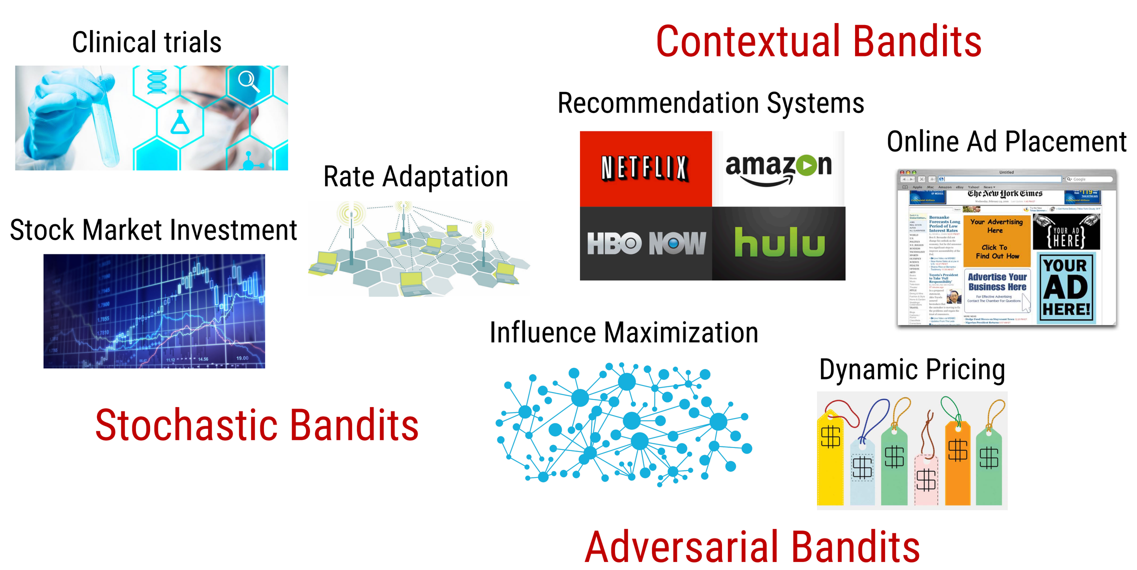
Learning to Control Renewal Processes with Bandit Feedback

Semih Cayci [1] Atilla Eryilmaz [1] R. Srikant [2]

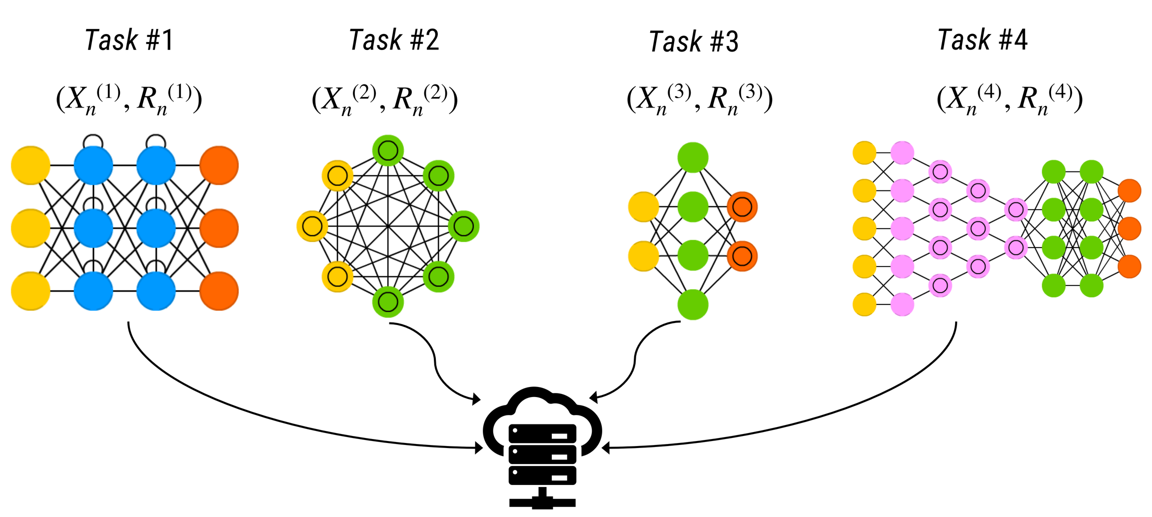
[1]ECE, The Ohio State University [2]CSL and ECE, UIUC

Introduction

In traditional bandit models, each arm pull takes a unit time → Violated in many real-life applications.



Example: Task scheduling, free-trial strategy, etc.



Task k yields $X_n^{(k)}$ completion time and $R_n^{(k)}$ reward after the completion. $(X_n^{(k)}, R_n^{(k)})$ unknown at the time of scheduling, unknown statistics, potentially heavy-tailed.

Objective: Maximize the expected cumulative reward in a given time interval $[0, \tau]$.

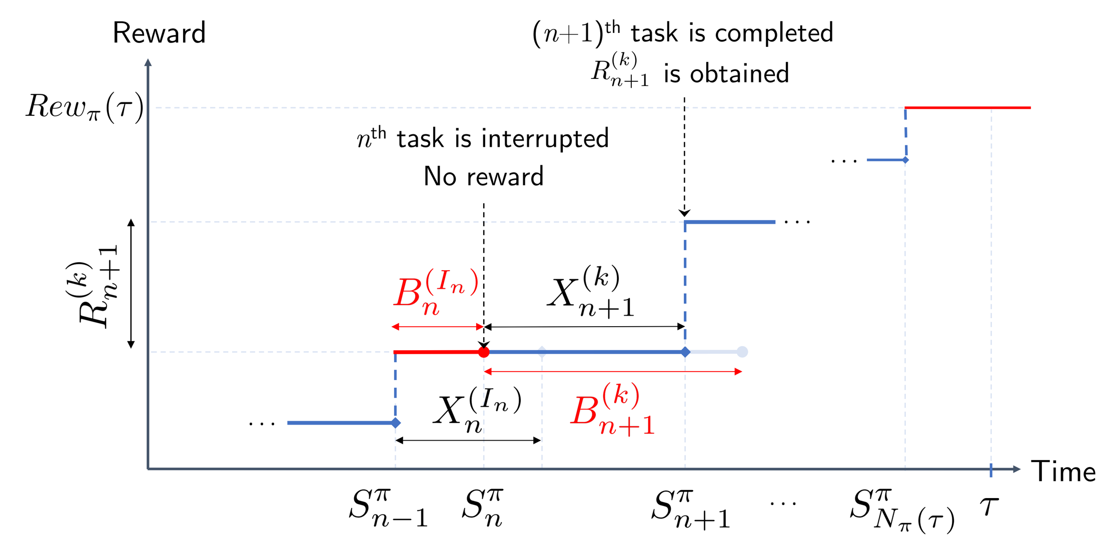
New dilemma: Complete an ongoing task vs. interrupt & switch for a possibly more rewarding one?

Bandits with Interrupts - `BwI`

We consider a K-armed bandit model.

- Arm $k \leftrightarrow (X_n^{(k)}, R_n^{(k)}) \overset{iid}{\sim} F_k$
- Heavy-tailed time and reward: For $\gamma_0 > 0$,
$$\max\{\mathbb{E}[(X_1^{(k)})^{1+\gamma_0}], \mathbb{E}[(R_1^{(k)})^{1+\gamma_0}]\} < \infty$$
- Interrupt an ongoing task if it takes "too long" time, reject the reward of that task.
- Censored bandit feedback:
$$\pi_n = (k, b) \Rightarrow \left(X_n^{(k)} \wedge b, R_n^{(k)} \mathbb{I}_{\{X_n^{(k)} \le b\}}\right)$$
for an interrupt time $b \in \mathcal{B} \subset \mathbb{R}_+$

Main Problem



Minimize $\overline{Reg}_\pi(\tau) = \mathbb{E}[Rew_{\pi^{\text{opt}}}(\tau)] - \mathbb{E}[Rew_\pi(\tau)]$

Optimal Policy and Approximations

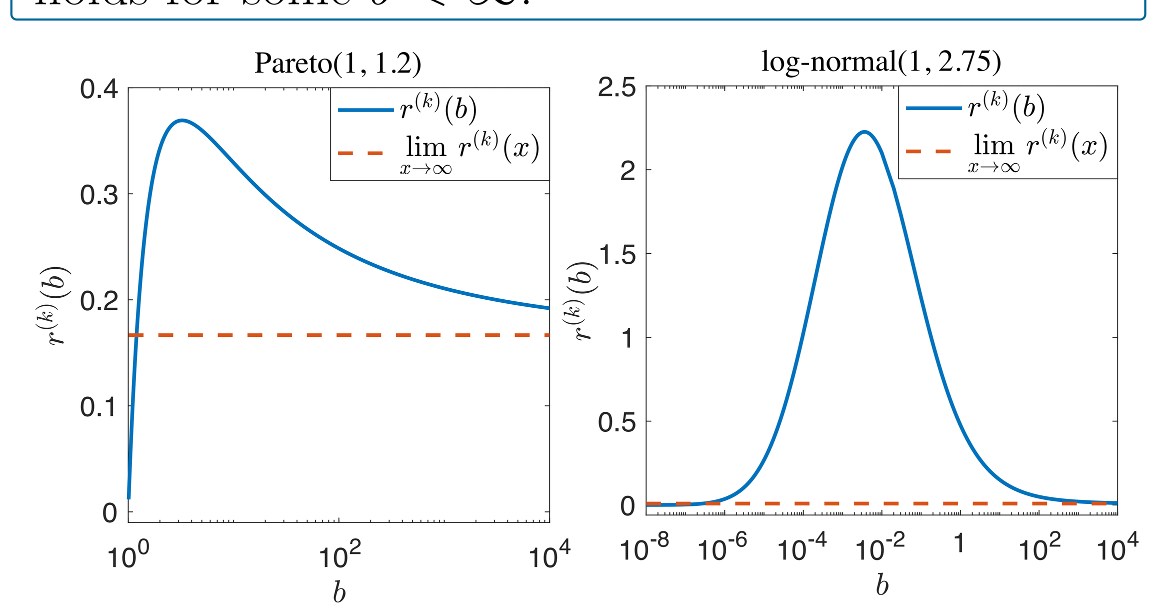
The problem is NP-hard even if all statistics are known [1] ⇒ Approximation algorithms

Renewal reward rate: For any (k, b):
$$r^{(k)}(b) = \frac{\mathbb{E}[R_1^{(k)} \mathbb{I}_{\{X_1^{(k)} \le b\}}]}{\mathbb{E}[X_1^{(k)} \wedge b]}$$

Optimal interrupt time: For every k,
$$b_k^* = \sup\{b : r^{(k)}(b) \ge r^{(k)}(b'),\ b' \in \mathcal{B}\}$$

Proposition 1. (Finite Interrupt Time)

Interruption is optimal, i.e., $b_k^* < \infty$, iff
$$\mathbb{E}[X_1^{(k)} - b | X_1^{(k)} > b] > \mathbb{E}[X_1^{(k)}],$$
holds for some $b < \infty$.



Optimal static policy: At every epoch n, pull (k, b) with largest reward rate:
$$\pi_n^* = \arg\max_{(k,b)} r^{(k)}(b)$$

Proposition 2. (Optimality Gap for π^*)

For any $\tau > 0$, the optimality gap for π^*:
$$\mathbb{E}[Rew_{\pi^{\text{opt}}}(\tau)] - \mathbb{E}[Rew_{\pi^*}(\tau)] \le O(1).$$
Thus, π^* is asymptotically optimal as $\tau \to \infty$.

`UCB-BwI` Algorithm

We consider a finite but arbitrary
$$\mathcal{B} = \{b_1, b_2, \ldots, \underbrace{b_L = \infty}_{\text{no interrupt}}\}.$$

- **Strategy:** For each (k, b), use upper confidence bounds for $r^{(k)}(b)$ as a surrogate.
- **Challenge:** $X_n^{(k)}$ and $R_n^{(k)}$ can be heavy-tailed.

The first candidate is *empirical reward rate*:
$$\hat{r}_s^{(k)}(b) = \frac{\sum_{i=1}^{s} R_i^{(k)} \mathbb{I}_{\{X_i^{(k)} \le b\}}}{\sum_{i=1}^{s} (X_i^{(k)} \wedge b)} \xrightarrow[s \to \infty]{} r^{(k)}(b), \text{ a.s.}$$

Convergence rate is polynomial, not exponential:
$$\mathbb{P}\left(\hat{r}_s^{(k)}(b_L) \le r^{(k)}(b_L) - \Delta_0(\epsilon)\right) = O\left(\frac{1}{s^\gamma \epsilon^{1+\gamma}}\right).$$

- **Robust median-of-means estimation:** For $w = \lfloor 8\log(e^{\frac{1}{8}}\delta^{-1}) \wedge \frac{s}{2} \rfloor$ and $m = \lfloor \frac{s}{w} \rfloor$,
$$M(U_{1:s}) \triangleq med\left\{\frac{1}{m}\sum_{i=1}^{m} U_i, \ldots, \frac{1}{m}\sum_{i=(w-1)m+1}^{wm} U_i\right\}.$$
Median boosts the performance of the weak empirical estimator [2].
$$\bar{r}_s^{(k)}(b) = \frac{M\left(R_i^{(k)} \mathbb{I}_{\{X_i^{(k)} \le b\}} : i \le s\right)}{M\left(X_i^{(k)} \wedge b : i \le s\right)}.$$
Exponential convergence rate is achieved despite heavy-tails.

- **Information structure:** Feedback for (k, b_j) is available for (k, b_l) if $l \le j$. Boosted convergence

Algorithm: `UCB-BwI`

At epoch $(n+1)$, $s_{k,l}$ observations for (k, b_l). Then, `UCB-BwI` makes a decision as follows:
$$\left(I_{n+1}, B_{n+1}^{(I_{n+1})}\right) \in \arg\max_{(k,b_l)} \left\{\bar{r}_{n,s_{k,l}}^{(k)}(b_l) + \frac{(1+r)\epsilon_{n,s_{k,l}}}{\mu + \epsilon_{n,s_{k,l}}}\right\}$$
where $r \ge r^{(k)}(b)$, $\mu \le \mathbb{E}[X_1^{(k)} \wedge b]$ for all k, b and
$$\epsilon_{n,s} = \beta\left[\frac{\log\left(2e^{\frac{1}{8}}(n+1)^4\right)}{s}\right]^{\frac{\gamma}{1+\gamma}},$$
for $\gamma = \min\{\gamma_0, 1\}$ and some $\beta > 0$.

Performance Analysis

- **Regret upper bound for `UCB-BwI`:**

Theorem 1. (Regret Bound for `UCB-BwI`)

Regret under `UCB-BwI` for any $\tau > 0$ satisfies:
$$\overline{Reg}_\pi(\tau) \le \sum_{k : d^{(k)} > 0} C_k \log(\tau) + O(KL),$$
where
$$C_k = \mathcal{O}\left(\left(\frac{1}{d_{min}^{(k)}}\right)^{\frac{1}{\gamma}} + \left(\frac{1}{d^{(k)}}\right)^{\frac{1}{\gamma}}\right),$$
$d_{min}^{(k)} \le r^{(k)}(b_k^*) - r^{(k)}(b_l)$ and $d^{(k)} = r^* - r^{(k)}(b_k^*)$

Info. structure ⇒ $\log(\tau)$ term independent of L.

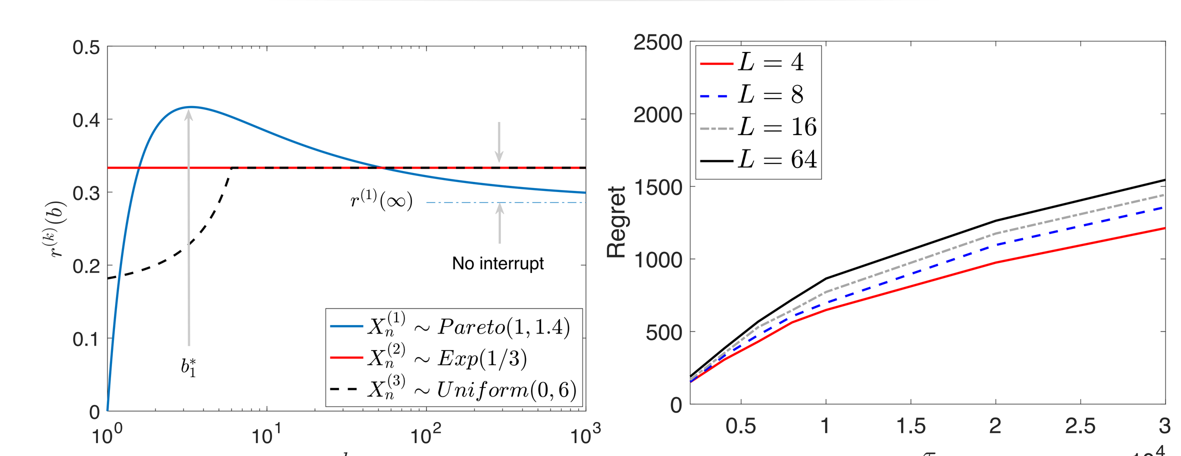
- **Regret lower bound:**

Theorem 2. (Regret Lower Bound)

Under any "good" policy π that makes only $o(n^\alpha)$ suboptimal decisions in n epochs, we have
$$\overline{Reg}_\pi(\tau) = \Omega\left(K\log(\tau)\right)$$

Matching bounds for `UCB-BwI` ⇒ order optimality

Numerical Results



(a) Tails are important: Gains by interruption for heavy-tailed $X_n^{(k)}$

(b) Low regret despite large $\mathcal{B}$: Exploiting information structure

Conclusions

We incorporated time dimension into bandits.

- Heavy-tailed completion time ⇒ Interrupt + novel dynamics
- There is an underlying information structure from temporal dynamics.
- `UCB-BwI`: $\Theta(K\log(\tau) + KL)$ regret

References

[1] B. C. Dean et al. Approximating the stochastic knapsack problem: The benefit of adaptivity. In *IEEE FOCS*, 2004.

[2] S. Bubeck et al. Bandits with heavy tail. *IEEE ToIT*, 2013.

[3] A. Badanidiyuru et al. Bandits with knapsacks. In *IEEE FOCS*, 2013.