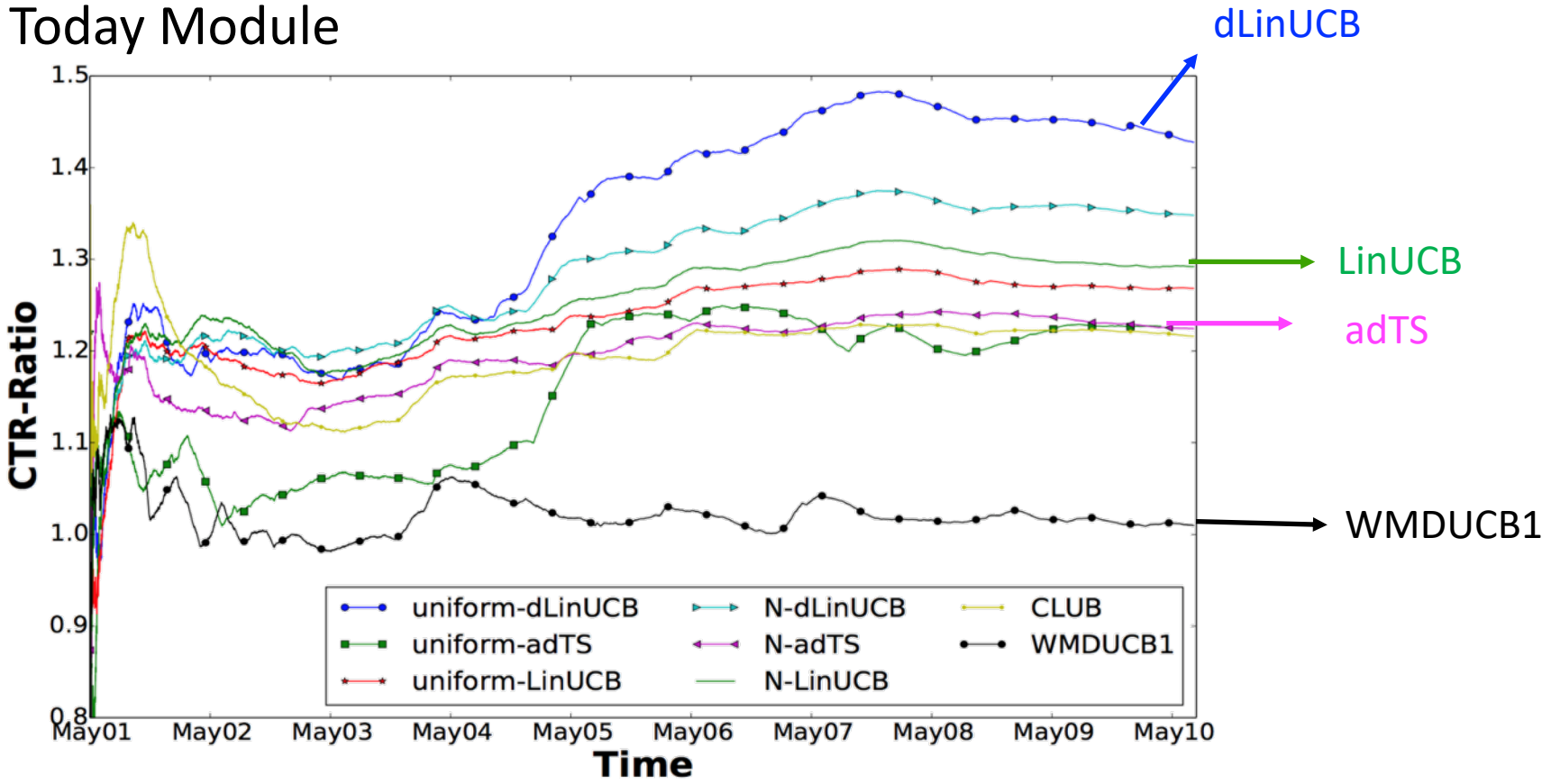
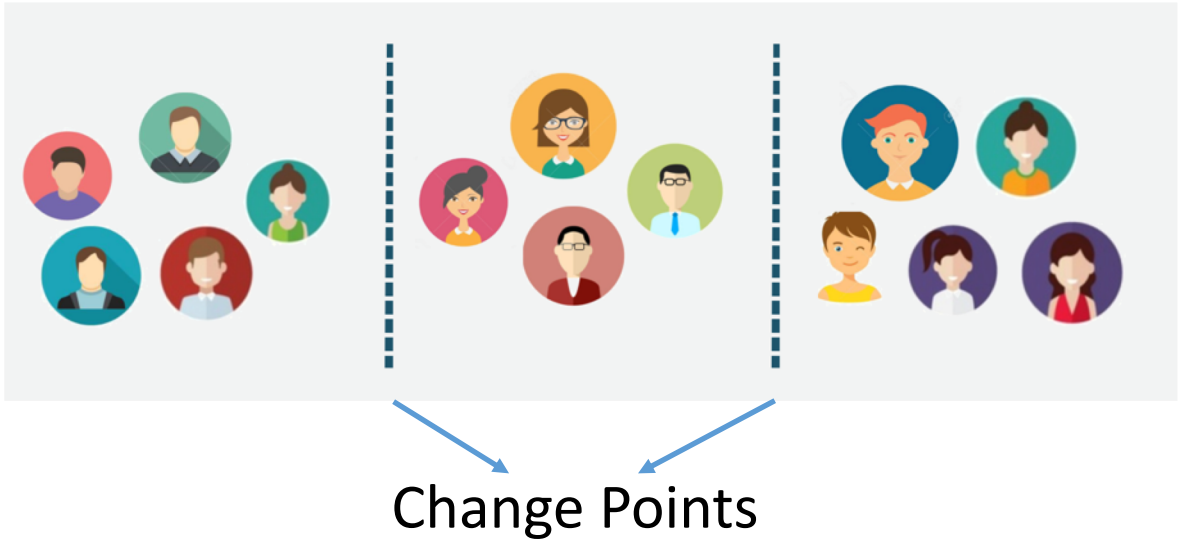


Figure 4. Real time CTR ratio in 10 days on Yahoo dataset

Empirical evaluations

- LastFM dataset
  - Simulate a non-stationary environment by attaching different users' observations[]

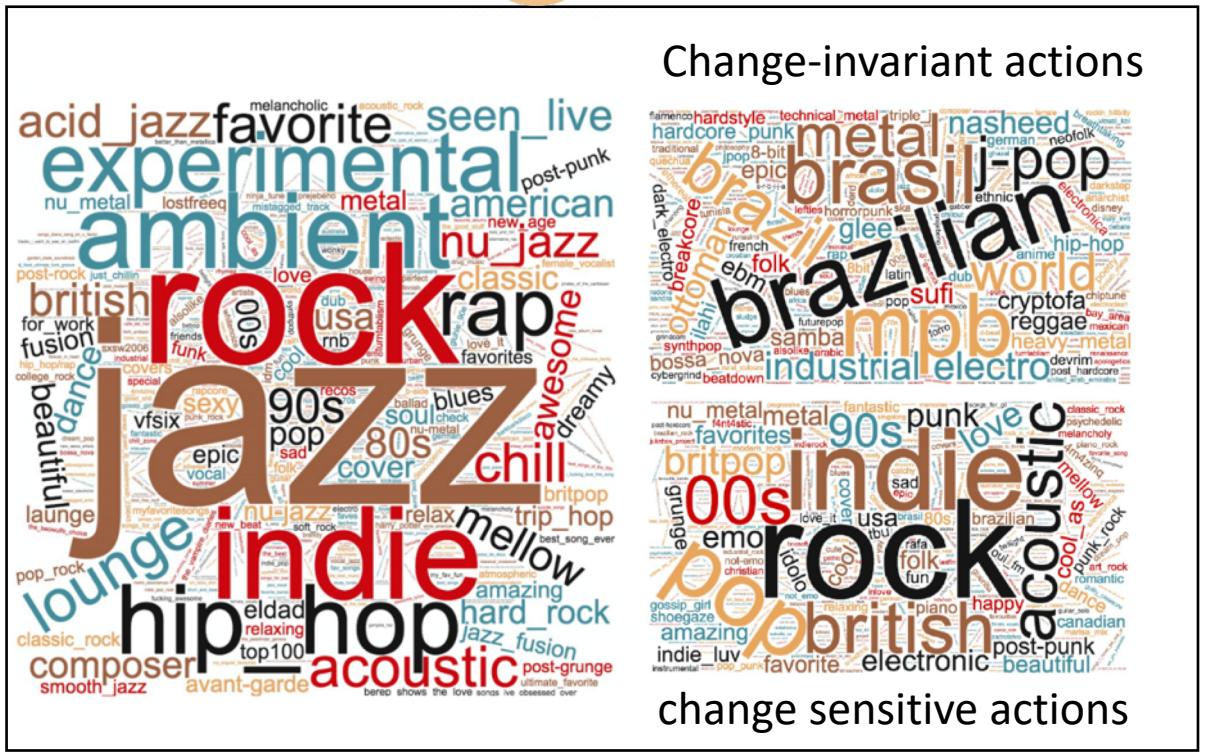


Empirical evaluations

Figure 5. Word cloud of tags from high reward actions in the four identified environments by dLinUCB on LastFM dataset.

Empirical evaluations

Learner 1



Learner 2

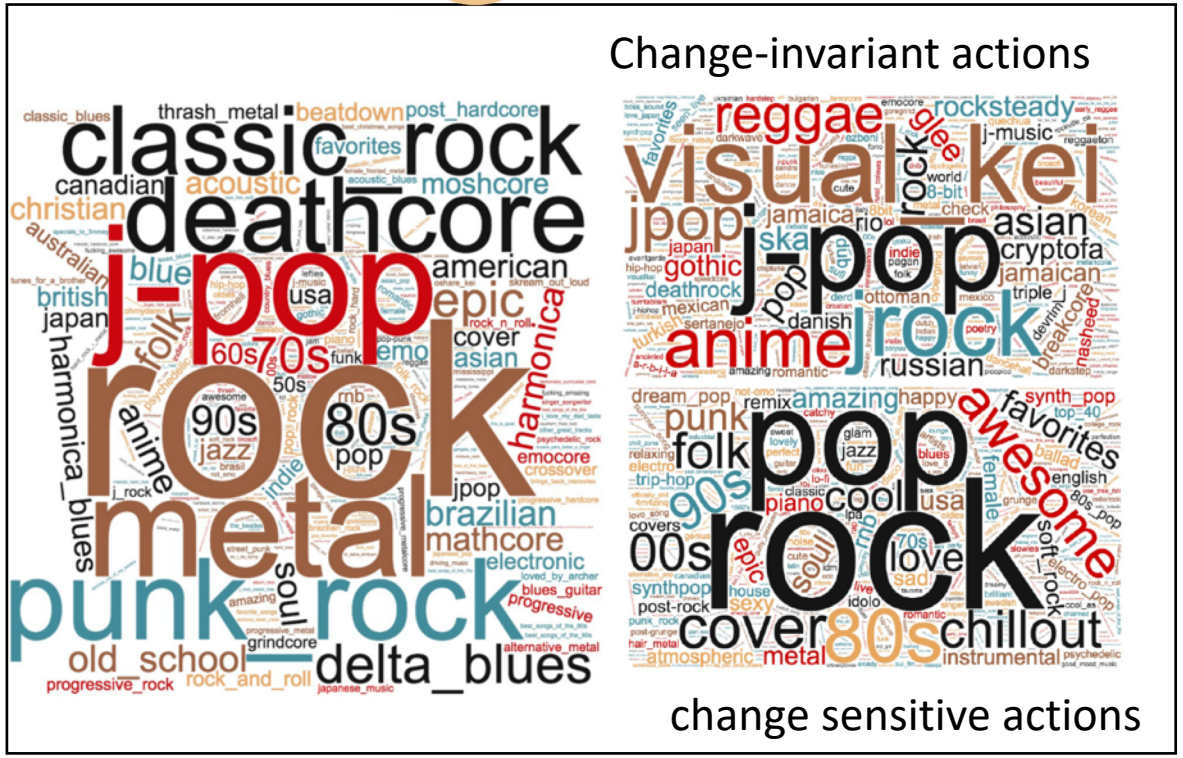


Figure 5. Word cloud of tags from high reward actions in the four identified environments by dLinUCB on LastFM dataset.