Lecture 12. Stochastic Bandits

Advanced Optimization (Fall 2023)

Peng Zhao

zhaop@lamda.nju.edu.cn

Nanjing University

Outline

- Multi-Armed Bandits
  - Explore-Then-Exploit
  - Upper Confidence Bound
- Linear Bandits
  - LinUCB Algorithm
  - Generalized Linear Bandits
- Advanced Topics

Stochastic Multi-Armed Bandit (MAB)

- MAB: A player is facing K arms. At each time t, the player pulls one arm $a \in [K]$ and then receives a reward $r_t(a) \in [0,1]$:

Arm 1	$r_1(1)$	$r_2(1)$	0.6	$r_4(1)$	$r_5(1)$
Arm 2	1	$r_2(2)$	$r_3(2)$	0.2	$r_5(2)$
Arm 3	$r_1(3)$	0.7	$r_3(3)$	$r_4(3)$	0.3

- Stochastic:

  Each arm $a \in [K]$ has an unknown distribution $\mathcal{D}_a$ with mean $\mu(a)$, such that rewards $r_1(a), r_2(a), \ldots, r_T(a)$ are i.i.d samples from $\mathcal{D}_a$.

Stochastic MAB: Formulation

At each round $t = 1, 2, \cdots$

(1) the player first chooses an arm $a_t \in [K]$;

(2) and then environment reveals a reward $r_t(a_t) \in [0, 1]$;

(3) the player updates the model by the pair $(a_t, r_t(a_t))$.

- The goal is to minimize the ***(pseudo)-regret***:

$$\mathbb{E}[\mathrm{Regret}_T] = \max_{a \in [K]} \mathbb{E}\left[\sum_{t=1}^{T} r_t(a) - \sum_{t=1}^{T} r_t(a_t)\right] = T\mu(a^*) - \sum_{t=1}^{T} \mu(a_t)$$

where $a^* = \arg\max_{a \in [K]} \mu(a)$ is the best arm in the sense of expectation.

Deploying Exp3 to Stochastic MAB

- Stochastic MAB is a special case of Adversarial MAB

⇨ Directly deploying Exp3 for stochastic MAB achieves

Theorem 1. *Suppose that* $\forall t \in [T]$ *and* $i \in [K], 0 \leq \ell_t(i) \leq 1$, *then Exp3 with learning rate* $\eta = \sqrt{(\ln K)/(TK)}$ *guarantees*

$$\mathbb{E}[\text{Regret}_T] = \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(a_t)\right] - \min_{a \in [K]} \sum_{t=1}^{T} \ell_t(a) \leq \mathcal{O}\left(\sqrt{TK \ln K}\right),$$

where the expectation is taken on the randomness of the algorithm.

⇨ Not yet exploit benign *stochastic* assumption…. *instance-dependent analysis*

Regret Decomposition

- For stochastic MAB, a natural characterization of the arms:

  (i) Suboptimality gap: $\Delta_a = \mu(a^*) - \mu(a)$;

  (ii) Number of times arm a is pulled in t rounds: $n_t(a) = \sum_{\tau=1}^{t} \mathbf{1}\{a_\tau = a\}$.

- Regret can be reformulated as

$$\mathbb{E}[\text{Regret}_T] = \max_{a \in [K]} \mathbb{E}\left[\sum_{t=1}^{T} r_t(a) - \sum_{t=1}^{T} r_t(a_t)\right] = T\mu(a^*) - \sum_{t=1}^{T} \mu(a_t)$$

$$= \sum_{a \in [K]} (\mu(a^*) - \mu(a_t)) n_T(a) = \sum_{a \in [K]} \Delta_a n_T(a)$$

A Natural Solution

- **Explore-then-Exploit (ETE):**
  (1) Do explore for the first T_0 round by pulling each arm for T_0/K times;
  (2) Do exploit for the rest $T - T_0$ round by always pulling $\widehat{a} = \arg\max_{a \in [K]} \widehat{\mu}_{T_0}(a)$.

Theorem 1. *Suppose that $\forall t \in [T]$ and $a \in [K], 0 \le r_t(a) \le 1$, then ETE with exploration period T_0 guarantees*

$$\mathbb{E}[\text{Regret}_T] \le \sum_{a \in [K]} \left(\frac{T_0}{K} + 2T \exp\left(-\frac{2T_0 \Delta_a^2}{K} \right) \right) \Delta_a .$$

Proof of ETE Regret Bound

Proof. $\mathbb{E}[\text{Regret}_T] = \sum_{a \in [K]} \Delta_a n_T(a)$

Exploration Exploitation

$$n_T(a) = T_0/K + (T - T_0)\Pr\{\widehat{a} = a\}$$

$$\leq T_0/K + (T - T_0)\Pr\{\widehat{\mu}_{T_0}(a) \geq \widehat{\mu}_{T_0}(a^*)\}$$

or $\widehat{\mu}_{T_0}(a) \leq \frac{\mu(a)+\mu(a^*)}{2} \leq \widehat{\mu}_{T_0}(a^*)$

$$\leq T_0/K + (T - T_0)\Pr\left\{\widehat{\mu}_{T_0}(a) \geq \frac{\mu(a)+\mu(a^*)}{2} \cup \widehat{\mu}_{T_0}(a^*) \leq \frac{\mu(a)+\mu(a^*)}{2}\right\}$$

$$\leq T_0/K + (T - T_0)\left(\Pr\left\{\widehat{\mu}_{T_0}(a) \geq \frac{\mu(a)+\mu(a^*)}{2}\right\} + \Pr\left\{\widehat{\mu}_{T_0}(a^*) \leq \frac{\mu(a)+\mu(a^*)}{2}\right\}\right)$$

Union bound $\Pr\{X \cup Y\} \leq \Pr\{X\} + \Pr\{Y\}$

Proof of ETE Regret Bound

Proof. $n_T(a) \le T_0/K + (T - T_0)\left(\Pr\left\{\widehat{\mu}_{T_0}(a) \ge \frac{\mu(a)+\mu(a^*)}{2}\right\} + \Pr\left\{\widehat{\mu}_{T_0}(a^*) \le \frac{\mu(a)+\mu(a^*)}{2}\right\}\right)$

> **Hoeffding's inequality.** for $X_i \in [0,1], i \in [m], \bar{X} = \frac{1}{m}\sum_{i=1}^m X_i$, we have
>
> $$\Pr\left\{\bar{X} - \mathbb{E}[\bar{X}] \ge \epsilon\right\} \le \exp\left(-2m\epsilon^2\right);$$
> $$\Pr\left\{\bar{X} - \mathbb{E}[\bar{X}] \le -\epsilon\right\} \le \exp\left(-2m\epsilon^2\right).$$

$\Rightarrow \Pr\left\{\widehat{\mu}_{T_0}(a) \ge \frac{\mu(a)+\mu(a^*)}{2}\right\} = \Pr\left\{\widehat{\mu}_{T_0}(a) \ge \mu(a) + \Delta_a\right\} \le \exp\left(-\frac{2T_0\Delta_a^2}{K}\right)$ $\quad \Delta_a = \mu(a^*) - \mu(a)$

$\Rightarrow \Pr\left\{\widehat{\mu}_{T_0}(a^*) \le \frac{\mu(a)+\mu(a^*)}{2}\right\} = \Pr\left\{\widehat{\mu}_{T_0}(a^*) \le \mu(a^*) + \Delta_a\right\} \le \exp\left(-\frac{2T_0\Delta_a^2}{K}\right)$

$\Rightarrow \text{Regret}_T = \sum_{a\in[K]} \Delta_a n_T(a) \le \sum_{a\in[K]} \left(\frac{T_0}{K} + 2T\exp\left(-\frac{2T_0\Delta_a^2}{K}\right)\right)\Delta_a$ $\square$

Issue of ETE

Theorem 1. *Suppose that* $\forall t \in [T]$ *and* $a \in [K], 0 \leq r_t(a) \leq 1$, *then ETE with explore period* T_0 *guarantees*

$$\mathbb{E}[\text{Regret}_T] \leq \sum_{a \in [K]} \left(\frac{T_0}{K} + 2T \exp\left(-\frac{2T_0 \Delta_a^2}{K} \right) \right) \Delta_a.$$

- Need to tune T_0

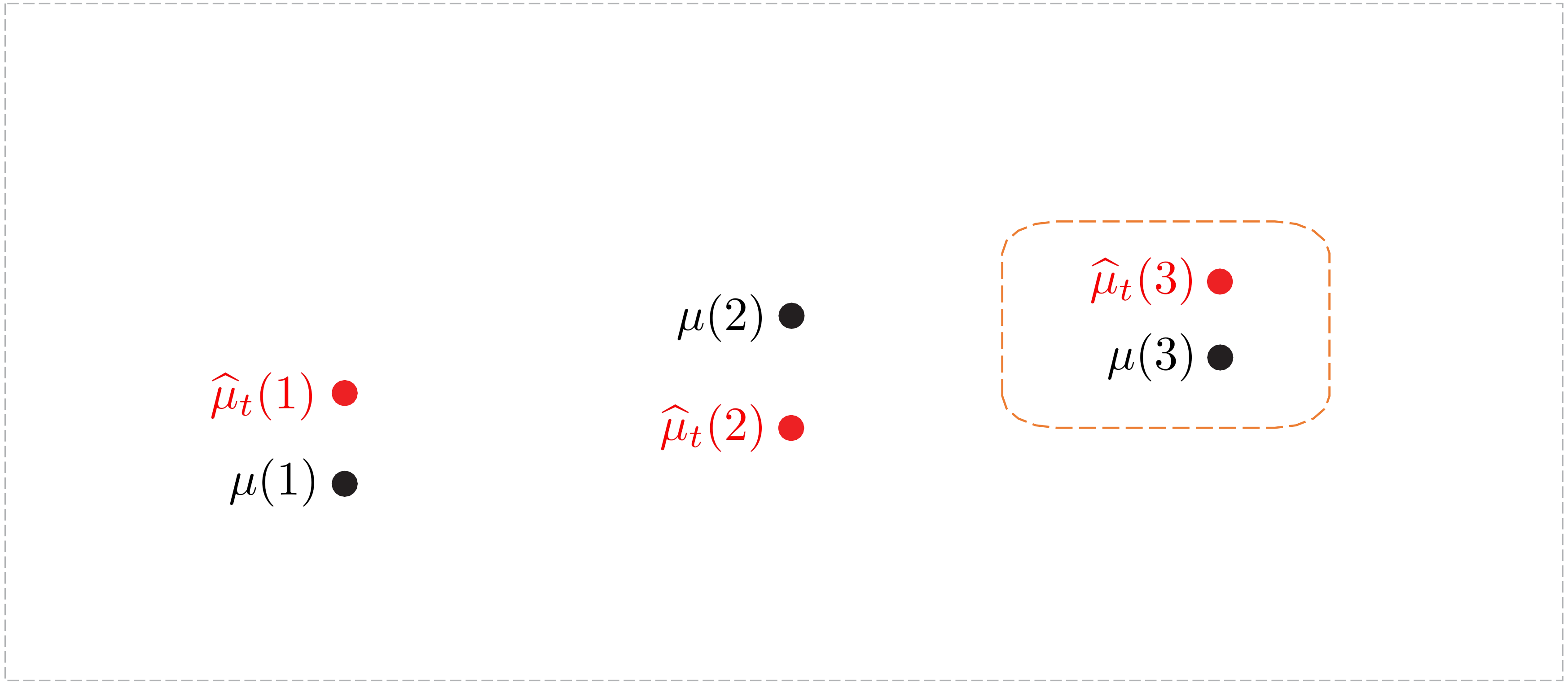
  Tune T_0 with prior of suboptimality gap Δ_a: $\mathbb{E}[\text{Regret}_T] = \widetilde{\mathcal{O}}(\sqrt{T})$

  Tune T_0 without prior of suboptimality gap Δ_a: $\mathbb{E}[\text{Regret}_T] = \widetilde{\mathcal{O}}(T^{2/3})$

$\Longrightarrow$ Solution: do explore and exploit adaptively.

Upper Confidence Bound

- ETE

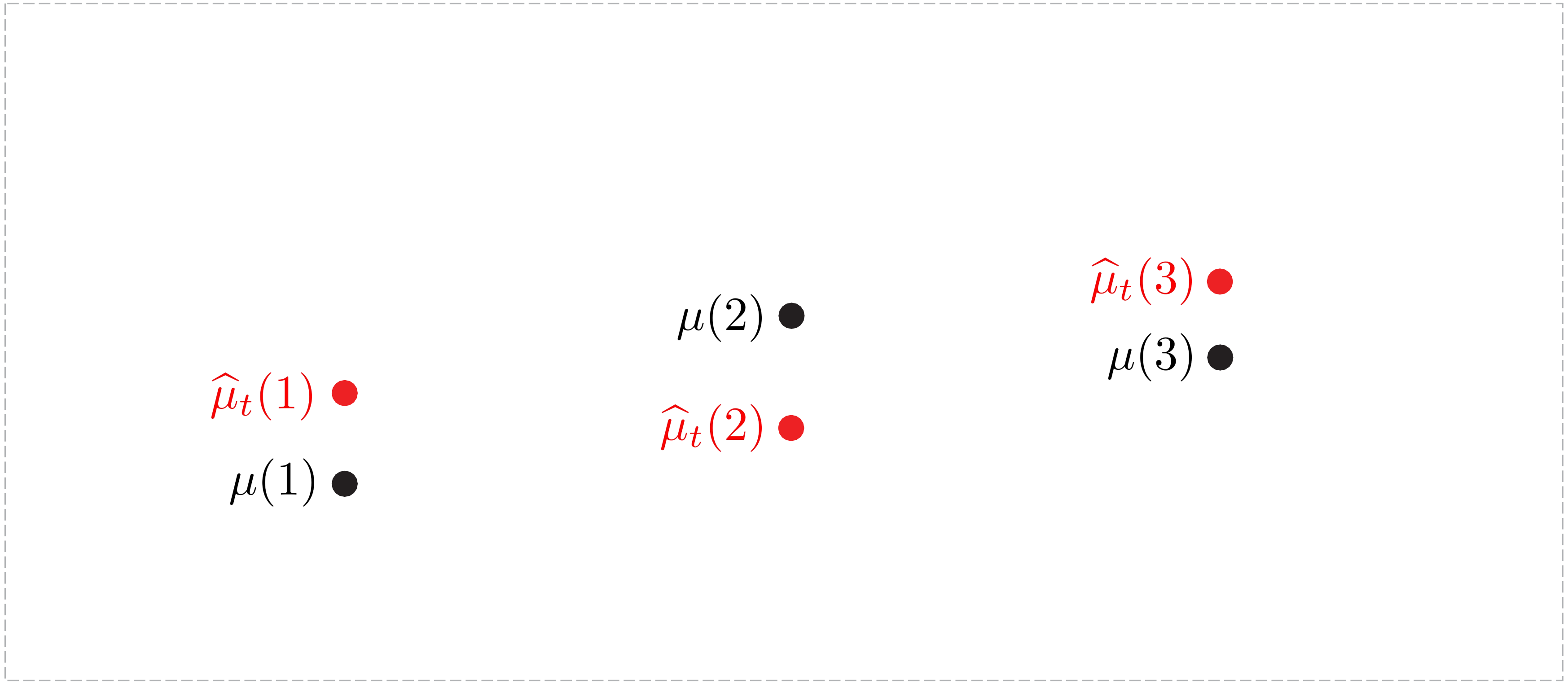


Relying on the estimate of the previous T_0 rounds.

There is no way to revise the estimate!

Upper Confidence Bound

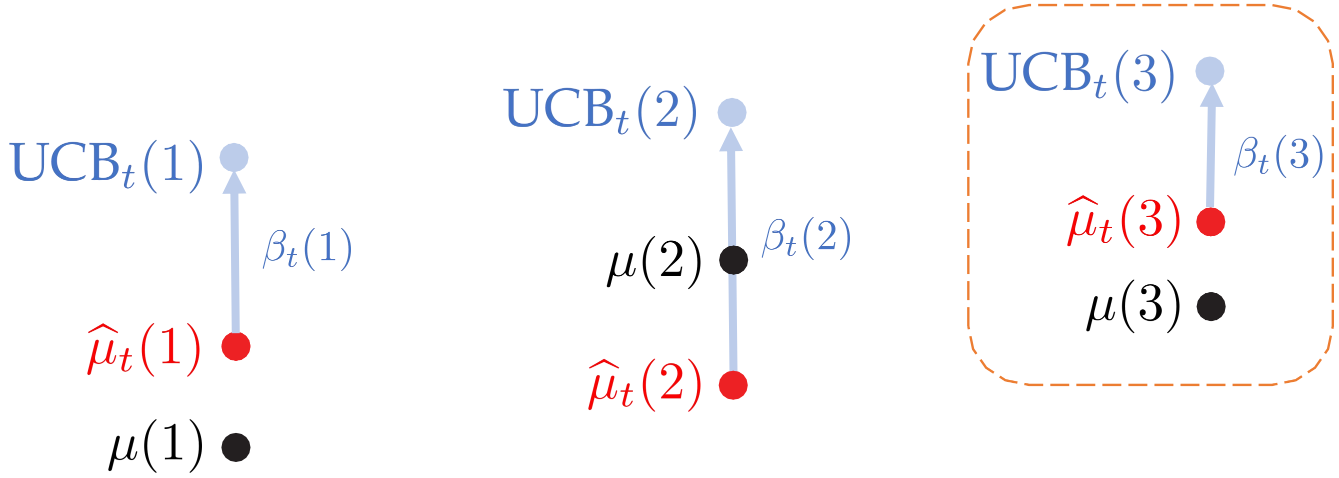
- UCB



Upper Confidence Bound

- UCB

With high probability $\mu(a) \le \mathrm{UCB}_t(a) = \widehat{\mu}_t(a) + \beta_t(a)$

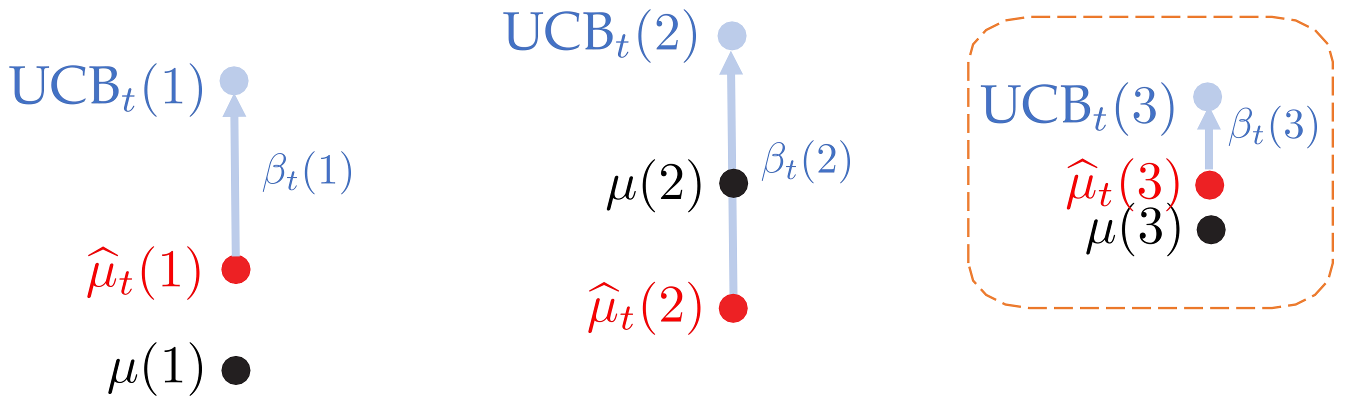


Optimism in Face of Uncertainty: $a_t = \arg\max_{a \in [K]} \mathrm{UCB}_t(a)$

Upper Confidence Bound

- UCB

With high probability $\mu(a) \leq \text{UCB}_t(a) = \widehat{\mu}_t(a) + \beta_t(a)$

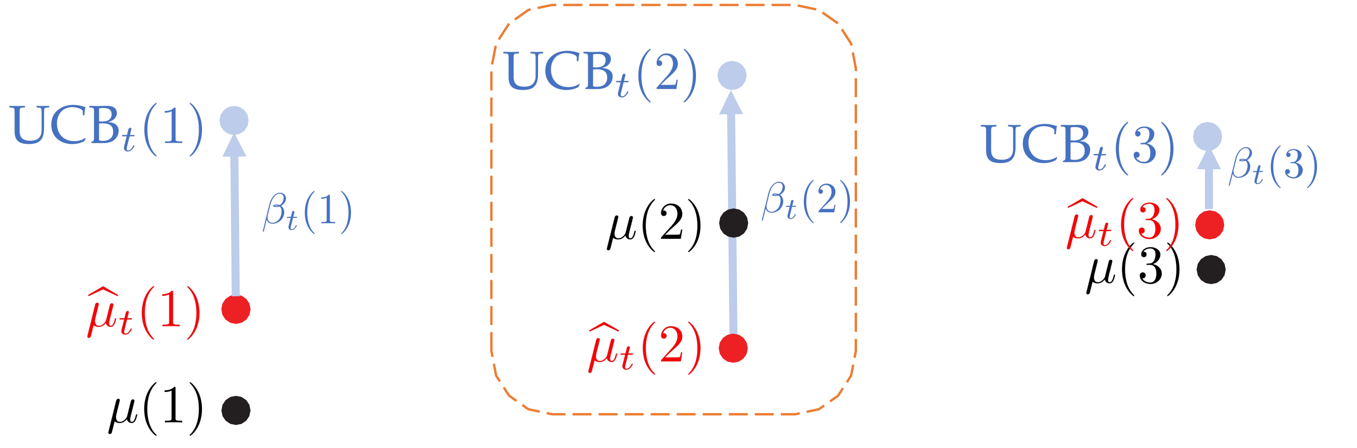


Optimism in Face of Uncertainty: $a_t = \arg\max_{a\in[K]} \text{UCB}_t(a)$

Upper Confidence Bound

- UCB

With high probability $\mu(a) \leq \mathrm{UCB}_t(a) = \widehat{\mu}_t(a) + \beta_t(a)$

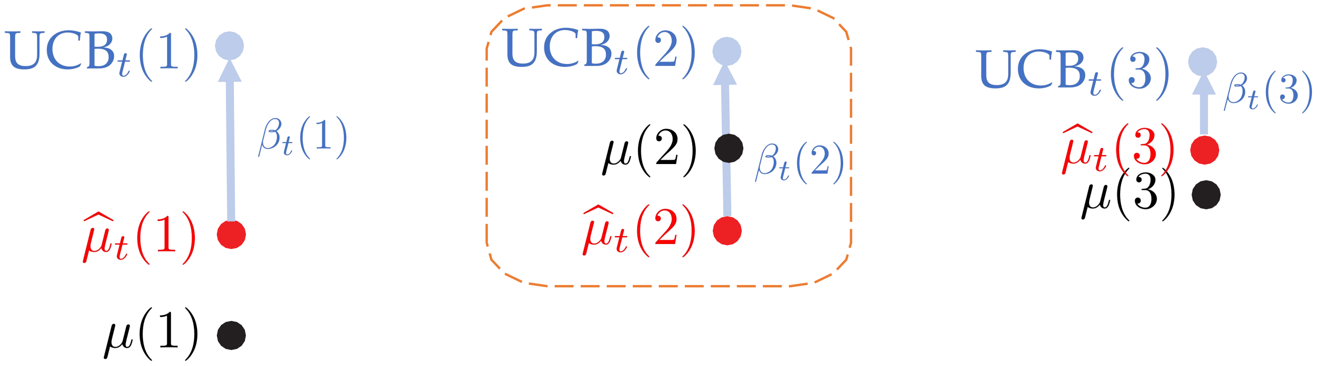


Optimism in Face of Uncertainty: $a_t = \arg\max_{a \in [K]} \mathrm{UCB}_t(a)$

Upper Confidence Bound

- UCB

With high probability $\mu(a) \leq \mathrm{UCB}_t(a) = \widehat{\mu}_t(a) + \beta_t(a)$

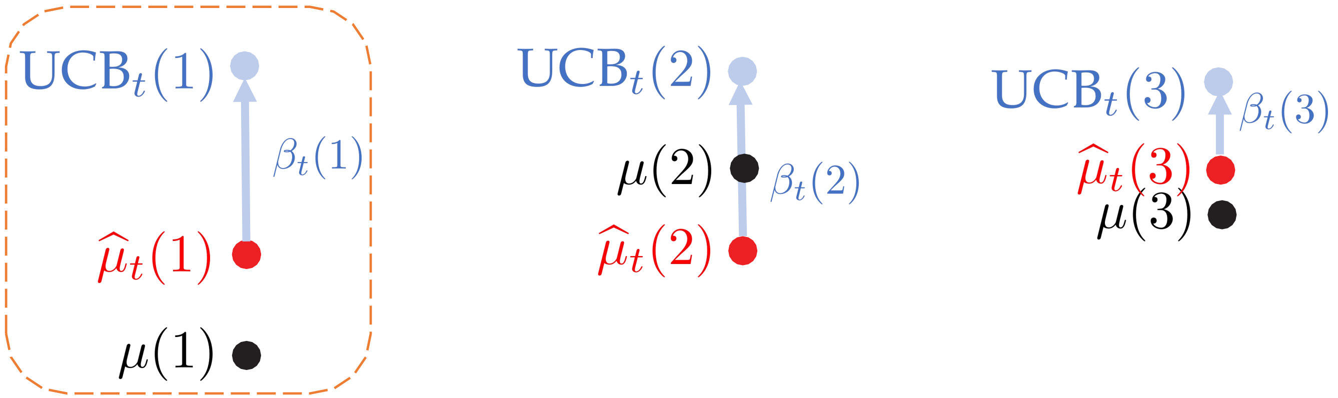


Optimism in Face of Uncertainty: $a_t = \arg\max_{a \in [K]} \mathrm{UCB}_t(a)$

Upper Confidence Bound

- UCB

With high probability $\mu(a) \leq \mathrm{UCB}_t(a) = \widehat{\mu}_t(a) + \beta_t(a)$

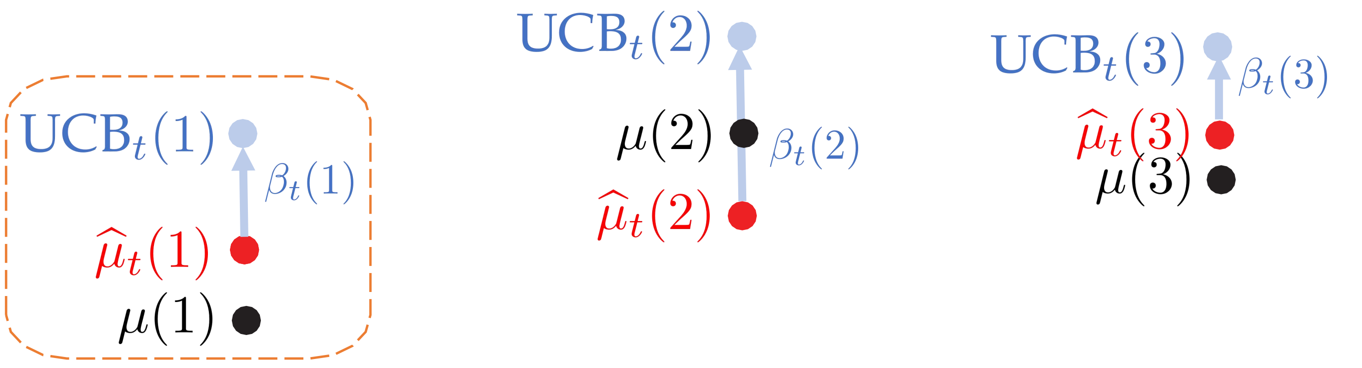


Optimism in Face of Uncertainty: $a_t = \arg\max_{a \in [K]} \mathrm{UCB}_t(a)$

Upper Confidence Bound

- UCB

With high probability $\mu(a) \leq \mathrm{UCB}_t(a) = \widehat{\mu}_t(a) + \beta_t(a)$

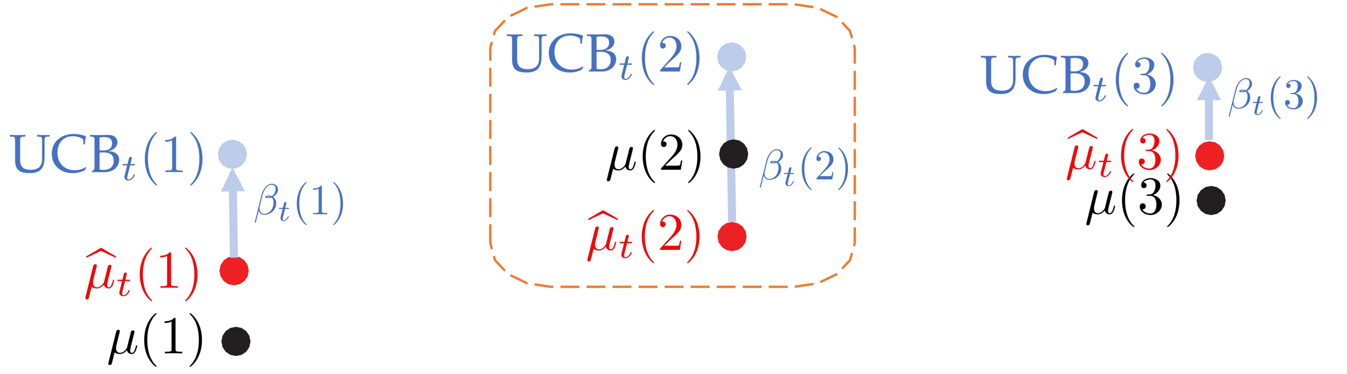


Optimism in Face of Uncertainty: $a_t = \arg\max_{a \in [K]} \mathrm{UCB}_t(a)$

Upper Confidence Bound

- UCB

With high probability $\mu(a) \leq \text{UCB}_t(a) = \widehat{\mu}_t(a) + \beta_t(a)$



Optimism in Face of Uncertainty: $a_t = \arg\max_{a \in [K]} \text{UCB}_t(a)$

A large UCB means uncertainty or good arm.

Choosing the largest UCB means either exploring or exploiting.

UCB Algorithm: Formulation

UCB Algorithm

At each round $t = 1, 2, \cdots$

(1) Choose arm $a_t = \arg\max_{a \in [K]} \mathrm{UCB}_{t-1}(a)$

(2) Observe reward r_t and update the estimation $\widehat{\mu}_t$

(3) Update upper confidence bounds UCB_t by new estimation

- Estimation: empirical average

$$\widehat{\mu}_t(a) = \frac{1}{n_t(a)} \sum_{\tau=1}^{t} \mathbf{1}\{a_\tau = a\} r_\tau(a)$$

- UCB construction: Hoeffding's inequality

Construct UCB

Lemma 1 (Estimation error). *With probability at least $1 - 2K/T$, we have,*

$$\forall a \in [K], t \in [T], |\mu(a) - \widehat{\mu}_t(a)| \leq \sqrt{\frac{\ln 1/\delta}{n_t(a)}}.$$

So we have $\mu(a) \leq UCB_t(a) \triangleq \widehat{\mu}_t(a) + \sqrt{\frac{\ln T}{n_t(a)}}$

Proof. For each arm a, by Hoeffding inequality and union bound, we have

$$\Pr\left\{|\mu(a) - \widehat{\mu}_t(a)| \leq \sqrt{\frac{\ln 1/\delta}{2n_t(a)}}\right\} \geq 1 - 2\delta$$

$$\Pr\left\{\bar{X} - \mathbb{E}[\bar{X}] \geq \epsilon\right\} \leq \exp\left(-2m\epsilon^2\right)$$
$$\Pr\left\{\bar{X} - \mathbb{E}[\bar{X}] \leq -\epsilon\right\} \leq \exp\left(-2m\epsilon^2\right)$$

Further more, by union bound again and let $\delta = 1/T^2$,

$$\Pr\left\{\forall a \in [K], t \in [T], |\mu(a) - \widehat{\mu}_t(a)| \leq \sqrt{\frac{\ln T}{n_t(a)}}\right\} \geq 1 - 2\frac{K}{T} \qquad \square$$

UCB: Distribution-Dependent Bound

Theorem 2 (Distribution-dependent). *Suppose that* $\forall t \in [T]$ *and* $a \in [K]$, $0 \le r_t(a) \le 1$, *then with probability at least* $1 - 2K/T$, *UCB satisfies*

$$\mathbb{E}[\mathrm{Regret}_T] \le \sum_{a:\Delta_a > 0} \frac{4 \ln T}{\Delta_a} + \Delta_a.$$

Proof. $\mathbb{E}[\mathrm{Regret}_T] = \sum_{a \in [K]} \Delta_a n_T(a)$

With probability at least $1 - 2K/T$

$$\begin{aligned}
\Delta_{a_t} = \mu(a^*) - \mu(a_t) &\le \mathrm{UCB}_{t-1}(a^*) - \mu(a_t) && \forall a \in [K], \mu(a) \le \mathrm{UCB}_t(a) \\
&\le \mathrm{UCB}_{t-1}(a_t) - \mu(a_t) && a_t = \arg\max_{a \in [K]} \mathrm{UCB}_{t-1}(a) \\
&\le 2\sqrt{\frac{\ln T}{n_{t-1}(a_t)}} && \mu(a) \le \mathrm{UCB}_t(a) \triangleq \widehat{\mu}_t(a) + \sqrt{\frac{\ln T}{n_t(a)}}
\end{aligned}$$

Proof of UCB Regret Bound

Proof. $\Delta_{a_t} \leq 2\sqrt{\frac{\ln T}{n_{t-1}(a_t)}}$

Let t be the last time a is selected, then with probability at least $1 - 2K/T$,

$$\Delta_a \leq 2\sqrt{\frac{\ln T}{n_{t-1}(a)}} = 2\sqrt{\frac{\ln T}{n_T(a) - 1}}$$

$$\Rightarrow \quad n_T(a) \leq 4\frac{\ln T}{\Delta_a^2} + 1$$

$$\mathbb{E}[\text{Regret}_T] = \sum_{a\in[K]} \Delta_a n_T(a) \leq \sum_{a:\Delta_a>0} \Delta_a \left(4\frac{\ln T}{\Delta_a^2} + 1\right) = \sum_{a:\Delta_a>0} 4\frac{\ln T}{\Delta_a} + \Delta_a$$

$\square$

UCB: Distribution-Dependent Bound

Theorem 2 (Distribution-dependent). *Suppose that for all $t \in [T]$ and $a \in [K]$, $0 \le r_t(a) \le 1$, then with probability at least $1 - 2K/T$, UCB satisfies*

$$\mathbb{E}[\text{Regret}_T] \le \sum_{a:\Delta_a > 0} \frac{4 \ln T}{\Delta_a} + \Delta_a.$$

- Smaller the Δ_a, larger the regret. Its harder to distinguish the optimal arm from the suboptimal one.
- However, tiny Δ_a should not lead to larger regret. Always pick arm a should just lead to $\mathbb{E}[\text{Regret}_T] = \Delta_a T$.

$$\Longrightarrow \mathbb{E}[\text{Regret}_T] \le \min\left\{ \max_{a \in [K]} \Delta_a T, \sum_{a:\Delta_a > 0} \frac{4 \ln T}{\Delta_a} + \Delta_a \right\}$$

UCB: Distribution-Free Bound

Theorem 3 (Distribution-free). *Suppose that for all $t \in [T]$ and $a \in [K]$, $0 \le r_t(a) \le 1$, then UCB satisfies*

$$\mathbb{E}[\mathrm{Regret}_T] \le 2\sqrt{TK \ln T} + \sum_{a \in [K]} \Delta_a = \mathcal{O}\left(\sqrt{TK \log T}\right)$$

Proof.

$$\mathbb{E}[\mathrm{Regret}_T] = \sum_{a \in [K]} \Delta_a n_T(a) = \sum_{a : \Delta_a < \Delta} \Delta_a n_T(a) + \sum_{a : \Delta_a \ge \Delta} \Delta_a n_T(a)$$

$n_T(a) \le 4\frac{\ln T}{\Delta_a^2} + 1$

$$\le T\Delta + \sum_{a : \Delta_a \ge \Delta} \Delta_a \left(4\frac{\ln T}{\Delta_a^2} + 1\right) \le T\Delta + 4\frac{K \ln T}{\Delta} + \sum_{a \in [K]} \Delta_a$$

$$\le 2\sqrt{TK \ln T} + \sum_{a \in [K]} \Delta_a$$

Choosing $\Delta = 2\sqrt{K \ln T / T}$ $\square$

Upper Bound and Lower Bound

Theorem 3 (Distribution-free). *Suppose that for all $t \in [T]$ and $a \in [K]$, $0 \le r_t(a) \le 1$, then UCB satisfies*

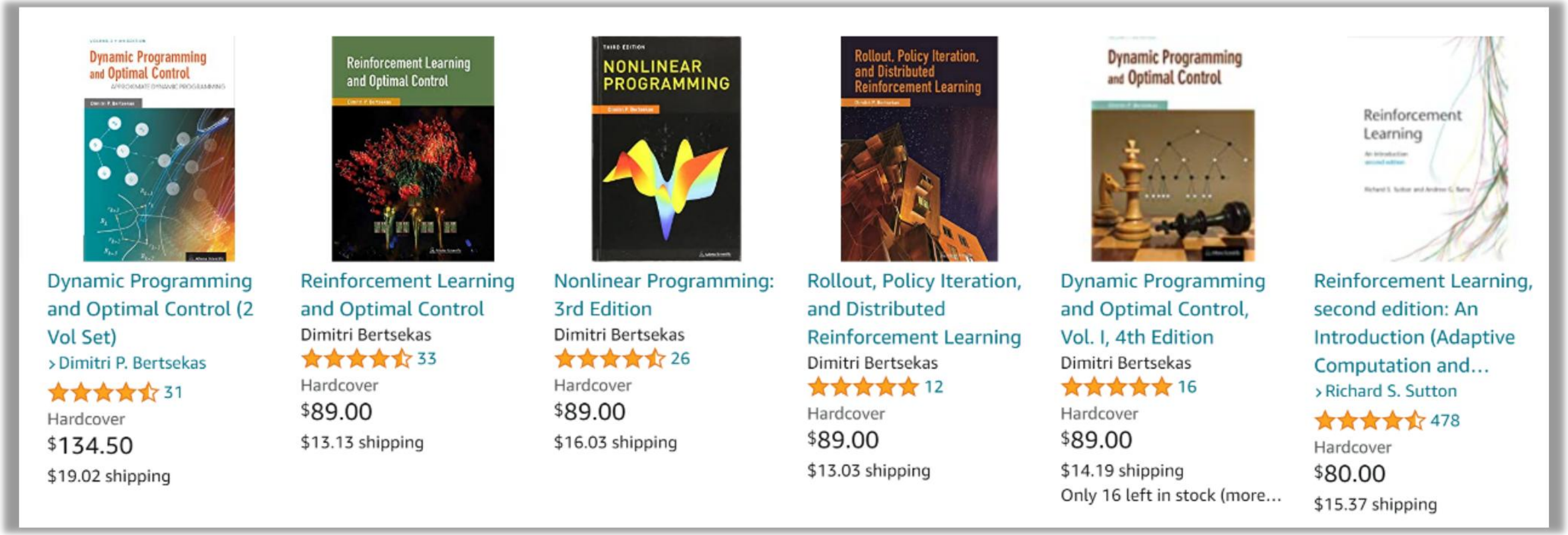
$$\mathbb{E}[\text{Regret}_T] \le 2\sqrt{TK \ln T} + \sum_{a \in [K]} \Delta_a = \mathcal{O}\left(\sqrt{TK \log T}\right)$$

Theorem 4 (Lower Bound for MAB). *For any bandit algorithm $\mathcal{A}$, there exists a sequence of loss vectors such that*

$$\inf_{\mathcal{A}} \sup_{\boldsymbol{\ell}_1, \ldots, \boldsymbol{\ell}_T} \mathbb{E}\left[\text{Regret}_T\right] = \Omega(\sqrt{TK})$$

Stochastic Linear Bandits

- A ubiquitous problem in real life:



  - Each arm represent a book and has side information;
  - Arm set could be very large or even infinite.

Stochastic LB: Formulation

Each arm is represented as a feature vector $\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^d$

At each round $t = 1, 2, \cdots$

(1) the player first chooses an arm X_t from arm set $\mathcal{X}$;

(2) and then environment reveals a reward $r_t \in \mathbb{R}$.

	Multi-Armed Bandits	**Linear Bandits**
Arm set	finite arm set $[K]$	infinite arm set $\mathcal{X} = \{\|\mathbf{x}\|_2 \leq L\}$
Model	$\mathbb{E}[r(a)] = \mu(a)$ $\forall t \in [T], a \in [K], r_t(a) \in [0, 1]$	$r_t = X_t^\top \theta_* + \eta_t$ $\quad\quad \mu(\mathbf{x}) = \mathbf{x}^\top \theta_*$ η_t: sub-Gaussian noise
Regret	$\mathbb{E}[\text{Regret}_T] = T \max_{a \in [K]} \mu(a) - \sum_{t=1}^{T} \mu(a_t)$	$\mathbb{E}[\text{Regret}_T] = T \max_{\mathbf{x} \in \mathcal{X}} \mathbf{x}^\top \theta_* - \sum_{t=1}^{T} X_t^\top \theta_*$

Deploying UCB to Linear Bandits

- Linear Bandits is a special case of MAB with infinite arm:

⇨ Why not directly deploy UCB to address Linear Bandits?

Theorem 3 (Distribution-free). *Suppose that* $\forall t \in [T]$ *and* $a \in [K]$, $0 \leq r_t(a) \leq 1$, *then UCB satisfies*

$$\mathbb{E}[\mathrm{Regret}_T] \leq 2\sqrt{TK \ln T} + \sum_{a \in [K]} \Delta_a = \mathcal{O}\left(\sqrt{TK \ln T}\right)$$

Infinite arm set ($K \to \infty$) leads to meaningless regret guarantee!

⇨ Not yet exploit the addition *contextual feature information*…

LinUCB Algorithm: Formulation

LinUCB Algorithm

At each round $t = 1, 2, \cdots$

(1) Select $X_t = \arg\max_{\mathbf{x}\in\mathcal{X}} \mathrm{UCB}_{t-1}(\mathbf{x})$

(2) Observe reward r_t and update the estimation $\widehat{\mu}_t$

(3) update upper confidence bounds UCB_t by new estimation

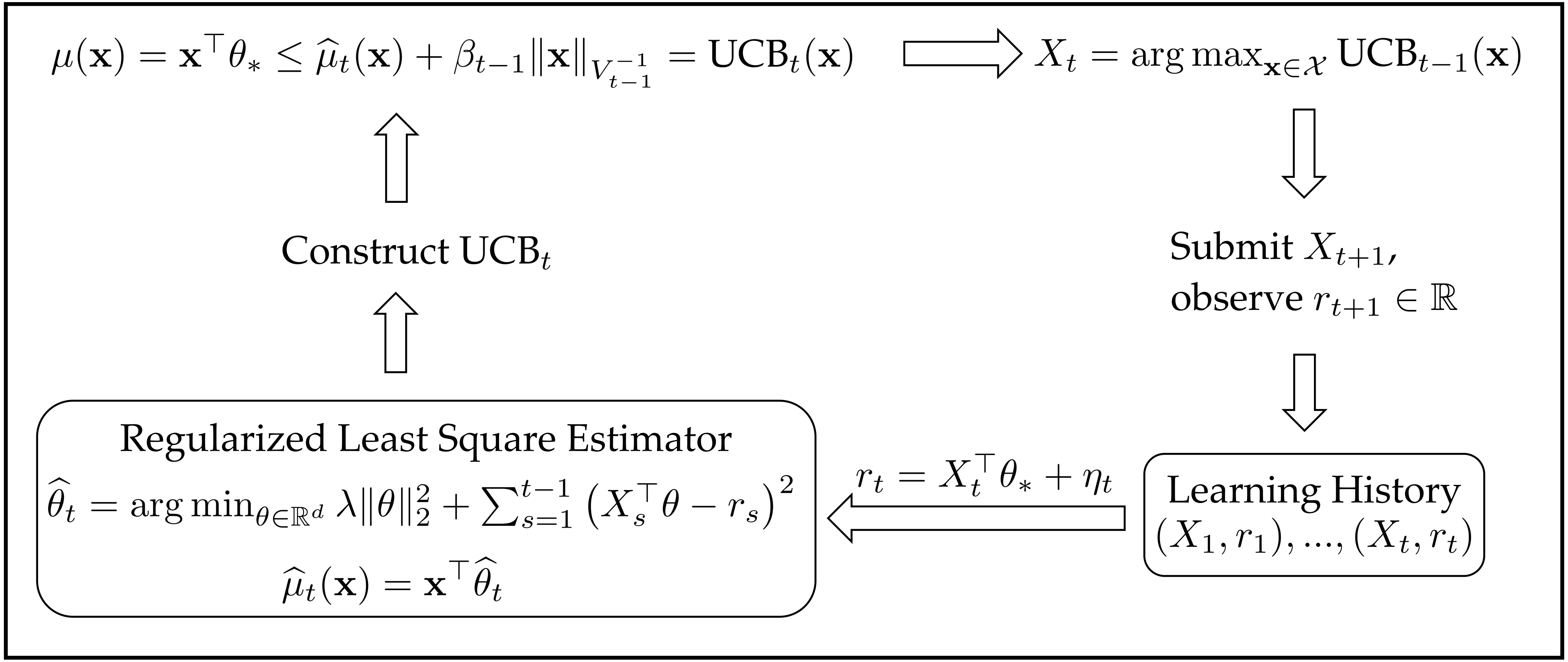
- Estimation: regularized least square (linear regression)

$$\widehat{\theta}_t = \arg\min_{\theta\in\mathbb{R}^d} \lambda\|\theta\|_2^2 + \sum_{s=1}^{t-1}\left(X_s^\top\theta - r_s\right)^2$$

Closed form: $\widehat{\theta}_t = V_{t-1}^{-1}\left(\sum_{s=1}^{t-1} r_s X_s\right), V_{t-1} = \lambda I + \sum_{s=1}^{t-1} X_s X_s^\top$

LinUCB Algorithm

Optimism in Face of Uncertainty



LinUCB Algorithm

- UCB for stochastic MAB

  (1) estimate $\mu(a)$ by average estimation;

  (2) construct upper confidence bound for $\mu(a)$ by concentration inequalities.

- UCB for stochastic LB (LinUCB)
  - More information can be used to estimate expected reward.

UCB estimation

$$\widehat{\mu}_t(a) = \frac{1}{n_t(a)} \sum_{\tau=1}^{t} \mathbf{1}\{a_\tau = a\} r_\tau(a)$$

LinUCB estimation

$$\widehat{\theta}_t = \arg\min_{\theta \in \mathbb{R}^d} \lambda \|\theta\|_2^2 + \sum_{s=1}^{t-1} \left(X_s^\top \theta - r_s\right)^2$$

$$\widehat{\mu}_t(\mathbf{x}) = \mathbf{x}^\top \widehat{\theta}_t$$

Construct UCB

Lemma 2 (Estimation error). *For any* $\mathbf{x} \in \mathcal{X}, \delta \in (0,1)$*, with probability at least* $1-\delta$*, the following holds for all* $t \in [T]$

$$\left|\mathbf{x}^\top(\widehat{\theta}_t - \theta)\right| \leq \beta_{t-1}\|\mathbf{x}\|_{V_{t-1}^{-1}}, \quad \textit{where } \beta_{t-1} = R\sqrt{2\log\left(\frac{1}{\delta}\right) + d\log\left(1 + \frac{(t-1)L^2}{\lambda d}\right)} + \sqrt{\lambda}S$$

So we have $\mu(\mathbf{x}) \leq UCB_t(\mathbf{x}) \triangleq \widehat{\mu}_t(\mathbf{x}) + \beta_{t-1}\|\mathbf{x}\|_{V_{t-1}^{-1}}$

Proof.

$$\begin{aligned}
\widehat{\theta}_t - \theta_* &= V_{t-1}^{-1}\left(\sum_{s=1}^{t-1} r_s X_s\right) - \theta_* \\
&= V_{t-1}^{-1}\left(\sum_{s=1}^{t-1}\left(X_s^\top\theta_* + \eta_s\right)X_s\right) - V_{t-1}^{-1}\left(\lambda I_d + \sum_{s=1}^{t-1} X_s X_s^\top\right)\theta_* \\
&= V_{t-1}^{-1}\left(\sum_{s=1}^{t-1}\eta_s X_s - \lambda\theta_*\right)
\end{aligned}$$

$\widehat{\theta}_t = V_{t-1}^{-1}\left(\sum_{s=1}^{t-1} r_s X_s\right)$

$V_{t-1} = \lambda I + \sum_{s=1}^{t-1} X_s X_s^\top$

Proof of Estimation Error Bound

Proof. $\widehat{\theta}_t - \theta_* = V_{t-1}^{-1}\left(\sum_{s=1}^{t-1} \eta_s X_s - \lambda\theta_*\right)$ $\quad V_{t-1} = \lambda I + \sum_{s=1}^{t-1} X_s X_s^\top$

$$\left|\mathbf{x}^\top\left(\widehat{\theta}_t - \theta_*\right)\right| \leq \|\mathbf{x}\|_{V_{t-1}^{-1}} \left\|\widehat{\theta}_t - \theta_*\right\|_{V_{t-1}}$$

Cauchy-Schwarz inequality: $|a^\top b| \leq \|a\|\|b\|_*$

$$\leq \|\mathbf{x}\|_{V_{t-1}^{-1}} \left(\left\|\sum_{s=1}^{t-1} \eta_s X_s\right\|_{V_{t-1}^{-1}} + \|\lambda\theta_*\|_{V_{t-1}^{-1}}\right)$$

Core difficulty: The actions $\{X_s\}_{s=1,\ldots,t}$ are neither fixed nor independent but are intricately correlated via the rewards $\{r_s\}_{s=1,\ldots,t}$

Self-Normalized Concentration

Theorem 4 (Self-normalized concentration for Vector-Valued Martingales). *Let $\{F_t\}_{t=0}^{\infty}$ be a filtration. Let $\{\eta_t\}_{t=0}^{\infty}$ be a real-valued stochastic process such that η_t is F_t-measurable and η_t is conditionally R-sub-Gaussian for some $R \geq 0$ i.e.*

$$\forall \lambda \in \mathbb{R}, \mathbb{E}\left[\exp(\lambda \eta_t) | X_{1:t}, \eta_{1:t-1}\right] \leq \exp\left(\frac{\lambda^2 R^2}{2}\right).$$

Let $\{X_t\}_{t=1}^{\infty}$ be an $\mathbb{R}^d$-valued stochastic process such that X_t is F_{t-1}-measurable. Assume that V is a $d \times d$ positive definite matrix. For any $t \geq 0$, define

$$V_t = V_0 + \sum_{s=1}^{t} X_s X_s^\top, \qquad S_t = \sum_{s=1}^{t} \eta_s X_s.$$

Then, for any $\delta > 0$, with probability at least $1-\delta$, for all $t \geq 0$,

$$\|S_t\|_{V_t^{-1}}^2 \leq 2R^2 \log\left(\frac{\det(V_t)^{\frac{1}{2}} \det(V_0)^{-\frac{1}{2}}}{\delta}\right).$$

Proof of Estimation Error Bound

Proof. $$\left|\mathbf{x}^\top\left(\widehat{\theta}_t-\theta_*\right)\right| \le \|\mathbf{x}\|_{V_{t-1}^{-1}}\left(\left\|\sum_{s=1}^{t-1}\eta_s X_s\right\|_{V_{t-1}^{-1}}+\|\lambda\theta_*\|_{V_{t-1}^{-1}}\right)$$

Theorem 4 (Self-normalized concentration). *For any $\delta\in(0,1)$, with probability at least $1-\delta$, for all $t\ge 0$,*

$$\|S_t\|^2_{V_t^{-1}} \le 2R^2\log\left(\frac{\det(V_t)^{\frac{1}{2}}\det(V_0)^{-\frac{1}{2}}}{\delta}\right).$$

$$\operatorname{Tr}(V_t)=\operatorname{Tr}(\lambda I)+\operatorname{Tr}\left(\sum_{s=1}^{t}X_sX_s^\top\right)\le\lambda d+tL^2 \qquad V_t=\lambda I+\textstyle\sum_{s=1}^{t}X_sX_s^\top$$

$$\det(V_t)=\prod_{i=1}^{d}\lambda_i\le\left(\frac{\sum_{i=1}^{d}\lambda_i}{d}\right)^d=\left(\frac{\operatorname{Tr}(V_t)}{d}\right)^d\le\left(\frac{\lambda d+tL^2}{d}\right)^d$$

$$\det(V_0)=\det(\lambda I)=\lambda^d \qquad V_0=\lambda I$$

Proof of Estimation Error Bound

Proof. $\left|\mathbf{x}^\top \left(\widehat{\theta}_t - \theta_*\right)\right| \leq \|\mathbf{x}\|_{V_{t-1}^{-1}} \left(\left\|\sum_{s=1}^{t-1} \eta_s X_s\right\|_{V_{t-1}^{-1}} + \|\lambda\theta_*\|_{V_{t-1}^{-1}}\right)$

$$\left\|\sum_{s=1}^{t-1} \eta_s X_s\right\|_{V_{t-1}^{-1}} \leq \sqrt{2R^2 \log\left(\frac{\det(V_t)^{\frac{1}{2}} \det(V_0)^{-\frac{1}{2}}}{\delta}\right)} \leq \sqrt{2R^2 \log\left(\frac{1}{\delta}\left(\frac{\lambda d + (t-1)L^2}{\lambda d}\right)^{\frac{d}{2}}\right)}$$

$$= R\sqrt{2\log\left(\frac{1}{\delta}\right) + d\log\left(1 + \frac{tL^2}{\lambda d}\right)}$$

$$\det(V_t) \leq \left(\frac{\lambda d + tL^2}{d}\right)^d$$

$$\det(V_0) = \lambda^d$$

$$\|\lambda\theta_*\|_{V_{t-1}^{-1}} \leq \frac{1}{\sqrt{\lambda_{\min}(V_{t-1})}} \|\lambda\theta_*\|_2 \leq \frac{1}{\sqrt{\lambda}} \|\lambda\theta_*\|_2 \leq \sqrt{\lambda} S$$

$$\left|\mathbf{x}^\top \left(\widehat{\theta}_t - \theta_*\right)\right| \leq \|\mathbf{x}\|_{V_{t-1}^{-1}} \left(R\sqrt{2\log\left(\frac{1}{\delta}\right) + d\log\left(1 + \frac{tL^2}{\lambda d}\right)} + \sqrt{\lambda} S\right)$$

□

LinUCB: Regret Bound

Theorem 5. *Let $\lambda = d$, the regret of LinUCB is bounded with probability at least $1 - 1/T$, by*

$$\mathbb{E}[\text{Regret}_T] \leq \widetilde{\mathcal{O}}\left(d\sqrt{T}\right)$$

Proof. Let $X_* \triangleq \arg\max_{\mathbf{x} \in \mathcal{X}} \mathbf{x}^\top \theta_*$, each of the following holds with probability at least $1 - \delta$,

$$\forall t \in [T], X_*^\top \theta_* \leq X_*^\top \widehat{\theta}_t + \beta_{t-1} \|X_*\|_{V_{t-1}^{-1}}$$

$$\forall t \in [T], X_t^\top \theta_* \geq X_t^\top \widehat{\theta}_t - \beta_{t-1} \|X_t\|_{V_{t-1}^{-1}}$$

With probability at least $1 - 2\delta$,

$$\begin{aligned} \forall t \in [T], X_*^\top \theta_* - X_t^\top \theta_* &\leq X_*^\top \widehat{\theta}_t - X_t^\top \widehat{\theta}_t + \beta_{t-1} \left(\|X_*\|_{V_{t-1}^{-1}} + \|X_t\|_{V_{t-1}^{-1}}\right) \\ &\leq 2\beta_{t-1} \|X_t\|_{V_{t-1}^{-1}}, \qquad X_*^\top \widehat{\theta}_t + \beta_{t-1} \|X_*\|_{V_{t-1}^{-1}} \leq X_t^\top \widehat{\theta}_t + \beta_{t-1} \|X_t\|_{V_{t-1}^{-1}} \end{aligned}$$

LinUCB: Regret Bound

Proof. With probability at least $1-2\delta$, $\forall t \in [T], X_*^\top \theta_* - X_t^\top \theta_* \le 2\beta_{t-1} \|X_t\|_{V_{t-1}^{-1}}$.

$$\mathbb{E}[\text{Regret}_T] = \sum_{t=1}^{T} \left(X_*^\top \theta_* - X_t^\top \theta_* \right) \le 2\beta_T \sum_{t=1}^{T} \|X_t\|_{V_{t-1}^{-1}} \le 2\beta_T \sqrt{T \sum_{t=1}^{T} \|X_t\|_{V_{t-1}^{-1}}^2}$$

Lemma 4 (Elliptical Potential Lemma). *For any sequence* $\{X_1, \ldots, X_T\} \in \mathbb{R}^{d\times T}$*, suppose* $V_0 = \lambda I, V_t = V_{t-1} + X_t X_t^\top$*, and* $\|X_t\|_2 \le L$*, then*

$$\sum_{t=1}^{T} \|X_t\|_{V_t^{-1}}^2 \le d \log\left(1 + \frac{L^2 T}{\lambda d}\right)$$

proofed in Lecture 6

$$\mathbb{E}[\text{Regret}_T] \le 2\beta_T \sqrt{T \sum_{t=1}^{T} \|X_t\|_{V_{t-1}^{-1}}^2} \le 2\beta_T \sqrt{Td \log\left(1 + \frac{L^2 T}{\lambda d}\right)}$$

LinUCB: Regret Bound

Proof. With probability at least $1 - 2\delta$, $\mathbb{E}[\mathrm{Regret}_T] \leq 2\beta_T \sqrt{Td \log\left(1 + \frac{L^2 T}{\lambda d}\right)}$

$$\mathbb{E}[\mathrm{Regret}_T] \leq 2\beta_T \sqrt{Td \log\left(1 + \frac{L^2 T}{\lambda d}\right)} \quad \beta_t = R\sqrt{2\log\left(\frac{1}{\delta}\right) + d\log\left(1 + \frac{tL^2}{\lambda d}\right)} + \sqrt{\lambda} S$$

$$\leq 2\left(R\sqrt{2\log\left(\frac{1}{\delta}\right) + d\log\left(1 + \frac{TL^2}{\lambda d}\right)} + \sqrt{\lambda} S\right)\sqrt{Td \log\left(1 + \frac{L^2 T}{\lambda d}\right)}$$

Let $\delta = 1/2T$, then with probability at least $1 - 1/T$,

$$\mathbb{E}[\mathrm{Regret}_T] \leq 2\left(R\sqrt{2\log\left(\frac{T}{2}\right) + d\log\left(1 + \frac{TL^2}{\lambda d}\right)} + \sqrt{\lambda} S\right)\sqrt{Td \log\left(1 + \frac{L^2 T}{\lambda d}\right)}$$

$$= \widetilde{\mathcal{O}}(d\sqrt{T}) \qquad \square$$

Improved Algorithms for Linear Stochastic Bandits

Yasin Abbasi-Yadkori
abbasiya@ualberta.ca
Dept. of Computing Science
University of Alberta

Dávid Pál
dpal@google.com
Dept. of Computing Science
University of Alberta

Csaba Szepesvári
szepesva@ualberta.ca
Dept. of Computing Science
University of Alberta

Abstract

We improve the theoretical analysis and empirical performance of algorithms for the stochastic multi-armed bandit problem and the linear stochastic multi-armed bandit problem. In particular, we show that a simple modification of Auer's UCB algorithm (Auer, 2002) achieves with high probability constant regret. More importantly, we modify and, consequently, improve the analysis of the algorithm for the for linear stochastic bandit problem studied by Auer (2002), Dani et al. (2008), Rusmevichientong and Tsitsiklis (2010), Li et al. (2010). Our modification improves the regret bound by a logarithmic factor, though experiments show a vast improvement. In both cases, the improvement stems from the construction of smaller confidence sets. For their construction we use a novel tail inequality for vector-valued martingales.

1 Introduction

Linear stochastic bandit problem is a sequential decision-making problem where in each time step we have to choose an action, and as a response we receive a stochastic reward, expected value of which is an unknown linear function of the action. The goal is to collect as much reward as possible over the course of n time steps. The precise model is described in Section 1.2.

Several variants and special cases of the problem exist differing on what the set of available actions is in each round. For example, the standard stochastic d-armed bandit problem, introduced by Robbins (1952) and then studied by Lai and Robbins (1985), is a special case of linear stochastic bandit problem where the set of available actions in each round is the standard orthonormal basis of $\mathbb{R}^d$. Another variant, studied by Auer (2002) under the name "linear reinforcement learning", and later in the context of web advertisement by Li et al. (2010), Chu et al. (2011), is a variant when the set of available actions changes from time step to time step, but has the same finite cardinality in each step. Another variant dubbed "sleeping bandits", studied by Kleinberg et al. (2008), is the case when the set of available actions changes from time step to time step, but it is always a subset of the standard orthonormal basis of $\mathbb{R}^d$. Another variant, studied by Dani et al. (2008), Abbasi-Yadkori et al. (2009), Rusmevichientong and Tsitsiklis (2010), is the case when the set of available actions does not change between time steps but the set can be an almost arbitrary, even infinite, bounded subset of a finite-dimensional vector space. Related problems were also studied by Abe et al. (2003), Walsh et al. (2009), Dekel et al. (2010).

In all these works, the algorithms are based on the same underlying idea—the *optimism-in-the-face-of-uncertainty* (OFU) principle. This is not surprising since they are solving almost the same problem. The OFU principle elegantly solves the exploration-exploitation dilemma inherent in the problem. The basic idea of the principle is to maintain a confidence set for the vector of coefficients of the linear function. In every round, the algorithm chooses an estimate from the confidence set and an action so that the predicted reward is maximized, i.e., estimate-action pair is chosen optimistically. We give details of the algorithm in Section 2.

1

Improved algorithms for linear stochastic bandits

Authors: Yasin Abbasi-Yadkori, Csaba Szepesvári, David Pal

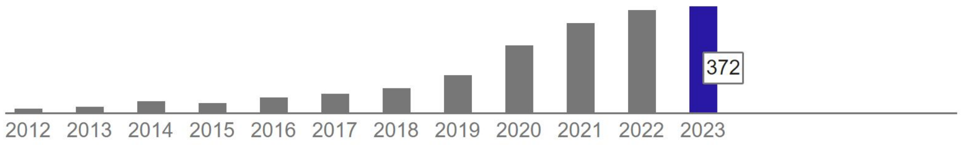
Publication date: 2011

Conference: Advances in Neural Information Processing Systems

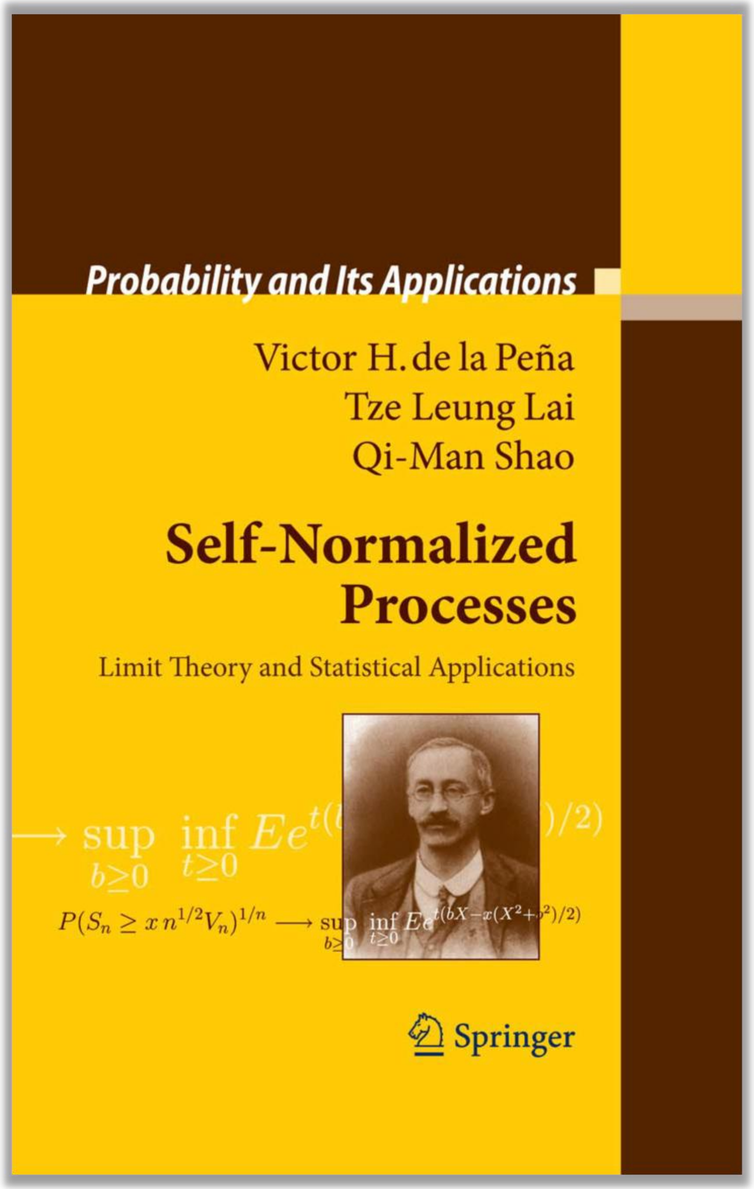
Pages: 2312-2320

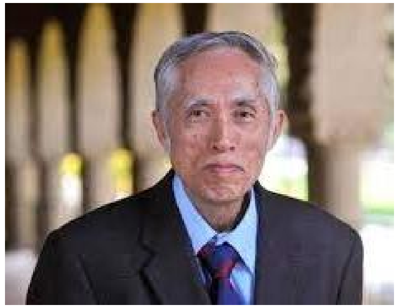
Description: We improve the theoretical analysis and empirical performance of algorithms for the stochastic multi-armed bandit problem and the linear stochastic multi-armed bandit problem. In particular, we show that a simple modification of Auer's UCB algorithm (Auer, 2002) achieves with high probability constant regret. More importantly, we modify and, consequently, improve the analysis of the algorithm for the for linear stochastic bandit problem studied by Auer (2002), Dani et al.(2008), Rusmevichientong and Tsitsiklis (2010), Li et al.(2010). Our modification improves the regret bound by a logarithmic factor, though experiments show a vast improvement. In both cases, the improvement stems from the construction of smaller confidence sets. For their construction we use a novel tail inequality for vector-valued martingales.

Total citations: Cited by 1726



Yasin Abbasi-Yadkori, David Pal, and Csaba Szepesvari. Improved algorithms for linear stochastic bandits. In Advances in Neural Information Processing Systems 24 (NIPS), pages 2312–2320, 2011.





Self-Normalized Processes: Limit theory and Statistical Applications

Victor H. de la Pena, Tze Leung Lai, and Qi-Man Shao

Probability and Its Applications Series. Springer. 2009.

Generalized Linear Bandits (GLB)

Extension: want to model Non-linear reward.

- Generalized linear model: $r_t = \mu(X_t^\top \theta_*) + \eta_t$
  - Link function $\mu : \mathbb{R} \mapsto \mathbb{R}$ k_μ-Lipschitz

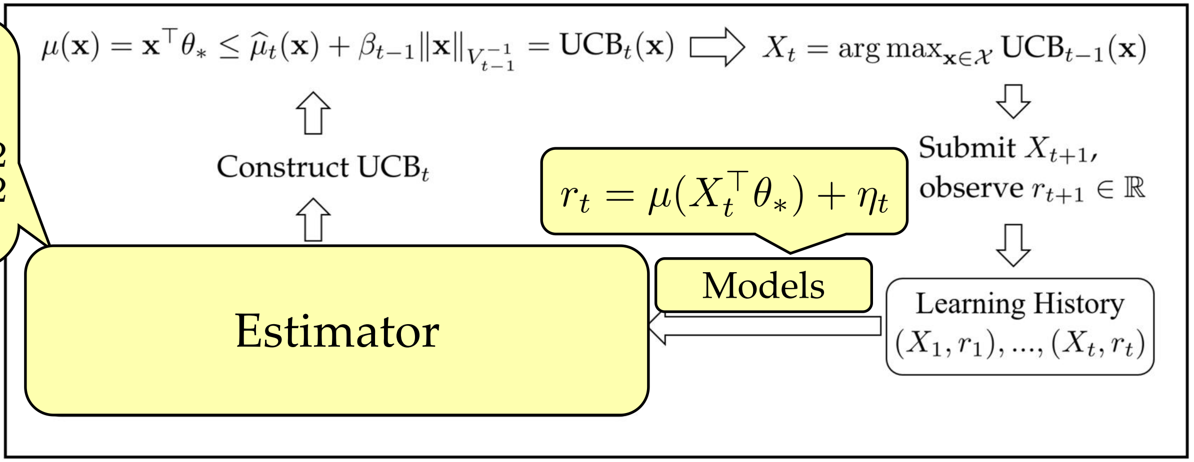
$$c_\mu = \inf_{\{\|\theta\|_2 \le S, \mathbf{x} \in \mathcal{X}\}} \dot{\mu}\left(\theta^{\mathrm{T}} \mathbf{x}\right) > 0$$

  Special cases: linear model: $\mu(x) = x$, logistic model: $\mu(x) = \frac{1}{1+\exp(-x)}$

- GLM-UCB

Maximum quasi-likelihood estimator

$$\hat{\theta}_t = \underset{\theta \in \mathbb{R}^d}{\arg\min} - \sum_{s=1}^{t-1} \log \mathbb{P}_\theta\left(r_{s+1} \mid X_s\right) + \frac{\lambda}{2} c_\mu \|\theta\|_2^2$$



Advanced Topic: Best of Both Worlds

- Best of adversarial MAB: $\mathbb{E}[\text{Regret}_T] = \max_{a\in[K]} \mathbb{E}\left[\sum_{t=1}^{T} r_t(a) - \sum_{t=1}^{T} r_t(a_t)\right] \le \mathcal{O}\left(\sum_{a:\Delta_a>0} \frac{\ln T}{\Delta_a}\right)$

- Best of stochastic MAB: $\mathbb{E}[\text{Regret}_T] = \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(a_t)\right] - \min_{a\in[K]} \sum_{t=1}^{T} \ell_t(a) \le \mathcal{O}\left(\sqrt{TK}\right)$

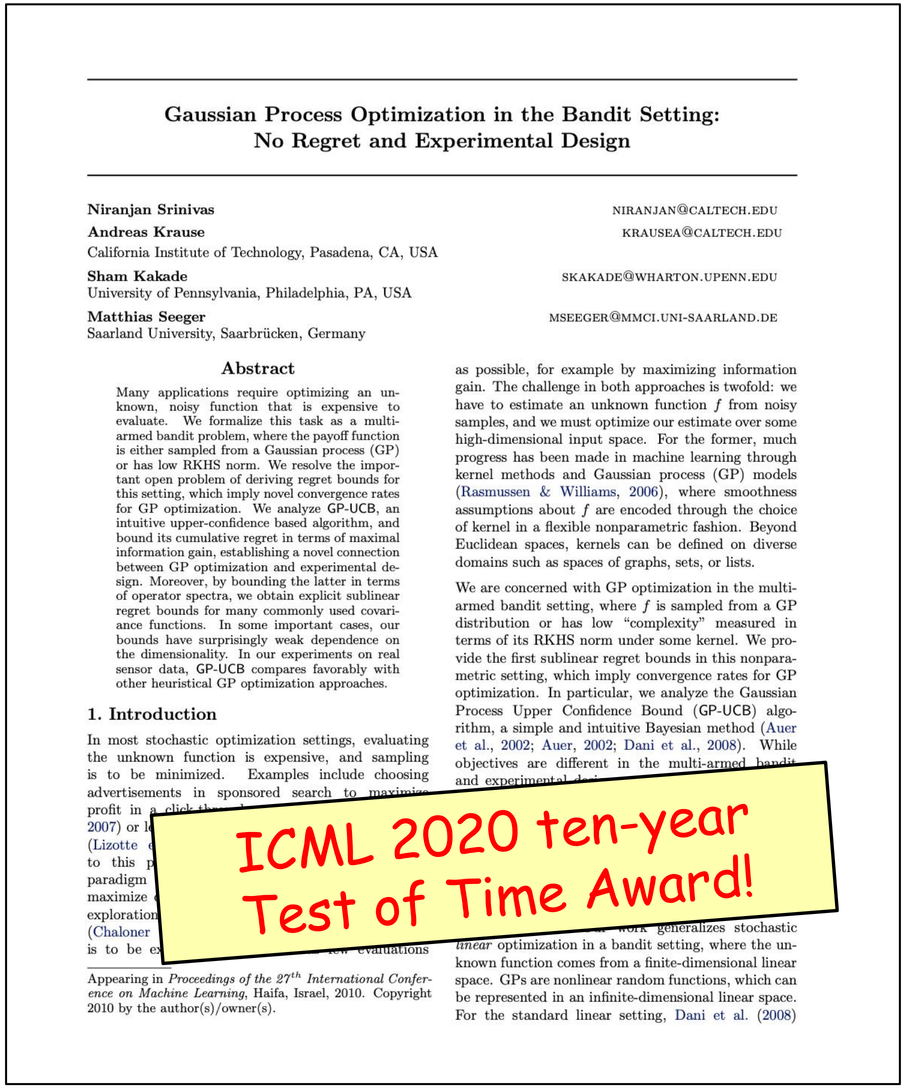
Can one algorithm achieve the ***best of both worlds***, without knowing whether the world is stochastic or adversarial?

- UCB: can get almost linear regret under adversarial setting.
- Exp3: can't have adaptive regret bound in stochastic case.1

⇨ Using OMD with ***Tsallis entropy*** regularizer.

Reference: Julian Zimmert, Yevgeny Seldin. An Optimal Algorithm for Stochastic and Adversarial Bandits. AISTATS 2019.

Advanced Topic: Bayesian Optimization



Gaussian Process Optimization in the Bandit Setting: No Regret and Experimental Design

Niranjan Srinivas NIRANJAN@CALTECH.EDU
Andreas Krause KRAUSEA@CALTECH.EDU
California Institute of Technology, Pasadena, CA, USA
Sham Kakade SKAKADE@WHARTON.UPENN.EDU
University of Pennsylvania, Philadelphia, PA, USA
Matthias Seeger MSEEGER@MMCI.UNI-SAARLAND.DE
Saarland University, Saarbrücken, Germany

Abstract

Many applications require optimizing an unknown, noisy function that is expensive to evaluate. We formalize this task as a multi-armed bandit problem, where the payoff function is either sampled from a Gaussian process (GP) or has low RKHS norm. We resolve the important open problem of deriving regret bounds for this setting, which imply novel convergence rates for GP optimization. We analyze GP-UCB, an intuitive upper-confidence based algorithm, and bound its cumulative regret in terms of maximal information gain, establishing a novel connection between GP optimization and experimental design. Moreover, by bounding the latter in terms of operator spectra, we obtain explicit sublinear regret bounds for many commonly used covariance functions. In some important cases, our bounds have surprisingly weak dependence on the dimensionality. In our experiments on real sensor data, GP-UCB compares favorably with other heuristic GP optimization approaches.

1. Introduction

In most stochastic optimization settings, evaluating the unknown function is expensive, and sampling is to be minimized. Examples include choosing advertisements in sponsored search to maximize profit in a click-through ... 2007) or l... (Lizotte ... to this p... paradigm ... maximize ... exploration ... (Chaloner ... is to be ex... few evaluations

Appearing in *Proceedings of the 27th International Conference on Machine Learning*, Haifa, Israel, 2010. Copyright 2010 by the author(s)/owner(s).

as possible, for example by maximizing information gain. The challenge in both approaches is twofold: we have to estimate an unknown function f from noisy samples, and we must optimize our estimate over some high-dimensional input space. For the former, much progress has been made in machine learning through kernel methods and Gaussian process (GP) models (Rasmussen & Williams, 2006), where smoothness assumptions about f are encoded through the choice of kernel in a flexible nonparametric fashion. Beyond Euclidean spaces, kernels can be defined on diverse domains such as spaces of graphs, sets, or lists.

We are concerned with GP optimization in the multi-armed bandit setting, where f is sampled from a GP distribution or has low "complexity" measured in terms of its RKHS norm under some kernel. We provide the first sublinear regret bounds in this nonparametric setting, which imply convergence rates for GP optimization. In particular, we analyze the Gaussian Process Upper Confidence Bound (GP-UCB) algorithm, a simple and intuitive Bayesian method (Auer et al., 2002; Auer, 2002; Dani et al., 2008). While objectives are different in the multi-armed bandit and experimental design ... work generalizes stochastic *linear* optimization in a bandit setting, where the unknown function comes from a finite-dimensional linear space. GPs are nonlinear random functions, which can be represented in an infinite-dimensional linear space. For the standard linear setting, Dani et al. (2008)

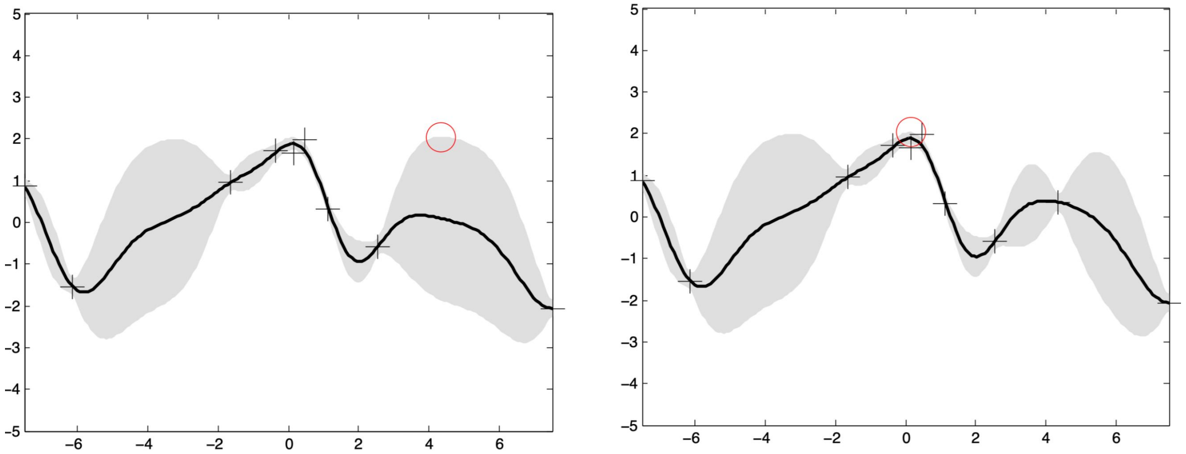
ICML 2020 ten-year Test of Time Award!

Reward function: $r_t = f(X_t) + \eta_t$

$f(\mathbf{x})$ belongs to RKHS with $k(\mathbf{x}, \mathbf{x}') = \sum_{m=1}^{|\mathcal{H}|} \varphi_m(\mathbf{x})\varphi_m(\mathbf{x}')$

Rewrite $f(x) = \sum_{m=1}^{|\mathcal{H}|} \theta_m \varphi_m(x) = \varphi(x)^\top \theta$

$\Rightarrow \quad r_t = \varphi(X_t)^\top \theta + \eta_t$ ***Linear bandits in RKHS***



Iteration t *Iteration t + 1*

[Gaussian Process Optimization in the Bandit Setting: No Regret and Experimental Design.](#) ICML 2010.

Advanced Topic: Linear MDPs

Linear MDPs

- Exists feature map $\phi : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}^d$
  - Such that:

$$r_h(s,a) = \theta_h^\star \cdot \phi(s,a), \quad P_h(\cdot|s,a) = \mu_h^\star \phi(s,a), \forall h$$

- Implies a low-rank assumption in large-MDP case

(Jin et al., 2020) Provably efficient reinforcement learning with linear function approximation 7

UCB-VI for Linear MDPs

- In every round:
  1. Run Ridge regression for estimating the model

$$\widehat{\mu}_h^n = \mathrm{argmin}_{\mu \in \mathbb{R}^{|\mathcal{S}| \times d}} \sum_{i=0}^{n-1} \left\| \mu\phi(s_h^i, a_h^i) - \delta(s_{h+1}^i) \right\|_2^2 + \lambda \|\mu\|_F^2.$$

$$\widehat{\mu}_h^n = \sum_{i=0}^{n-1} \delta(s_{h+1}^i)\phi(s_h^i, a_h^i)^\top (\Lambda_h^n)^{-1}$$

  2. Construct the exploration bonuses

$$b_h^n(s,a) = \beta\sqrt{\phi(s,a)^\top (\Lambda_h^n)^{-1} \phi(s,a)},$$

  3. Run optimistic value iterations, and update greedy policy

11

[Yu-Xiang Wang's course CS292F Lecture 10 Exploration IV: Linear MDP]

History bits

- Bandit problems were introduced for the clinical trial design by **William R. Thompson** in an article published in 1933 [Thompson, 1933].

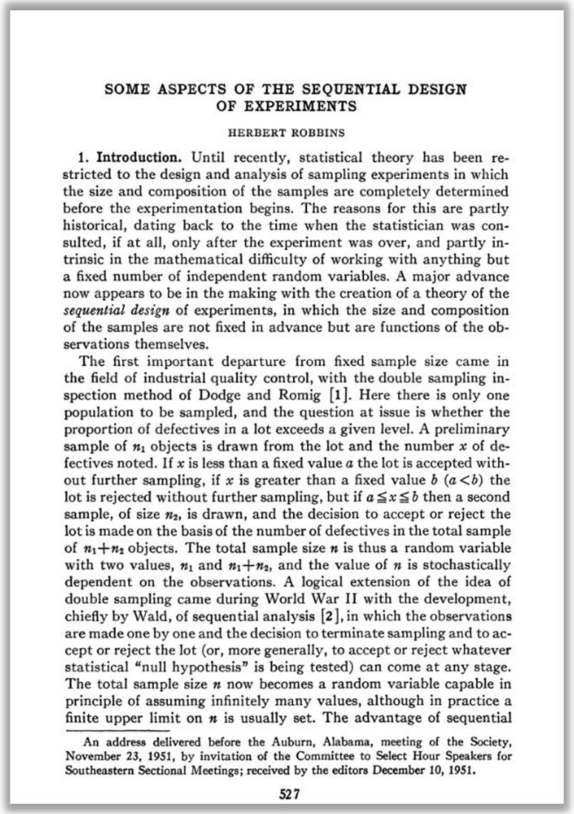
ON THE LIKELIHOOD THAT ONE UNKNOWN PROBABILITY EXCEEDS ANOTHER IN VIEW OF THE EVIDENCE OF TWO SAMPLES.

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

- Thompson Sampling (TS) was originally described in this paper but has been largely ignored by the artificial intelligence community.
- TS was subsequently rediscovered numerous times independently in the context of reinforcement learning.

History bits

- Bandit problems were later formally restated in a short but influential paper [Robbins, 1952] and further developed in [Lai and Robbins, 1985].



SOME ASPECTS OF THE SEQUENTIAL DESIGN OF EXPERIMENTS

HERBERT ROBBINS

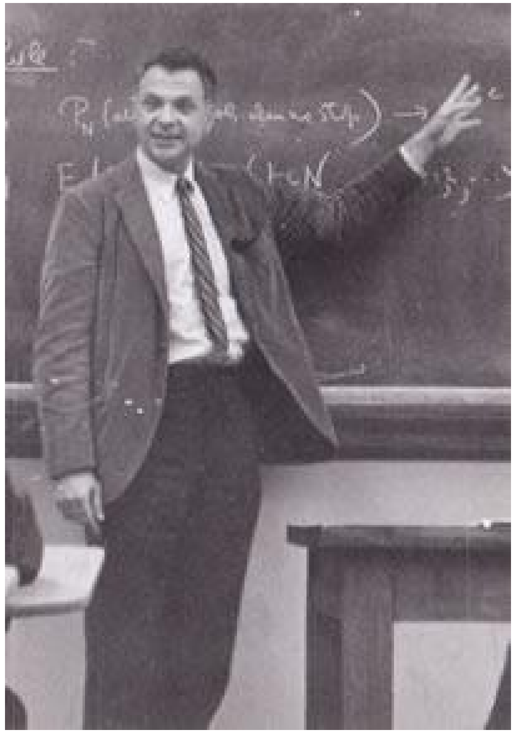
1. **Introduction.** Until recently, statistical theory has been restricted to the design and analysis of sampling experiments in which the size and composition of the samples are completely determined before the experimentation begins. The reasons for this are partly historical, dating back to the time when the statistician was consulted, if at all, only after the experiment was over, and partly intrinsic in the mathematical difficulty of working with anything but a fixed number of independent random variables. A major advance now appears to be in the making with the creation of a theory of the *sequential design* of experiments, in which the size and composition of the samples are not fixed in advance but are functions of the observations themselves.

The first important departure from fixed sample size came in the field of industrial quality control, with the double sampling inspection method of Dodge and Romig [1]. Here there is only one population to be sampled, and the question at issue is whether the proportion of defectives in a lot exceeds a given level. A preliminary sample of n_1 objects is drawn from the lot and the number x of defectives noted. If x is less than a fixed value a the lot is accepted without further sampling, if x is greater than a fixed value b $(a<b)$ the lot is rejected without further sampling, but if $a \leqq x \leqq b$ then a second sample, of size n_2, is drawn, and the decision to accept or reject the lot is made on the basis of the number of defectives in the total sample of n_1+n_2 objects. The total sample size n is thus a random variable with two values, n_1 and n_1+n_2, and the value of n is stochastically dependent on the observations. A logical extension of the idea of double sampling came during World War II with the development, chiefly by Wald, of sequential analysis [2], in which the observations are made one by one and the decision to terminate sampling and to accept or reject the lot (or, more generally, to accept or reject whatever statistical "null hypothesis" is being tested) can come at any stage. The total sample size n now becomes a random variable capable in principle of assuming infinitely many values, although in practice a finite upper limit on n is usually set. The advantage of sequential

An address delivered before the Auburn, Alabama, meeting of the Society, November 23, 1951, by invitation of the Committee to Select Hour Speakers for Southeastern Sectional Meetings; received by the editors December 10, 1951.

527

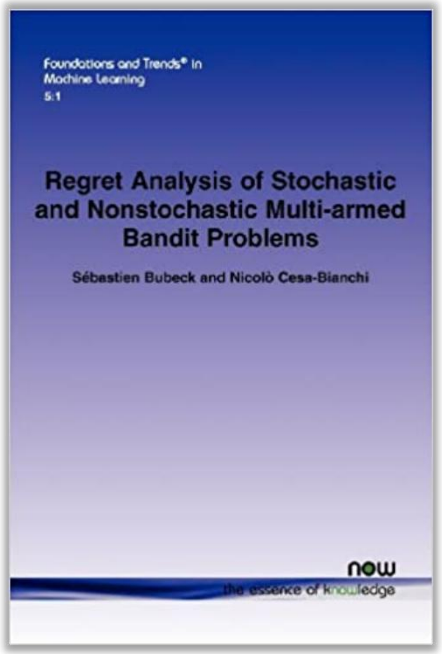
H. Robbins. Some aspects of the sequential design of experiments. Bulletin of the American Mathematical Society, 58(5):527–535, 1952.

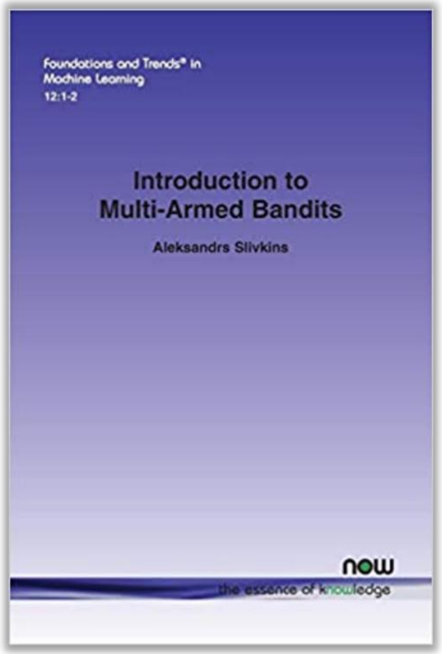


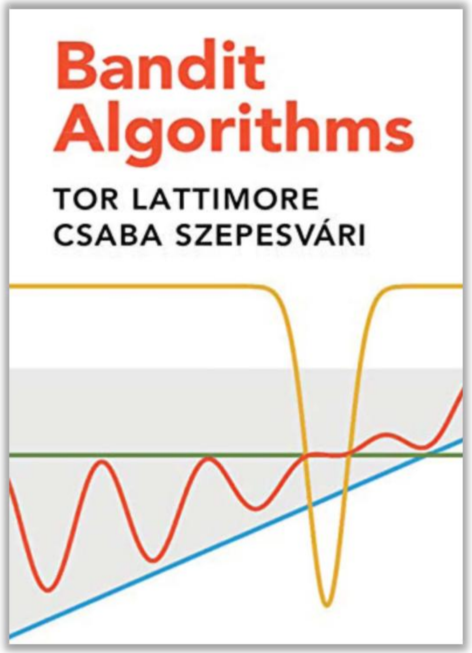
Herbert Ellis Robbins (1915 -- 2001)

History bits

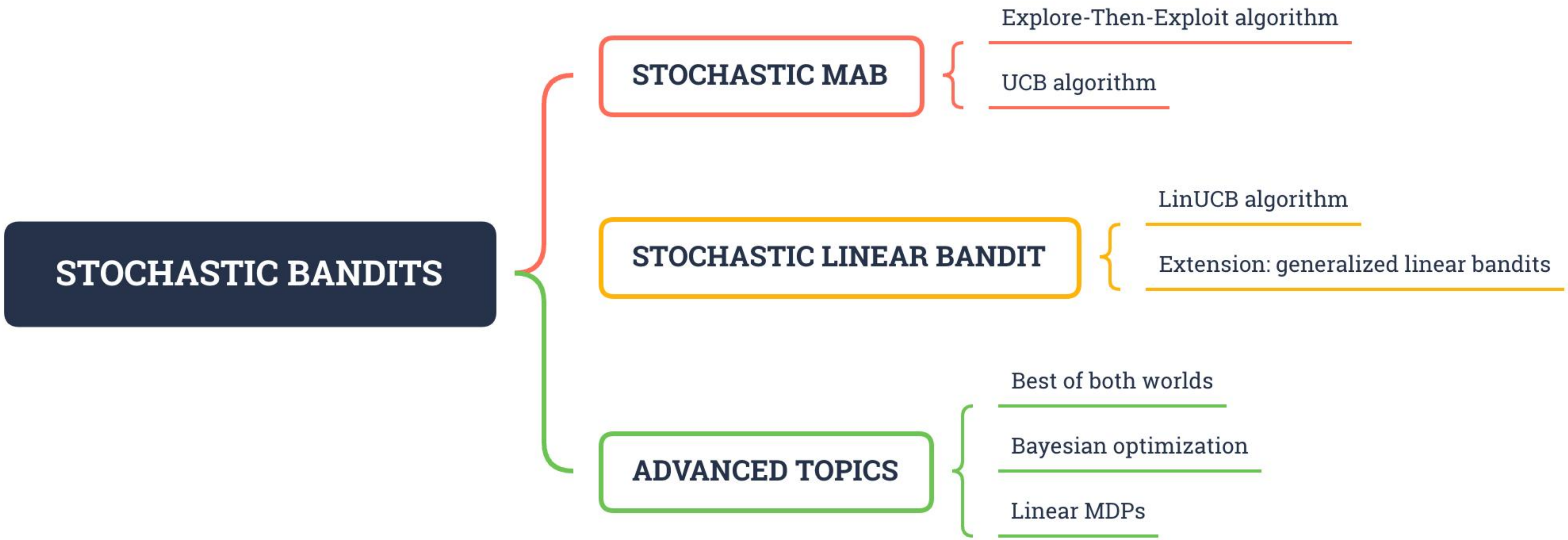
- Techniques developed in bandit problems have been applied in many areas, including machine learning, statistics, operational research, and information theory [Bubeck and Cesa-Bianchi, 2012; Slivkins, 2019; Lattimore and Szepesvári, 2020].







Summary


STOCHASTIC BANDITS
STOCHASTIC MAB
Explore-Then-Exploit algorithm
UCB algorithm
STOCHASTIC LINEAR BANDIT
LinUCB algorithm
Extension: generalized linear bandits
ADVANCED TOPICS
Best of both worlds
Bayesian optimization
Linear MDPs