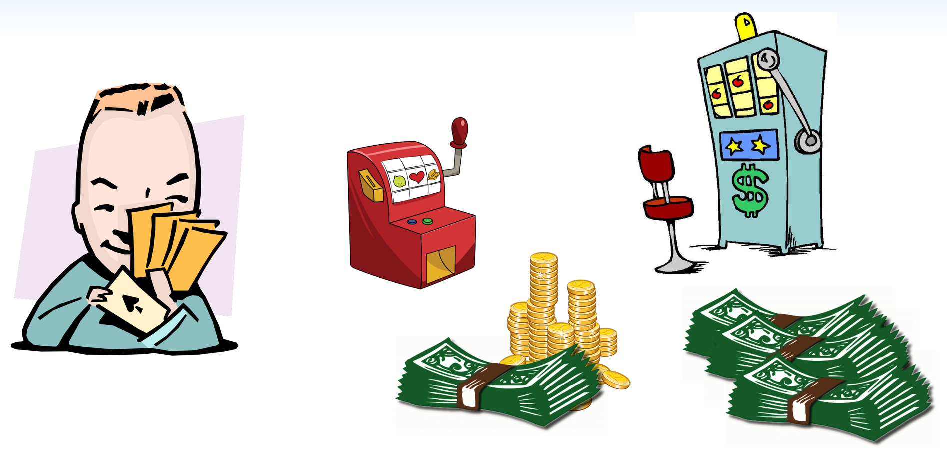


Blocking Bandits

Soumya Basu#, Rajat Sen*, Sujay Sanghavi*,# and Sanjay Shakkottai#
#The University of Texas at Austin, *Amazon

Blocking Bandits Model

Arms:	1	2	...	K
Mean Rewards:	μ_1	μ_2	...	μ_K
Fixed Delays:	D_1	D_2	...	D_K

μ_i unknown
D_i known

Each time arm i is played, arm i is blocked for the next $(D_i - 1)$ time steps

Objective: Maximize the expected reward in T time slots

Unit Delay: $\forall i, D_i = 1 \equiv$ Multi armed bandit problem

Applications

Job scheduling with Maximum QoS

- Arms are **servers/machines**
- Each timeslot one **homogeneous** task arrives
- Server i has delay D_i and quality of service (QoS) μ_i
  (Service time varies across servers)

Hard System Constraints on Inter Action Distance

Ad Placement with Gap Constraint

- Arms are **users/subscribers**
- Each timeslot one **homogeneous** ad needs to be placed
- User i requires a gap of D_i and mean CTR of μ_i
  (Avoid annoyance, engagement time)

Existing Approaches

Existing Methods are Computationally Intractable!

Combinatorial Semi-Bandits

- Take decisions for a block of time and observe all rewards
- Approaches [Y. Gai et al. 12, B. Kveton et al. 14, …]
- Block length = $lcm(\{D_i: i = 1 \text{ to } K\})$

Online Markov Decision Processes (MDP)

- MDP with known transitions, unknown random reward
- Approaches [P. Auer et al. 07, A. Tewari et al. 08, G. Neu et al. 09, A Zimin et al. 13,…]
- State Space = $\prod_{i \in [K]} D_i$, Horizon = $lcm(\{D_i: i = 1 \text{ to } K\})$

Offline Optimization

- The mean rewards of the arms $(\boldsymbol{\mu_i})$ are known
- **Blocking Constraint**: Each D_i blocks at most one play of arm i
- **Optimal Expected Reward** $(\mathbf{E}[R])$**:** $\mathbf{OPT} = \max_{\{a_t:\, t \le T\}\ s.t. (*)\ holds} \sum_{t=1}^{T} \mu_{a_t}$

Combinatorial optimization problem across timeslots

Result 1: NO pseudo-polynomial time algorithm given randomized Exponential Time Hypothesis holds

Greedy Algorithm

At each time, Play the Available Arm with Highest μ_i

Bad News: There are instances where Greedy achieves **3/4-th** of the optimal reward

Result 2 : Greedy is (1-1/e) Optimal

Online Optimization

- The mean rewards of the arms $(\boldsymbol{\mu_i})$ are **unknown**

α-Regret: ($\alpha \times \mathrm{E}[R]$ of OPT - $\mathrm{E}[R]$ of Online Alg)

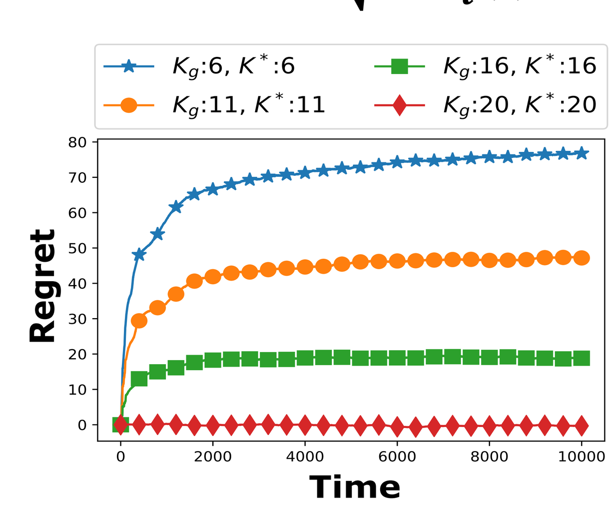
UCB-Greedy Algorithm

At time t, Play the Available Arm with Highest $ucb_i(t)$

- Empirical mean of arm i at time t, $\widehat{\mu_i}(t)$
- Number of times arm arm i played at time **t**, $N_i(t)$
- UCB of arm i at time t, $ucb_i(t) = \widehat{\mu_i}(t) + \sqrt{\left(\frac{8\, log t}{N_i(t)}\right)}$

Synthetic Experiments

- Bernoulli Reward with Fixed Mean
- Greedy plays arm $1 \text{ to } K_g$
- $K^* = \min\{\mathbf{i}: \sum_{j=1 \text{ to } i} D_j^{-1} \ge 1\}$



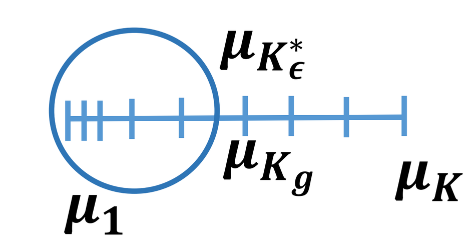
Performance Guarantees

- Sorted Means $\mu_1 \ge \mu_2 \ge \cdots \ge \mu_K$, **Gap** $\Delta_{i,j} = \mu_i - \mu_j$
- Greedy plays arm $1 \text{ to } K_g$
- Arms to cover $(1-\epsilon)$, $K_\epsilon^* = \min\{\mathbf{i}: \sum_{j=1 \text{ to } i} D_j^{-1} \ge 1-\epsilon\}$

Result 3: (1-1/e)-Regret of UCB-Greedy equals

$$O\left(\frac{1}{\epsilon}\, log\left(\frac{1}{\epsilon}\right)\right) + \frac{32\, K_g (K - K_\epsilon^*)}{min_{\{i=K_\epsilon^* \text{ to } K_g\}}\, \Delta_{i,i+1}}\, log(T)$$

These Gaps do not influence the regret bound

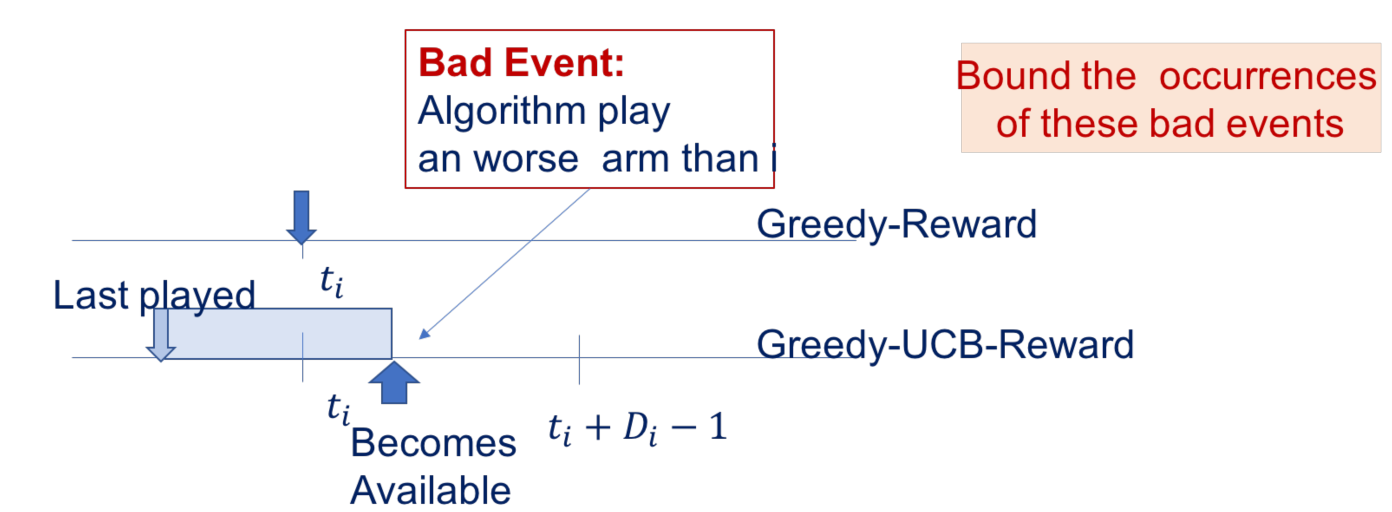


Result 4: Lower Bound $\frac{(K-K_g)}{\Delta_{K_g, K_g+1}}\, log(T) + O(1)$

Techniques: Coupling and Free Exploration

- Decision sets of Greedy and UCB-Greedy do not converge

Couple Each Arm Separately!



Free explore: Due to blocking of higher ranked arms, each arm $i \in [1, K_\epsilon^*]$ played $\ge cT$ times up to time T

Future Work

- **Stochastic Unknown Delay**
- **Multi-type Extension:**

In each time slot an i.i.d. type is chosen by nature.
For each type j, arm i has delay D_{ij} and reward μ_{ij}