Demystifying Online Clustering of Bandits: Enhanced Exploration Under Stochastic and Smoothed Adversarial Contexts

Zhuohua Li

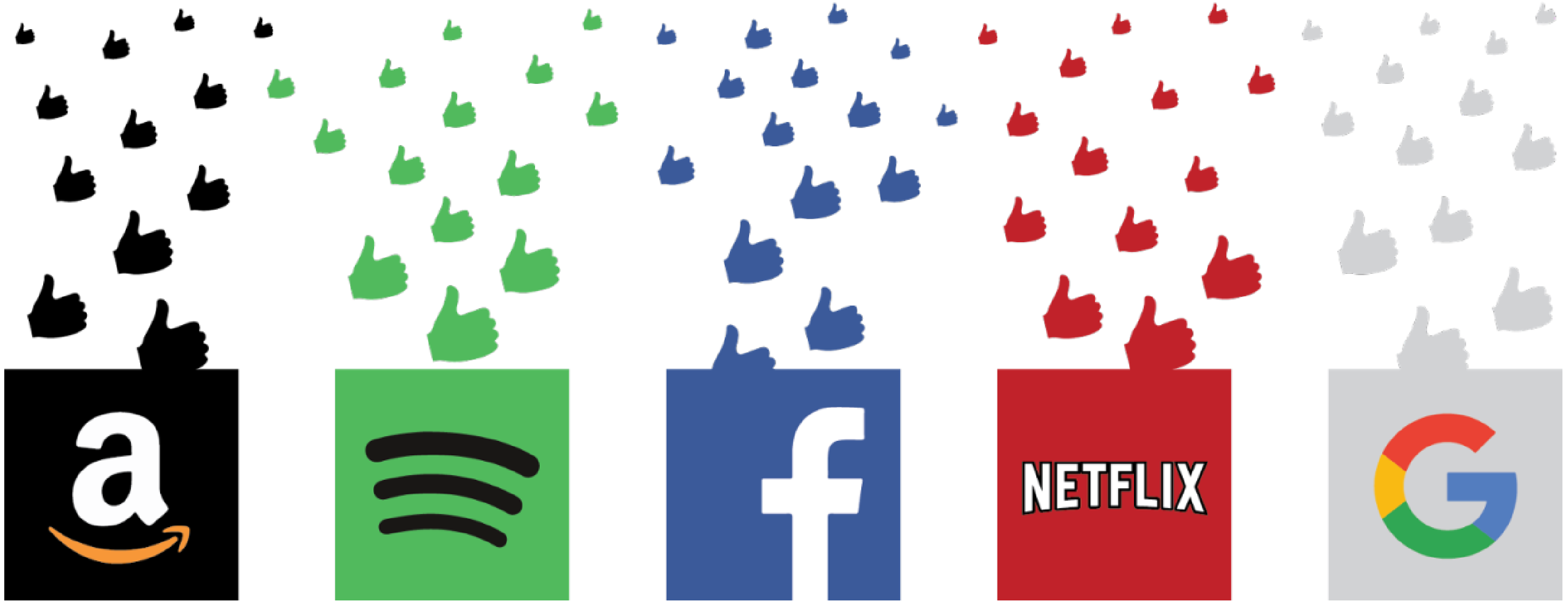
The Chinese University of Hong Kong

Online Learning with Full Information Feedback

- Online learning: learning from data **sequentially**
- Can observe feedback of every action

Online Learning with Bandit Feedback

- Can only observe feedback for the selected action



Multi-armed Bandits [Thompson (1933)]

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

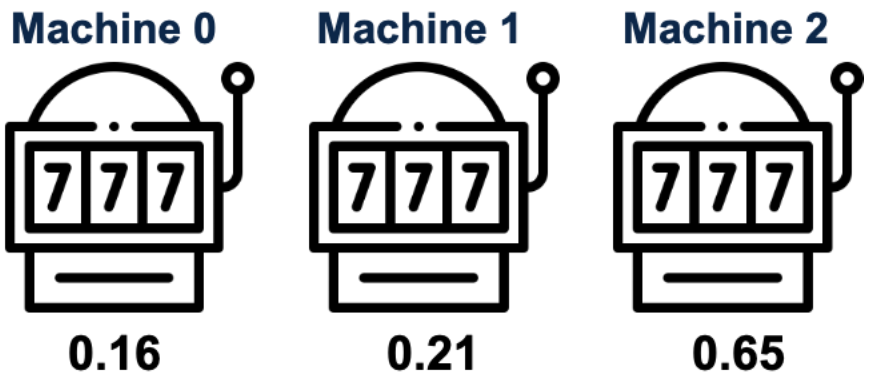
- Which arm should we choose next?

Setting: Finite-armed Stochastic Bandits

Ads/products/movies/news

Click rates/profits

- There are K arms
  - Each arm a has an unknown reward distribution v_a with unknown mean μ_a
  - The best arm is $a^* = \operatorname{argmax}_a \mu_a$



- At each time t
  - The learning agent selects an arm a_t
  - Observes the reward $X_t \sim v_{a_t}$

Bandit feedback

Objective

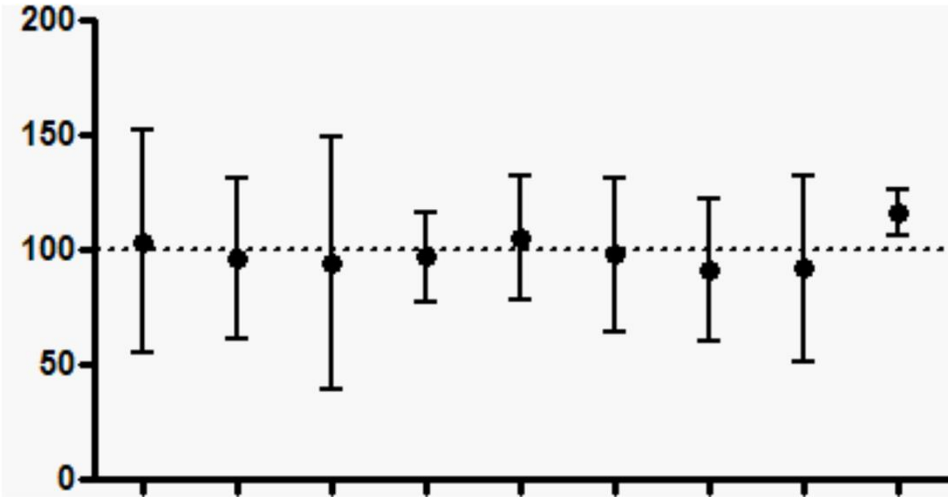
- Minimize the regret in T rounds:

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

- Balance the trade-off between exploration and exploitation
  - Exploration: Select arms that have not been tried much before
  - Exploitation: Select arms that yield good results so far
- Smaller order of T in $R(T)$ is better

UCB-Upper Confidence Bound [Auer et al. (2002)]

- Let $N_a(t)$ be the selection times of arm a till round t
- The sample mean of arm a is $\hat{\mu}_a(t) = \frac{\sum_{i=1}^{t} X_i \, \mathbb{I}\{a_t = a\}}{N_a(t)}$
- By Hoeffding's inequality, with high probability:



$$\mu_a \in \left[\hat{\mu}_a(t) - \sqrt{\frac{2\log t}{N_a(t)}}, \hat{\mu}_a(t) + \sqrt{\frac{2\log t}{N_a(t)}}\right]$$

- Algorithm: Select arm a_t with:

$$a_t = \operatorname{argmax}_a \hat{\mu}_a(t) + \sqrt{\frac{2\log t}{N_a(t)}}$$

- Regret: $R(T) = O(\sqrt{T})$

Linear Bandits

A **time-varying** set of ads/products/movies/news

- At each time t:
  - The learning agent receives $\mathcal{A}_t \subset \mathbb{R}^d$
  - Selects an arm $a_t \in \mathcal{A}_t$
  - Receives a random reward:

Bandit feedback with linear structure

$$r_t = \theta^T a_t + \eta_t$$

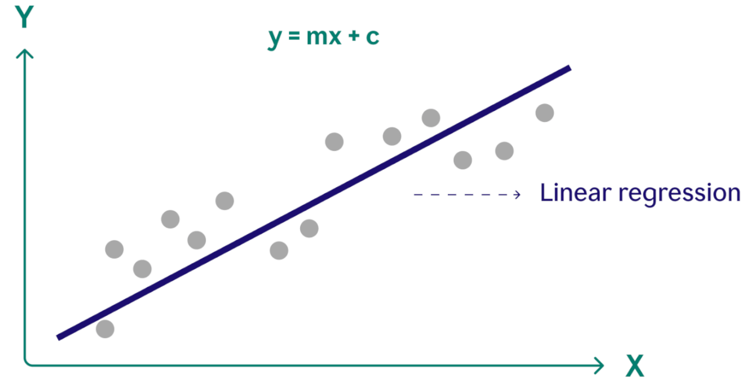
Noise term

for some fixed but unknown vector $\theta \in \mathbb{R}^d$

LinUCB Algorithm [Li et al. (2010)]

- Given the observed feedback till time t:

$$\{(a_1, r_1), (a_2, r_2), \dots, (a_t, r_t)\}$$



- Perform ordinary least squares:

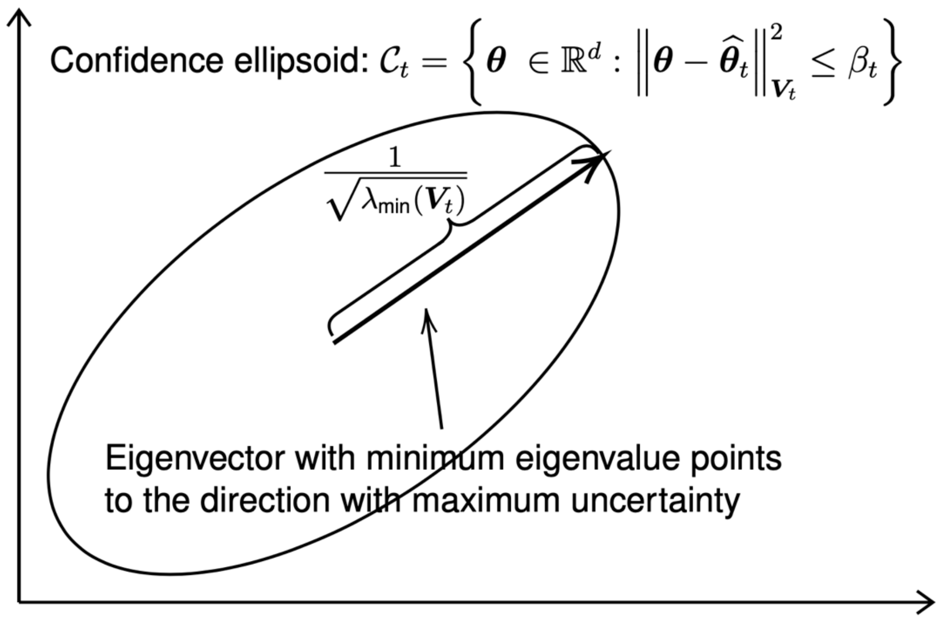
$$V_t = \sum_{s=1}^{t} a_s a_s^T + \lambda I, \qquad b_t = \sum_{s=1}^{t} r_s \, a_s$$

$$\hat{\theta}_t = V_t^{-1} b_t$$

LinUCB Algorithm [Li et al. (2010)]

- With high probability:

$$\left\|\theta - \hat{\theta}_t\right\|_{V_t}^2 = \left(\theta - \hat{\theta}_t\right)^T V_t\left(\theta - \hat{\theta}_t\right) \leq \beta_t$$



- Regret: $R(T) = O\left(\sqrt{T}\right)$

Simultaneous Inference and Regret Minimization: Is It Possible?

- Is there any algorithm such that:
  1. Can estimate θ precisely
  2. Achieve $O(\sqrt{T})$ regret
- Clustering of bandits problem [Gentile et al. (2014)]:
  - We have multiple vectors to estimate: $\theta_1, \theta_2, \ldots, \theta_N$, some of them are the same (i.e., in the same cluster)
  - The authors prove that **simultaneous clustering and regret minimization** is possible if arms are sampled from a distribution X such that:
    1. $\lambda_{\min}(\mathbb{E}[XX^T]) = \lambda_x$
    2. For any unit vector $z \in \mathbb{R}^d$, $\mathrm{Var}[(z^T X)^2] \leq \frac{\lambda_x^2}{8 \log 4K}$

A Long-standing Open Problem

- The assumption used in [Gentile et al. (2014)] is very restrictive, and it is unknown how to eliminate it
- Some studies (e.g., [Amani et al. (2019)]) do not use this assumption, but get deteriorated regret of $O\left(T^{\frac{2}{3}}\right)$
- Open Problem:

  Can we achieve $O\left(\sqrt{T}\right)$ regret without using the assumptions?

Restrictive Assumptions

- In fact, it is known that the assumptions in [Gentile et al. (2014)] do not even hold

Proposition 1. *Suppose that* $\epsilon < 1/27$ *and* $\rho > 0$. *There does not exist a probability measure* μ *on* $\mathbb{R}^d$ *such that when* X *has law* μ *the following hold:*

(a) $\mathbb{E}[XX^\top] \succeq \rho I$; *and*

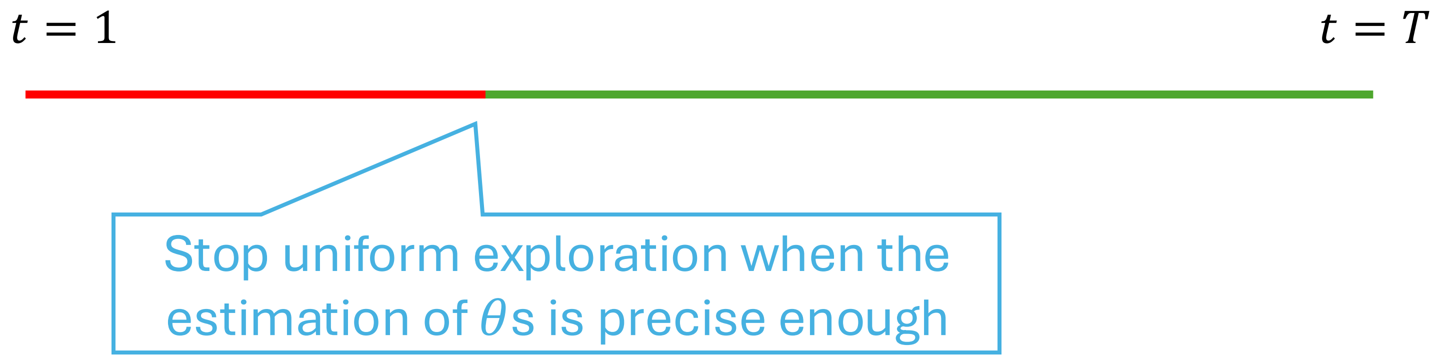
(b) $\mathbb{V}[\langle X, \eta \rangle^2] \leq \epsilon \rho^2$ *for all unit vectors* η.

Our Contribution 1

- We propose an algorithm that achieves $O(\sqrt{T})$ regret if the minimum gap γ between clusters is known

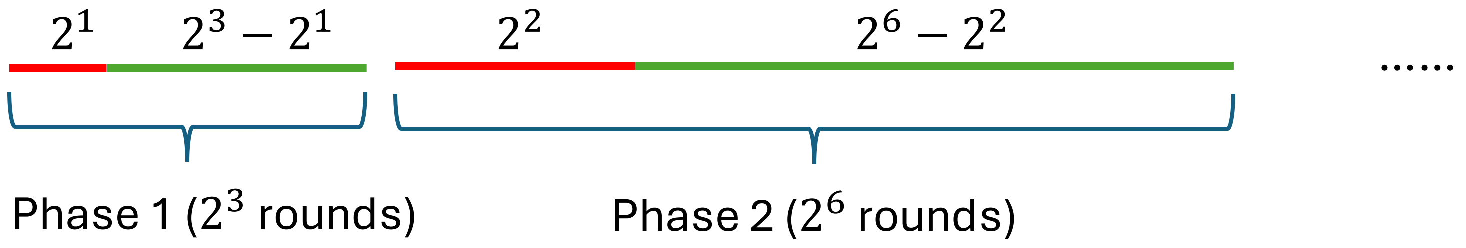
$$\left\|\theta_i - \theta_j\right\|_2 \geq \gamma \text{ for any } i \text{ and } j \text{ from different clusters}$$

- The idea is to incorporate an appropriate amount of **uniform exploration** into the **UCB strategy**



Our Contribution 2

- When γ is **unknown**, we design a phase-based algorithm that also achieves $O(\sqrt{T})$ regret
- The idea is to split the time horizon into phases:



- Conclusion: **Yes, we can achieve simultaneous inference and regret minimization!**