Fast and Regret-Optimal Best Arm Identification: Fundamental Limits and Low-Complexity Algorithms

Qining Zhang, Ph.D. Candidate

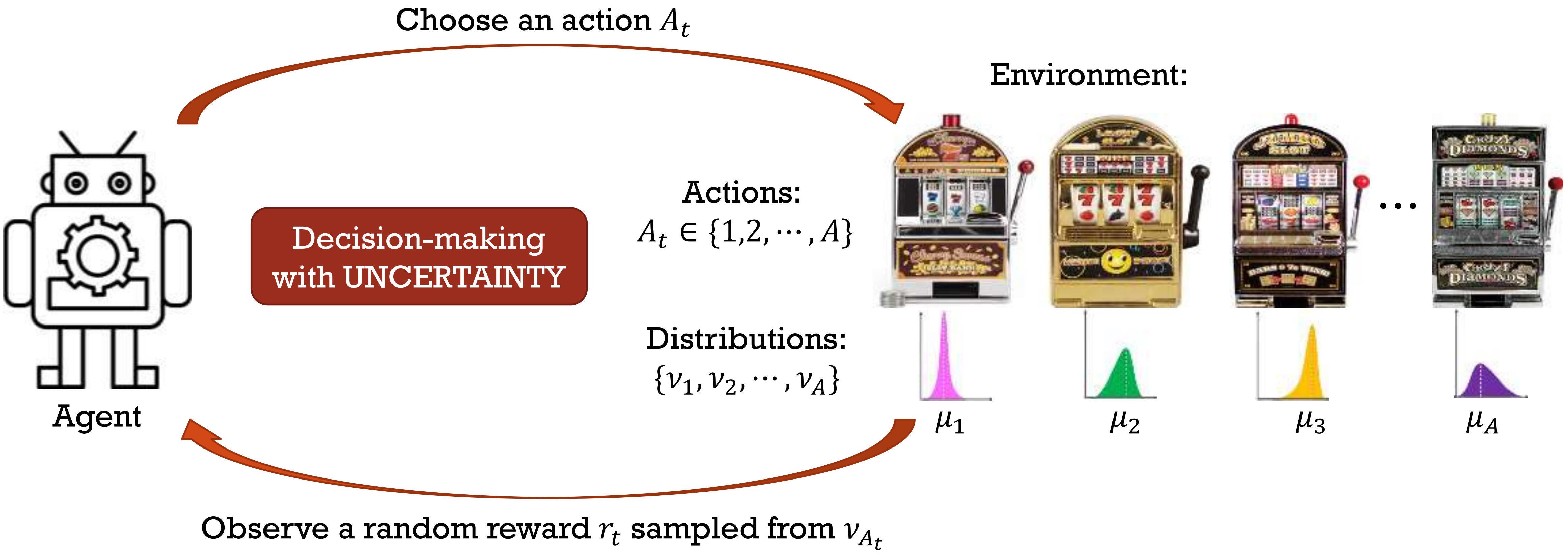
EECS, University of Michigan, Ann Arbor

Joint work with my advisor Lei Ying (Michigan)

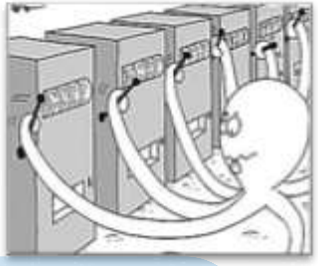
NeurIPS 2023

Multi-armed Bandits

- Online decision making for T slots.



Classic Views of MAB



- Regret Minimization (Lai and Robbins. 1985)
  - Exploration v.s. Exploitation

$$\mathrm{Reg}_\mu(T) = T\mu_{a^*} - \mathbb{E}_\mu\left[\sum_{t=1}^{T} \mu_{A_t}\right]$$

Does not commit to any arm

- Best Arm Identification (Garivier, et al. 2016)
  - Sample Complexity

$$\min \tau \textbf{ s.t. } \Pr(\hat{a}_\tau = a^*) \geq 1 - \delta$$

Over-exploration of suboptimal arms

- Online MAB (Auer et al. 2002)
  - Optimism

UCB bonus

$$A_t = \mathrm{argmax}_a \, \bar{r}_{t-1}(a) + \sqrt{\frac{c\log(T)}{N_{t-1}(a)}}$$

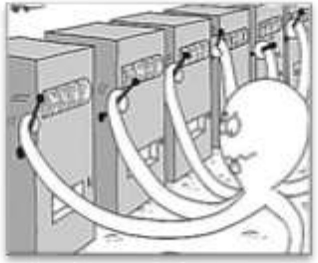
- Offline MAB (Rashidinejad et al. 2021)
  - Pessimism

LCB penalty

$$\hat{a} = \mathrm{argmax}_a \, \bar{r}_N(a) - \sqrt{\frac{c\log(N)}{N(a)}}$$

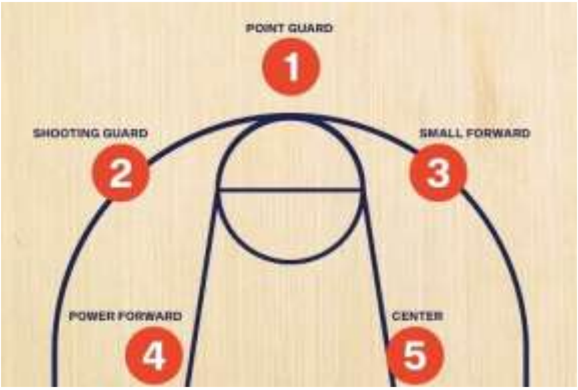
What is the fundamental difference between online and offline data?

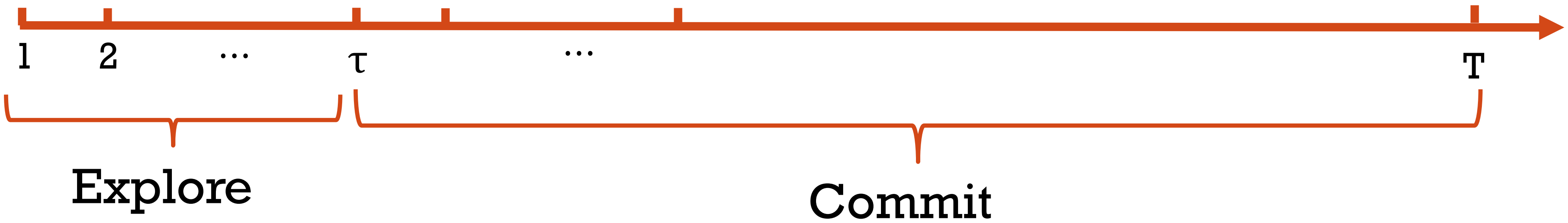
What Happens in Real-World Applications



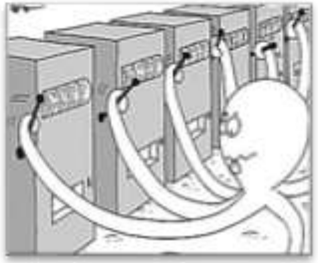
Clinic trials

Career choices





Regret Optimal Best Arm Identification

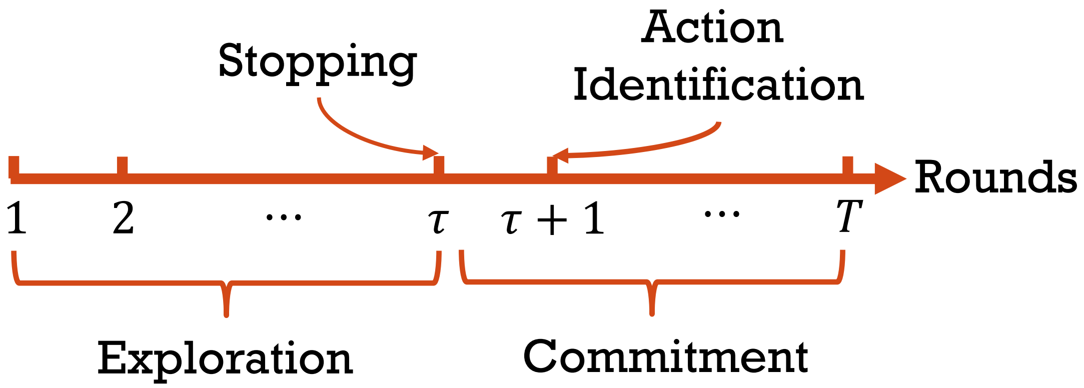


- Two Goals
  - Optimal cumulative regret.
  - Commit to optimal action quickly.
- Three Components
  - Exploration
  - Stopping
  - Action Identification

$$\min_{\pi \in \Pi_{\mathrm{RO}}} \mathbb{E}[\tau] \text{ such that } \Pr(\hat{a} \neq a^*) = \mathcal{O}(T^{-1}),$$

where

$$\Pi_{\mathrm{RO}} = \left\{ \pi : \limsup_{T \to \infty} \frac{\mathrm{Reg}_{\mu}^{\pi}(T)}{\log T} = \sum_{a \neq a^*} \frac{\Delta_a}{\mathrm{KL}(\mu_a, \mu_a^*)} \right\}.$$



Can we design an algorithm for ROBAI?
What are the fundamental limits?

EOCP: Explore Optimistically then Commit Pessimistically

- Exploration
  - Modified-UCB.

$$A_t = \text{argmax}_a \, \bar{r}_{t-1}(a) + \sqrt{\frac{2l}{N_{t-1}(a)}}$$

- Action Identification
  - Modified-LCB algorithm.

$$\hat{a} = \text{argmax}_a \, \bar{r}_{t-1}(a) - \sqrt{\frac{2l}{N_{t-1}(a)}}$$

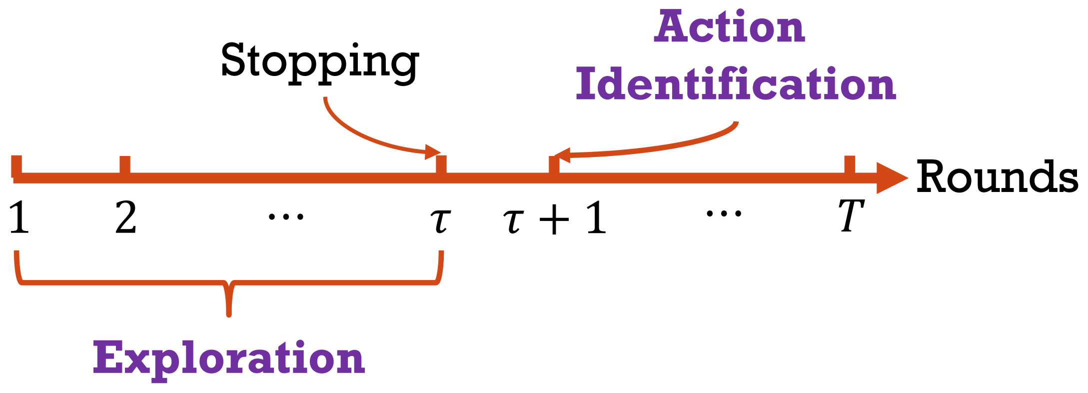
- Stopping
  - Pre-determined Stopping
  - Adaptive Stopping

Stop when $t = \frac{8Al}{\Delta_{\min}^2}$

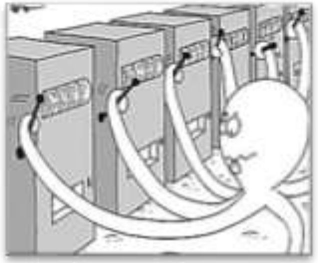
Lower bound on the minimum reward gap

Stop when:

$$\max_a \min_{a'} N_t(a) - lN_t(a') > 1$$



Main Results



- Pre-determined Stopping Time:

Theorem 1. (Pre-determined)
If we choose $l = \log(T) + c_0\sqrt{\log(T)}$, the regret of EOCP is bounded:

$$\mathrm{Reg}_{\mu}^{\mathrm{EOCP}}(T) \leq \sum_{a:\Delta_a>0} \frac{2}{\Delta_a} \log T + o(\log T).$$

And $\mathrm{SCC}_{\mu}^{\mathrm{EOCP}}(T) = \mathcal{O}\left(\Delta_{\min}^{-2} \log T\right)$.

- Adaptive Stopping Time:

Constant Optimal Regret

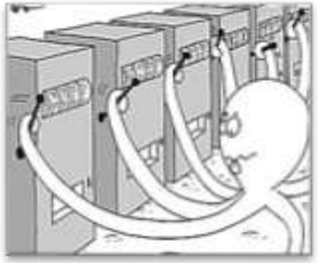
Theorem 2. (Adaptive)
If we choose $l = \log(T) + c_0\sqrt{\log(T)}$, the regret of EOCP-UG is bounded:

$$\mathrm{Reg}_{\mu}^{\mathrm{EOCP}}(T) \leq \sum_{a:\Delta_a>0} \frac{2}{\Delta_a} \log T + o(\log T).$$

And $\mathrm{SCC}_{\mu}^{\mathrm{EOCP}}(T) = \mathcal{O}\left(\Delta_{\min}^{-2} \log^2 T\right)$.

Sample Complexity Loss

Fundamentality



- Commitment Time Limits for Regret Optimal Algorithms

Theorem 3 (Informal).
For 2-armed Gaussian bandit, for any algorithm π with regret is $\mathcal{O}(\log^c T)$ away from optimal, in pre-determined setting:

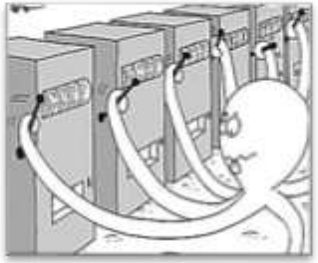
$$\mathrm{SCC}_\mu^\pi(T) = \Omega\left(\frac{\log(T)}{\Delta^2}\right),$$

in adaptive setting:

$$\mathrm{SCC}_\mu^\pi(T) = \Omega\left(\frac{\log^{2-c}(T)}{\Delta^2}\right).$$

- EOCP matches the LB when Δ is known a priori.

Comparison to Literature



Bandit Algorithm	Regret	Sample Complexity	Confidence
UCB(Auer, et al. 2002)	$\frac{2}{\Delta}\log(T)$	T	N/A
TS(Thompson. 1933)	$\frac{2}{\Delta}\log(T)$	T	N/A
BAI-ETC(Garivier, et al. 2016)	$\frac{4}{\Delta}\log(T)$	$\mathcal{O}\left(\frac{\log(T)}{\Delta^2}\right)$	$\tilde{\mathcal{O}}(T^{-1})$
EOCP(Ours)	$\frac{2}{\Delta}\log(T)$	$\mathcal{O}\left(\frac{\log(T)}{\Delta^2}\right)$	$\mathcal{O}(T^{-1})$
EOCP-UG(Ours)	$\frac{2}{\Delta}\log(T)$	$\mathcal{O}\left(\frac{\log^2(T)}{\Delta^2}\right)$	$\mathcal{O}(T^{-1})$
KL-EOCP(Ours)	$\frac{\Delta}{\mathrm{KL}(\mu_2,\mu_1)}\log T$	$\mathcal{O}\left(\frac{\log(T)}{\mathrm{KL}(\mu_2,\mu_1)}\right)$	$\mathcal{O}(T^{-1})$
Lower Bound (Gaussian)	$\frac{2}{\Delta}\log(T)$	$\mathcal{O}\left(\frac{\log(T)}{\Delta^2}\right)$	$\mathcal{O}(T^{-1})$
Lower Bound (General)	$\frac{\Delta}{\mathrm{KL}(\mu_2,\mu_1)}\log T$	$\mathcal{O}\left(\frac{\log(T)}{\mathrm{KL}(\mu_2,\mu_1)^2}\right)$	$\mathcal{O}(T^{-1})$

Thanks!
Questions?

https://arxiv.org/abs/2309.00591