Budget-Constrained Learning and Optimization with Bandit Feedback

A Controlled First-Passage Approach

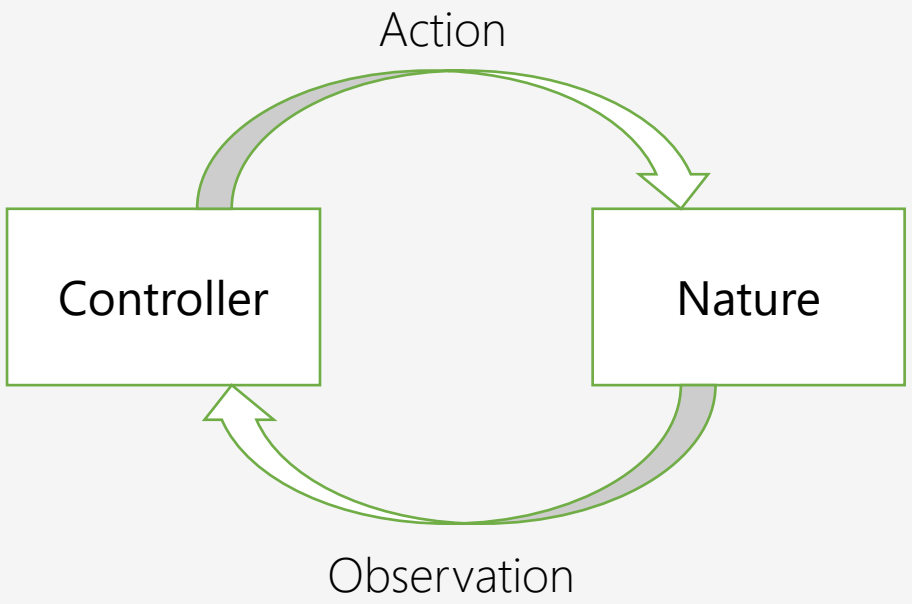
Presenter: Semih Çaycı

Co-authors: Atilla Eryilmaz, R. Srikant, Swati Gupta

October 9, 2020

Exploration-Exploitation Tradeoff

Learning from experience



Exploration-Exploitation Dilemma

Bandit feedback: Observe only what you choose

Limited knowledge → Dilemma:

- Choose the best one according to your knowledge?
- Make an exploration for a potentially better choice?

Multi-armed bandits: Prominent mathematical model

Applications: Robotics, medical trials, online ads, routing, …

What are we investigating?

- Time dimension: as required by many applications
- Very general model for applications
- Novel dynamics + new dilemma

Outline

- Budget-Constrained Bandits (Basic Setup)

  C., Eryilmaz, Srikant, "Budget-Constrained Bandits over General Cost and Reward Distributions", AISTATS 2020

- Continuous-Time Bandits with Controlled Interrupts/Restarts

  C., Eryilmaz, Srikant, "Learning to Control Renewal Processes with Bandit Feedback", SIGMETRICS 2019

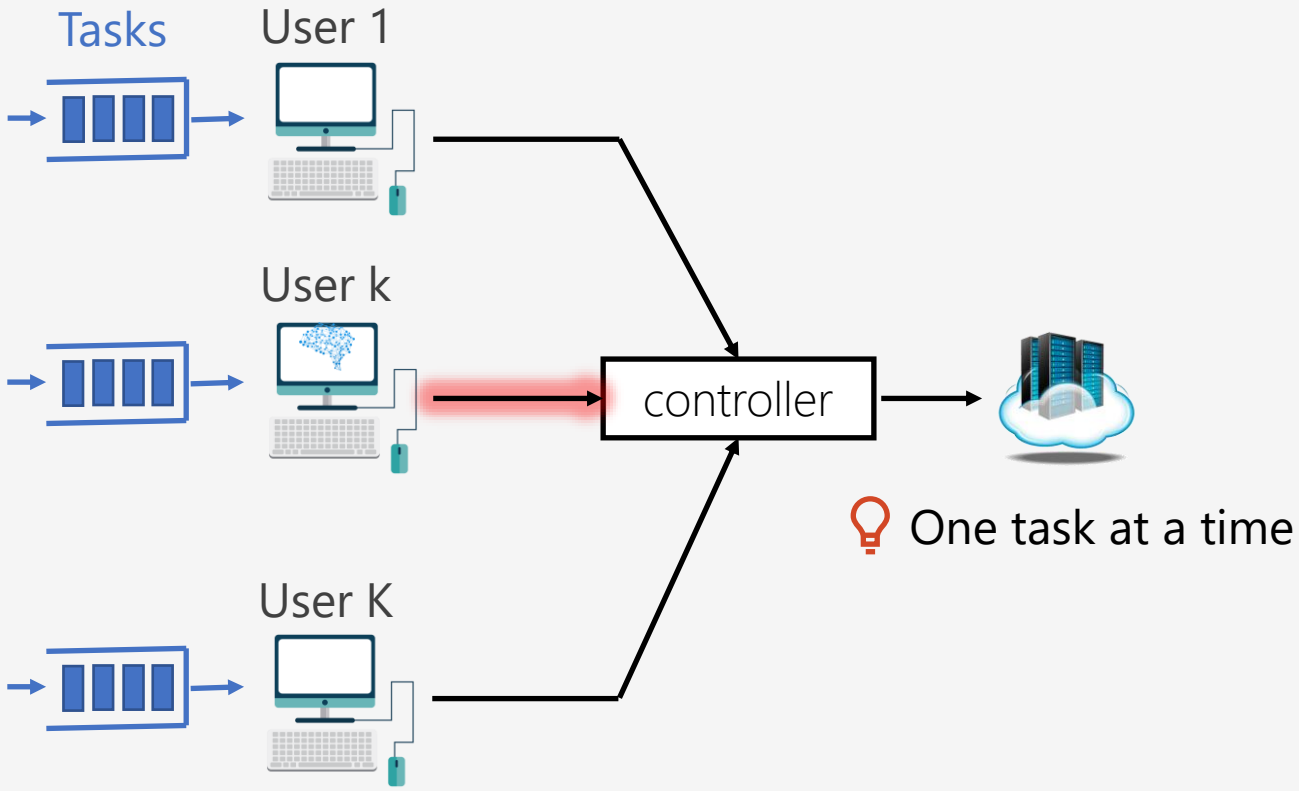
  C., Eryilmaz, Srikant, "Continuous-Time Multi-Armed Bandits with Controlled Restarts", submitted to JMLR

- Conclusions and Future Work

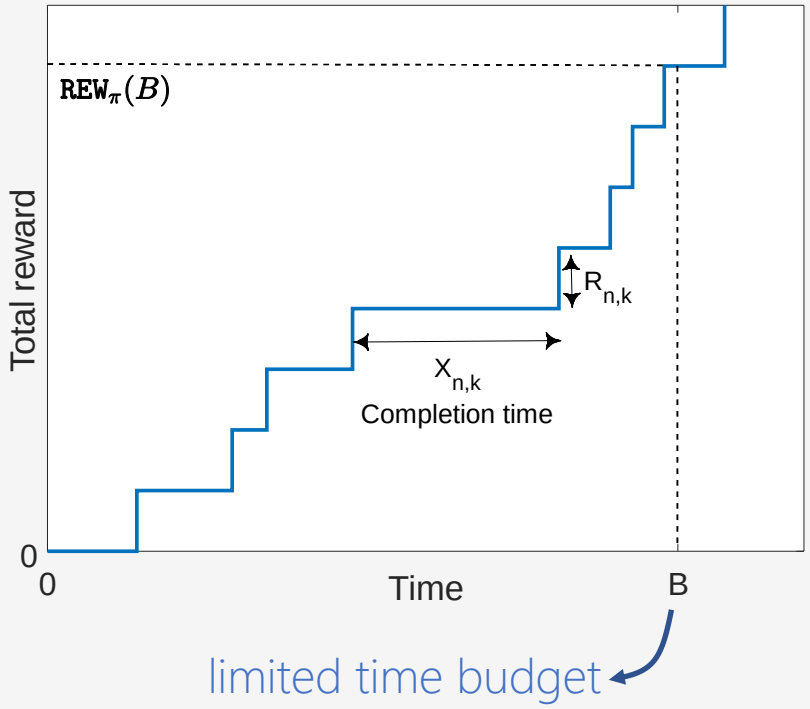
Outline

- Budget-Constrained Bandits over General Cost and Reward Distributions
- Continuous-Time Bandits with Controlled Interrupts-Restarts
- Conclusions and Future Work

Example: Single-Server Task Scheduling



One task at a time



limited time budget

- Random, type-dependent completion time
- Random reward at the end

} Heavy-tailed & positively correlated (Harchol-Balter, '99)

- Objective: Learn to maximize the expected total reward in [0, B] (Throughput maximization)

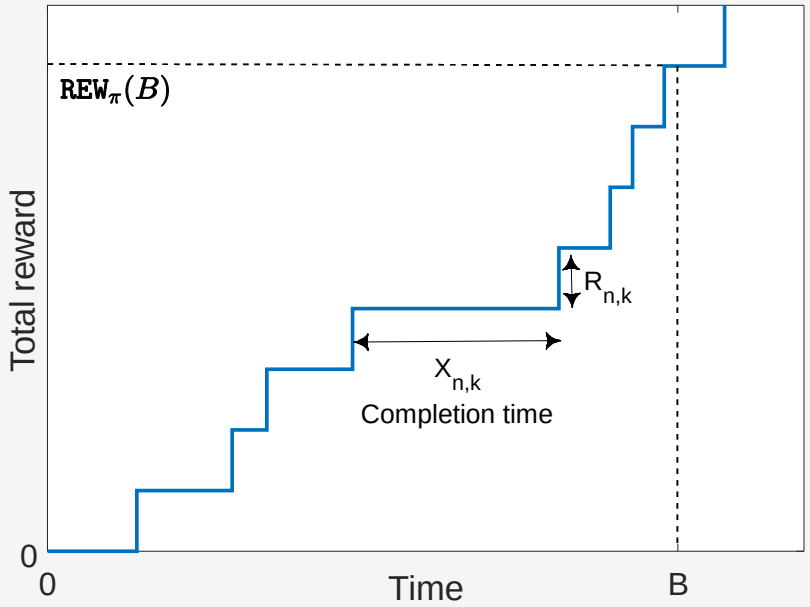
System Model: Controlled First-Passage Formulation

- **Arm k:** $(X_{n,k}, R_{n,k})$ iid from an unknown distribution

  Positive drift: $\mathbb{E}[X_{1,k}] > 0$ (Not necessarily $X_{n,k} > 0$ a.s.)

- **Number of pulls:** $N_\pi(B) = \inf\left\{n : \sum_{t=1}^{n} X_{t,I_t} > B\right\}$ (random)

- **Total reward:** $\mathrm{REW}_\pi(B) = \sum_{n=1}^{N_\pi(B)} R_{n,I_n}$



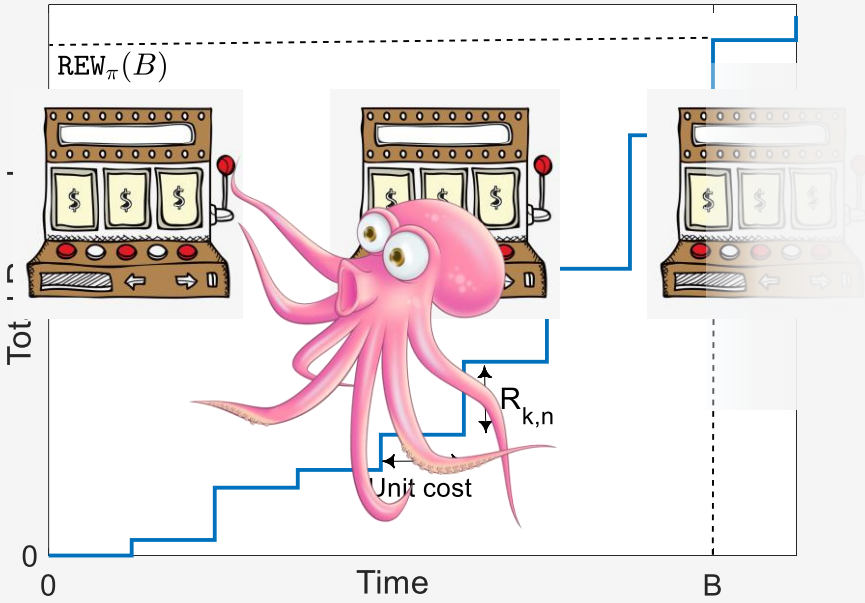
- **Objective:** $\max_\pi \mathbb{E}[\mathrm{REW}_\pi(B)] \equiv \min_\pi \{\underbrace{\mathrm{OPT}(B) - \mathbb{E}[\mathrm{REW}_\pi(B)]}_{\mathrm{Reg}_\pi(B)}\}$

- **Statistics:**
  - Cost $X_{n,k}$ and reward $R_{n,k}$ can be positively correlated.
  - $X_{n,k}$ and $R_{n,k}$ have unbounded support, potentially heavy-tailed

Simpler Case: Classical Stochastic Bandit Problem

- Arm k: $(1, R_{n,k})$ iid from an unknown distribution

$$X_{n,k} = 1, \forall k, n$$ (Unit cost for all arms)



- Number of pulls: $N_\pi(B) = B$ (deterministic)

- Total reward: $\mathrm{REW}_\pi(B) = \sum_{n=1}^{N_\pi(B)} R_{n,I_n}$

- Objective: $\max_\pi \mathbb{E}[\mathrm{REW}_\pi(B)] \equiv \min_\pi \{\underbrace{\mathrm{OPT}(B) - \mathbb{E}[\mathrm{REW}_\pi(B)]}_{\mathrm{Reg}_\pi(B)}\}$

- Optimal offline algorithm: $k^* = \arg\max_k \mathbb{E}[R_{1,k}]$

Budget-Constrained Bandit Problem / Bandits with Knapsacks

1. Stochastic setting: (Badanadiyuru et al., '13; Agrawal & Devanur, '14; Xia et al., '15)
2. Adversarial setting: (Immorlica et al., '19)
3. Contextual bandits: (Agrawal & Devanur, '16; Gyorgy et al. '07)

   See (Slivkins, '19) for a thorough literature review.

Contributions:

- First tight (almost-matching) regret bounds in this setting in a very general setting.
- Understanding the effect of correlation, variability and tail distributions.
- Achievability of O(log B) regret: moments of order $p > 2$

General Budget-Constrained Learning Problems

- Arm k: $(X_{n,k}, R_{n,k})$ iid from an unknown distribution

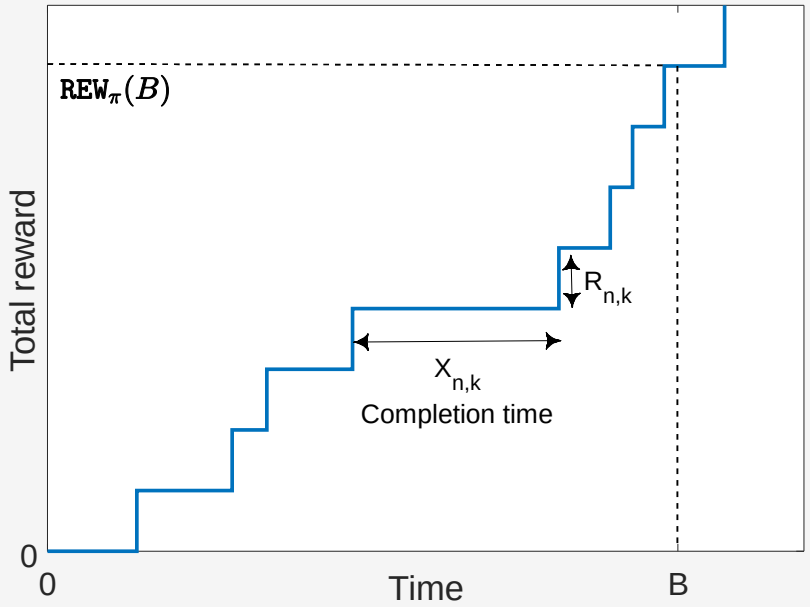
  Positive drift: $\mathbb{E}[X_{1,k}] > 0$ (Not necessarily $X_{n,k} > 0$ a.s.)

- Number of pulls: $N_\pi(B) = \inf\left\{n : \sum_{t=1}^{n} X_{t,I_t} > B\right\}$ (FPT)

- Total reward: $\mathrm{REW}_\pi(B) = \sum_{n=1}^{N_\pi(B)} R_{n,I_n}$

- Optimal offline policy?

? $\arg\max_k \mathbb{E}[R_{1,k}]$ $\arg\max_k \mathbb{E}\left[\frac{R_{1,k}}{X_{1,k}}\right]$ $\arg\max_k \frac{\mathbb{E}[R_{1,k}]}{\mathbb{E}[X_{1,k}]}$



Approximation of the Oracle

1. $\max_{\pi} \mathbb{E}[\mathtt{REW}_{\pi}(B)]$ Unbounded knapsack problem → PSPACE-hard (Papadimitriou, '96)

2. Static approximation: Persistently pull arm k

Renewal theory: $\mathbb{E}[\mathtt{REW}_{\pi}(B)] = \frac{\mathbb{E}[R_{1,k}]}{\mathbb{E}[X_{1,k}]} \cdot B + o(B) = r_k \cdot B + o(B)$ (Gut, '09)

- $o(B)$: O(1) if $Var(X_k)$ exists (Lorden Ineq); $O(B^{1/q})$ if $E|X_k|^q$ exists (Teugels Ineq)
- r_k: reward rate (per unit cost)

Optimal static policy: $\pi_n^{st} = \underset{k}{\operatorname{argmax}}\ r_k$ for all n

Theorem 1. (Optimality Gap)

If $\mathbb{E}[|R_{k,1}|^p] < \infty$ and $\mathbb{E}[|X_{k,1}|^p] < \infty$ for all arms for p > 2, then we have:

$$\mathtt{OPT}(B) - \mathbb{E}[\mathtt{REW}_{\pi^{st}}(B)] = O(1)$$

Bounded optimality gap and asymptotic optimality for π^{st} even for unbounded X_k as $B \to \infty$

Can be used as a benchmark algorithm for online learning purposes

Approximation of the Oracle

1. $\max_{\pi} \mathbb{E}[\mathtt{REW}_{\pi}(B)]$ Unbounded knapsack problem → PSPACE-hard (Papadimitriou, '96)

2. Static approximation: Persistently pull arm k

Renewal theory: $\mathbb{E}[\mathtt{REW}_{\pi}(B)] = \frac{\mathbb{E}[R_{1,k}]}{\mathbb{E}[X_{1,k}]} \cdot B + o(B) = r_k \cdot B + o(B)$ (Gut, '09)

- $o(B)$: O(1) if $Var(X_k)$ exists (Lorden Ineq); $O(B^{1/q})$ if $E|X_k|^q$ exists (Teugels Ineq)
- r_k: reward rate (per unit cost)

Optimal static policy: $\pi_n^{st} = \operatorname{argmax}_k r_k$ for all n

Theorem 1. (Optimality Gap)

If $\mathbb{E}[|R_{k,1}|^p] < \infty$ and $\mathbb{E}[|X_{k,1}|^p] < \infty$ for all arms for p > 2, then we have:

$$\mathtt{OPT}(B) - \mathbb{E}[\mathtt{REW}_{\pi^{st}}(B)] = O(1)$$

$\mathtt{REW}_{\pi}(B) \le r^* \cdot \left(B + \mathbb{E}[S^{\pi}_{N_{\pi}(B)} - B]\right)$ under any policy π (by generalized Wald's identity + MCT)

$\mathtt{REW}_{\pi^{st}}(B) \ge r^* \cdot B$ by elementary renewal theorem (generalized to stopped random walks (Gut, '09))

UCB-B1 Algorithm: Jointly Gaussian Case

1. $\pi_n^{st} = \underset{k}{\text{argmax}}\ r_k$ "Optimism in the face of uncertainty" principle
2. Simple case: Jointly Gaussian + known 2nd-order moments

$$\text{rad}_s(X, \delta) = \sqrt{\frac{2\text{Var}(X)\log(\delta^{-1})}{s}}$$ by Hoeffding inequality

Empirical estimation

$$\widehat{\mathbb{E}}_s[X_k] = \frac{1}{s}\sum_{t=1}^{s} X_{t,k}$$

$$\widehat{r}_{k,s} = \frac{\widehat{\mathbb{E}}_s[R_k]}{\widehat{\mathbb{E}}_s[X_k]}$$

Concentration inequality for reward rate

$$\left|\widehat{r}_{k,s} - r_k\right| \le \Psi_{k,s}\left(\text{rad}_s(X_k, \delta), \text{rad}_s(R_k, \delta)\right)$$ w.h.p.

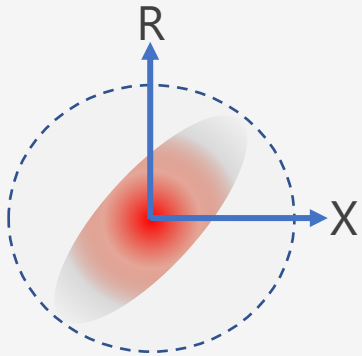
where $\Psi_{k,s}(x, y) = \dfrac{y + \widehat{r}_{k,s} \cdot x}{\left(\widehat{\mathbb{E}}_s[X_k] - x\right)_+}$

increasing in x and y

Uncertainty $\propto$ Confidence radius

UCB–B1 Algorithm: Jointly Gaussian Case

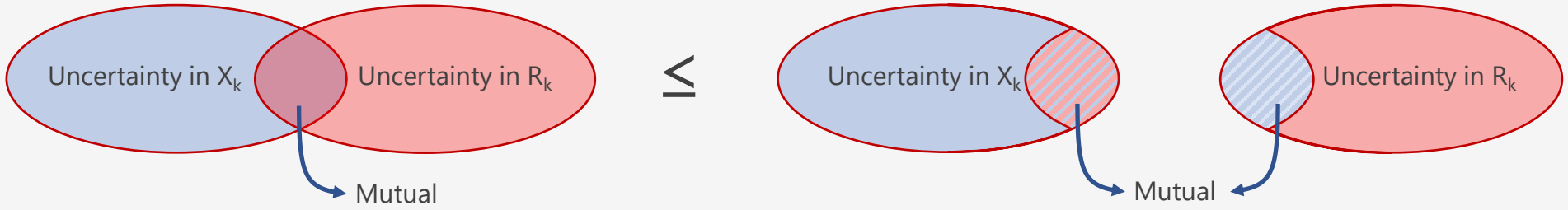
1 Joint estimation problem: Estimate separately (ignore correlation)



$\{X_{i,k} : 1 \le i \le s\} \rightarrow \widehat{\mathbb{E}}_s[X_k]$

$\{R_{i,k} : 1 \le i \le s\} \rightarrow \widehat{\mathbb{E}}_s[R_k]$

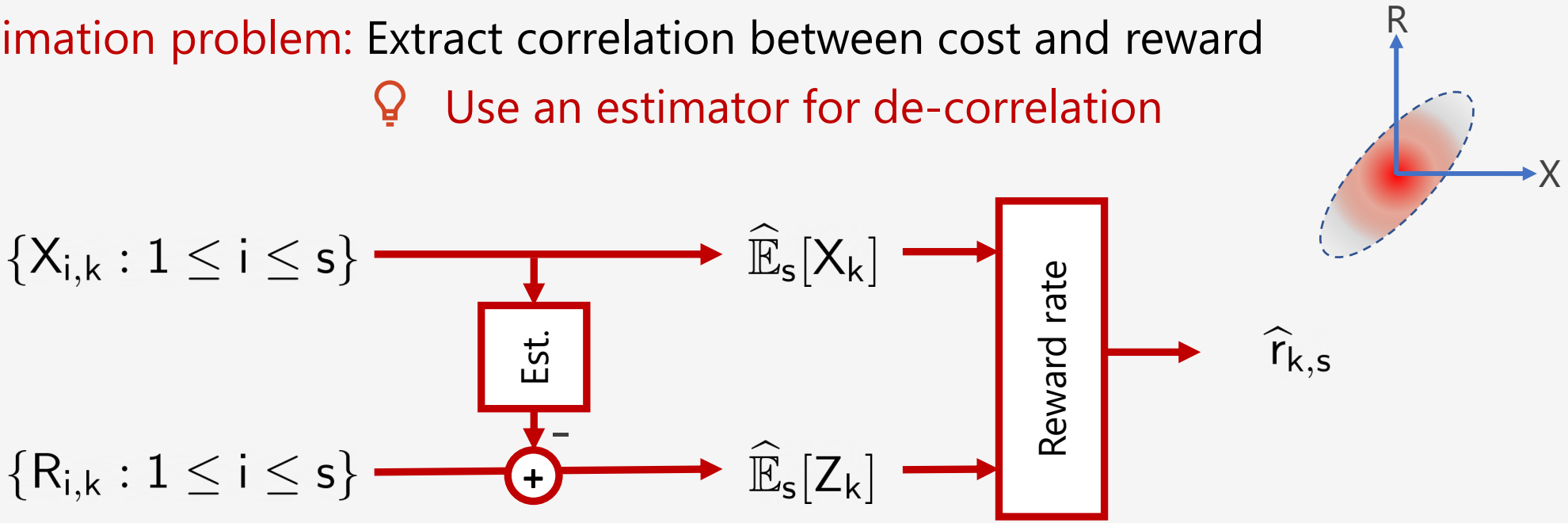
Reward rate $\rightarrow \widehat{r}_{k,s}$

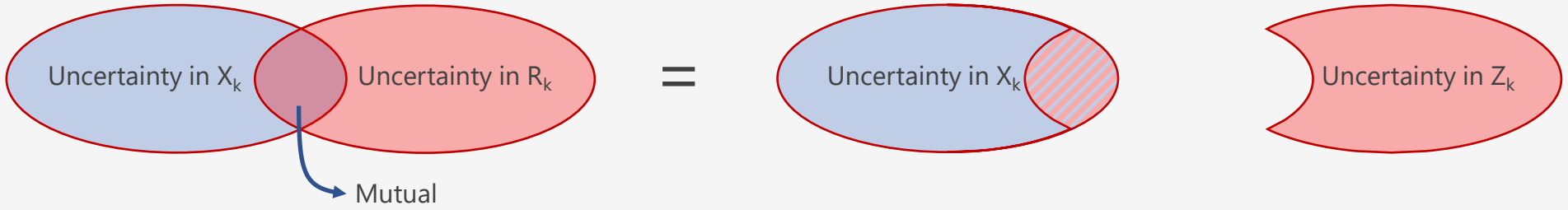


UCB–B1 Algorithm: Jointly Gaussian Case

2 **Joint estimation problem:** Extract correlation between cost and reward

Use an estimator for de-correlation

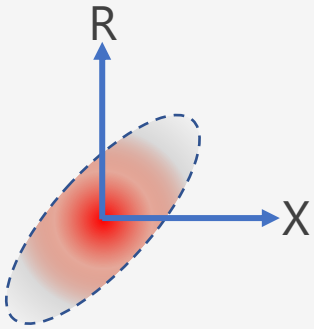




UCB-B1 Algorithm: Jointly Gaussian Case

Q How to extract the correlation?

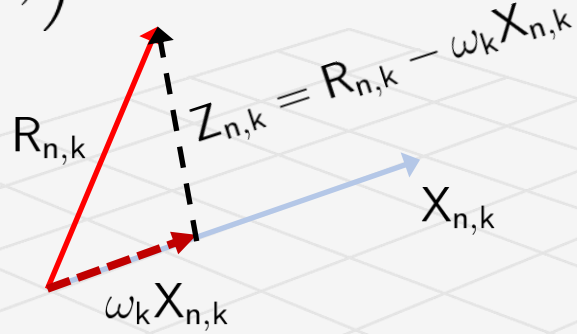
A Linear MMSE Estimator: $\omega_k = \arg\min_{\omega\in\mathbb{R}} \mathrm{Var}(R_{k,1} - \omega X_{k,1}) = \dfrac{\mathrm{Cov}(X_{k,1}, R_{k,1})}{\mathrm{Var}(X_{k,1})}$



Idea: Minimize the variance of the residual term $Z_{k,n} = R_{k,n} - \omega_k X_{k,n}$

Result: Mean estimation with smaller variance $\mathtt{rad}_s(Z_k, \delta) \le \mathtt{rad}_s(R_k, \delta)$

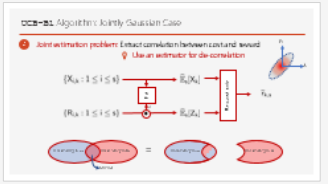
$$\begin{aligned}\left|\hat{r}_{k,s} - r_k\right| &\le \Psi_{k,s}\Big(\mathtt{rad}_s(X_k, \delta), \mathtt{rad}_s(Z_k, \delta)\Big) \\ &\le \Psi_{k,s}\Big(\mathtt{rad}_s(X_k, \delta), \mathtt{rad}_s(R_k, \delta)\Big)\end{aligned}$$



Smaller confidence radius ⇨ Lower regret

UCB-B1 Algorithm: Jointly Gaussian Case

$$\texttt{UCB-B1}: I_n \in \arg\max_{k\in[K]} \left\{ \hat{r}_{k,T_k(n)} + \Psi_{k,T_k(n)}\left(\texttt{rad}_{T_k(n)}(X_k, n^{-\alpha}), \texttt{rad}_{T_k(n)}(Z_k, n^{-\alpha})\right)\right\}$$



Theorem 2. (Regret Upper Bound for UCB-B1)

Let $\sigma_k^2 = \mathrm{Var}(Z_{k,1}) + (r^* - \omega_k)^2 \mathrm{Var}(X_k)$ and $\Delta_k = r^* - r_k$. Then, we have:

$$\mathrm{Reg}_{\pi^{B1}}(B) \le C_1 \sum_{k:\Delta_k>0} \frac{\sigma_k^2}{\Delta_k \mathbb{E}[X_k]} \log(B) + O(1) \quad \text{for} \quad C_1 = 10.5\alpha$$

- Classical stochastic MAB regret bounds for $\mathrm{Var}(X_{1,k}) = 0$.
- Exploiting the correlation: $O\left(\log(B) \sum_{k:\Delta_k>0} \frac{\mathrm{Cov}(X_{1,k}, R_{1,k})}{\Delta_k}\right)$ gain.

UCB-B1 Algorithm: Jointly Gaussian Case

$$\texttt{UCB-B1}: I_n \in \arg\max_{k\in[K]} \left\{ \hat{r}_{k,T_k(n)} + \Psi_{k,T_k(n)}\left(\texttt{rad}_{T_k(n)}(X_k, n^{-\alpha}), \texttt{rad}_{T_k(n)}(Z_k, n^{-\alpha})\right)\right\}$$

Theorem 2. (Regret Upper Bound for UCB-B1)

Let $\sigma_k^2 = \mathrm{Var}(Z_{k,1}) + (r^* - \omega_k)^2 \mathrm{Var}(X_k)$ and $\Delta_k = r^* - r_k$. Then, we have:

$$\mathrm{Reg}_{\pi^{B1}}(B) \le C_1 \sum_{k:\Delta_k>0} \frac{\sigma_k^2}{\Delta_k \mathbb{E}[X_k]} \log(B) + O(1) \quad \text{for} \quad C_1 = 10.5\alpha$$

Challenge: Random cost → Random stopping time and asynchronous decisions

$$\texttt{REG}_\pi(B) \le \sum_{k:\Delta_k>0} \mathbb{E}\left[T_k(n_0(B))\right] \underbrace{\mathbb{E}[X_{1,k}]\Delta_k}_{\text{Regret per stage}} + O(1)$$

$n_0(B) = 2B/\mu_{\min} \ge N_\pi(B)$

? How tight?

Regret Lower Bounds: Jointly Gaussian Case

(Regret Lower Bound for Gaussian Case)

Let $(X_{k,n}, R_{k,n}) \sim \mathcal{N}(\mu_k, \Sigma_k)$ with known $\Sigma_k > 0$. Then,

$$\liminf_{B\to\infty} \frac{\mathrm{Reg}_\pi(B)}{\log(B)} \geq \sum_{k:\Delta_k>0} \frac{\sigma_k^2}{\mathbb{E}[X_{k,1}]\Delta_k}$$

$$\sigma_k^2 = \mathrm{Var}(Z_{k,1}) + (r^* - \omega_k)^2 \mathrm{Var}(X_k)$$

Recall: $\mathrm{Reg}_{\pi^{B1}}(B) \leq C_1 \sum_{k:\Delta_k>0} \frac{\sigma_k^2}{\Delta_k \mathbb{E}[X_k]} \log(B) + O(1)$

- **Optimal** regret up to a universal constant C_1 for `UCB-B1`.
- Reflects the impact of variability and correlation on the regret.

Regret Lower Bounds: General Case

Let $(X_{k,n}, R_{k,n}) \sim P_{\theta_k}$ for $\theta_k \in \Theta_k$

Information geometry: For any r > 0,

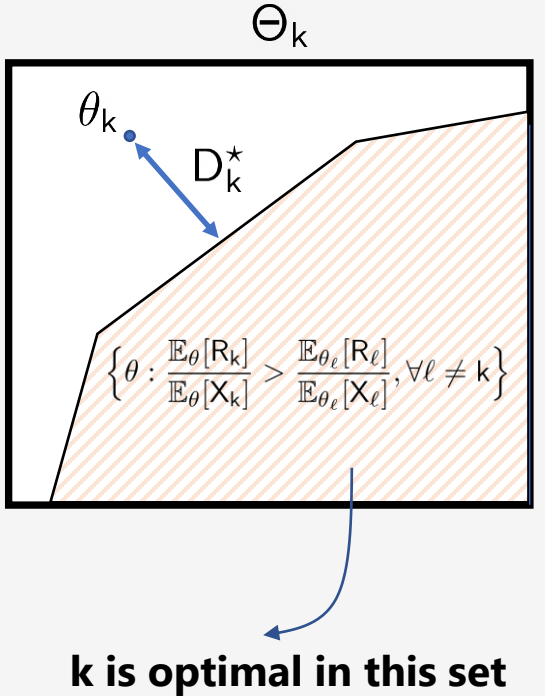
$$D_k^*(r) = \min_{\theta \in \Theta_k} D(P_{\theta_k} || P_\theta) \text{ s.t. } \frac{\mathbb{E}_\theta[R_{k,1}]}{\mathbb{E}_\theta[X_{k,1}]} \geq r \quad \text{(M-projection)}$$

Theorem 3. (Regret Lower Bound)

Let $D_k^\star = D_k^*(\max_k r_k)$. Then, for any uniformly good policy π:

$$\liminf_{B \to \infty} \frac{\mathrm{Reg}_\pi(B)}{\log(B)} \geq \frac{1}{2} \sum_{k:\Delta_k > 0} \frac{\mathbb{E}[X_{k,1}]\Delta_k}{D_k^\star}$$

if all moments of order p > 2 exist. $\frac{1}{D_k^\star} \propto CoV(X_k)$ $D_k^\star \propto (r^* - r_k)$



UCB–B1 Algorithm: Bounded Cost and Reward

Known 2nd-order moments, $X_{n,k} \in [0, M_X],\ R_{n,k} \in [0, M_R]$ Hoeffding inequality – only $Vol(\text{supp}(X_k))$

Bernstein inequality – $Var(X_k)$

$$\texttt{rad}_s(X_k, \delta) = \frac{2M_X \log(\delta^{-1})}{3s} + \sqrt{\frac{2\text{Var}(X_{n,k}) \log(\delta^{-1})}{s}}$$

$$\text{Reg}_{\pi^{B1}}(B) \leq C_1 \sum_{k:\Delta_k>0} \left(\frac{\sigma_k^2}{\Delta_k \mathbb{E}[X_k]} + (M_R + r_k M_X) + \frac{M_X \Delta_k}{2} \right) \log(B) + O(1)$$

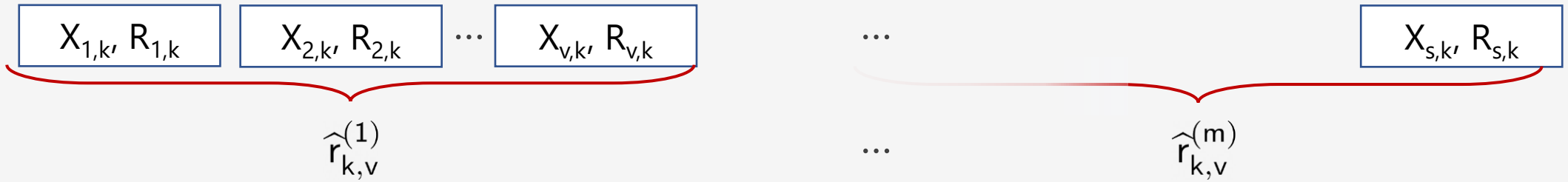
$$\sigma_k^2 = \text{Var}(Z_{k,1}) + (r^* - \omega_k)^2 \text{Var}(X_k)$$

- M_X and M_R dependence is inevitable
- Higher regret as $\min(M_X, M_R)$ increases
- Defect of the empirical estimator (Bubeck, 2012)

UCB–M1 for Heavy-Tailed Cost and Reward

- Empirical estimation fails for HT: polynomial not exponential convergence rate
- Median-based robust rate estimation (Nemirovski & Yudin, '83; Bubeck et al., '13)

  Idea: Divide the data into chunks and take the median – exploit the correlation inside the chunks



$\bar{r}_{k,s} = \text{median}\{\widehat{r}_{k,v}^{(1)}, \ldots, \widehat{r}_{k,v}^{(m)}\} \to r_k$ exponentially fast if $m = \lceil 3.5 \log(\delta^{-1}) \rceil + 1$

- UCB–M1: If $\mathbb{E}[|R_{k,1}|^p] < \infty$ and $\mathbb{E}[|X_{k,1}|^p] < \infty$ for all arms for $p > 2$,

$$\text{Reg}_{\pi^{M1}}(B) \leq \sum_{k:\Delta_k > 0} C_{M1} \frac{\sigma_k^2}{\mathbb{E}[X_{1,k}]\Delta_k} \log(B) + O(1)$$

- $C_{M1} > C_1$: Price of generality

UCB-B2: Using Empirical Estimates

- What if you do not know the second-order moments?
- Using empirical estimates in UCB-B1 → UCB-B2

$$\widehat{\mathbb{V}}_s(X_k) = \frac{1}{s}\sum_{i=1}^{s}\left(X_{k,i} - \widehat{\mathbb{E}}_s[X_k]\right)^2 \to \mathrm{Var}(X_{k,1})$$

$$\widehat{\omega}_{k,s} = \frac{\sum_{i=1}^{s}\left(X_{k,i} - \widehat{\mathbb{E}}_s[X_k]\right)\left(R_{k,i} - \widehat{\mathbb{E}}_s[R_k]\right)}{\sum_{i=1}^{s}\left(X_{k,i} - \widehat{\mathbb{E}}_s[X_k]\right)^2} \to \omega_k$$

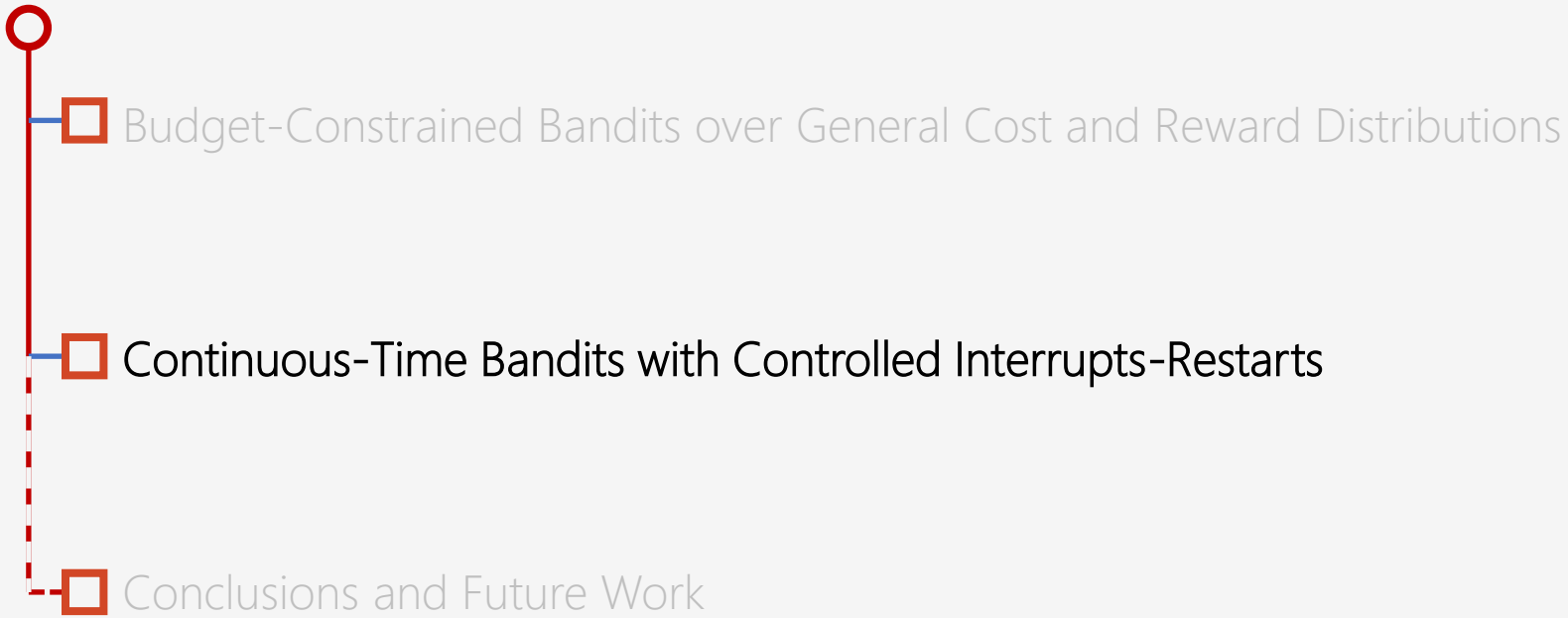
Random-design least squares

- Non-asymptotic analysis of using these empirical estimates: kurtosis
- Small Δ_k → UCB-B1

Conclusions

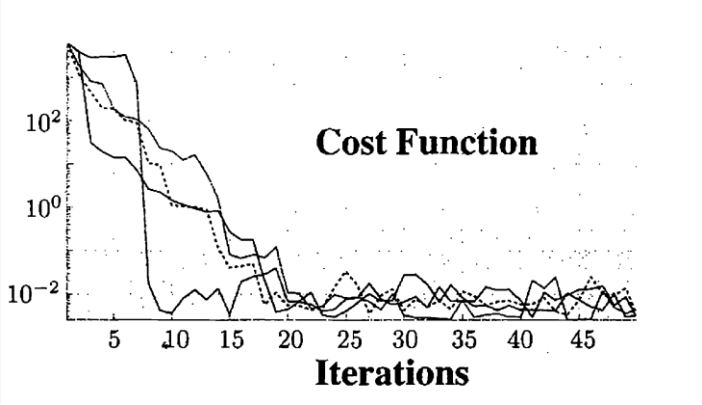
1. Regret lower bound: Problem-dependent fundamental performance limit
2. Algorithms with tight problem-dependent regret bounds
3. Optimality: Optimal regret up to a constant factor in the Gaussian case
4. Achievability: O(log B) regret if p-moments exist for $p > 2$

Outline


Budget-Constrained Bandits over General Cost and Reward Distributions
Continuous-Time Bandits with Controlled Interrupts-Restarts
Conclusions and Future Work

Example: Early Stopping

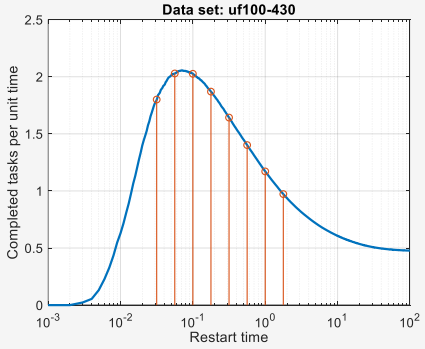
Early Stopping in SGD



Source: G. Strang, "Linear Algebra and Learning from Data"

- Drops very fast, then fluctuates indefinitely
- Early stopping dilemma: At any time, should the controller:
  - Stop the process, make a new trial
  - Wait until convergence (which may never happen)
  - Too early (miss the solution) vs. too late (waste time)

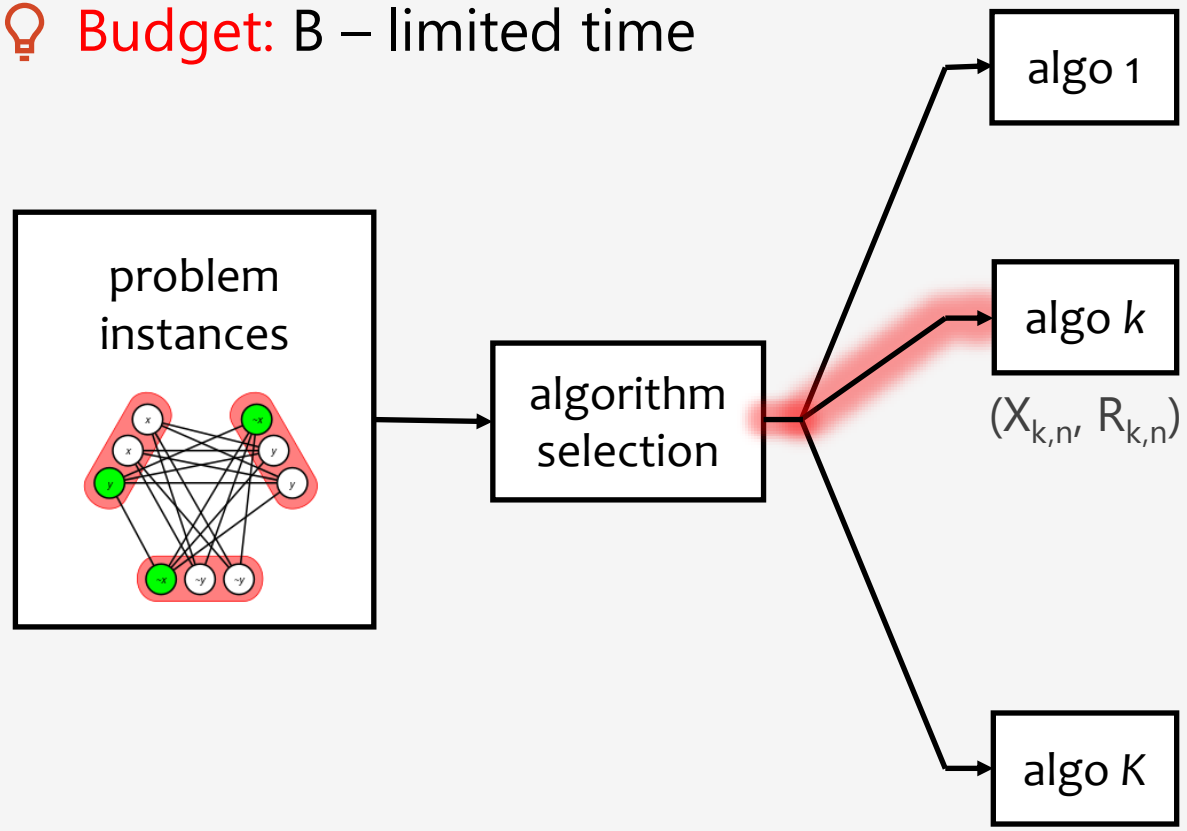
Restart in k-SAT Problem



- Solve a Boolean equation: $f(x_1, x_2, \ldots, x_n) = 0,\ x_i \in \{0, 1\}$
- GD-like WalkSAT: Choose x(t) at random, flip it, check if it improves
- Solution time is heavy-tailed.
- If it takes too long, HALT, RESTART with a new random assignment

Example: Algorithm Portfolio with Early Stopping

- Budget: B – limited time



- K hyperparameter choices
- Each choice: Different time and success
- Objective:

Find optimal (hyperparameter, stopping rule)

to maximize success in [0, B]

- Wanted:

Adaptive learning algorithms for restart strategies

Problem Setting

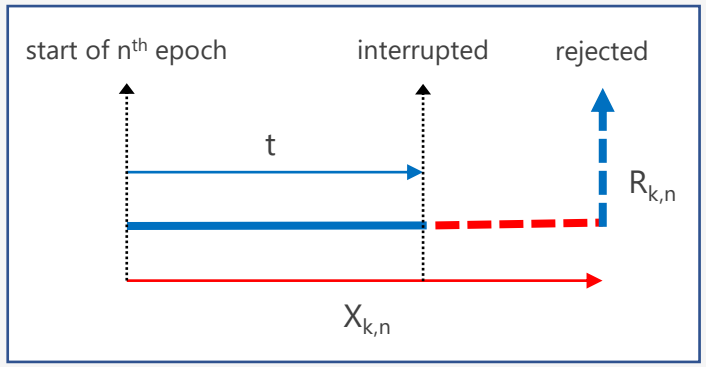
1. K-armed bandit setting: Arm k ↔ $(X_{k,n}, R_{k,n})$ (renewal reward process)
2. Controlled interrupts: Interrupt a task if it is not completed in t time units
3. Actions: $\pi_n = (I_n, \tau_n) \in [K] \times \mathbb{T}$ for $\mathbb{T} \subset \mathbb{R}_+ \cup \{\infty\}$
4. Feedback: Right-censored feedback

$$U_{k,n}(t) = \min\{X_{k,n}, t\}$$

$$V_{k,n}(t) = R_{k,n}\mathbb{I}\{X_{k,n} \le t\}$$

$$r_k(t) = \frac{\mathbb{E}[V_{k,1}(t)]}{\mathbb{E}[U_{k,1}(t)]} = \frac{\mathbb{E}[R_{k,1}\mathbb{I}\{X_{k,1} \le t\}]}{\mathbb{E}[\min\{X_{k,1}, t\}]}$$

5. Objective: $\max_{\pi} \mathbb{E}[\mathrm{REW}_{\pi}(B)]$
6. Dilemmas:
   1. Exploration or exploitation
   2. Interrupt an ongoing task or wait until the end



Optimal Interrupt Time

1. Consider the largest possible set of interrupt times: $\mathbb{T} = \mathbb{R}_+ \cup \{\infty\}$ (No interrupt)
2. Optimal interrupt time: $t_k^* = \arg\max_{t \in \mathbb{T}} r_k(t)$

Theorem 4. (Optimal Interrupt Time)

Interruption is optimal, i.e., $t_k^* < \infty$ if there exists $t \in \mathbb{T}$ s. t.

$$\frac{\mathbb{E}[R_{k,1} | X_{k,1} > t]}{\mathbb{E}[X_{k,1} - t | X_{k,1} > t]} < \frac{\mathbb{E}[R_{k,1}]}{\mathbb{E}[X_{k,1}]}$$

Mean residual reward rate < reward rate of a fresh cycle at any time t

Optimal Interrupt Time

Easier to interpret if $X_{k,n}$ and $R_{k,n}$ are independent.

Corollary. (Optimal Interrupt Time for Independent Time and Reward)

Interruption is optimal, i.e., $t_k^* < \infty$ if there exists $t \in \mathbb{T}$ s. t.

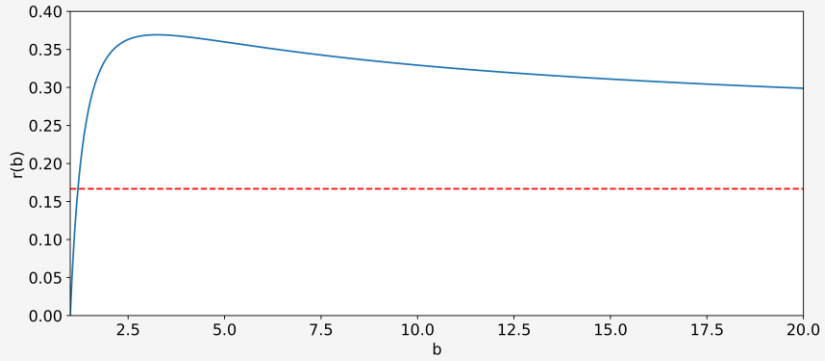
$$\mathbb{E}[X_{k,1} - t | X_{k,1} > t] > \mathbb{E}[X_{k,1}]$$

- Mean residual life > mean completion time of a fresh epoch at t

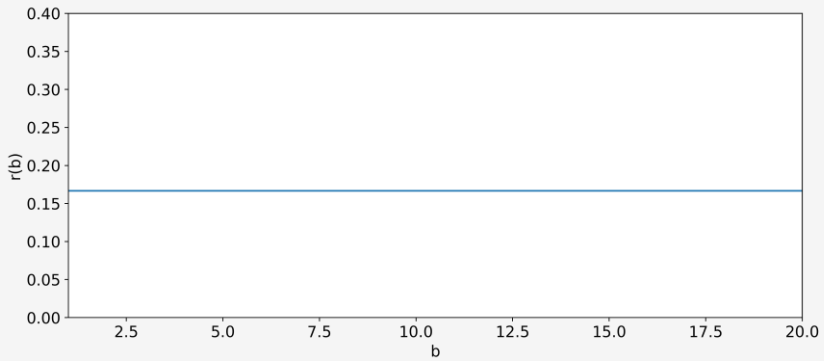
- All **heavy-tailed** completion time distributions require interruption.
- Some **light-tailed** distributions require interruption: Hyperexponential
- Exponential distribution is the barrier case: Memoryless property

Optimal Interrupt Time: Independent R and X

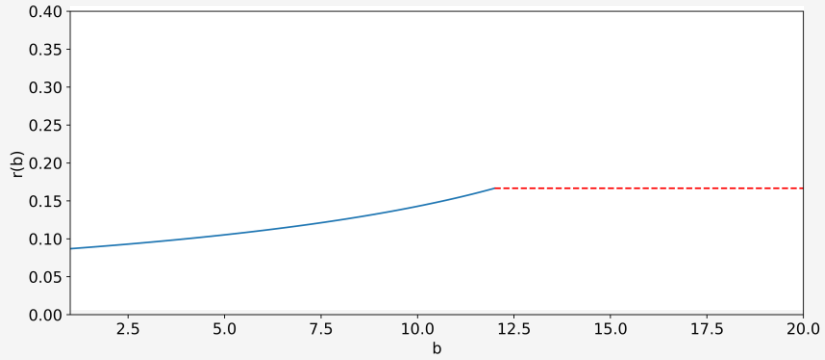
Reward rate as a function of interrupt times



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



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



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

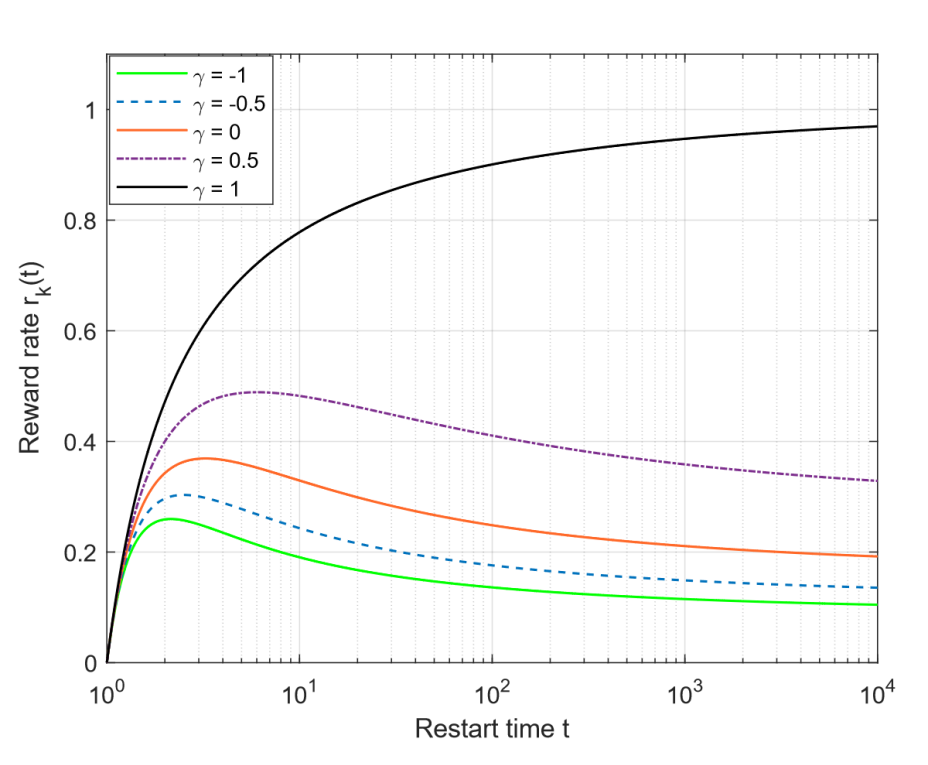
- All have the same mean → **Higher-order moments**

  O(B) regret without learning about the tail distribution.

- In the classic setting → **Only mean values are important**

Optimal Interrupt Time: Correlated R and X

Reward rate as a function of interrupt times



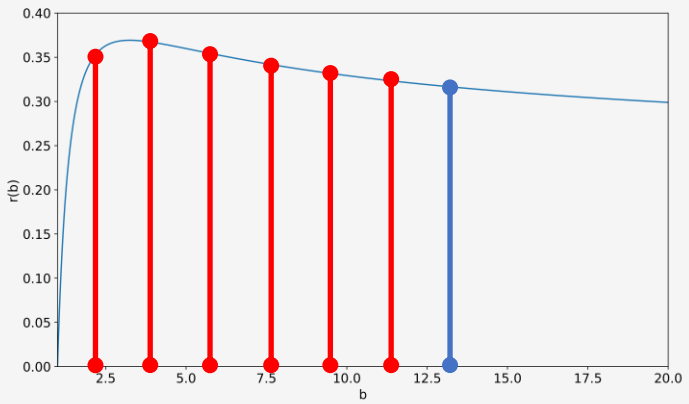
$X_n \sim Pareto(1, 1.2)$

$R_n = X_n^{\gamma}$

- $\gamma < 1$: Interruption is optimal
- $\gamma \geq 1$: Wait until the end

UCB-BwI Algorithm

Information structure:



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

UCB-RB Algorithm: Performance Analysis

1. Regret upper bounds for UCB-RB: If $|\mathbb{T}| = L$, then

$$\mathrm{Reg}_{\pi^R}(B) = O\big(K\log(B) + KL\big)$$

Interesting result: $|[K] \times \mathbb{T}| = KL$ means O(KL log(B)) regret without info structure

2. Regret lower bound: For any policy, $\mathrm{Reg}_{\pi}(B) = \Omega\big(K\log(B)\big)$

UCB-RB is **order-optimal** in all parameters K, L, B.

3. Unimodal & Holder-continuous $r_k(t)$ and $\mathbb{T} = [a, b]$: $O\left(\sqrt{B\log(B)}\right)$ regret.

Unimodal bandit optimization with information structure (Combes et al., '14)

Application: Boosting WalkSAT for Random 3-SAT Problems

- WalkSAT: Stochastic local search to solve k-SAT problem (Papadimitriou, 1991)
- How does it work? $f(x_1, x_2, \ldots, x_m) = \underbrace{(x_1 \vee \neg x_3 \vee x_7)}_{\text{Clause}} \wedge (x_2 \vee \neg x_4 \vee \neg x_{12}) \wedge \ldots$

  1. Random initial assignment: $x_i \sim \text{Ber}(1/2),\ i = 1, 2, \ldots, m$
  2. Until finding an assignment s.t. f(x) = 1:

     Choose J at random, flip x_J

     - If # of satisfied clauses increases, keep the flipped x_J
     - Otherwise, revert back to the original value.

  } Boolean analogue of SGD

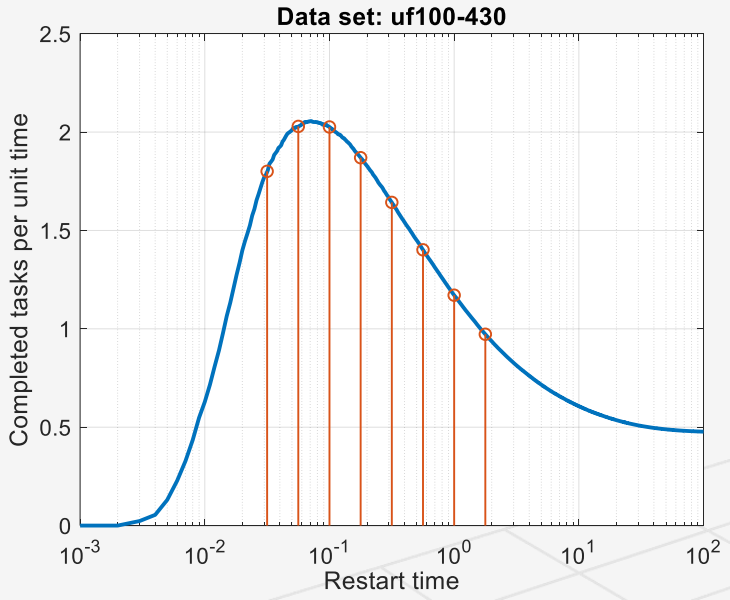
- Heavy-tailed run-time distribution (Selman et al., '04)

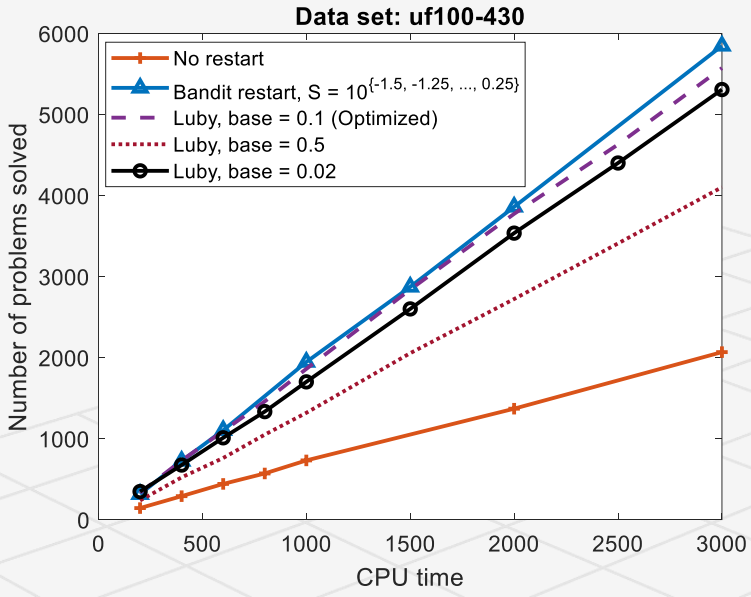
Application: Boosting WalkSAT for Random 3-SAT Problems

- Solver: WalkSAT v56 (source: https://gitlab.com/HenryKautz/Walksat)
- Data set: `uf100-430`

  Description: 1000 Uniform Random 3-SAT instances with 100 variables and 430 clauses

  source: SATLIB Benchmark https://www.cs.ubc.ca/~hoos/SATLIB/benchm.html
- Experiments: 10 trials from each instance. $R_n = 1$ for successful trials. Bootstrap resampling.
- Empirical data:

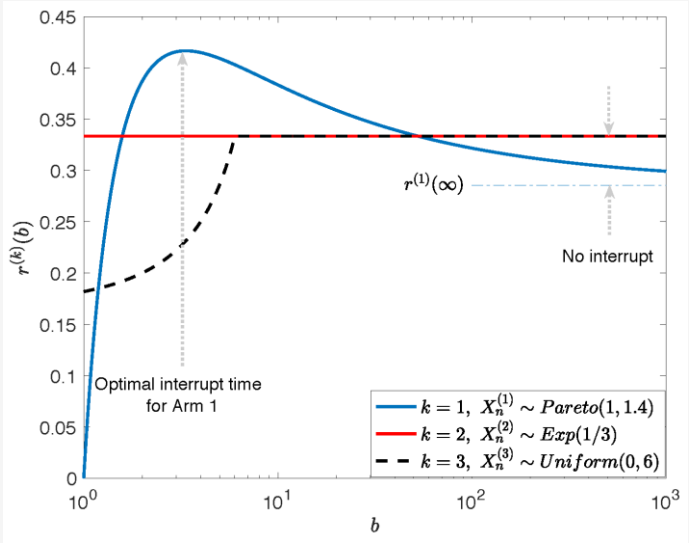




Performance Analysis: Numerical Results

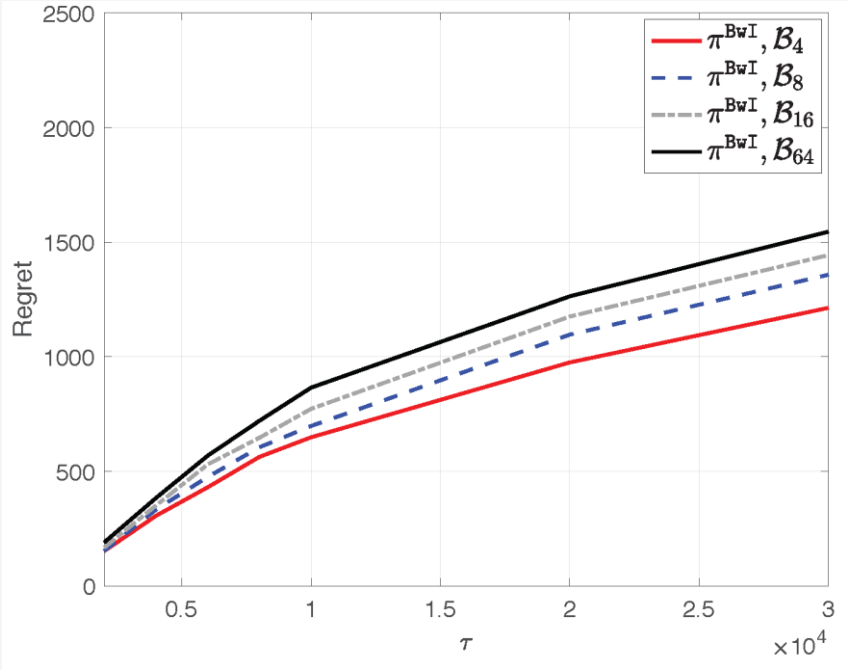
Task scheduling with heterogeneous statistics

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

- Incorporated random time-varying completion times with heterogeneous statistics.
- Principled approach to learn optimal restart strategies
- No-regret learning algorithm based on a flexible and low-complexity design paradigm.

Future work:

- Constrained convex optimization with bandit feedback based on Lyapunov optimization:

$$\max \sum_{k} U_k\left(\bar{r}_k^{\pi}(B)\right) \text{ s.t. } \frac{\mathbb{E}[\sum_{n=1}^{N_\pi(B)} Y_{m,n}^{\pi}]}{B} \leq c_m, \ m = 1, 2, \ldots, M$$

- Preliminary result: $\tilde{O}\left(\sqrt{1/B}\right)$ regret, $\tilde{O}\left(\sqrt{1/B}\right)$ constraint violation.
- Light-weight algorithm design paradigm for dynamical systems.
- Potential improvements: momentum methods (heavy-ball), Thompson sampling.

Thank you.

References

- M. Harchol-Balter et al., "*The Effect of Heavy-Tailed Job Size Distributions on Computer System Design*", ASA-IMS Conf. on App. of Heavy Tailed Distributions in Economics, Engineering and Statistics 1999
- A. Badanidiyuru et al., "*Bandits with Knapsacks*", IEEE FOCS 2013
- Y. Xia et al., "*Thompson Sampling for Budgeted Multi-Armed Bandits*", IJCAI 2015
- S. Agrawal et al., "*Bandits with Concave Rewards and Convex Knapsacks*", ACM EC 2014
- N. Immorlica et al., "*Adversarial Bandits with Knapsacks*", IEEE FOCS 2019
- A. Gyorgy et al., "*Continuous Time Associative Bandits*", IJCAI 2007
- S. Agrawal et al., "*Linear Contextual Bandits*", NeurIPS 2016
- G. Papadimitriou et al., "*The Complexity of Optimal Queueing Network Control*", MOOR 1999
- A. Slivkins, "*Introduction to Multi-Armed Bandits*", arXiv:1904.07272 2019