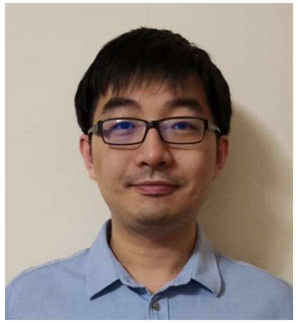
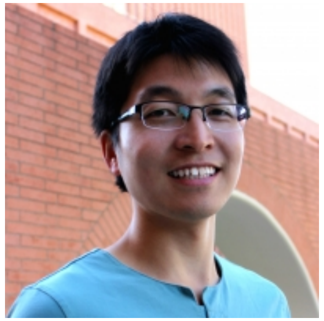
# Learning from a Learning User for Optimal Recommendations

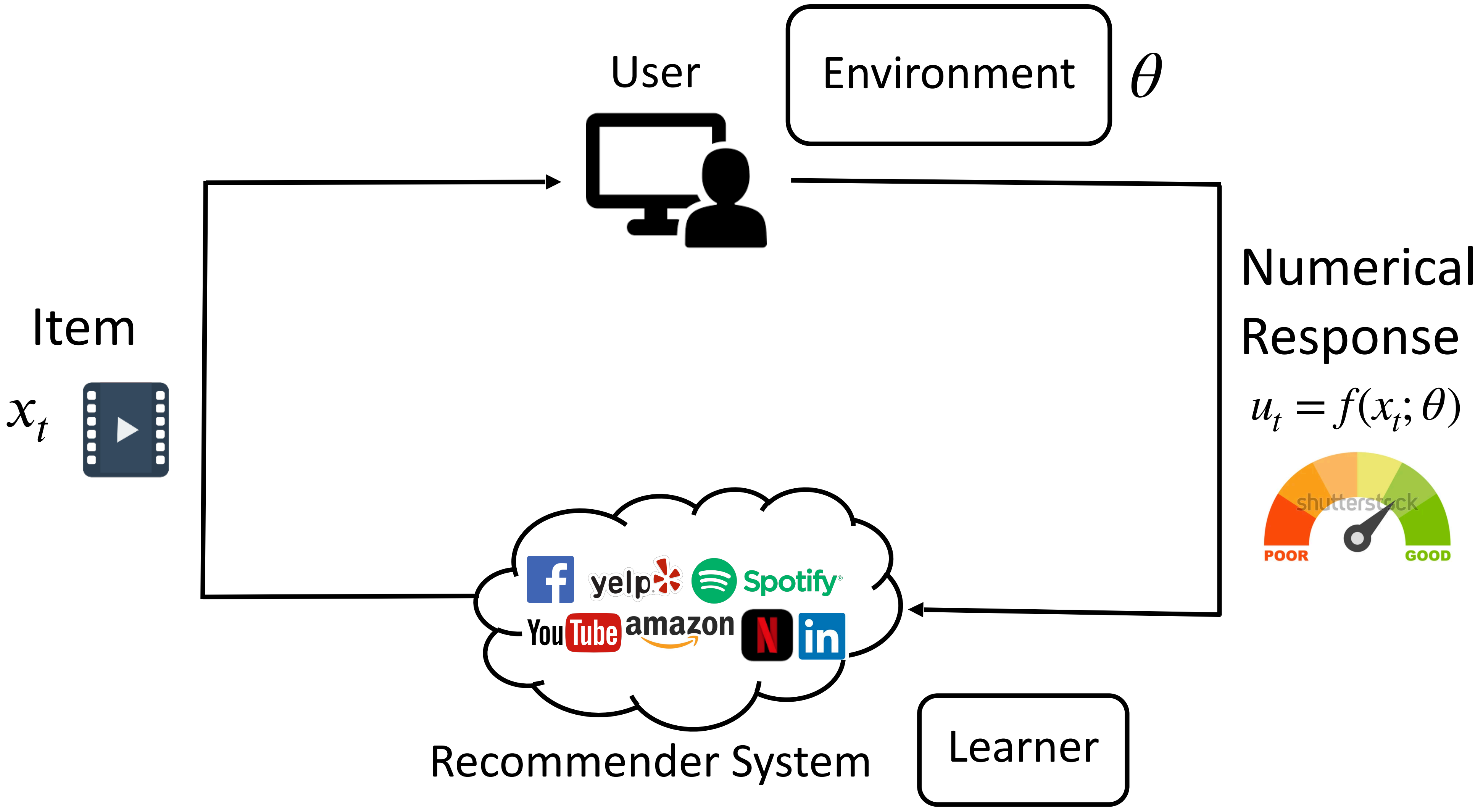



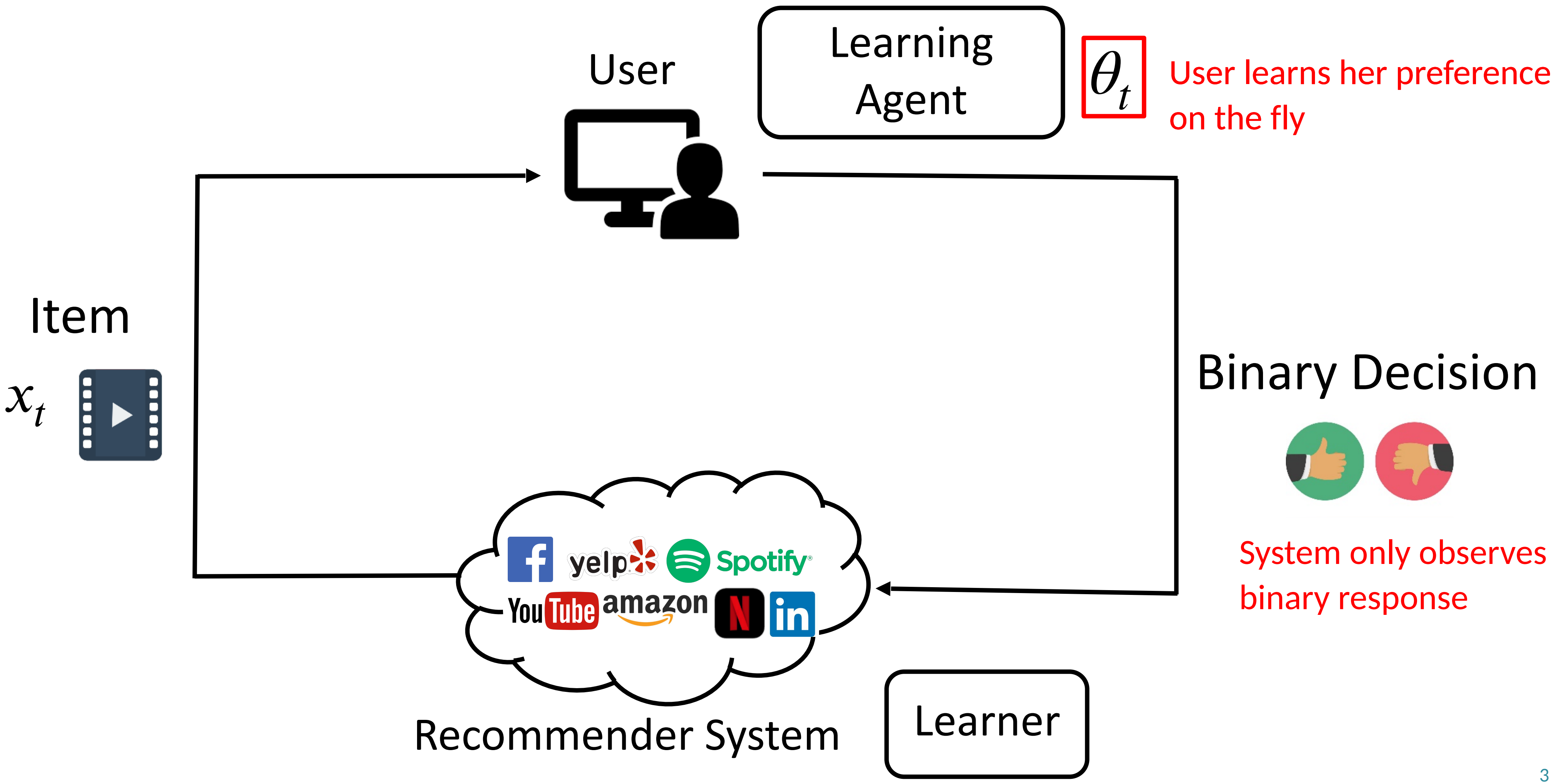
Fan Yao, Chuanhao Li, Denis Nekipelov, Hongning Wang, Haifeng Xu

IIS-1553568
IIS-2007492
IIS-1838615

# Classic View of Recommender Systems



# Our View of Recommender Systems

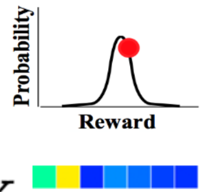


# Problem Formulation

- Online interaction in contextual bandit framework

Item set

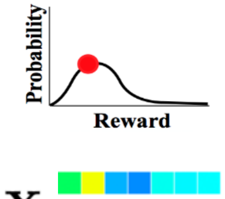
$$\mathcal{A} = \{\mathbf{x}_a \in \mathbb{R}^d\}_{i=1}^{K}$$



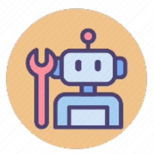

$\mathbf{x}_{a_1}$


$\mathbf{x}_{a_2}$

• • •




$\mathbf{x}_{a_k}$



recommendation

decision

$$\mathcal{H}_t = \{(\mathbf{x}_{a_i,t}, y_i = \{0,1\})\}_{i=1}^{t-1}$$

$$\boldsymbol{a}_{0,t}, \boldsymbol{a}_{1,t}$$

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

System observes binary choices

System's goal: identify the best arm for the user

User observes rewards

User's goal: no regret about her past decisions

# Modeling a learning user

- Example: a user running LinUCB

Linear reward assumption [APS11]:

$$\mathbf{E}[r_i] = \mathbf{x}_{a_i,t}^{\top}\boldsymbol{\theta}^*$$ Unknown to both user and system!

  - Run ridge regression on $\mathcal{H}_t = \{(\mathbf{x}_{a_i,t}, r_i)\}_{i=1}^{t-1}$ to estimate $\boldsymbol{\theta}_t$

$$\|\boldsymbol{\theta}_* - \boldsymbol{\theta}_t\|_{V_t} \leq O\left(\sqrt{d \log \frac{t}{\delta}}\right) \qquad V_t = V_0 + \sum_{s=1}^{t-1} \mathbf{x}_s \mathbf{x}_s^{\top}$$

  - Choose the item with the largest upper confidence bound:

$$\hat{r}_{i,t} = \theta_t^{\top}\mathbf{x}_{i,t} + \beta_t \|\mathbf{x}_{i,t}\|_{V_t^{-1}} \qquad \beta_t = O(\sqrt{\log t})$$

# Modeling a learning user

- Generalize user's learning behavior
  - Can use any algorithm $F$ on $\mathcal{H}_t = \{(\mathbf{x}_{a_i,t}, r_i)\}_{i=1}^{t-1}$ to estimate $\boldsymbol{\theta}_{t+1} = F(\mathcal{H}_t)$ such that

$$\|\boldsymbol{\theta}_* - \boldsymbol{\theta}_t\|_{V_t} \leq c_1 t^{\gamma_1} g(\delta)$$

$$\gamma_1 \in (0, \tfrac{1}{2})$$

Inaccuracy of the learning algorithm

  - Estimate rewards with arbitrary confidence level:

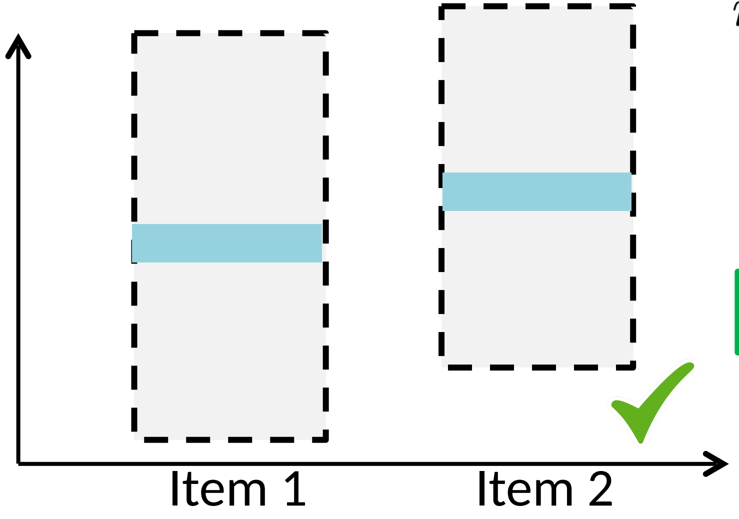
$$\hat{r}_{i,t} = \theta_t^\top \mathbf{x}_{i,t} + \beta_t^{(i)} \|\mathbf{x}_{i,t}\|_{V_t^{-1}}$$

$$\beta_t^{(i)} \in [-c_2 t^{\gamma_2}, c_2 t^{\gamma_2}]$$

Account for a wide range of user behaviors when facing uncertainty, including even irrational behaviors

# What can we learn from such a user?

- Revealed preference between the recommended items
  - A cutting hyperplane suggesting where the true model parameter is!




$$\hat{r}_{1,t} = \theta_t^\top \mathbf{x}_{1,t} + \beta_t^{(1)} \|\mathbf{x}_{1,t}\|_{V_t^{-1}} \leq \hat{r}_{2,t} = \theta_t^\top \mathbf{x}_{2,t} + \beta_t^{(2)} \|\mathbf{x}_{2,t}\|_{V_t^{-1}}$$

Our learning target!

$$\theta_*^\top (\mathbf{x}_{1,t} - \mathbf{x}_{2,t}) \leq \theta_t^\top (\mathbf{x}_{1,t} - \mathbf{x}_{2,t}) + (\theta_* - \theta_t)^\top (\mathbf{x}_{1,t} - \mathbf{x}_{2,t})$$

Cutting direction

$$\leq ct^\gamma \left( \|\mathbf{x}_{1,t}\|_{V_t^{-1}} + \|\mathbf{x}_{2,t}\|_{V_t^{-1}} + g(\delta) \|\mathbf{x}_{1,t} - \mathbf{x}_{2,t}\|_{V_t^{-1}} \right)$$

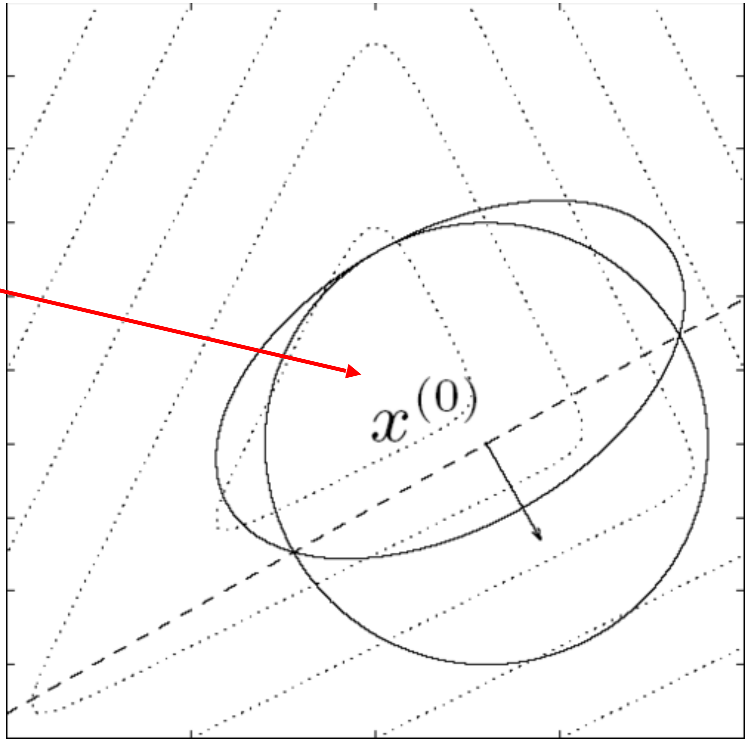
User's decision error

Ellipsoid method!

# Background: ellipsoid method

- An iterative optimization method [GLS81]
  - A classical method for linear programming
  - Polynomial time

Where the optimal solution locates

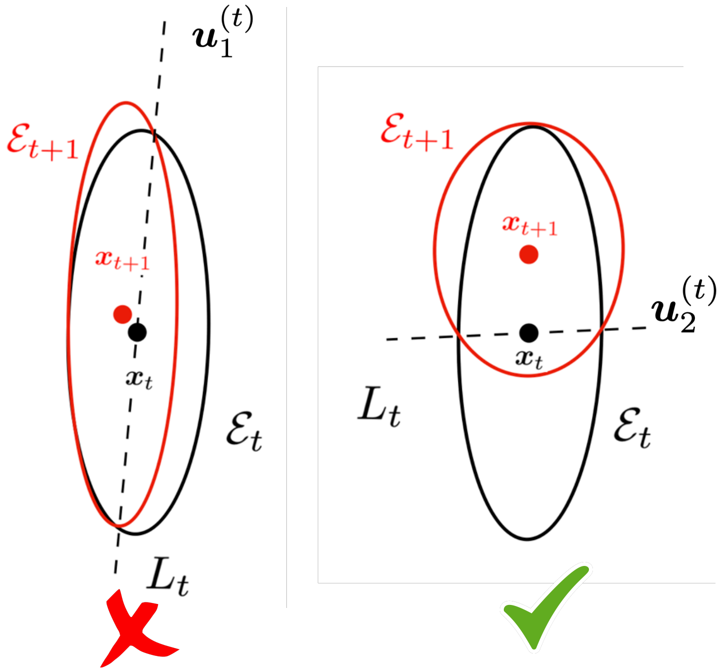


# Find a good cut

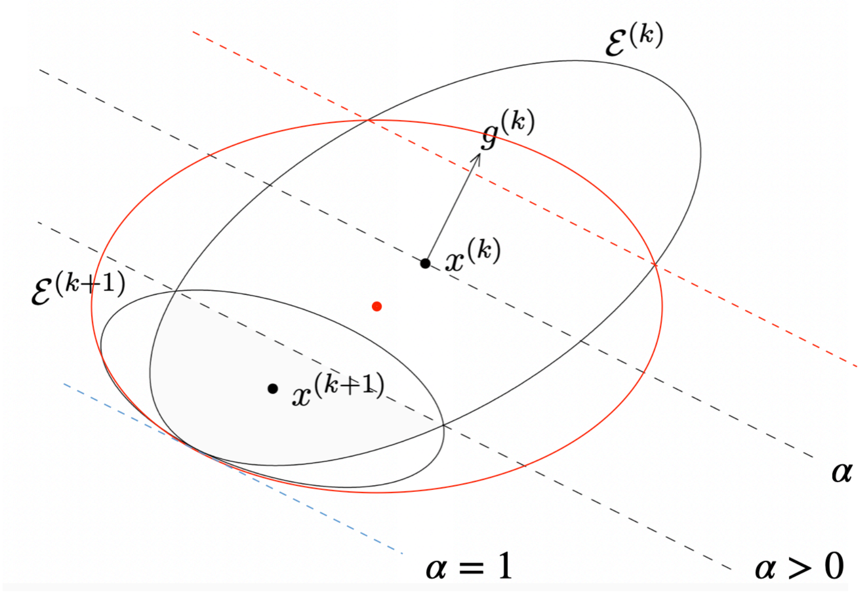
- A good cut = good direction + good depth

*Reduces uncertainty along all directions* (→ good direction)

*Shrinks the volume* (→ good depth)




good cutting direction




$\alpha < -\frac{1}{d}$ : $\mathrm{Vol}(\mathcal{E}') > \mathrm{Vol}(\mathcal{E})$ ✗

$\alpha = -\frac{1}{d}$ : $\mathrm{Vol}(\mathcal{E}') = \mathrm{Vol}(\mathcal{E})$ ✗

$-\frac{1}{d} < \alpha < 1$ : $\mathrm{Vol}(\mathcal{E}') < \mathrm{Vol}(\mathcal{E})$ ✓

good cutting depth

# Solution: Noise-Robust Active Ellipsoid Search

- Balancing three factors

Until we are ready

Improve our estimation

Prepare the user

1. (Cut) If $t \leq T_0$ and $\alpha_t \geq -\frac{1}{kd}$, cut $\mathcal{E}_t$ and update $(\mathbf{x}_t, P_t)$.
2. (Exploration) If $t \leq T_0$ and $\alpha_t > -\frac{1}{kd}$, make recommendations to ensure the user is exposed to the least explored directions in $V_t$.
3. (Exploitation) If $t > T_0$, recommend the empirically best arm to the user.

We emphasize the notion of strong regret:

$$R_T = \sum_{t=1}^{T} \theta_*^\top (2\mathbf{x}_* - \mathbf{x}_{1,t} - \mathbf{x}_{2,t})$$

# Regret analysis

- For proper choices of $T_0$ , with high probability, the regret of RAES is upper bounded by

*Difficulty of the learning problem increases with $\gamma$!*

$$O(d^2 T^{\frac{1}{2}+\gamma})$$

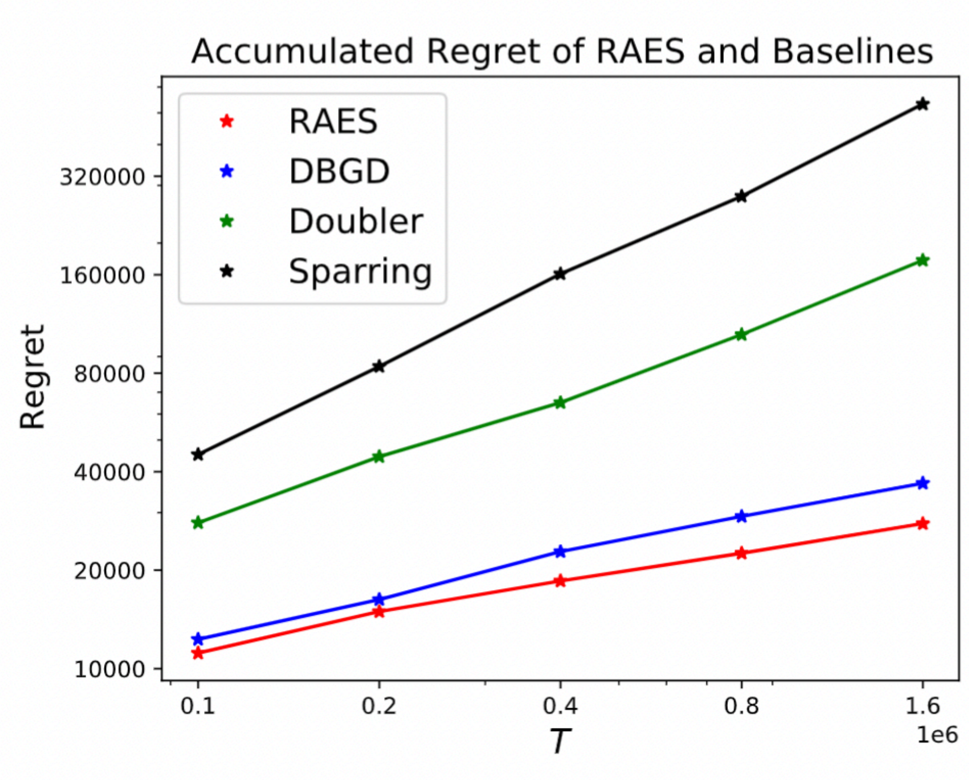
- The expected regret of any algorithm facing a learning user is at least

*At least as difficult as linear contextual bandit problems.*

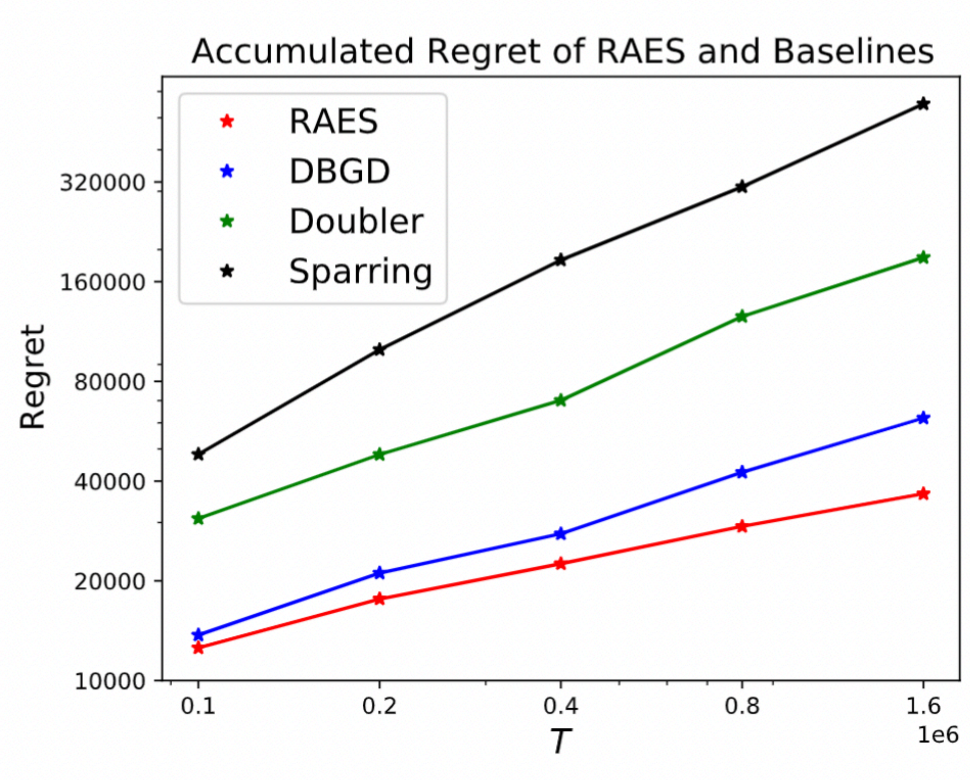
$$\Omega(d T^{\frac{1}{2}})$$

# Experiment results

$\gamma = 0.2, \qquad d = 20$




$V_0 = 100 I_d$




$V_0 = diag(1, \cdots, 1, 100, \cdots, 100)$

# Summary

- Learning from a learning user in a contextual setting is still possible
  - An efficient ellipsoid method to search for the ground-truth model parameters based on users' revealed preferences
  - Nearly optimal regret guarantee is provided
- Next Step: learning from strategic learning agents?
  - They can be cooperating or competing with each other
  - They might share distinct objectives

BanditLib: https://github.com/HCDM/BanditLib