

OPEN PROBLEM

CMABs with Ranking Feedback: play k arms, see *partial feedback* from a **Plackett-Luce ranking** with parameters θ , and minimize **top- k regret**.

Current Guarantees [1]

$$\text{Reg}_{\theta}(T) \lesssim \text{Regime}_{\theta}(T) + \text{BurnIn}_{\theta}(T)$$

$$\sup_{\theta} \min\{\text{BurnIn}_{\theta}(T), T\} = T$$

Research Question

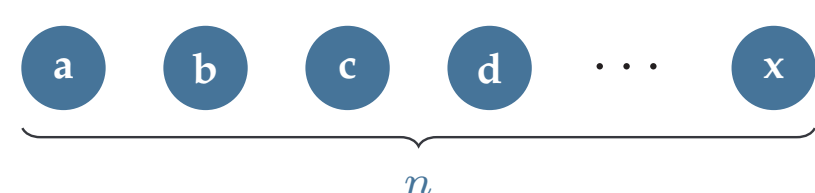
Can we achieve

$$\sup_{\theta} \text{Reg}_{\theta}(T) = \tilde{O}(\sqrt{T})?$$

SETTING

Items

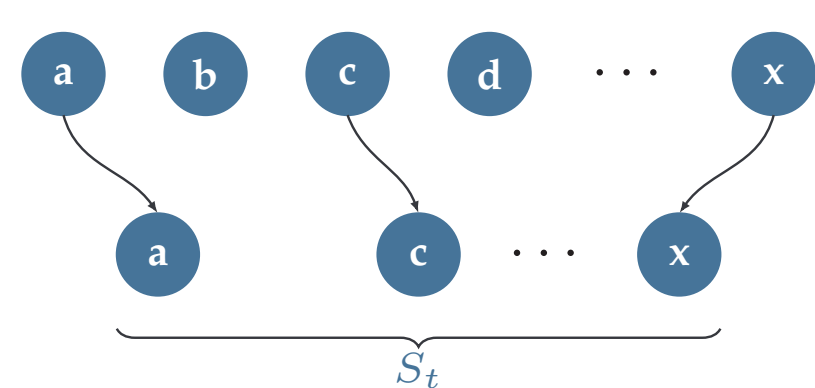
A set of **items** $[n]$
with PL parameters θ_i



Assortment

$$\text{Pull } S_t \subseteq [n]$$

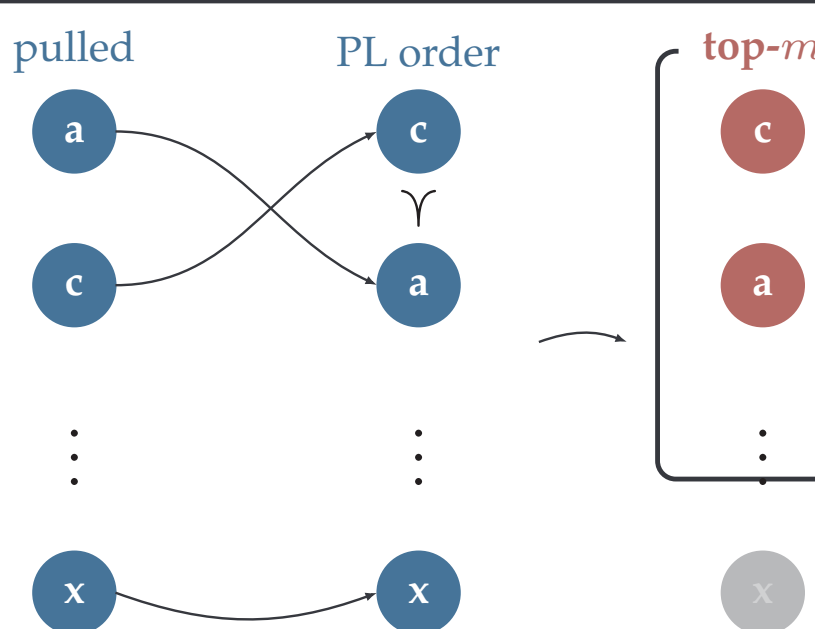
$$|S_t| = k$$



Feedback

Latent PL ranking
 $R_t \sim \text{PL}((\theta_i)_{i \in S_t})$

Observe top- m : $R_t^{(m)}$



Reward

$$\text{Gain } G_t = \frac{1}{k} \sum_{i \in S_t} \theta_i$$

Performance Index

$$\text{Reg}_{\theta}(T) = \mathbb{E}_{\theta} \left[\sum_{t=1}^T (G^* - G_t) \right]$$

G^* : gain of the **best** k items

REGRET LOWER BOUND

KL Decomposition

$$D_{\text{KL}}(P_{\theta} \| P_{\theta'}) \lesssim \frac{m}{k} \sum_i (\theta_i - \theta'_i)^2 \mathbb{E}_{\theta} N_i(T)$$

Pull count: $N_i(T) = \sum_{t=1}^T \mathbb{I}[i \in S_t]$

Parallel Games [2]

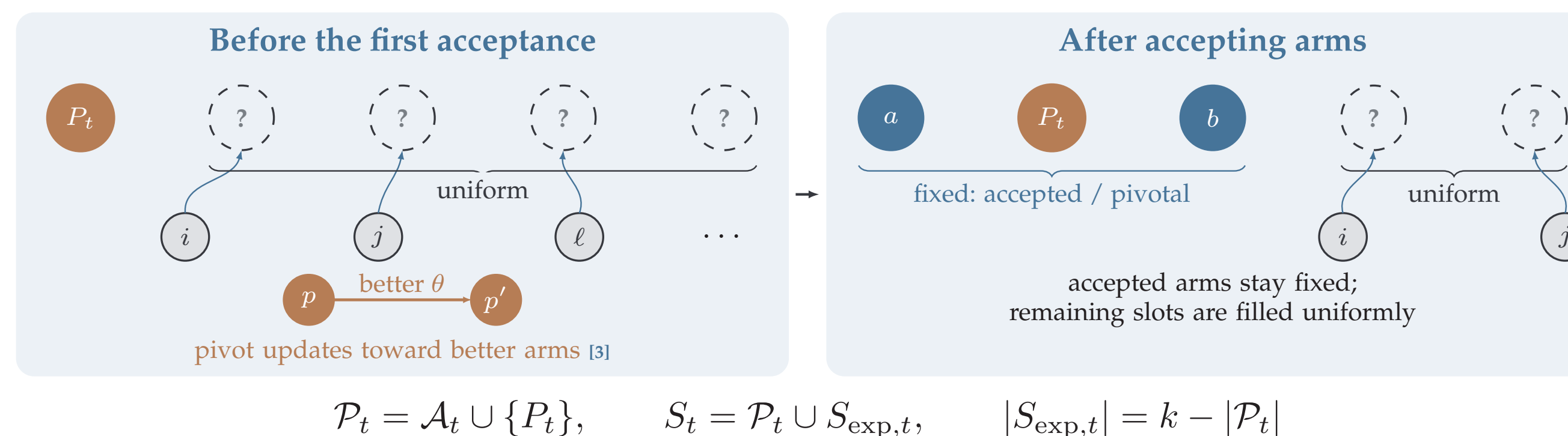
k bandit games $\mathcal{G}_1, \dots, \mathcal{G}_k$ in parallel, each with $q = n/k$ arms



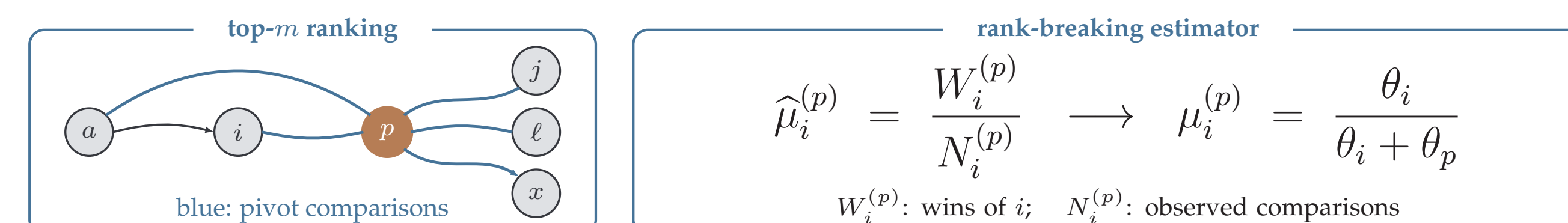
$$\sup_{\theta} \text{Reg}_{\theta}(T) = \Omega(\sqrt{T/m})$$

ALGORITHMIC SOLUTION: DYNAMIC PIVOTING

Elimination & Exploration



Rank Breaking w.r.t. a Pivot



Pivot Observability

Comparison frequency

$$f_{i|S}^{(p)} := \mathbb{P}_{\theta}(R^{(m)}(i) \neq R^{(m)}(p)) \gtrsim \frac{m}{k} (\theta_i + \theta_p)$$

poor i and poor $p \Rightarrow$ comparisons arrive slowly

Acceptance safeguard

$$\text{LCB}_i^{(P_t)} > \frac{1}{2}$$

$$\Rightarrow i \in \mathcal{A}_{t+1} \subseteq \mathcal{P}_{t+1}$$

prevents a stale, uninformative pivot

$$\sup_{\theta} \text{Reg}_{\theta}(T) = \tilde{O}(\sqrt{T/m})$$

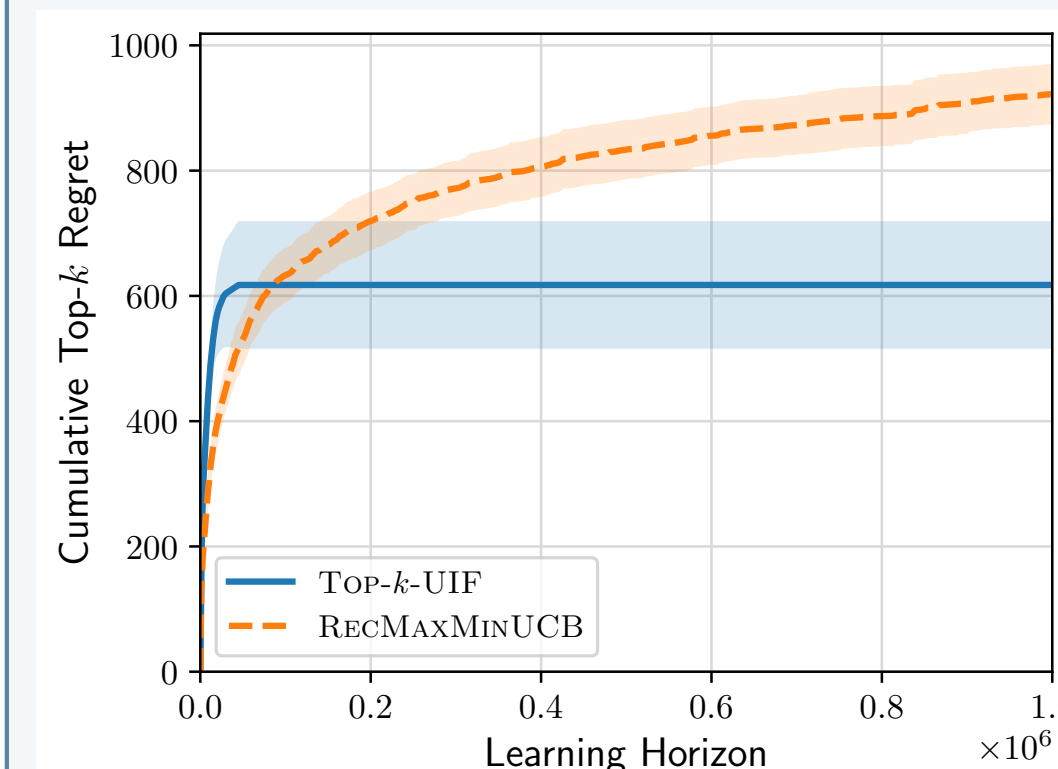
NUMERICAL SIMULATIONS

Comparison. RecMaxMinUCB [1] vs. TOP- k -UIF on synthetic PL instances.

Protocol. 30 independent trials to $T = 10^6$; curves and error bars show mean $\pm$ SEM.

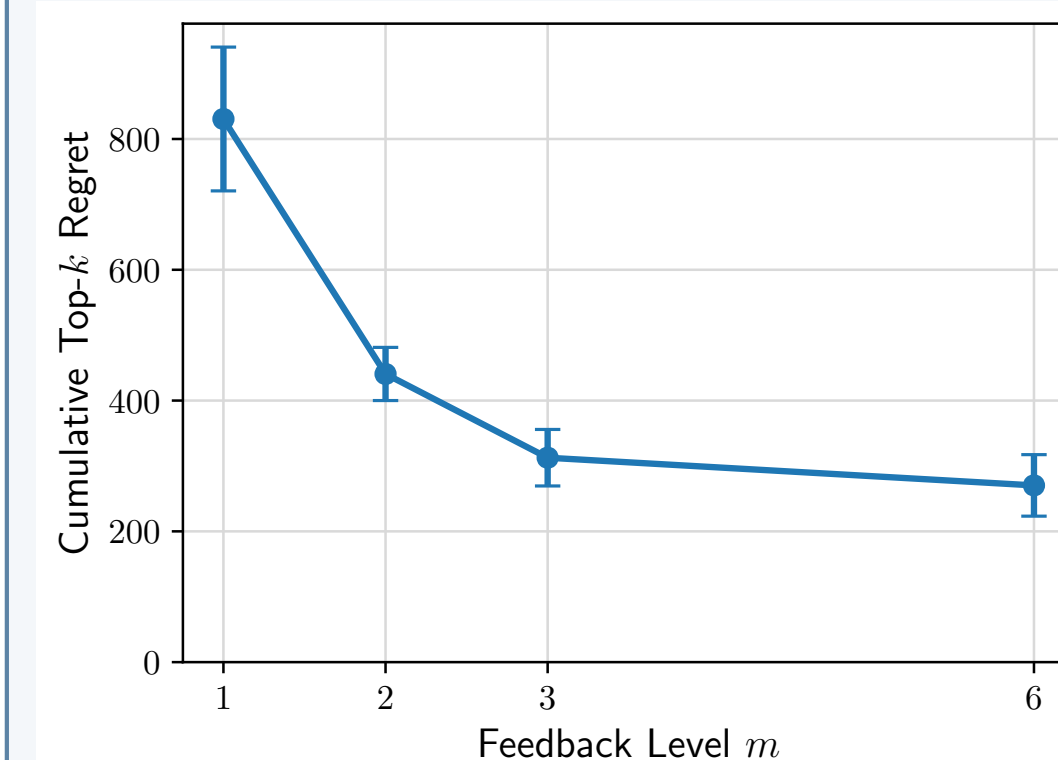
What changes? We track regret over time and vary m to assess richer feedback.

Regret across time



TOP- k -UIF plateaus; RecMaxMinUCB [1] keeps growing ($n = 10, k = m = 3$).

Benefit of top- m feedback



Final regret falls as m grows ($n = 16, k = 6$).

Empirical message. Larger m adds comparisons and lowers regret.

REFERENCES

- [1] Saha & Gopalan. *Combinatorial Bandits with Relative Feedback*. NeurIPS, 2019.
- [2] Audibert, Bubeck & Lugosi. *Regret in Online Combinatorial Optimization*. MOR, 2014.
- [3] Yue, Broder, Kleinberg & Joachims. *The K-Armed Dueling Bandits Problem*. COLT, 2009.