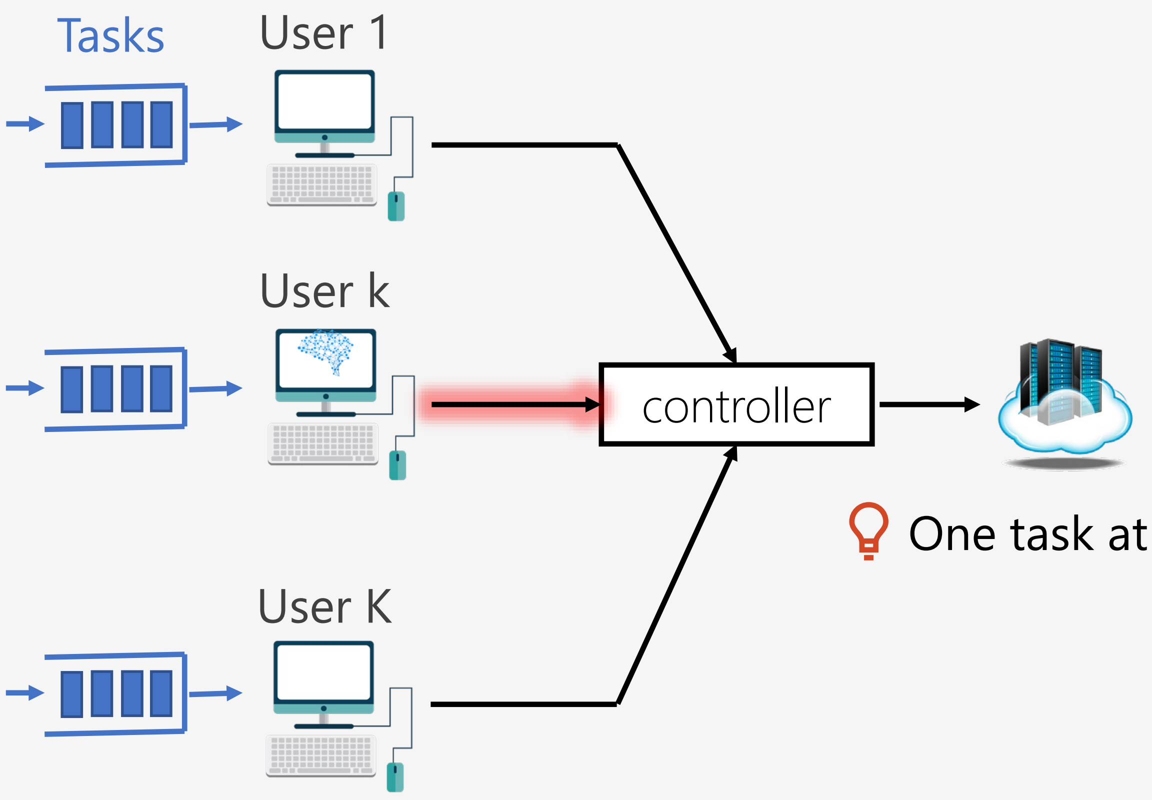
Budget-Constrained Bandits over General Cost and Reward Distributions

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

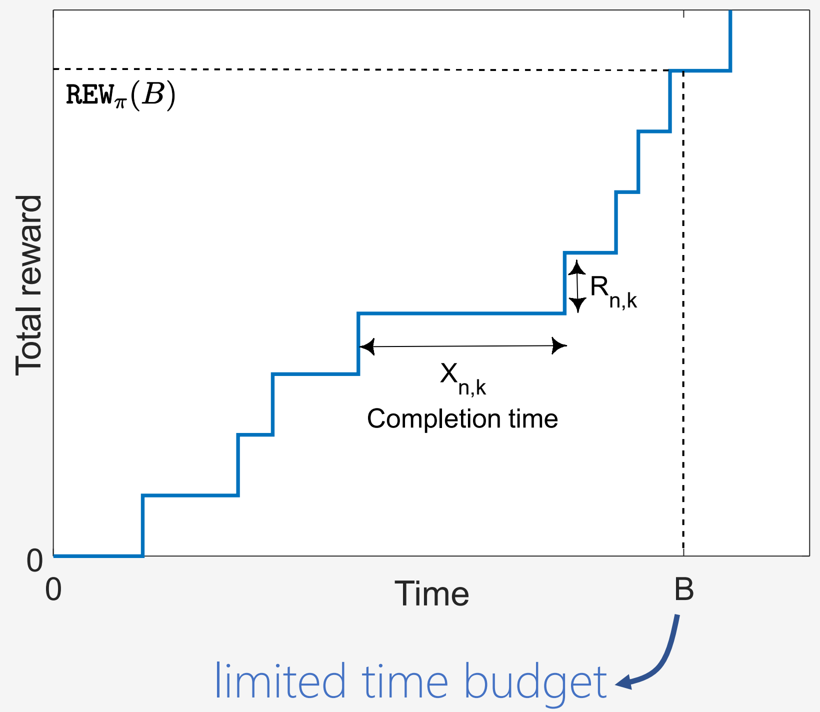
[1] The Ohio State University, ECE

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

Example: Single-Server Task Scheduling



- One task at a time



- 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]

Budget-Constrained Bandit Problem

Bandit problem with random $X_{n,k}$. Examples incl. dynamic pricing, adaptive routing.

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

Our contributions:

- Regret lower bound
- Algorithms that achieve tight (almost-matching) regret bounds
  - Positive correlation between cost and reward
  - Variability of cost and reward
- Empirical variance estimates for improved performance without prior knowledge
- Unbounded and potentially heavy-tailed cost and reward

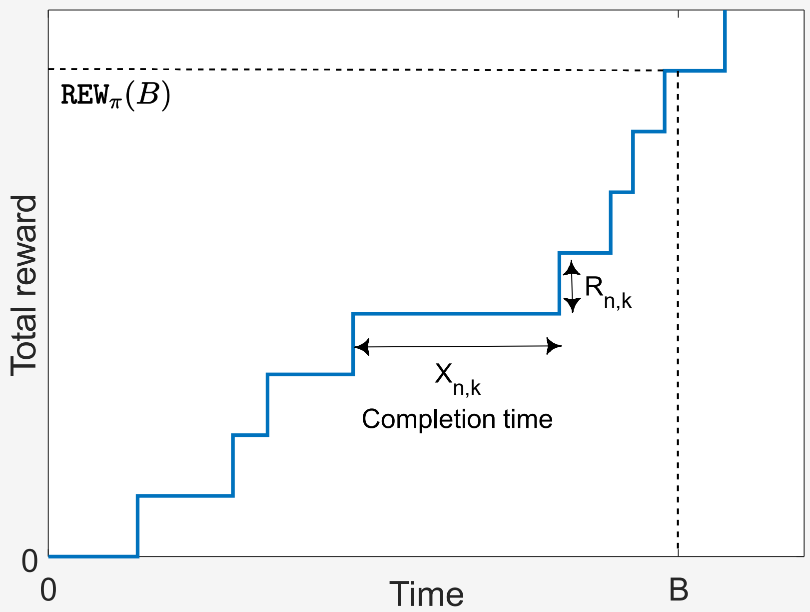
General Budget-Constrained Learning Problems

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

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



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

- **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

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)

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:

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

Independent of B
Depends on $\mathbb{E}[X_{k,1}]$
$\mathrm{Var}(X_{k,1})$

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

$\texttt{UCB-B1}$ Algorithm: Jointly Gaussian Case

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

$$\texttt{rad}_s(X, \delta) = \sqrt{\frac{2\mathrm{Var}(X)\log(\delta^{-1})}{s}}$$ 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

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

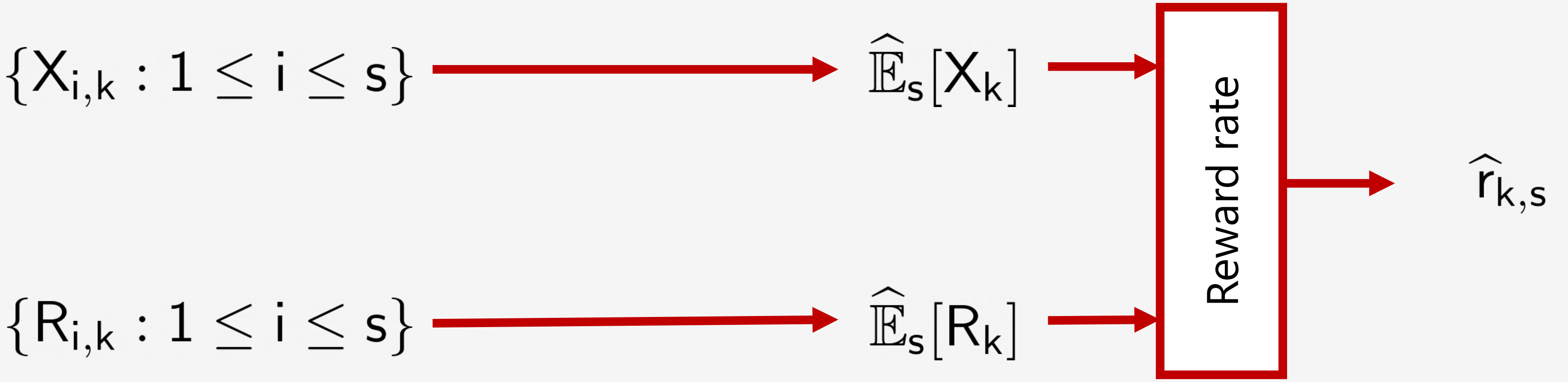
where $\Psi_{k,s}(x, y) = \dfrac{y + \widehat{r}_{k,s} \cdot x}{\left(\widehat{\mathbb{E}}_s[X_k] - x\right)_+}$

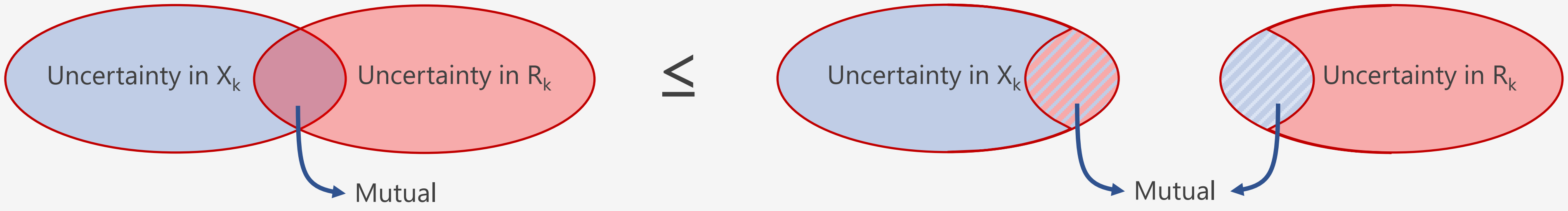
increasing in x and y

Uncertainty $\propto$ Confidence radius

`UCB-B1` Algorithm: Jointly Gaussian Case

1 Joint estimation problem: Estimate separately (ignore correlation)

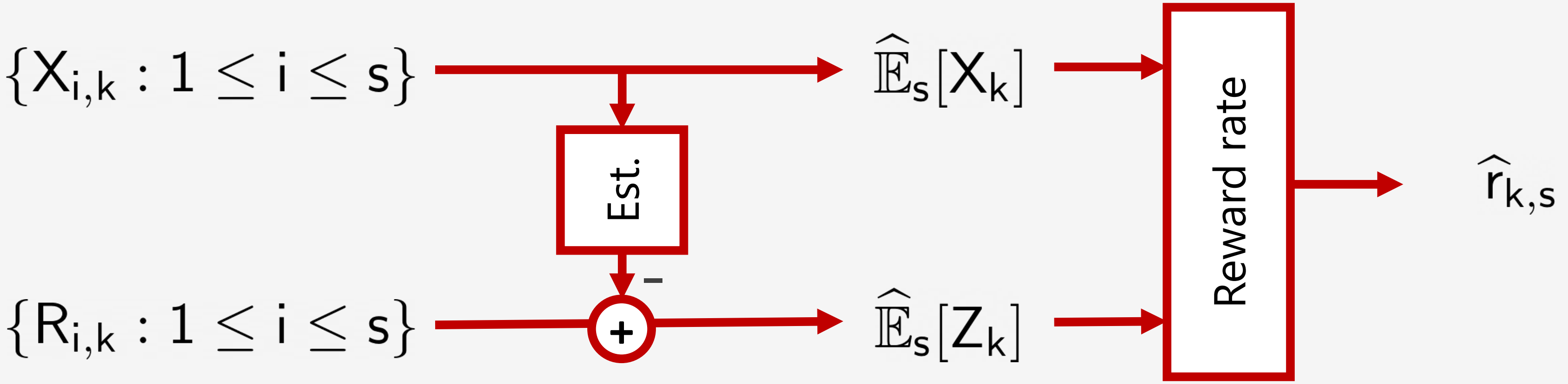


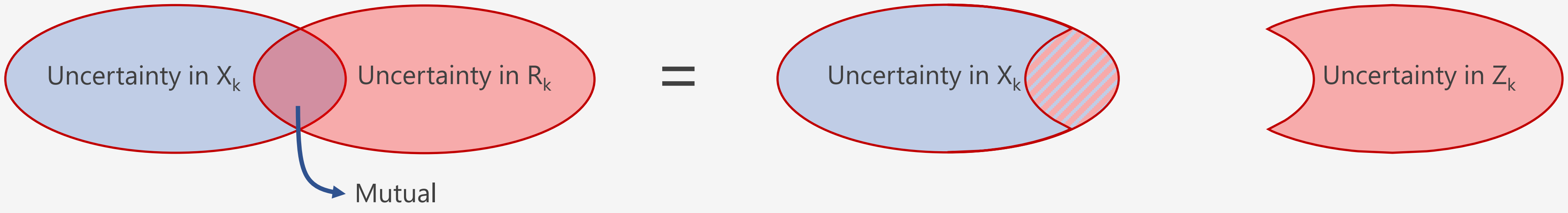


UCB-B1 Algorithm: Jointly Gaussian Case

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

Use an estimator for de-correlation

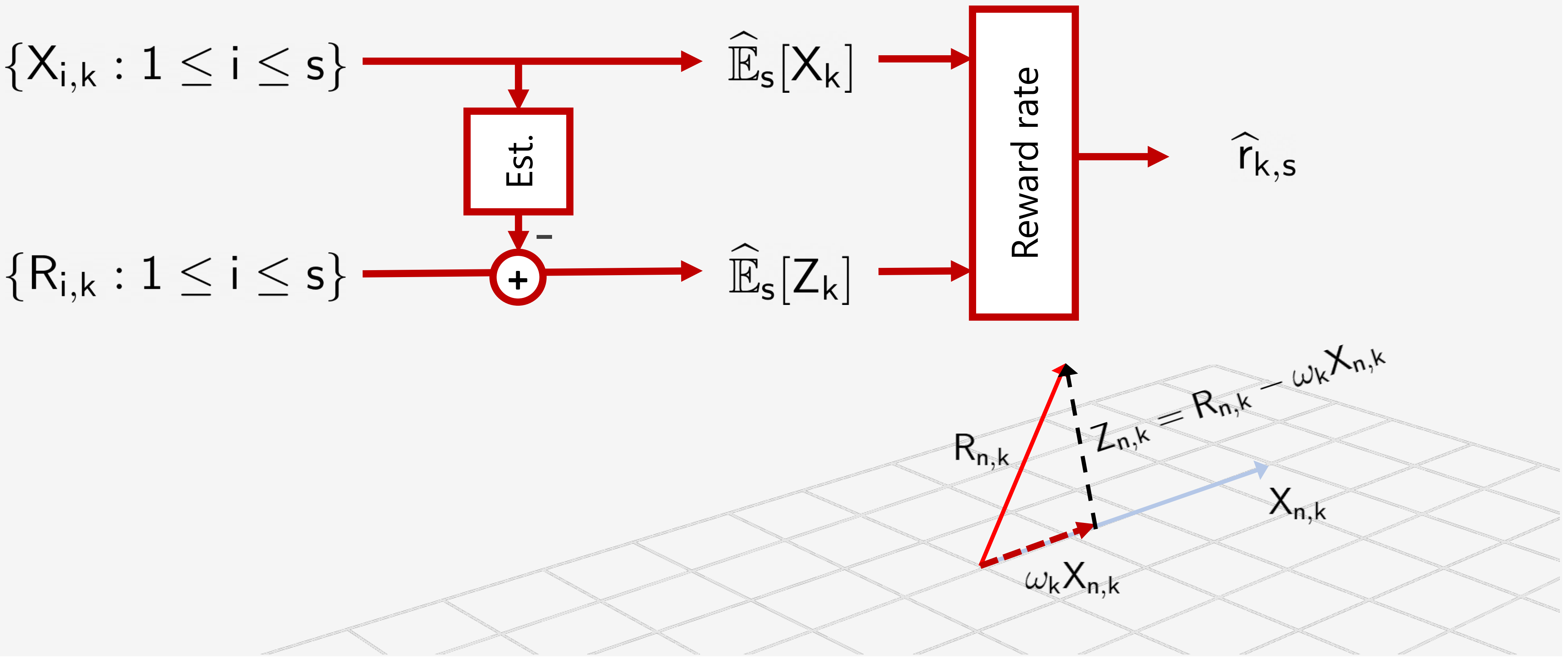




UCB-B1 Algorithm: Jointly Gaussian Case

1. Joint estimation problem: Extract correlation between cost and reward

Use an estimator for de-correlation



$\texttt{UCB-B1}$ Algorithm: Jointly Gaussian Case

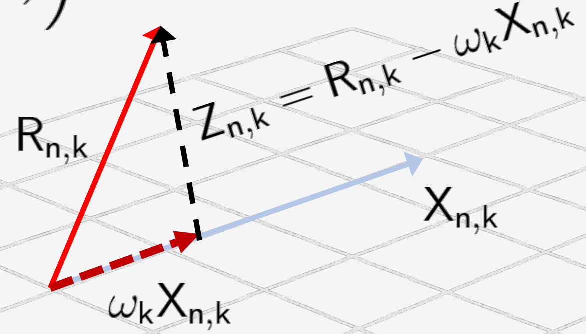
Q How to extract the correlation?

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 $\quad \texttt{rad}_s(Z_k, \delta) \leq \texttt{rad}_s(R_k, \delta)$

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



Smaller confidence radius $\Rightarrow$ Lower regret

UCB-B1 Algorithm: Jointly Gaussian Case

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

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:

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

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

- **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:

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

`UCB-B1` Algorithm: Bounded Cost and Reward

<u>Known</u> 2nd-order moments, $X_{n,k} \in [0, M_X],\ \ R_{n,k} \in [0, M_R]$

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

Bernstein inequality (tighter than Hoeffding)

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

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

- 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 <u>not</u> 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

$$\underbrace{\boxed{X_{1,k}, R_{1,k}}\ \boxed{X_{2,k}, R_{2,k}} \cdots \boxed{X_{v,k}, R_{v,k}}}_{\widehat{r}^{(1)}_{k,v}} \quad \cdots \quad \underbrace{\cdots \quad \boxed{X_{s,k}, R_{s,k}}}_{\widehat{r}^{(m)}_{k,v}}$$

$\bar{r}_{k,s} = \text{median}\{\widehat{r}^{(1)}_{k,v}, \ldots, \widehat{r}^{(m)}_{k,v}\} \to r_k$ 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

$\texttt{UCB-B2}$: Using Empirical Estimates

? What if you do not know the second-order moments?

Using empirical estimates in $\texttt{UCB-B1} \rightarrow \texttt{UCB-B2}$

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

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

Non-asymptotic analysis of using these empirical estimates: kurtosis

Small $\Delta_k \rightarrow \texttt{UCB-B1}$

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 $p > 2$

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*”, IJACI 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