A Unified Approach to Translate Classical Bandit Algorithms to Structured Bandits

Samarth Gupta

Carnegie Mellon University

ICASSP 2021
Virtual

Joint work with

Shreyas Chaudhari
Carnegie Mellon

Subhojyoti Mukherjee
Wisconsin-Madison

Gauri Joshi
Carnegie Mellon

Osman Yagan
Carnegie Mellon

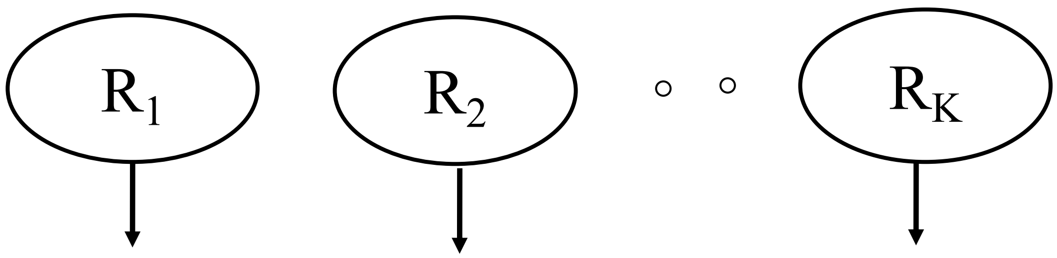
A unified approach to translate classical bandit algorithms to the structured bandit setting

https://arxiv.org/abs/1808.07576

S. Gupta, S. Chaudhari, S. Mukherjee, G. Joshi, and O. Yagan

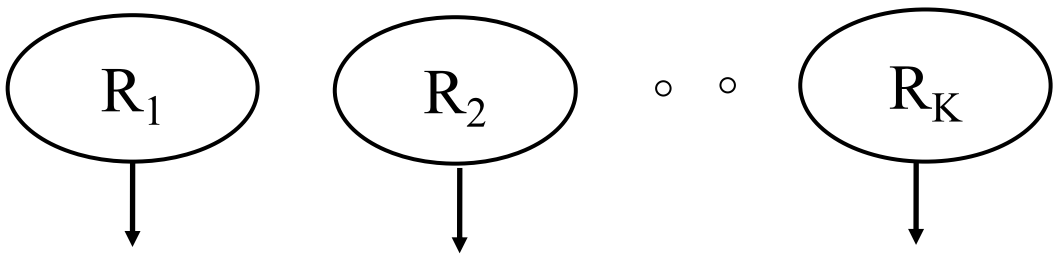
Funding/Support: NSF Eager, CyLab presidential fellowship, Carnegie Bosch Institute, Siebel Energy Institute

Classic Multi-armed Bandits



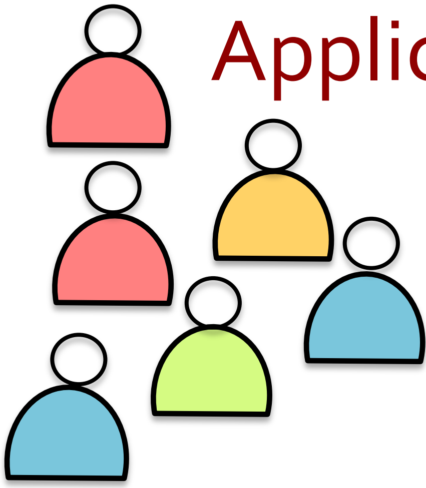
- Unknown reward distributions
- Goal: Maximize Cumulative Reward $\sum_{t=1}^{T} R_{k_t}$

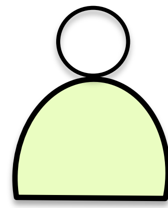
Classic Multi-armed Bandits



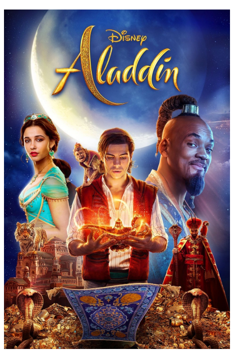
- Unknown reward distributions
- Equivalent Goal: Min. Cumulative Regret $\sum_{t=1}^{T}(R_{k_t^*} - R_{k_t})$

Application to recommendation system



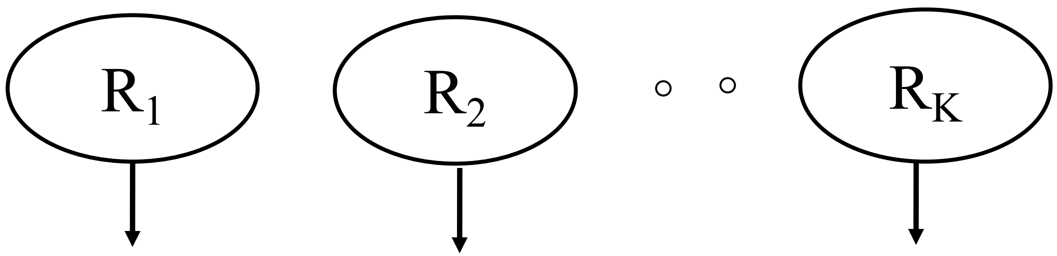


Anonymous user enters the system



Maximize cumulative reward by sequentially recommending available movies to entering users

Classic Multi-armed Bandits

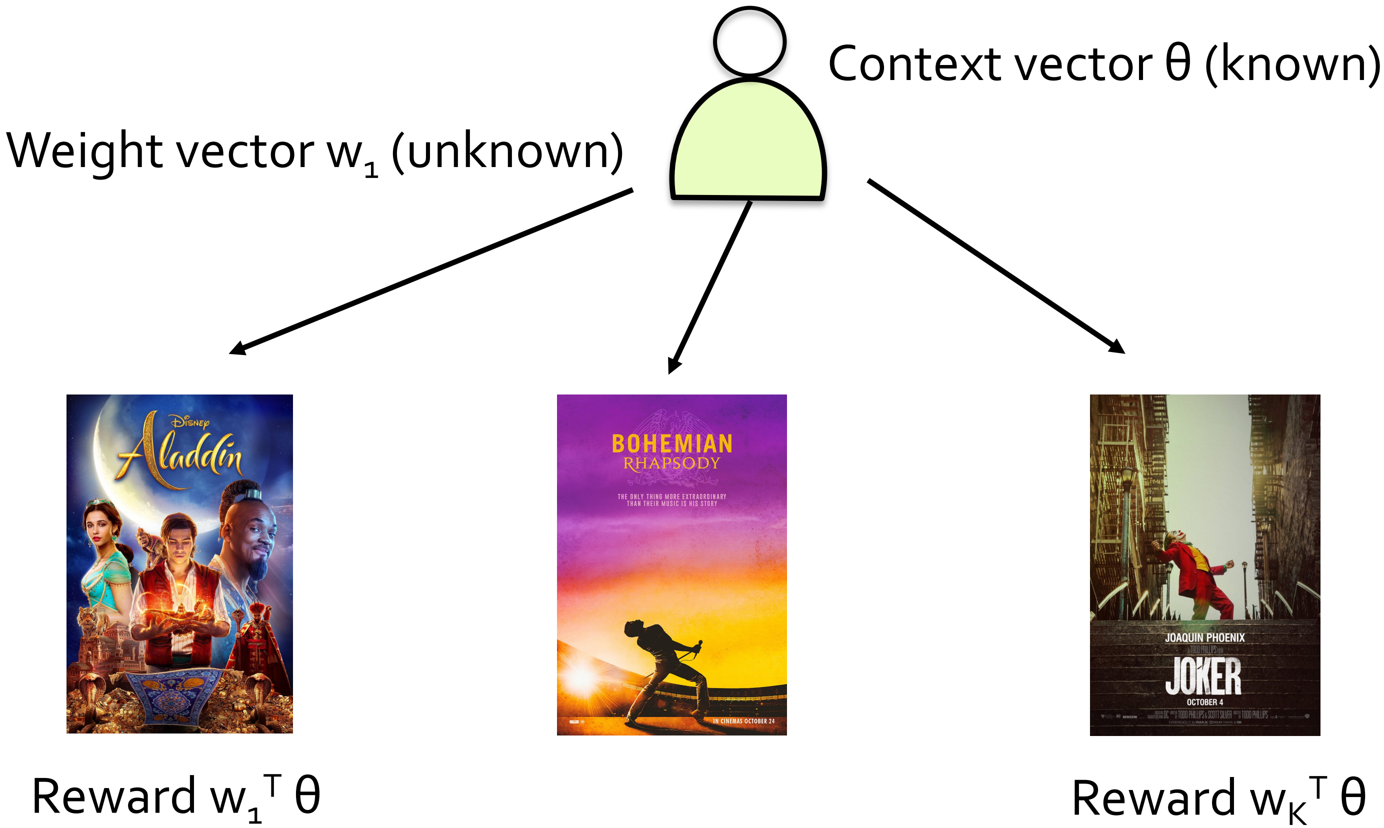


- Algorithms: UCB [Auer at el], Thompson Sampling [Thompson], KL-UCB [Bubeck et al], etc.
- Expected Regret is $\Theta((K-1)\log T)$

LIMITATION: Rewards assumed to be independent across arms

Variants for Personalized Recommendations

Contextual Bandits

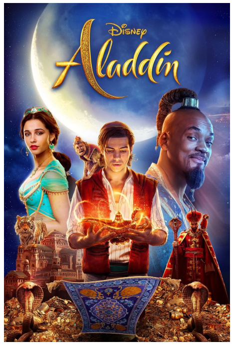


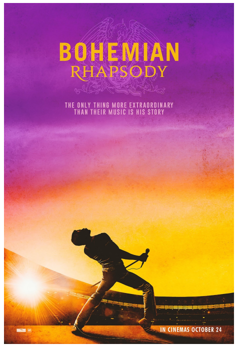
[Li et al, Agarwal et al, and many other works]

Variants for Personalized Recommendations

This work: Structured Bandits

Context vector θ (unknown, user not sign in)





Reward $\mu_1(\theta)$

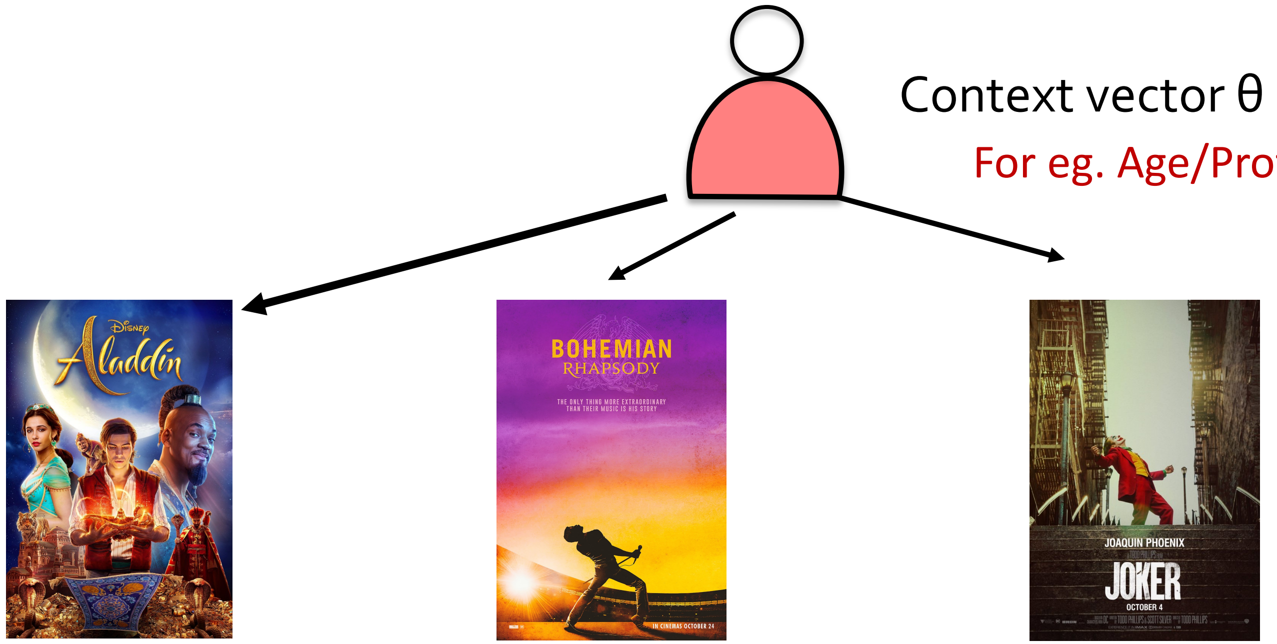
Reward $\mu_K(\theta)$

How do we know the mean reward functions μ(.)?

- Controlled user studies for different types of users
- Using contextual information from a previous campaign

Context vector θ (unknown)

For eg. Age/Profession



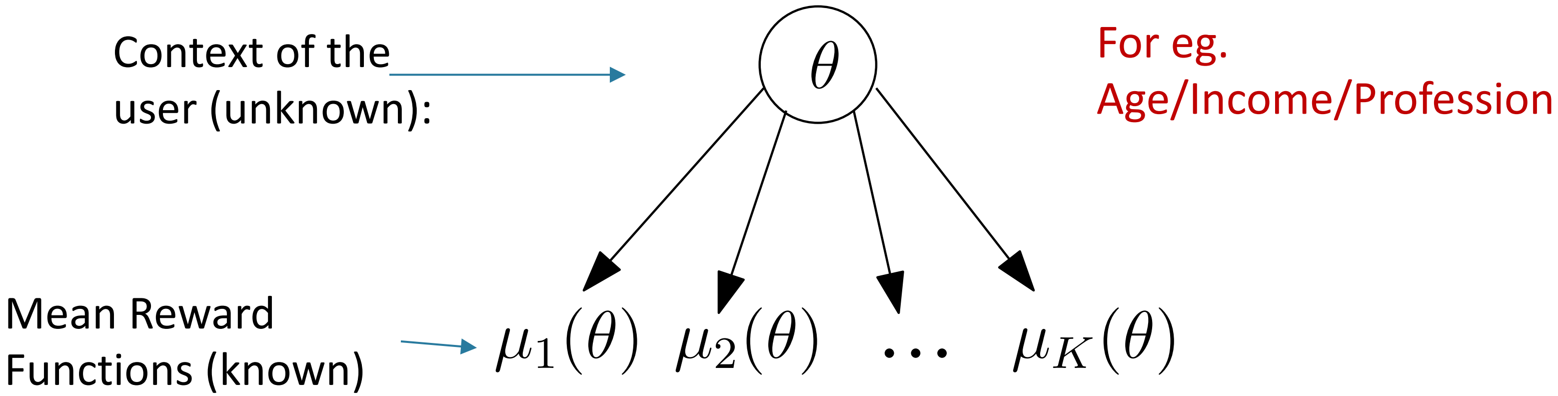
Reward $\mu_1(\theta)$

Reward $\mu_K(\theta)$

The Structured Bandit Framework

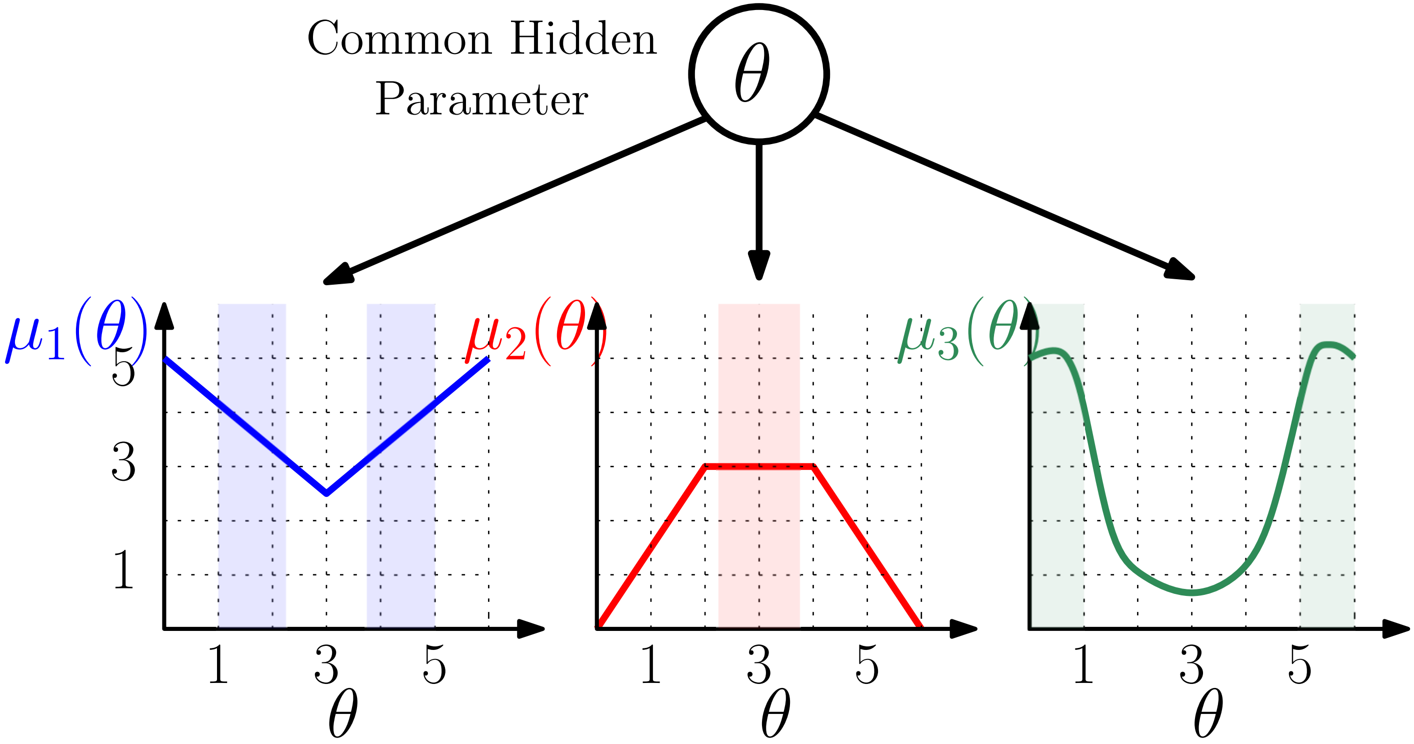
- There is a fixed unknown parameter lies θ^* in a known set Θ
- No restrictions on the reward functions $\mu_k(\theta)$
- θ can be continuous, or a vector

GOAL: Maximize cumulative reward



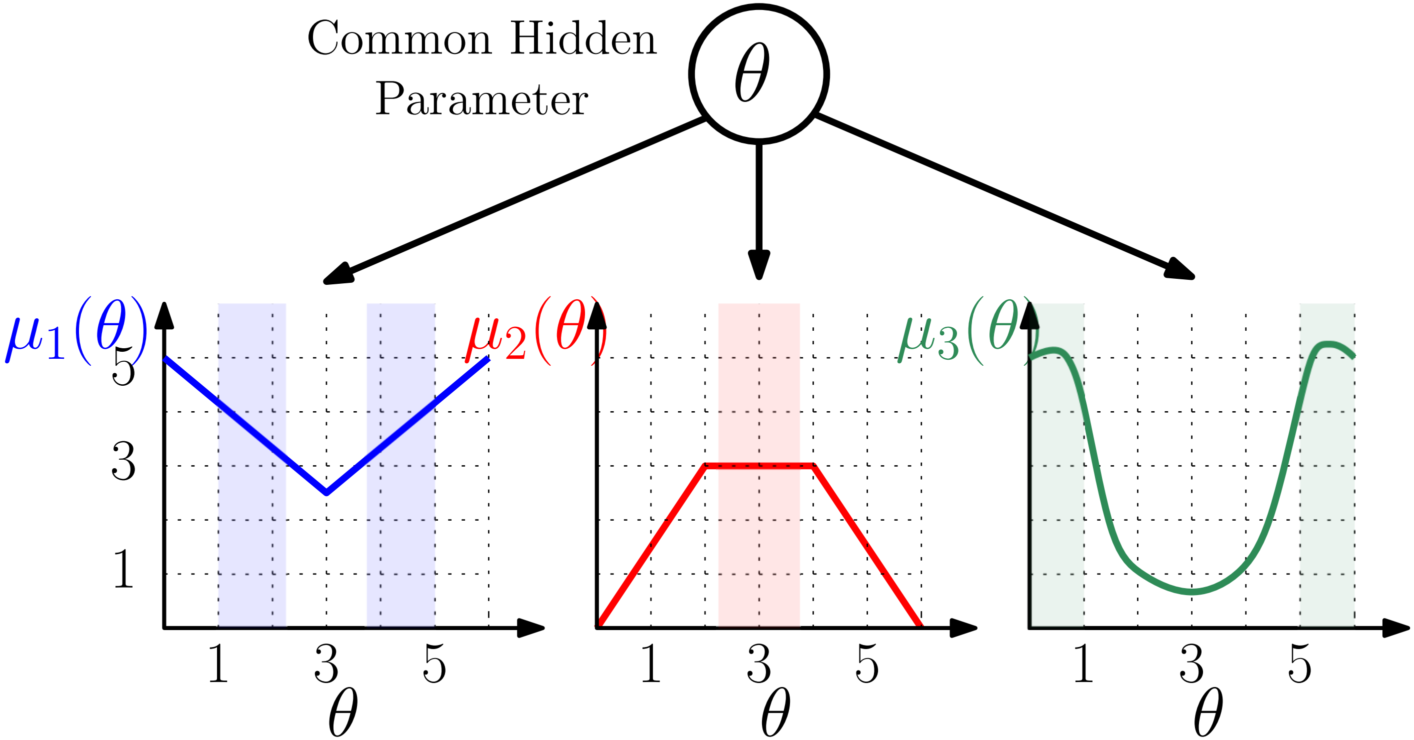
Example

Hidden context
$\theta^* = 3$



Example

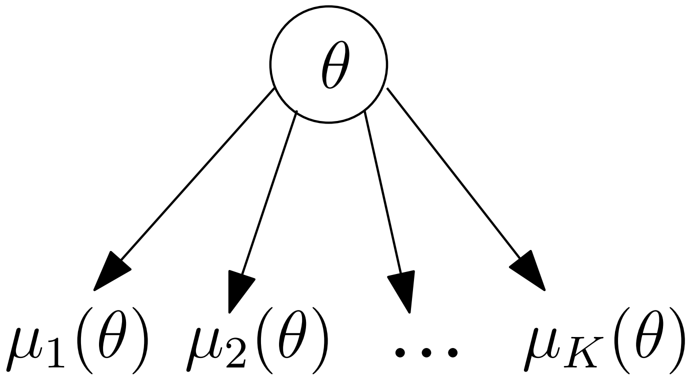
Hidden context
$\theta^* = 3$



Suppose we choose Arm 1: Receive a random reward with mean 2.5

Related Work

- Linear Bandits $\mu_k(\theta) = w_k^T \theta$
- GLM Bandits [Filippi et al] $\mu_k(\theta) = g(w_k^T \theta)$, invertible g
- Global and Regional Bandits [Atan et al, Wang et al], invertible $\mu_k(\theta)$
- Known conditional reward distributions [Combes et al 2017]
- Closest work: [Lattimore et al 2014], works for UCB

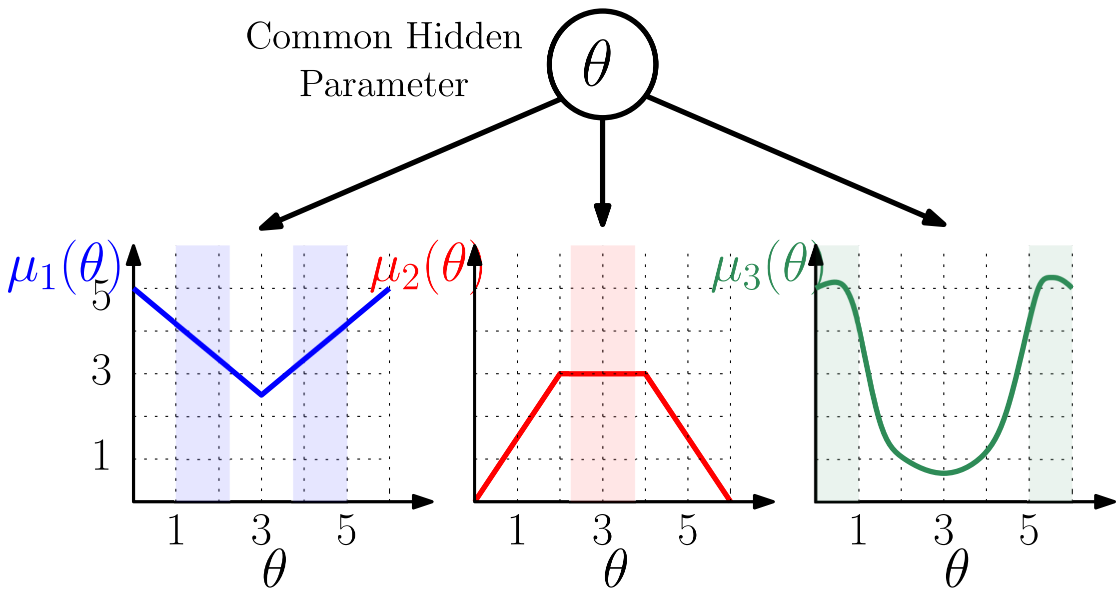


Overview of Our Algorithm

1) Estimating a confidence set $\widehat{\Theta}_t$ for θ^*

2) Remove $\widehat{\Theta}_t$-non-competitive Arms for step t

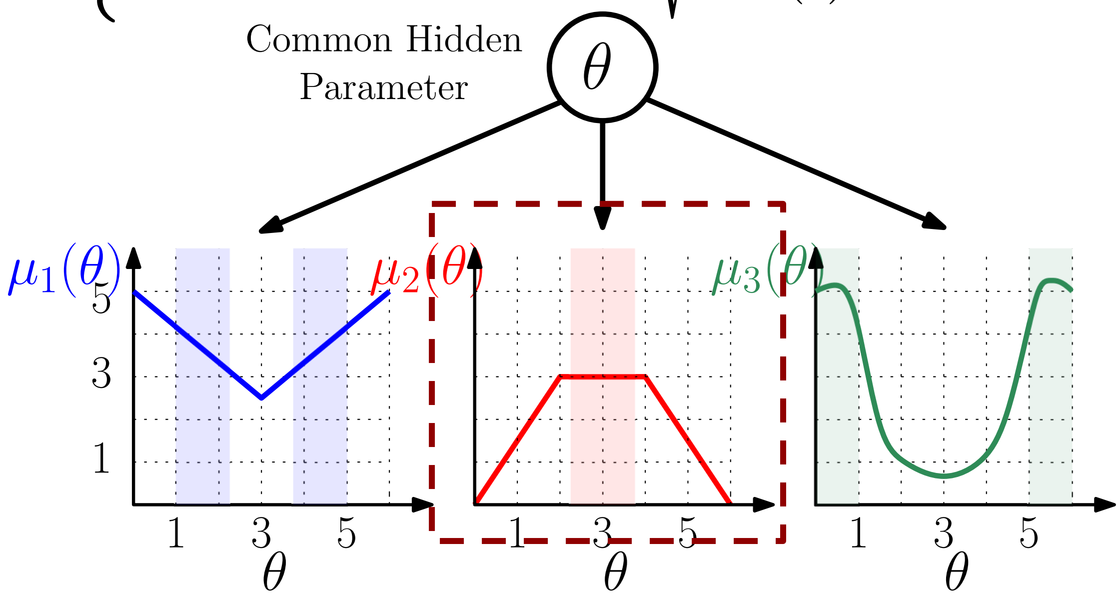
3) Play one of $\widehat{\Theta}_t$-competitive arms using any classic bandit algorithm



Step 1: Estimating a Confidence set $\widehat{\Theta}_t$

- Obtain the empirical mean $\hat{\mu}_k(t)$ of each arm k using its $n_k(t)$ samples until time
- The confidence set is constructed as follows

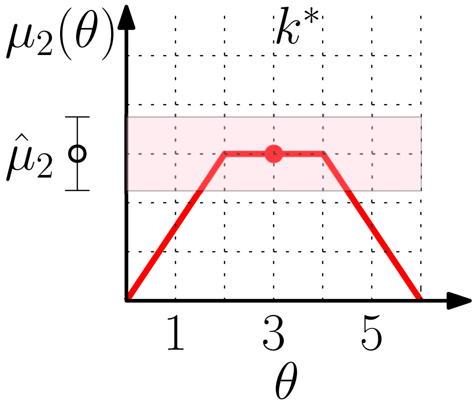
$$\hat{\Theta}_t = \left\{ \theta : |\hat{\mu}_k(t) - \mu_k(\theta)| \leq \sqrt{\frac{a \log t}{n_k(t)}}, \text{ for all } k \in [K] \right\}$$



Step 1: Estimating a Confidence set $\widehat{\Theta}_t$

- Obtain the empirical mean $\hat{\mu}_k(t)$ of each arm k using its $n_k(t)$ samples until time t
- The confidence set is constructed as follows

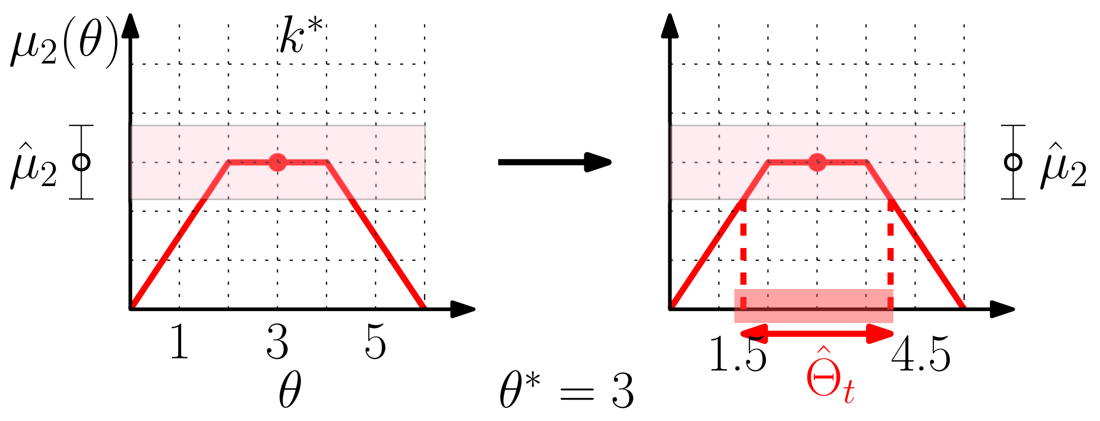
$$\hat{\Theta}_t = \left\{ \theta : |\hat{\mu}_k(t) - \mu_k(\theta)| \leq \sqrt{\frac{a \log t}{n_k(t)}}, \text{ for all } k \in [K] \right\}$$



Step 1: Estimating a Confidence set $\widehat{\Theta}_t$

- Obtain the empirical mean $\hat{\mu}_k(t)$ of each arm k using its $n_k(t)$ samples until time t
- The confidence set is constructed as follows

$$\hat{\Theta}_t = \left\{ \theta : |\hat{\mu}_k(t) - \mu_k(\theta)| \leq \sqrt{\frac{a \log t}{n_k(t)}}, \text{ for all } k \in [K] \right\}$$

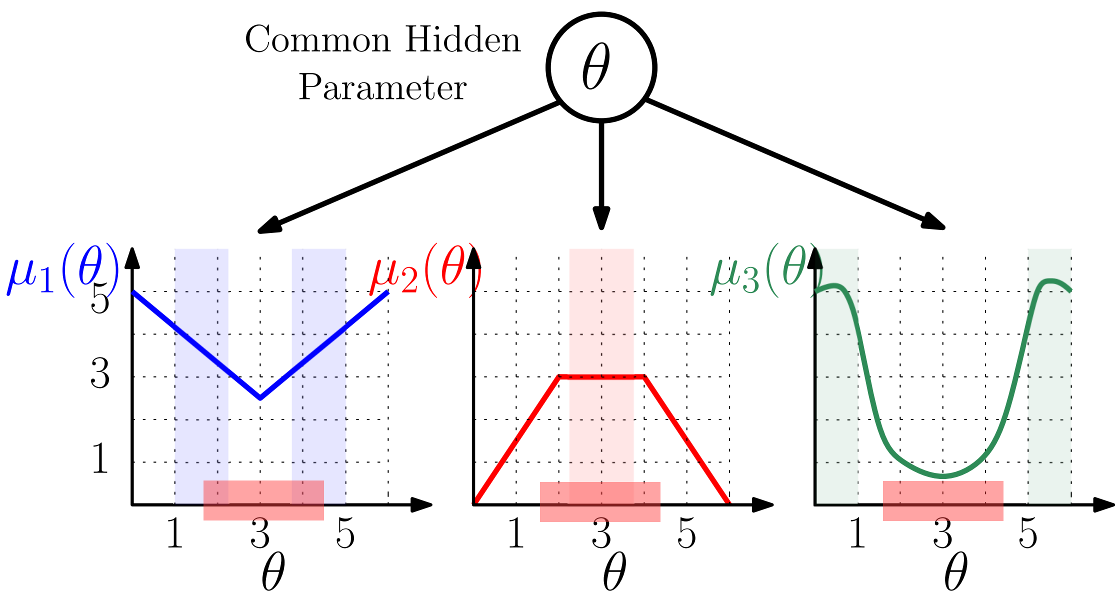


Step 2: Remove $\widehat{\Theta}_t$-non-competitive Arms

- For $\widehat{\Theta}_t$ = [1.5,4.5], then Arm 3 cannot be the best arm since

$$\mu_k(\theta) < \max_{l \in \{1,2\,..K\}} \mu_l(\theta) \quad \forall \theta \in \widehat{\Theta}_t$$

- We say that Arm 3 is $\widehat{\Theta}_t$*-non-competitive* and focus on arms 1 & 2

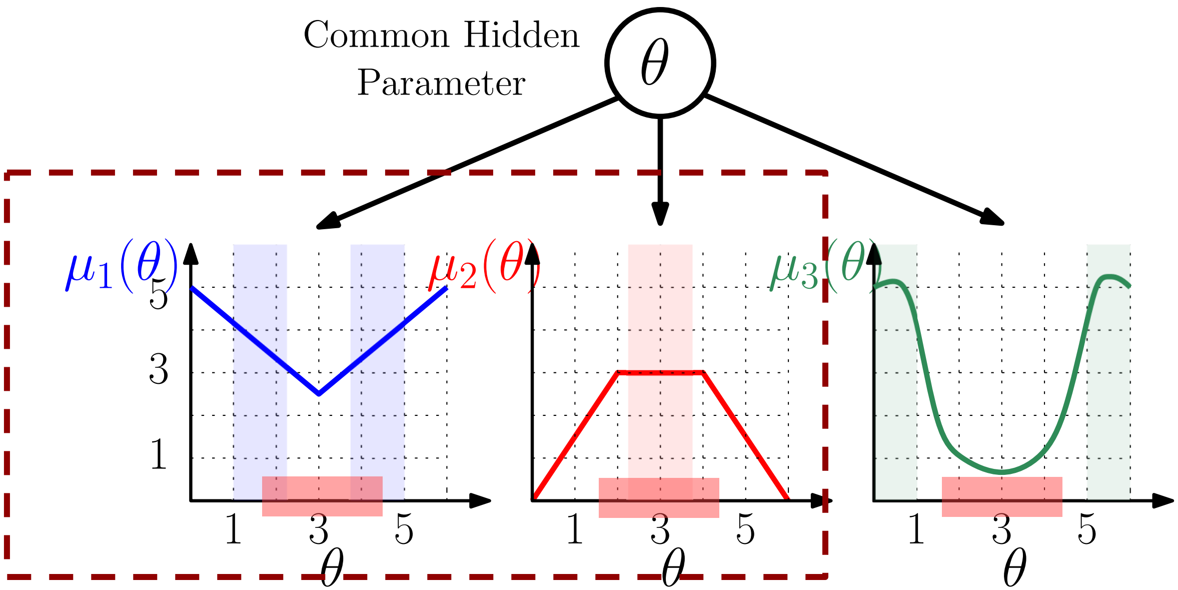


Step 2: Remove $\widehat{\Theta}_t$-non-competitive Arms

- For $\widehat{\Theta}_t$ = [1.5,4.5], then Arm 3 cannot be the best arm since

$$\mu_k(\theta) < \max_{l \in \{1,2\,..K\}} \mu_l(\theta) \quad \forall \theta \in \widehat{\Theta}_t$$

- We say that Arm 3 is $\widehat{\Theta}_t$-*non-competitive* and focus on arms 1 & 2

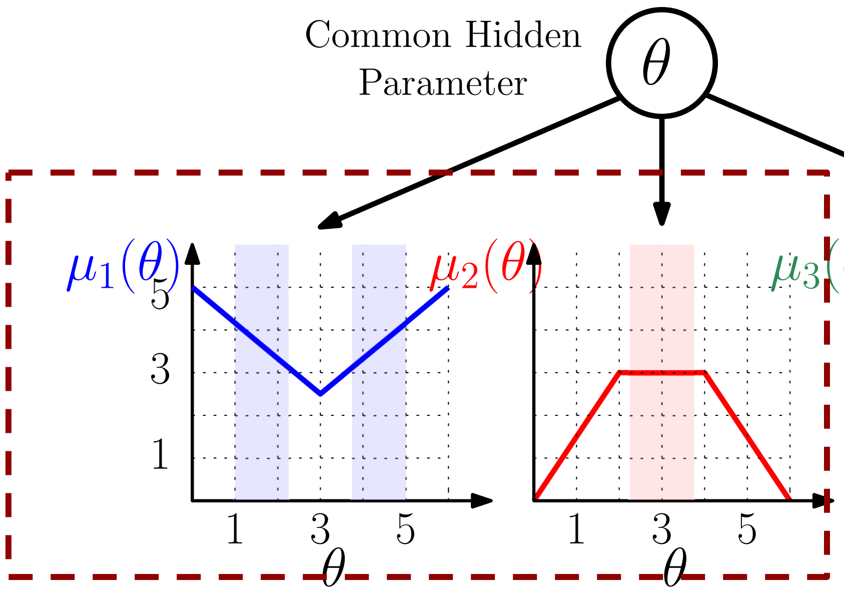


Step 2: Remove $\widehat{\Theta}_t$-non-competitive Arms

- For $\widehat{\Theta}_t$ = [1.5,4.5], then Arm 3 cannot be the best arm since

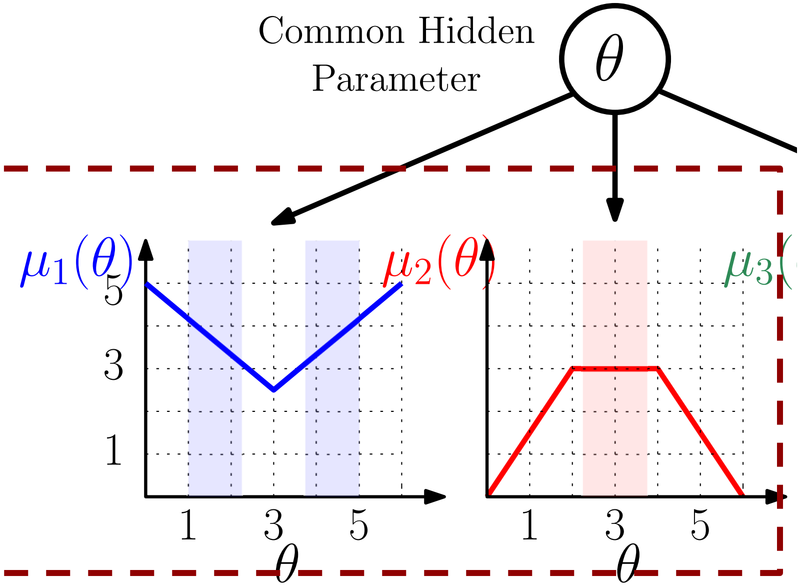
$$\mu_k(\theta) < \max_{l \in \{1,2\,..K\}} \mu_l(\theta) \quad \forall \theta \in \widehat{\Theta}_t$$

- We say that Arm 3 is *$\widehat{\Theta}_t$-non-competitive* and focus on arms 1 & 2



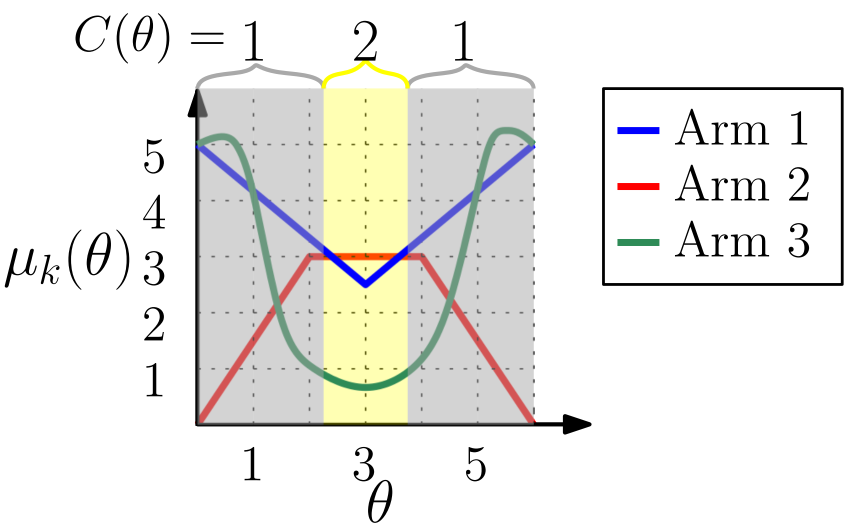
Step 3: Use any classic bandit algorithm

- Options: UCB, Thompson sampling, KL-UCB, etc.
- Previous work [Lattimore et al 2014] works only with a modified UCB arm-pulling scheme



What are competitive and non-competitive arms?

- Θ^* is the confidence set after pulling arm k^* infinitely many times
- An arm is non-competitive if it is Θ^* −non-competitive, that is, if there exists $\epsilon > 0$: $\mu_{k^*}(\theta) > \mu_k(\theta)\ \forall\{\ \theta\text{: } |\mu_{k^*}(\theta^*) - \mu_{k^*}(\theta)| < \epsilon\ \}$
- Number of competitive arms depends on the hidden θ



Regret Bound for UCB-C

Theorem 1: Expected pulls of Non-Competitive arms are bounded, i.e. $O(1)$

$$E[n_k(T)] \leq Kt_0 + \sum_{t=1}^{T} 2Kt^{1-\alpha} + K^3 \sum_{Kt_0}^{T} 6\left(\frac{t}{K}\right)^{2-\alpha}$$

$$= O(1)$$

Theorem 2: Expected pulls of Competitive arms are $O(\log T)$

$$E[n_k(T)] \leq \frac{8\alpha\sigma^2 \log T}{\Delta_k^2} + \frac{2\alpha}{\alpha - 2} + \sum_{t=1}^{T} 2Kt^{1-\alpha}$$

$$= O(\log T)$$

Regret Bound for UCB-C

Theorem 1: Expected pulls of Non-Competitive arms are bounded, i.e. $O(1)$

$$E[n_k(T)] \leq Kt_0 + \sum_{t=1}^{T} 2Kt^{1-\alpha} + K^3 \sum_{Kt}^{T} 6\left(\frac{t}{K}\right)^{2-\alpha}$$

$E[Reg(T)] \leq (C-1)$ O(log T) + O(1)

If C = 1, E[Reg(T)] = O(1)

$$E[n_k(T)] \leq \frac{8\alpha\sigma^2 \log T}{\Delta_k^2} + \frac{2\alpha}{\alpha-2} + \sum_{t=1} 2Kt^{1-\alpha}$$

$$= O(\log T)$$

Comparison with Classical Bandits

Regret upper bound of classic UCB/TS

$$Reg_{UCB}(T) = (K - 1) \times O(\log T)$$

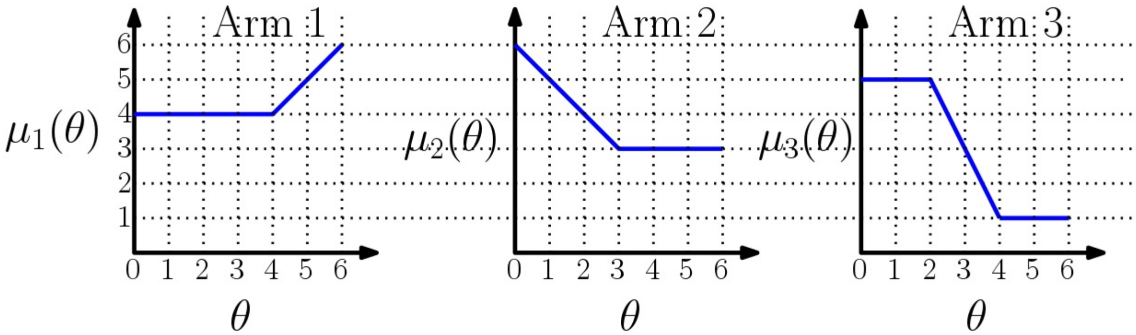
Each sub-optimal arm pulled $O(\log T)$ times

Regret upper bound for UCB-C

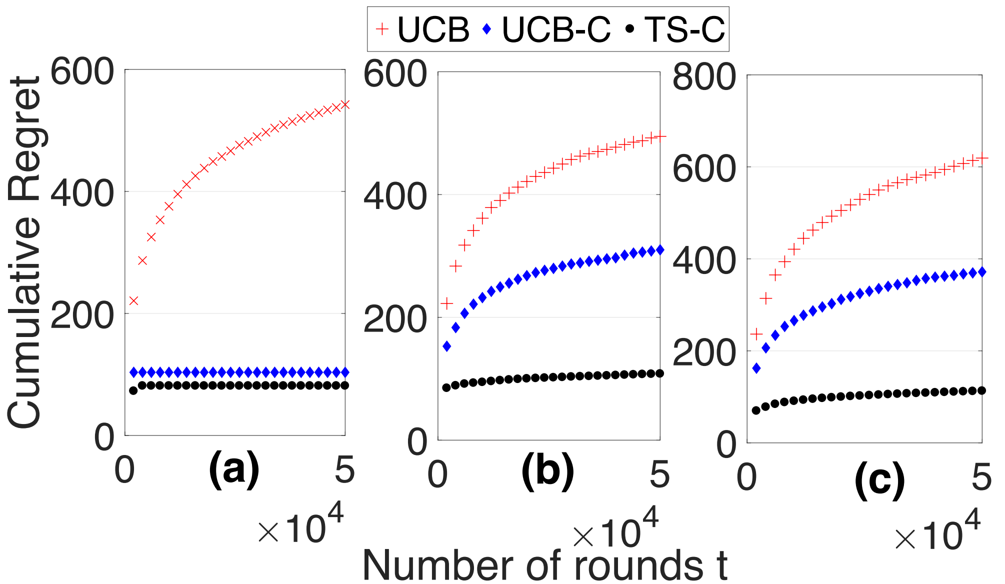
$$Reg_{UCB-C}(T) = (C - 1) \times O(\log T) + O(1)$$

Only C-1 competitive sub-optimal arms are pulled $O(\log T)$ times

Simulations



Rewards $\sim N(\mu_k(\theta^*), 4)$



$\theta^* = 0.5$ $\quad$ $\theta^* = 1.8$ $\quad$ $\theta^* = 2.8$

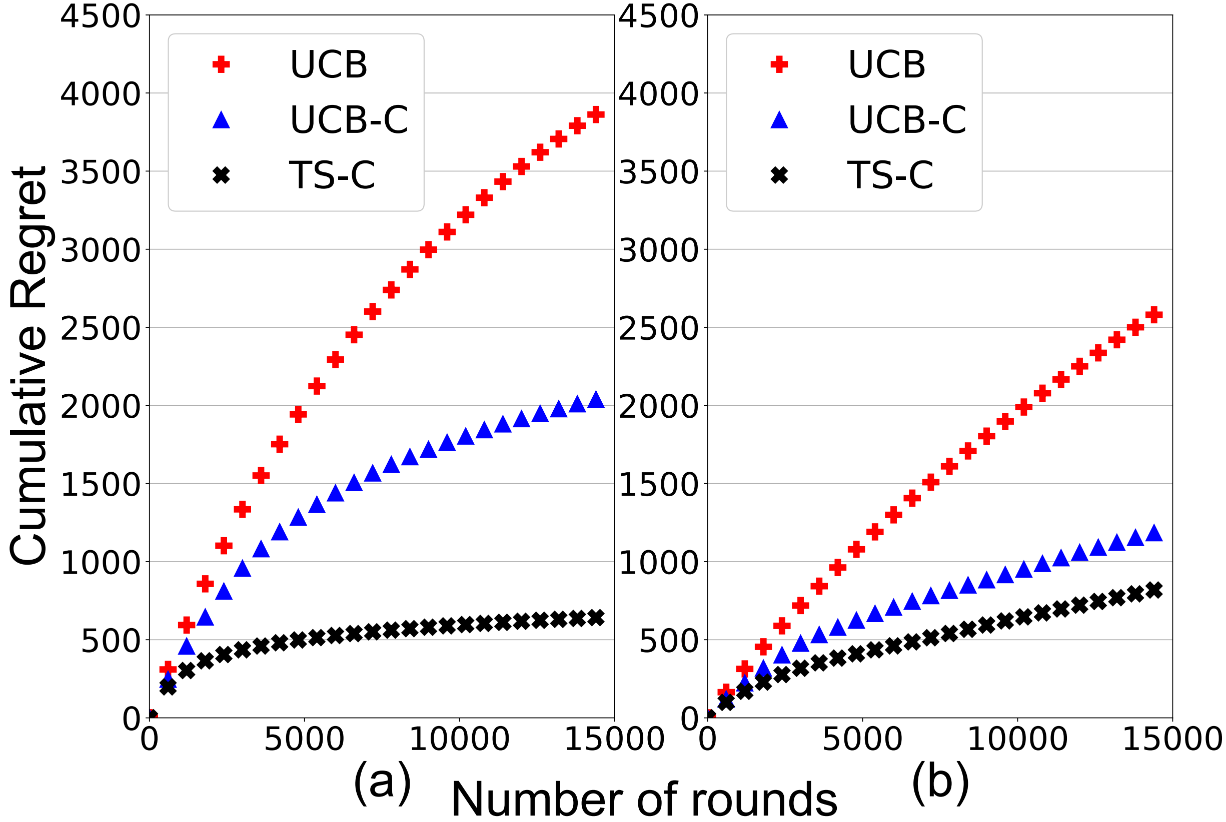
$C = 1$ $\quad$ $C = 2$ $\quad$ $C = 3$

Experiments on the MovieLens Dataset

- Dataset has 1M ratings for 3883 movies by 6040 users
- Movies have 18 different genres
- We classify users based on θ = (age, occupation) pair
- Mean rewards learnt on 50% of the dataset

GOAL: Find the right movie genre for an unknown user type

Experiments on MovieLens

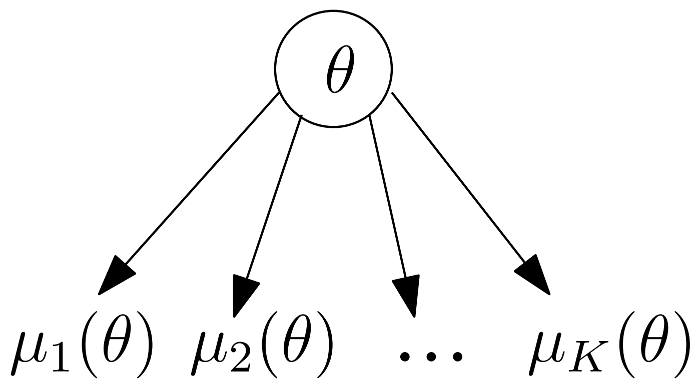


$\theta^* =$ 18-25 year old college students

$\theta^* =$ 25-34 year old executives

Key Takeaways

- Exp. Regret O ((C-1) log T), C is the no. of *competitive* arms
- Competitive sub-optimal arms are pulled O(log T) times, and non-competitive arms are pulled O(1) times
- When C = 1, we get bounded or O(1) regret!
- Allows us to use any classic bandit algorithm (UCB, TS, etc.)



Future Directions

- Best-Arm Identification problem in the structured setting
- Better use of informative arms that can help estimate θ