Learning to Control Renewal Processes with Bandit Feedback

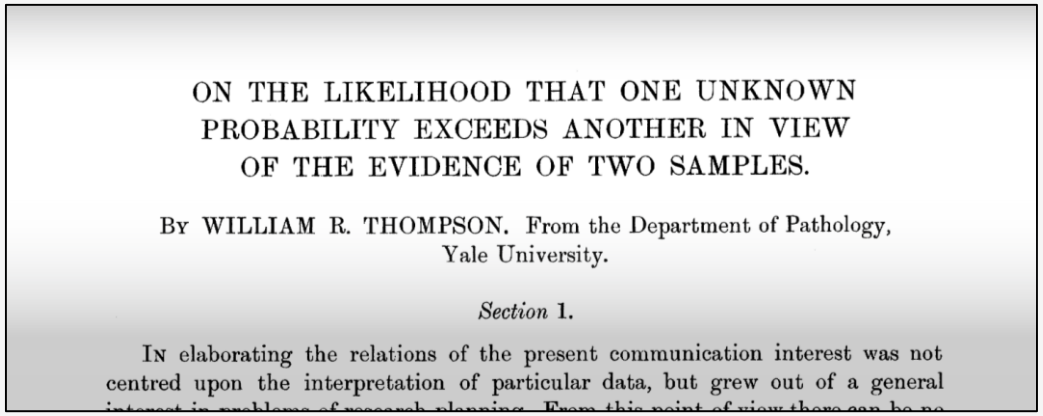
Semih Cayci[1], Atilla Eryilmaz[1], R. Srikant[2]

[1] The Ohio State University, ECE

[2] University of Illinois at Urbana-Champaign, CSL and ECE

Multi-Armed Bandits

Prominent model for exploration and exploitation dilemma since 1933.



ON THE LIKELIHOOD THAT ONE UNKNOWN PROBABILITY EXCEEDS ANOTHER IN VIEW OF THE EVIDENCE OF TWO SAMPLES.

By WILLIAM R. THOMPSON. From the Department of Pathology, Yale University.

Section 1.

In elaborating the relations of the present communication interest was not centred upon the interpretation of particular data, but grew out of a general

What are we investigating?

- Time dimension into multi-armed bandits as required by many applications
- Very general time and reward distributions including heavy-tailed distributions
- Novel dynamics + new dilemma

Multi-Armed Bandits: Applications

Traditional bandit models have a broad area of applications:

1. **Stochastic bandits** [Lai & Robbins, ‘85], [Auer, ‘02], [Agrawal, ‘11], [Bubeck, ‘12a]

   Heavy-tailed reward [Bubeck, ‘12b], [Zhao, ‘11]

2. **Contextual bandits** [Dudik, ‘11], [Slivkins, ‘14]

   Linear bandits [Dani, ‘08], [Abbasi-Yadkori, ‘11]

3. **Budgeted bandits** [Gyorgy, ‘07], [Tran-Tranh, ‘10], [Badanidiyuru, ‘13], [Combes, ‘15], [Xia, ‘15]

- Cost and reward in [0, 1]
- No interrupt mechanism
- Limited control

Recap: Classical Stochastic Multi-Armed Bandit Problem

In the classical stochastic bandit model:

Q: Is it always valid?

- **Discrete time:** Each arm pull takes a unit time for all arms.
- **Control:** Choose an arm I_n among {1, 2, …, K}
- **Goal:** Maximize cumulative reward in τ units of time.

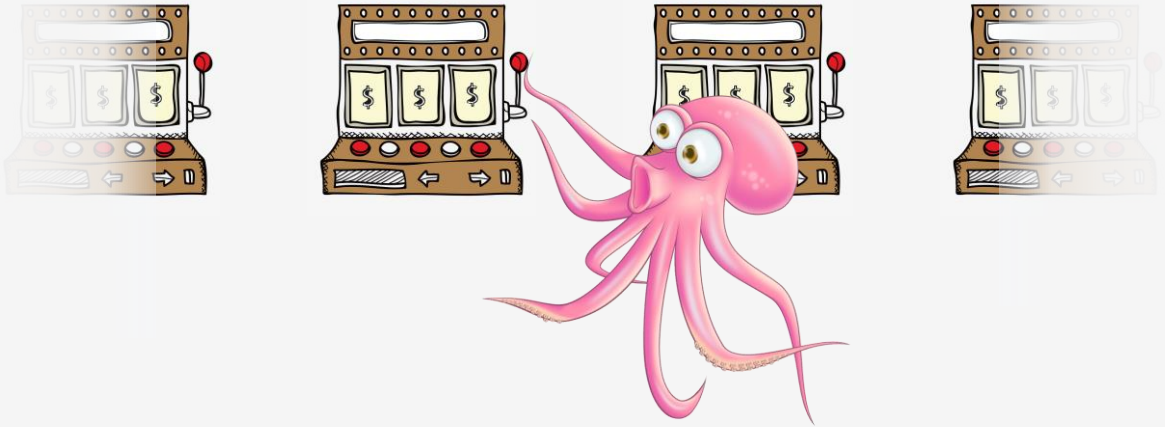
$$\mathbb{E}\big[\mathrm{Rew}_\pi(\tau)\big] = \mathbb{E}\Big[\sum_{n=1}^{\tau} R_n^{(I_n)}\Big]$$

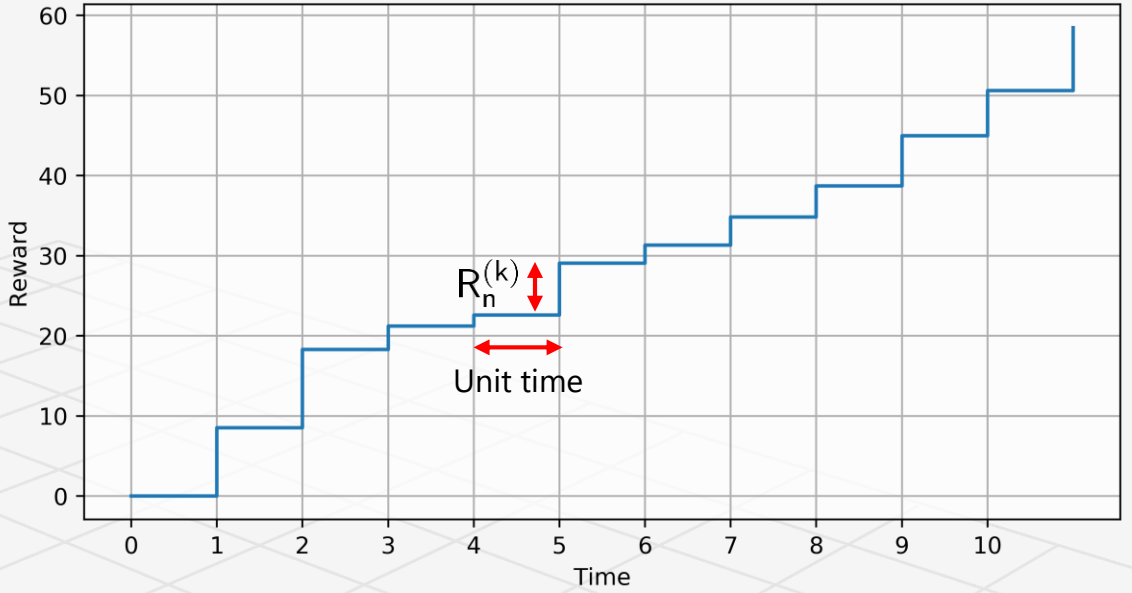
- **Optimal policy:** Arm with maximum expected reward.

$$\pi_n^{\mathrm{opt}} = \arg\max_k \mathbb{E}[R_n^{(k)}]$$

1st-order statistics

- **Learning:** Upper confidence bound (UCB), ϵ-greedy [Auer '02], Thompson sampling [Agrawal, '11], IDS [Russo, '14]





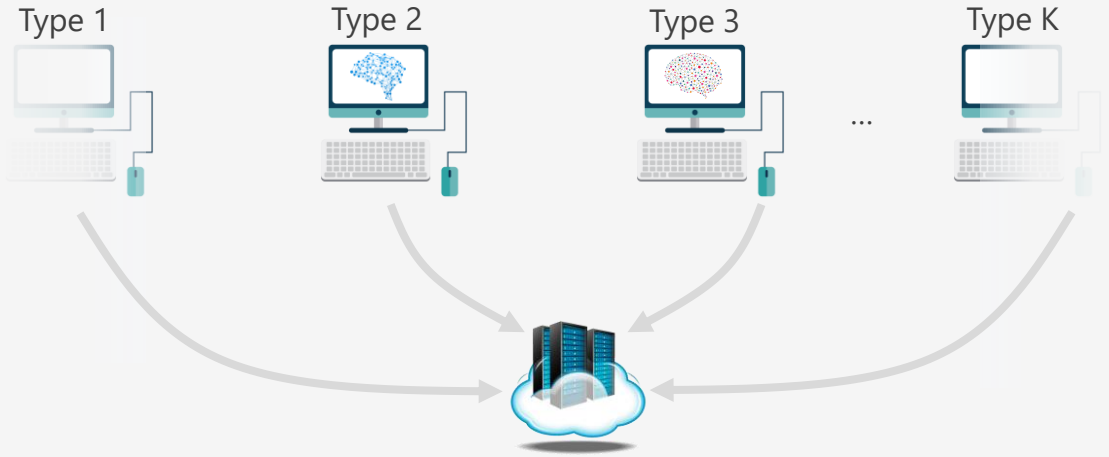
Time Dimension in Bandits: Task Scheduling Example

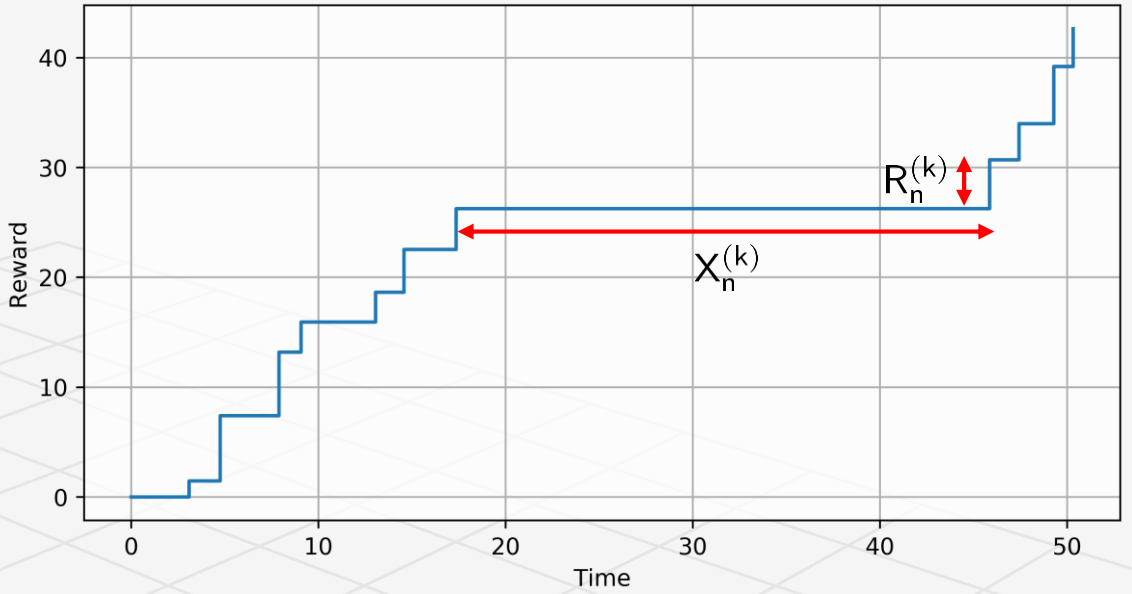
Non-clairvoyant task scheduling with K task types and single server.

- Type-k task takes a random **completion time** $X_n^{(k)}$.
- A random **reward** $R_n^{(k)}$ is obtained once the task is completed.
- For a **time budget** τ, maximize total reward in $[0, \tau]$
- Completion time might be **heavy-tailed** [Harchol-Balter, '99]
  - What if the completion time is "too long"? Should I wait?
- Optimal policy?

? $\arg\max_k \mathbb{E}[R_n^{(k)}]$ $\arg\max_k \mathbb{E}\left[\frac{R_n^{(k)}}{X_n^{(k)}}\right]$ $\arg\max_k \frac{\mathbb{E}[R_1^{(k)}]}{\mathbb{E}[X_1^{(k)}]}$

- Existing **MAB** models fall short for this application.



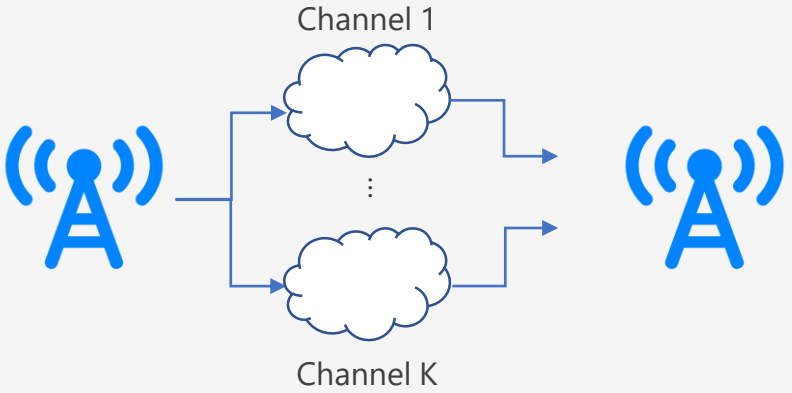


Time Dimension in Bandits: Applications

There are many other applications that have similar time-dependence. Here are some examples:

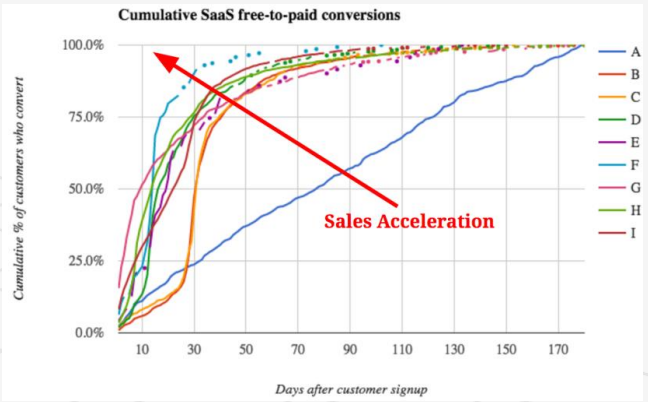
1 **Adaptive routing in telecommunications:**

- K parallel channels, transmit over one at a time, bandit feedback
- Transmission time of n^{th} packet over channel k: $X_n^{(k)}$ (heavy-tailed [Asmussen, '10])
- Goal: Maximize throughput within $[0, \tau]$, i.e., $R_n^{(k)} = 1$



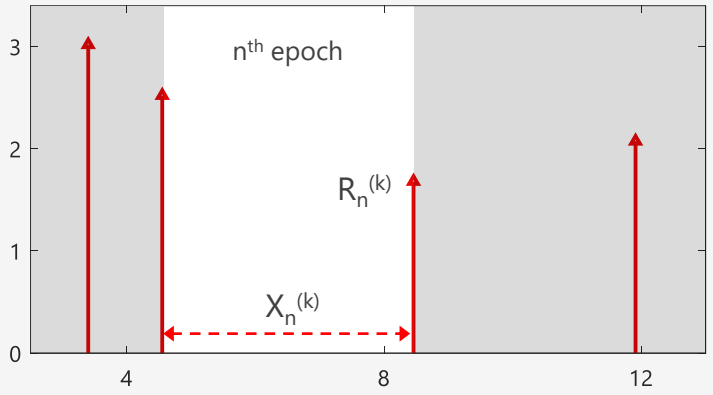
2 **Online free trial strategy for advertising:**

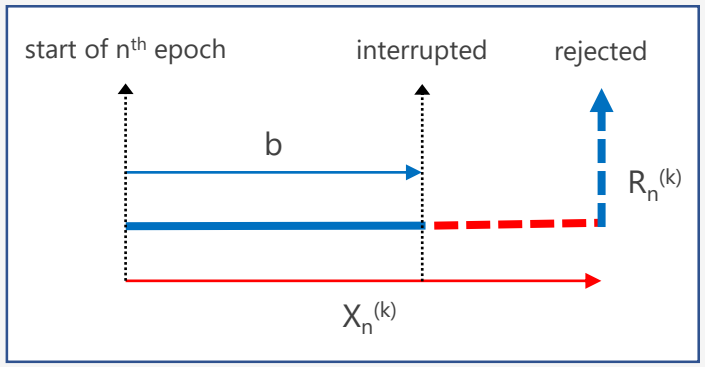
- Type-k user's free-to-paid conversion time: $X_n^{(k)}$
- Price of the product $R_n^{(k)}$
- **Free riders > Interrupts**

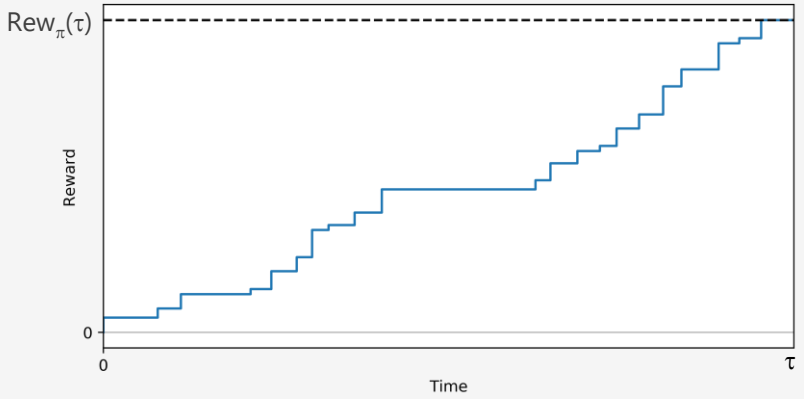


Data from madkudu.com

Bandits with Interrupts (BwI)







1 **Arms:** Arm k ↔ renewal reward process

$$\left(X_n^{(k)}, R_n^{(k)}\right) \sim F^{(k)}$$

There exists a $\gamma > 0$ such that:

$$\max\left\{\mathbb{E}\left[\left(X_n^{(k)}\right)^{1+\gamma}\right], \mathbb{E}\left[\left(R_n^{(k)}\right)^{1+\gamma}\right]\right\} < \infty$$

Remark: Variance might be infinite.

2 **Control:** If an arm pull / task takes too long time, the controller might interrupt.

$$\pi_n = \left(I_n, B_n^{(I_n)}\right) \in \mathcal{K} \times \mathcal{B}$$

New control: Interrupt time

Feedback: Censored. For a decision (k, b)

$$\left(X_n^{(k)} \wedge b, R_n^{(k)} \mathbb{I}_{\{X_n^{(k)} \le b\}}\right)$$

3 **Objective:** Minimize regret in $[0, \tau]$.

$$\overline{\mathrm{Reg}}_\pi(\tau) = \mathbb{E}[\mathrm{Rew}_{\pi^{\mathrm{opt}}}(\tau)] - \mathbb{E}[\mathrm{Rew}_\pi(\tau)]$$

where

Number of comp. tasks

$$\mathrm{Rew}_\pi(\tau) = \sum_{n=1}^{N_\pi(\tau)} R_n^{(I_n)} \mathbb{I}_{\{X_n^{(I_n)} \le B_n^{(I_n)}\}}$$

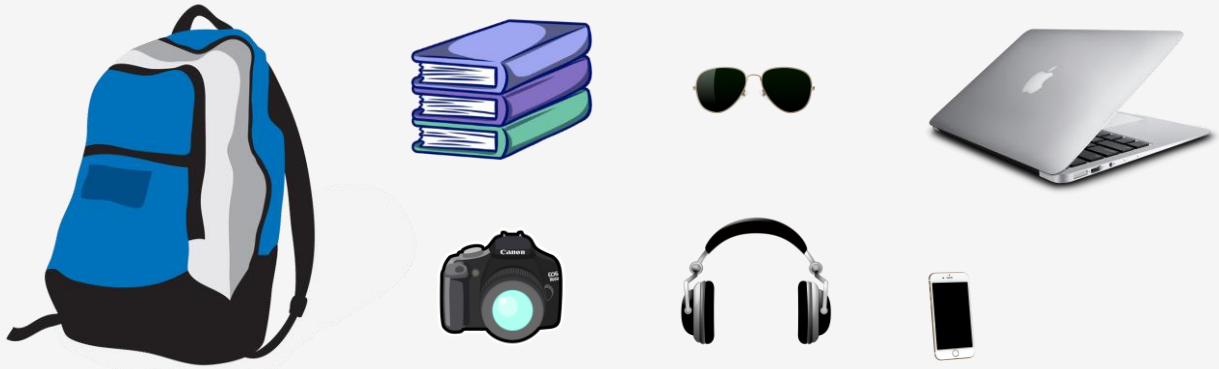
Classical dilemma: Explore **or** Exploit**?**

New dilemma: Should it interrupt an ongoing cycle and reject the reward for a more rewarding alternative**, or** wait and gather the reward**?**

Optimal Policy with Known Statistics: Complexity

`BwI` problem is NP-hard even if all statistics are known.

Wanted: Approximation algorithms



To find an approximation algorithm:

1. Consider a **renewal reward process** (X_n, R_n).

   By the **key renewal theorem**, we have:

$$\mathbb{E}\Big[\sum_{n=1}^{N(\tau)} R_n\Big] = \frac{\mathbb{E}[R_n]}{\mathbb{E}[X_n]} \cdot \tau + o(\tau)$$

   (See [Asmussen, '08] and [Gut, '09])

2. For an action (k, b), the **renewal reward rate** is defined as follows:

$$r^{(k)}(b) = \frac{\mathbb{E}[R_n^{(k)} \mathbb{I}_{\{X_n^{(k)} \le b\}}]}{\mathbb{E}[X_n^{(k)} \wedge b]}$$

Interpretation: Reward per unit time if (k, b) is persistently chosen.

Optimal Policy with Known Statistics: A Good Static Approximation

Approximation algorithm: Pull (k, b) with the largest reward rate persistently until the time budget is depleted.

$$\pi_n^* = \arg\max_{(k,b)} \; r^{(k)}(b)$$

for all n.

? How well does π^* approximate the optimal policy π^{opt} ?

The static policy π^* has some nice properties:

1. Low complexity, time-invariant.
2. The optimality gap with π^{opt} is **bounded** for all $\tau > 0$.

The optimal policy π^{opt} is NP-hard.

Challenge: $r^{(k)}(b)$ depends on the tails, not only mean.

Proposition 1. (Optimality Gap for π^*)

For any $\tau > 0$, the following inequality holds for the static policy π^* :

$$\mathbb{E}[\mathrm{Rew}_{\pi^{\mathrm{opt}}}(\tau)] - \mathbb{E}[\mathrm{Rew}_{\pi^*}(\tau)] = O(1),$$

Consequently, π^* is asymptotically optimal as $\tau \to \infty$.

✓ Appropriate for low-regret learning algorithms.

Optimal Policy with Known Statistics: Optimal Interrupt Time

Renewal reward rate for (k, b): $r^{(k)}(b) = \frac{\mathbb{E}[R_n^{(k)} \mathbb{I}_{\{X_n^{(k)} \le b\}}]}{\mathbb{E}[X_n^{(k)} \wedge b]}$

? Does interruption improve reward rate?

$$b_k^* = \max_{b \in \mathcal{B}} r^{(k)}(b)$$

Proposition 2. (Optimal Interrupt Time)

Interruption at a finite time is optimal, i.e., $b_k^* \neq \infty$, if and only if the following holds for some $b < \infty$:

$$\mathbb{E}[X_1^{(k)} - b | X_1^{(k)} > b] > \mathbb{E}[X_1^{(k)}]$$

Interruption is optimal if the **mean residual life** at some time **b** is larger than the mean completion time of a fresh new cycle.

For **correlated** $(X_n^{(k)}, R_n^{(k)})$, the condition becomes the following:

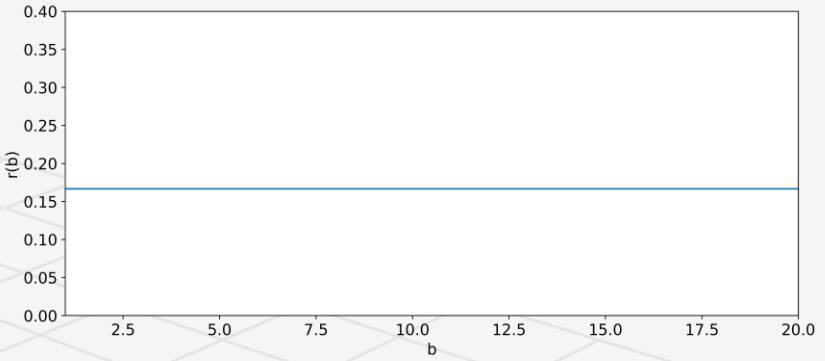
$$\frac{\mathbb{E}[R_1^{(k)} - b | X_1^{(k)} > b]}{\mathbb{E}[X_1^{(k)} - b | X_1^{(k)} > b]} < \frac{\mathbb{E}[R_1^{(k)}]}{\mathbb{E}[X_1^{(k)}]}$$

1. All **heavy-tailed** and some light-tailed completion time distributions lead to a finite optimal interrupt time.

2. Exponential distribution acts as a barrier case: **memoryless property** leads to indifference to interruption.

3. Most light-tailed distributions (folded Gaussian, uniform, logistic, gamma) have **decreasing MRL**, thus no interrupt is optimal.

$R_n = 1$, $X_n \sim Exp(1/6)$



All have the same mean → **Tails are important**

Optimal Policy with Known Statistics: Optimal Interrupt Time

Renewal reward rate for (k, b): $r^{(k)}(b) = \frac{\mathbb{E}[R_n^{(k)} \mathbb{I}_{\{X_n^{(k)} \le b\}}]}{\mathbb{E}[X_n^{(k)} \wedge b]}$

? Does interruption improve reward rate?

$$b_k^* = \max_{b \in \mathcal{B}} r^{(k)}(b)$$

Proposition 2. (Optimal Interrupt Time)

Interruption at a finite time is optimal, i.e., $b_k^* \neq \infty$, if and only if the following holds for some $b < \infty$:

$$\mathbb{E}[X_1^{(k)} - b | X_1^{(k)} > b] > \mathbb{E}[X_1^{(k)}]$$

Interruption is optimal if the **mean residual life** at some time **b** is larger than the mean completion time of a fresh new cycle.

For **correlated** $(X_n^{(k)}, R_n^{(k)})$, the condition becomes the following:

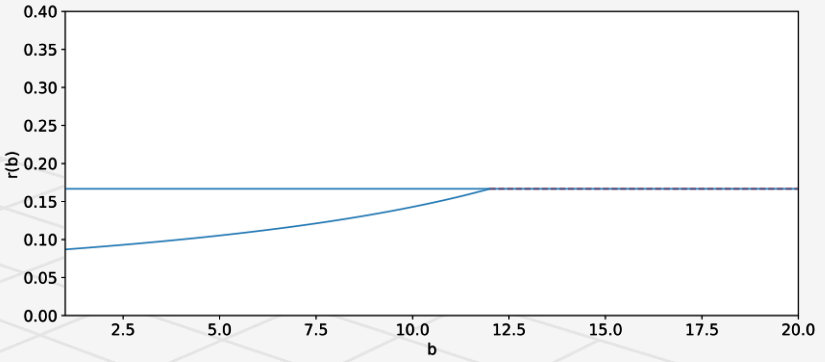
$$\frac{\mathbb{E}[R_1^{(k)} - b | X_1^{(k)} > b]}{\mathbb{E}[X_1^{(k)} - b | X_1^{(k)} > b]} < \frac{\mathbb{E}[R_1^{(k)}]}{\mathbb{E}[X_1^{(k)}]}$$

1. All **heavy-tailed** and some light-tailed completion time distributions lead to a finite optimal interrupt time.

2. Exponential distribution acts as a barrier case: **memoryless property** leads to indifference to interruption.

3. Most light-tailed distributions (folded Gaussian, uniform, logistic, gamma) have **decreasing MRL**, thus no interrupt is optimal.

$R_n = 1$, $X_n \sim$ *Uniform(0, 12)*



All have the same mean → **Tails are important**

Optimal Policy with Known Statistics: Optimal Interrupt Time

Renewal reward rate for (k, b): $r^{(k)}(b) = \frac{\mathbb{E}[R_n^{(k)} \mathbb{I}_{\{X_n^{(k)} \le b\}}]}{\mathbb{E}[X_n^{(k)} \wedge b]}$

? Does interruption improve reward rate?

$$b_k^* = \max_{b \in \mathcal{B}} r^{(k)}(b)$$

Proposition 2. (Optimal Interrupt Time)

Interruption at a finite time is optimal, i.e., $b_k^* \neq \infty$, if and only if the following holds for some $b < \infty$:

$$\mathbb{E}[X_1^{(k)} - b | X_1^{(k)} > b] > \mathbb{E}[X_1^{(k)}]$$

Interruption is optimal if the **mean residual life** at some time **b** is larger than the mean completion time of a fresh new cycle.

For **correlated** $(X_n^{(k)}, R_n^{(k)})$, the condition becomes the following:

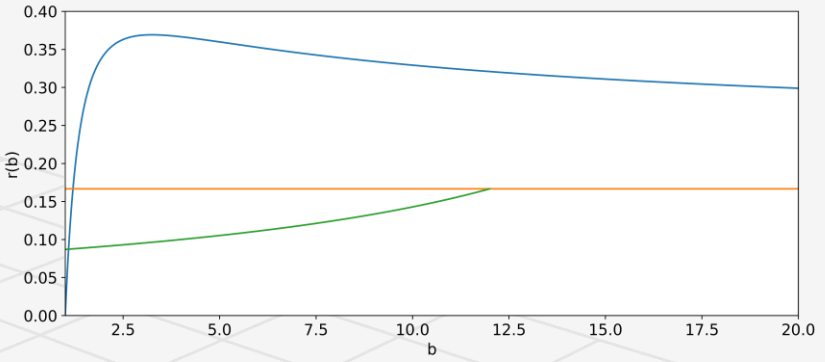
$$\frac{\mathbb{E}[R_1^{(k)} - b | X_1^{(k)} > b]}{\mathbb{E}[X_1^{(k)} - b | X_1^{(k)} > b]} < \frac{\mathbb{E}[R_1^{(k)}]}{\mathbb{E}[X_1^{(k)}]}$$

1. All **heavy-tailed** and some light-tailed completion time distributions lead to a finite optimal interrupt time.

2. Exponential distribution acts as a barrier case: **memoryless property** leads to indifference to interruption.

3. Most light-tailed distributions (folded Gaussian, uniform, logistic, gamma) have **decreasing MRL**, thus no interrupt is optimal.

$R_n = 1$, $X_n \sim$ *Pareto*(1, 1.2)



All have the same mean → **Tails are important**

Algorithm Design: Preliminaries

Objective: For all (k, b) actions, learn the reward rate:

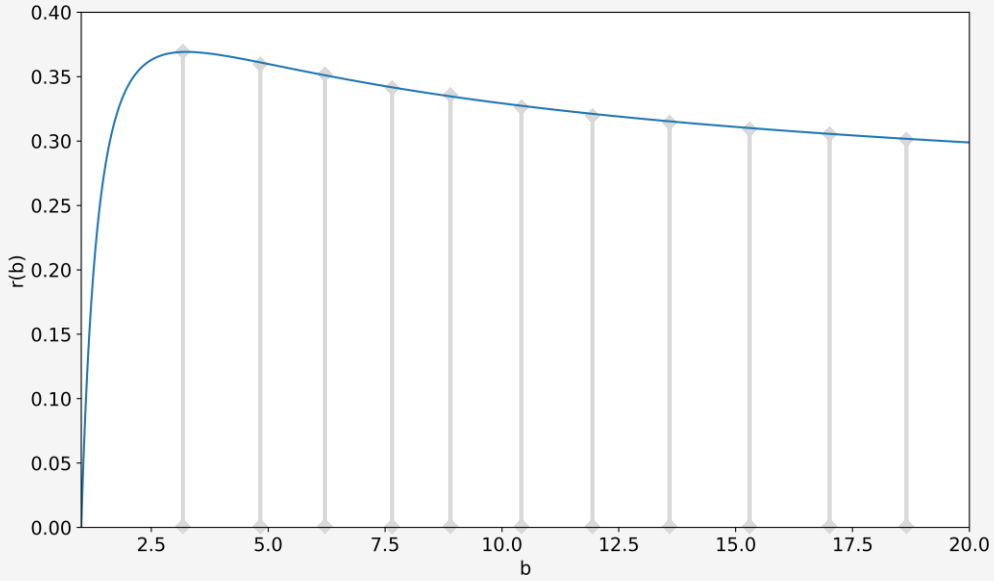
$$r^{(k)}(b) = \frac{\mathbb{E}[R_n^{(k)} \mathbb{I}_{\{X_n^{(k)} \le b\}}]}{\mathbb{E}[X_n^{(k)} \wedge b]}$$

Far beyond the first-order statistics unlike classical MAB.

Assumption: **Finite** but **arbitrary** set of interrupt times

$$\mathcal{B} = \{b_1, b_2, \ldots, b_L = \infty\}$$

Imposed by the nature: CPU scheduling, 1/3/7/14 days in free trials, etc.



Strategy: Use upper confidence bound as a surrogate for $r^{(k)}(b_l)$

1. Find **concentration inequalities for reward rate**.

   Need: Estimators that provide exponential convergence rate despite the heavy tails.

2. Exploit the **information structure** imposed by the temporal dynamics.

   There is an information sharing between different (k, b) actions based on observability.

3. Propose **`UCB-BwI`** for the online learning problem.

   Based on the concentration inequalities, exploit the information structure for low-regret.

Concentration Inequalities for Renewal Processes

Wanted: UCB for $r^{(k)}(b) = \dfrac{\mathbb{E}[R_n^{(k)} \mathbb{I}_{\{X_n^{(k)} \le b\}}]}{\mathbb{E}[X_n^{(k)} \wedge b]}$

First candidate: Empirical reward rate [Asmussen, '08], [Karlin, '83]

$$\hat{r}^{(k)}(b_l) = \frac{\sum_{i=1}^{s} R_i \mathbb{I}_{\{X_i^{(k)} \le b_l\}}}{\sum_{i=1}^{s} (X_i \wedge b_l)} \to r^{(k)}(b_l) \text{ a.s.}$$

Problem: Heavy-tailed completion time and reward

Convergence rate is **polynomial**, not exponential (Chebyshev)

$$\mathbb{P}\big(\hat{r}_s(b_L) \le r(b_L) - \Delta_0(\epsilon)\big) = O\Big(\frac{1}{s^{\gamma} \epsilon^{1+\gamma}}\Big)$$

for $\Delta_0(\epsilon) = \dfrac{(1+r)\epsilon}{\mu + \epsilon}$. Bound is tight [Catoni, '10, Bubeck '12].

Cause: Outliers due to the heavy tails pull the sample mean away from the ensemble mean.

Fix: Robust **median-of-means estimator** [Nemirovski, '82], [Bubeck, '12]

Given $\{U_i: i = 1, 2, \ldots, s\}$, $w = \lfloor 8 \log(e^{\frac{1}{8}} \delta^{-1}) \wedge \frac{s}{2} \rfloor$ and $m = \lfloor \frac{s}{w} \rfloor$

$$M(U_{1:s}) \triangleq \text{med}\Big\{ \frac{1}{m} \sum_{i=1}^{m} U_i, \ldots, \frac{1}{m} \sum_{i=(w-1)m+1}^{wm} U_i \Big\}$$

For each (k, b_l), the median-of-means estimator for the reward rate $r^{(k)}(b_l)$:

$$\bar{r}_s(b) = \frac{M\big(R_i \mathbb{I}_{\{X_i \le b\}} : i \le s\big)}{M\big(X_i \wedge b : i \le s\big)}$$

Proposition 3. (Conc. Ineq. for Renewal Reward Processes)

For any decision (k, b) and $\delta \in (0, 1)$,

$$\mathbb{P}\Big(\bar{r}_s^{(k)}(b) \le r^{(k)}(b) - \Delta\big(\epsilon(\delta)\big)\Big) \le \delta$$

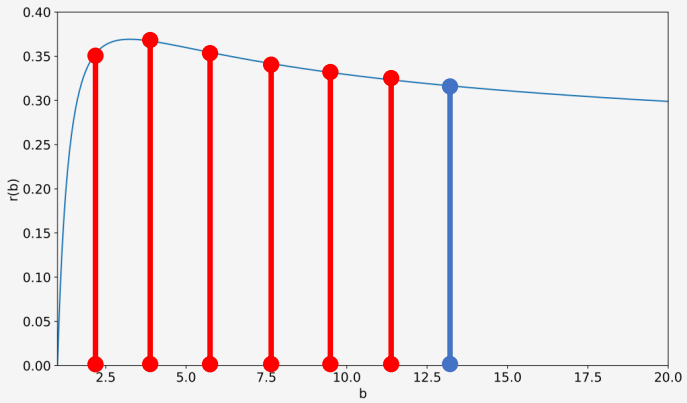
where $\epsilon(\delta) = \beta\Big(\dfrac{\log(\nu \delta^{-1})}{s}\Big)^{\frac{\gamma}{1+\gamma}}$

✓ Exponential convergence rate

Sub-Gaussian (optimal) accuracy-confidence tradeoff if $\gamma = 1$ [Bubeck, '12b]

UCB-BwI Algorithm

Information structure:



Information sharing

Feedback

- Is $X_n^{(k)}$ smaller than b?
- If so, values of $X_n^{(k)}$ and $R_n^{(k)}$

These questions can be answered for all $b' \le b$ by this feedback **[Observability]**

Formally, $Y(k, b') \in \sigma(Y(k, b)),\ \forall b' \le b$

Action (k, b) increases the sample size for all (k, b') such that $b' \le b$.

UCB-BwI Algorithm

At epoch (n+1), UCB-BwI Algorithm makes a decision as follows:

$$(I_{n+1}, B_{n+1}^{(I_{n+1})}) \in \arg\max_{(k,b_I)} \left\{ \bar{r}_{n,\overline{T}_I^{(k)}(n)}(b_I) + \frac{(1+r)\epsilon_{n,\overline{T}_I^{(k)}(n)}}{\mu + \epsilon_{n,\overline{T}_I^{(k)}(n)}} \right\}$$

where

$$\epsilon_{n,s} = \beta\left[\frac{\log\left(2e^{\frac{1}{8}}(n+1)^4\right)}{s}\right]^{\frac{\gamma}{1+\gamma}}$$

$\overline{T}_I^{(k)}(n)$ is the effective sample size → information structure

Performance Analysis: Regret Bounds for UCB-BwI

Let

$$d_{min}^{(k)} \le r^{(k)}(b_k^*) - \max_{1\le l\le L} r^{(k)}(b_l) \quad \text{(Suboptimality in k)}$$

$$d^{(k)} = r^* - \max_{1\le l\le L} r^{(k)}(b_l) \quad \text{(Suboptimality across k)}$$

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

Regret under UCB-BwI satisfies the following bound for all $\tau > 0$:

$$\overline{Reg}_\pi(\tau) \le \sum_{k:d^{(k)}>0} C_k \log\left(\frac{\tau}{\min_k \mathbb{E}[X_1^{(k)}]}\right) + O(K \times L)$$

where

$$C_k = \mathcal{O}\left(\left(\frac{1}{d_{min}^{(k)}}\right)^{\frac{1}{\gamma}} + \left(\frac{1}{d^{(k)}}\right)^{\frac{1}{\gamma}}\right)$$

- Logarithmic regret in τ.
- Action space: O(K x L) but C_k is O(K) as a result of info structure.

Proof Sketch: Asynchronous decisions + random number of trials

1. Dealing with asynchronous decisions: **regret rate**

$$\overline{Reg}_\pi(\tau) \le \left[\sum_{n=1}^{N_\pi(\tau)} \sum_{k,l} \mathbb{I}_{\{\pi_n=(k,b_l)\}}\left(r^* - r^{(k)}(b_l)\right)\mu_{max}\right] + Res$$

2. Dealing with randomness of $N_\pi(\tau)$: **High prob. upper bounds for $N_\pi(\tau)$**

$$\overline{Reg}_\pi(\tau) \le \sum_{k,l} \mathbb{E}[T_l^{(k)}(\bar{n})]\left(r^* - r^{(k)}(b_l)\right)\mu_{max} + \underbrace{K \times L \times r^* \mu_{max} \sum_{n>\bar{n}} \mathbb{P}(N_\pi(\tau) > n) + Res}_{O(1) \text{ if } \bar{n} = \frac{2\tau}{\mu}}$$

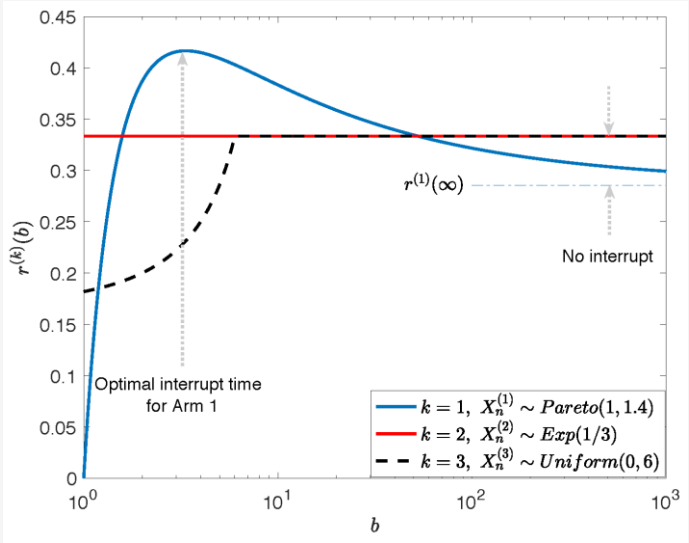
3. Number of suboptimal decisions in $2\tau/\mu$ epochs: Bandit analysis + information structure

- Matching lower bound of order $\Omega(K \log \tau)$ → **order optimality**

Performance Analysis: Numerical Results

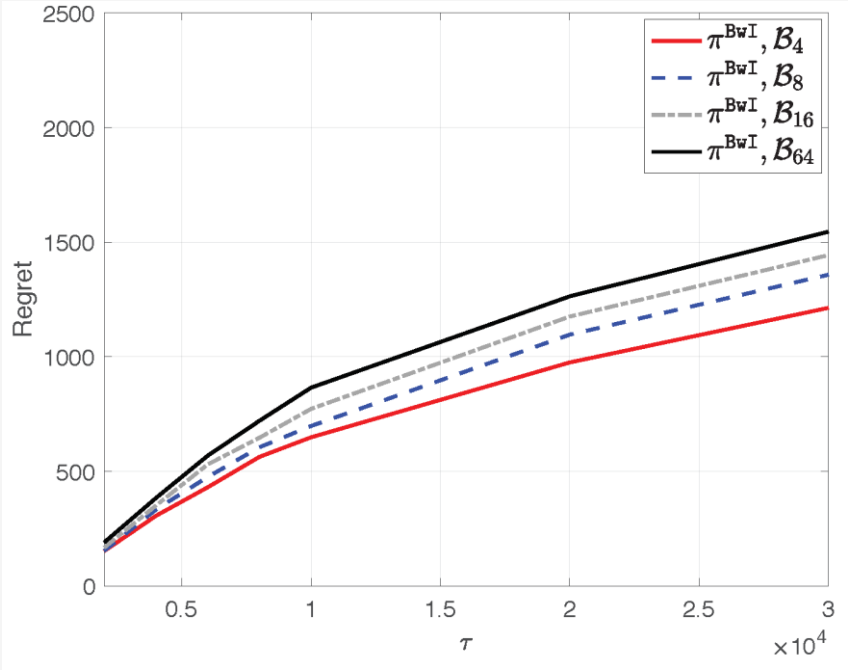
Task scheduling with heterogeneous statistics

1. If no interrupts, arms 2 and 3 are equally better than arm 1. **Optimal interruption → arm 1 optimal**



2. Monotone interrupt set

$$\mathcal{B}_{\mathrm{L}} = \{3, 6, \ldots, 3(\mathrm{L}-1)\}$$



- Logarithmic regret
- Despite **x48** expansion of the action space, only little increase in the regret as a result of the information structure

Conclusions

1. Presented `BwI` framework to incorporate time dimension into sequential learning.

2. General completion time and reward distributions: Interruption as a new control

3. Non-parametric learning algorithm `UCB-BwI` that achieves order optimality in all parameters τ, K, L