NJUAI 南京大学 人工智能学院
SCHOOL OF ARTIFICIAL INTELLIGENCE, NANJING UNIVERSITY

Lecture 12. Stochastic Bandits

Advanced Optimization (Fall 2024)

Peng Zhao

zhaop@lamda.nju.edu.cn

Nanjing University

Outline

- Multi-Armed Bandits
- Linear Bandits
- Advanced Topics

Part 1. Multi-Armed Bandits

- Problem Formulation
- Explore-Then-Exploit
- Upper Confidence Bound

Bandits

- Bandit problems
  - named after a ***one-armed bandit***
  - ***arm***: a colloquial term for a slot machine that is pulled to try to win
  - ***bandit***: comes from the idea that the machine is a "thief" that takes your money without offering a guaranteed return

- Multi-armed bandits
  - Context: there are multiple slot machines, each with its own probability of payout
  - Goal: the player (gambler) places her bets on a slot machine to maximize the total reward
  - **Exploration-Exploitation tradeoff**

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 $r_1(a), r_2(a), ..., r_T(a)$ are i.i.d samples from $\mathcal{D}_a$.

For conventional issue, we will use the "***reward language***" in stochastic bandits.

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***:

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

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

- Caveat: note the difference between *pseudo regret* and the *(expected) regret*.

Deploying Exp3 to Stochastic MAB

- Stochastic MAB is a special case of Adversarial MAB

⇨ Deploying Exp3 achieves the expected regret (though having gap to pseudo regret).

Theorem 1 (Upper Bound for Exp3). *Suppose that* $\forall t \in [T]$ *and* $a \in [K]$, $0 \le \ell_{t,a} \le 1$, *then Exp3 with learning rate* $\eta = \sqrt{(\ln K)/(TK)}$ *guarantees*

$$\mathbb{E}[\mathrm{Regret}_T] = \mathbb{E}\left[\sum_{t=1}^{T} \ell_{t,a_t}\right] - \min_{a\in[K]} \sum_{t=1}^{T} \ell_{t,a} \le \mathcal{O}\left(\sqrt{TK \log K}\right),$$

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

⇨ Not yet to exploit benign ***stochastic*** modeling….

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: $n_t(a) = \sum_{s=1}^{t} \mathbf{1}\{a_s = a\}$.

- Regret can be reformulated as

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

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

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* $a \in [K], 0 \leq r_t(a) \leq 1$, *then ETE with exploration period* T_0 *guarantees*

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

Proof of ETE Regret Bound

Proof. Note the regret decomposition: $\mathbb{E}[\mathrm{Regret}_T] = \sum_{a\in[K]} \Delta_a \cdot n_T(a)$

Below we estimate the random variable $n_T(a)$ for each $a \in [K]$.

Exploration Exploitation

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

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

pulling strategy $\widehat{a} = \arg\max_{a\in[K]} \widehat{\mu}_{T_0}(a)$

Note that when $\widehat{\mu}_{T_0}(a) \geq \widehat{\mu}_{T_0}(a^*)$ happens, it implies the one of the following two rare events must happen:

$$\widehat{\mu}_{T_0}(a) \geq (\mu(a) + \mu(a^*))/2, \text{ and } \widehat{\mu}_{T_0}(a^\star) \leq (\mu(a) + \mu(a^*))/2.$$

Otherwise, $\widehat{\mu}_{T_0}(a) < (\mu(a) + \mu(a^*))/2 < \widehat{\mu}_{T_0}(a^\star)$.

Proof of ETE Regret Bound

Proof. Note the regret decomposition: $\mathbb{E}[\mathrm{Regret}_T] = \sum_{a \in [K]} \Delta_a \cdot n_T(a)$

Below we estimate the random variable $n_T(a)$ for each $a \in [K]$.

pulling strategy $\widehat{a} = \arg\max_{a \in [K]} \widehat{\mu}_{T_0}(a)$

Exploration Exploitation

$$
\begin{aligned}
n_T(a) &= T_0/K + (T - T_0)\Pr\{\widehat{a} = a\} \\
&\le T_0/K + (T - T_0)\Pr\{\widehat{\mu}_{T_0}(a) \ge \widehat{\mu}_{T_0}(a^*)\} \\
&\le T_0/K + (T - T_0)\Pr\left\{\widehat{\mu}_{T_0}(a) \ge \frac{\mu(a) + \mu(a^*)}{2} \cup \widehat{\mu}_{T_0}(a^*) \le \frac{\mu(a) + \mu(a^*)}{2}\right\} \\
&\le T_0/K + (T - T_0)\left(\Pr\left\{\widehat{\mu}_{T_0}(a) \ge \frac{\mu(a) + \mu(a^*)}{2}\right\} + \Pr\left\{\widehat{\mu}_{T_0}(a^*) \le \frac{\mu(a) + \mu(a^*)}{2}\right\}\right)
\end{aligned}
$$

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

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 independent $X_i \in [0,1], i \in [m], \bar{X} = \frac{1}{m}\sum_{i=1}^m X_i$, we have

$$\Pr\left\{\bar{X} - \mathbb{E}[\bar{X}] \ge \epsilon\right\} \le \exp\left(-2m\epsilon^2\right);$$
$$\Pr\left\{\bar{X} - \mathbb{E}[\bar{X}] \le -\epsilon\right\} \le \exp\left(-2m\epsilon^2\right).$$

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

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

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

□

Issue of ETE

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*

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

- Need to tune T_0

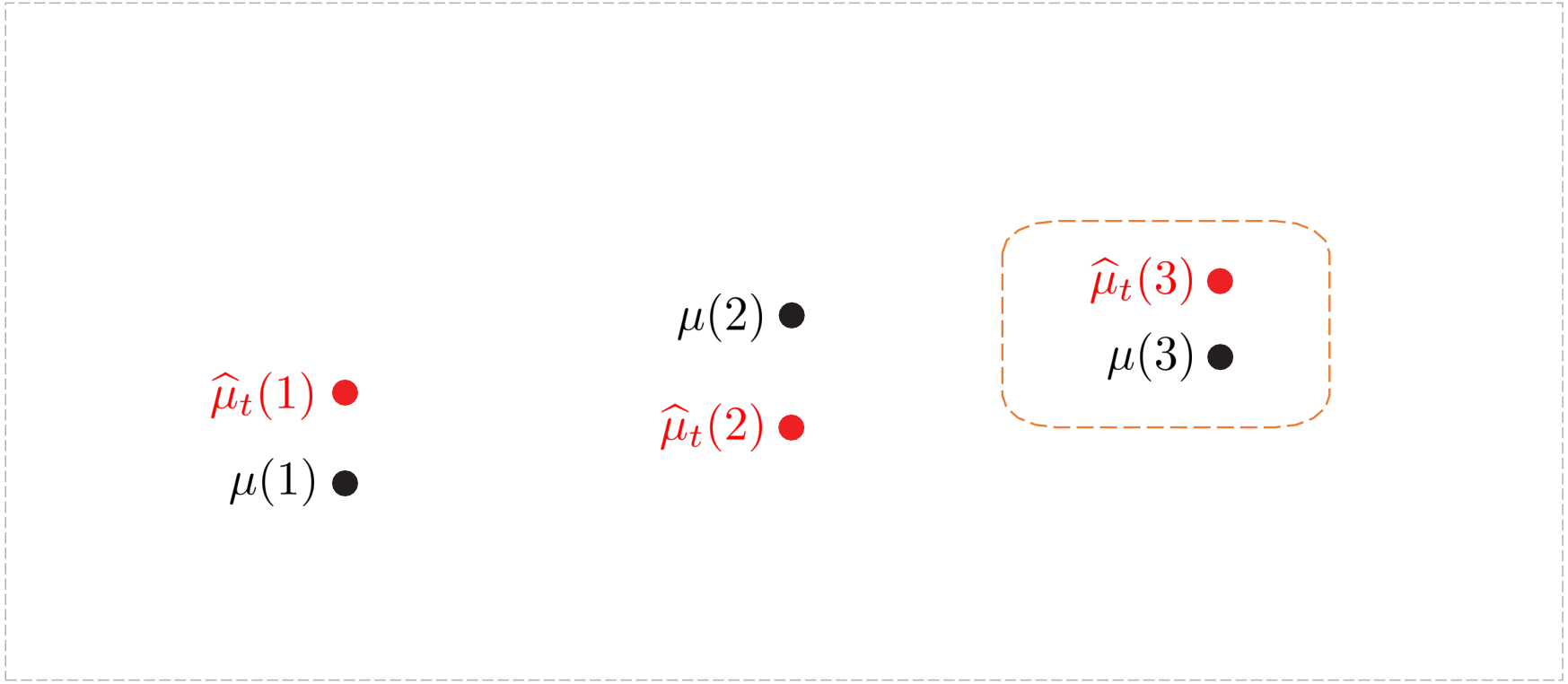
  Tune T_0 with prior of suboptimality gap Δ_a: $\mathbb{E}[\mathrm{Regret}_T] = \widetilde{\mathcal{O}}(\sqrt{T})$

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

⇨ Solution: do explore and exploit ***adaptively***.

Explore-then-Exploit (ETE)

- ETE

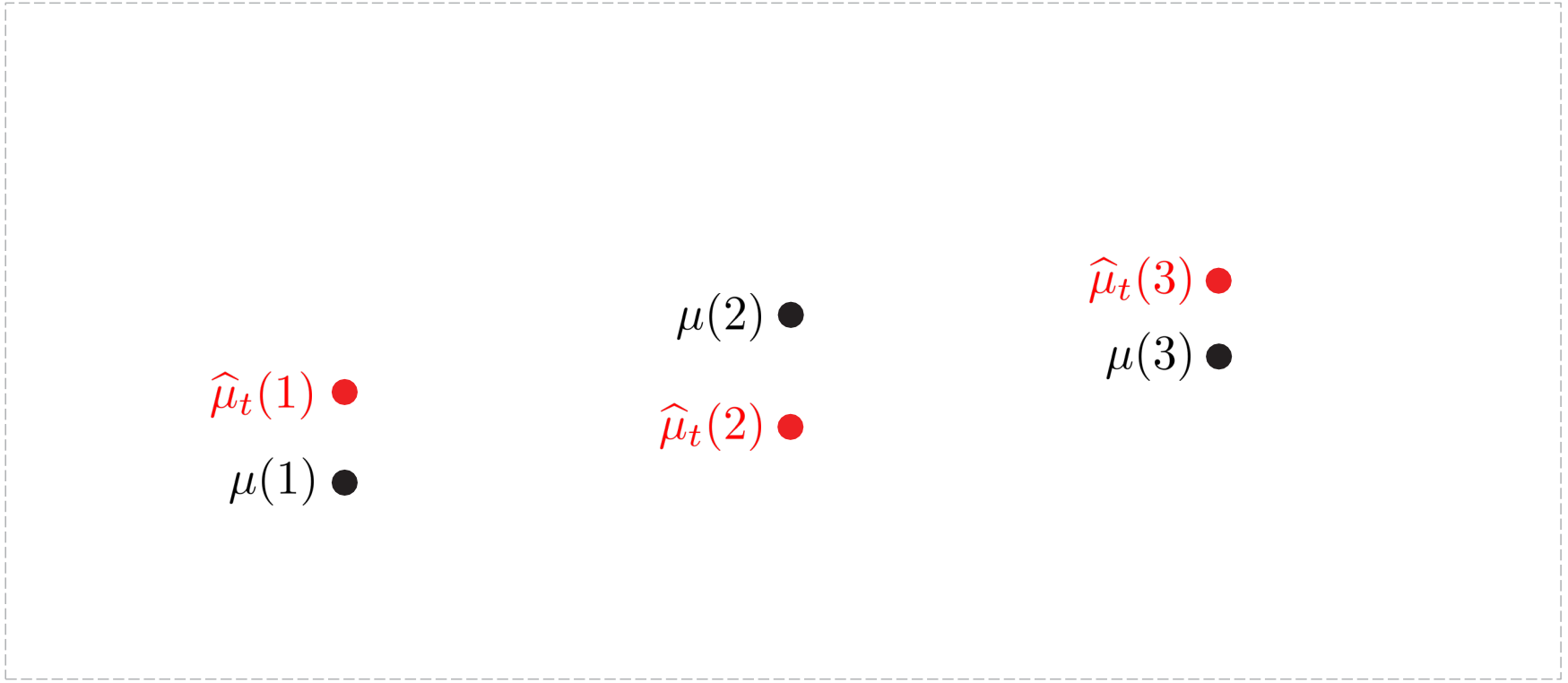


Relying on the estimate of the previous T_0 rounds.

There is no way to revise the estimate!

Upper Confidence Bound

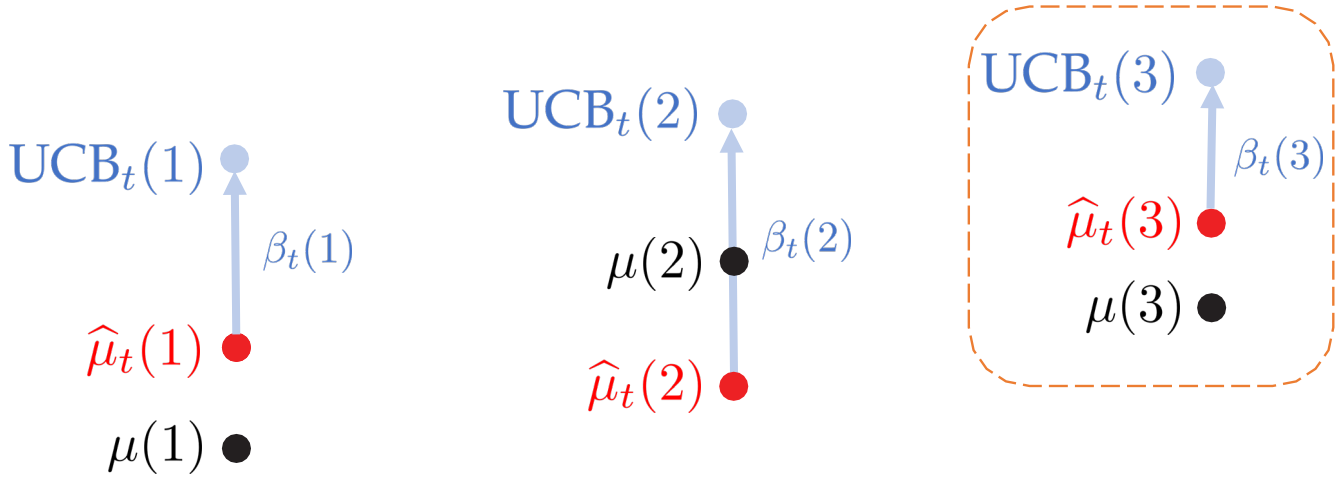
- UCB



Upper Confidence Bound

- UCB

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

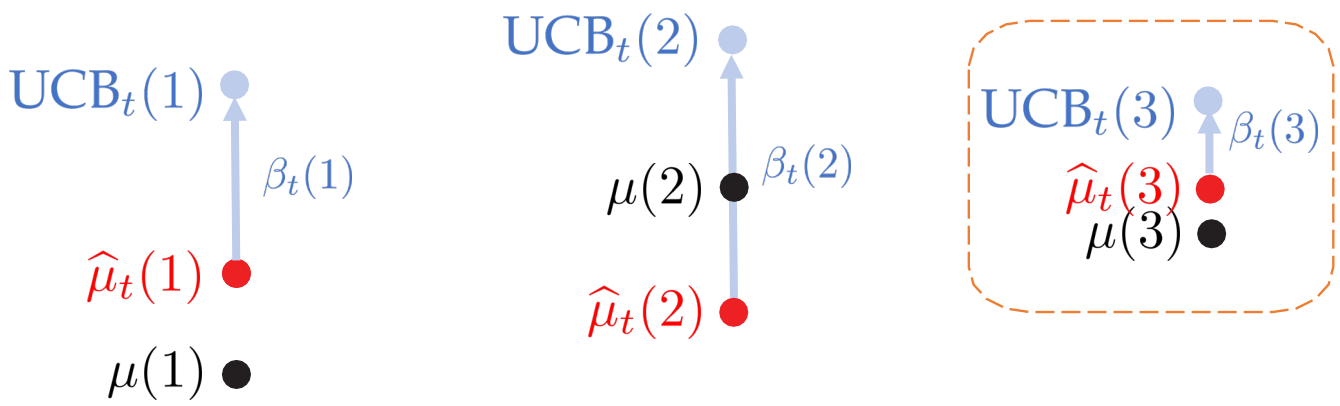


Optimism in Face of Uncertainty: $a_t = \arg\max_{a \in [K]} \mathrm{UCB}_t(a)$

Upper Confidence Bound

- UCB

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

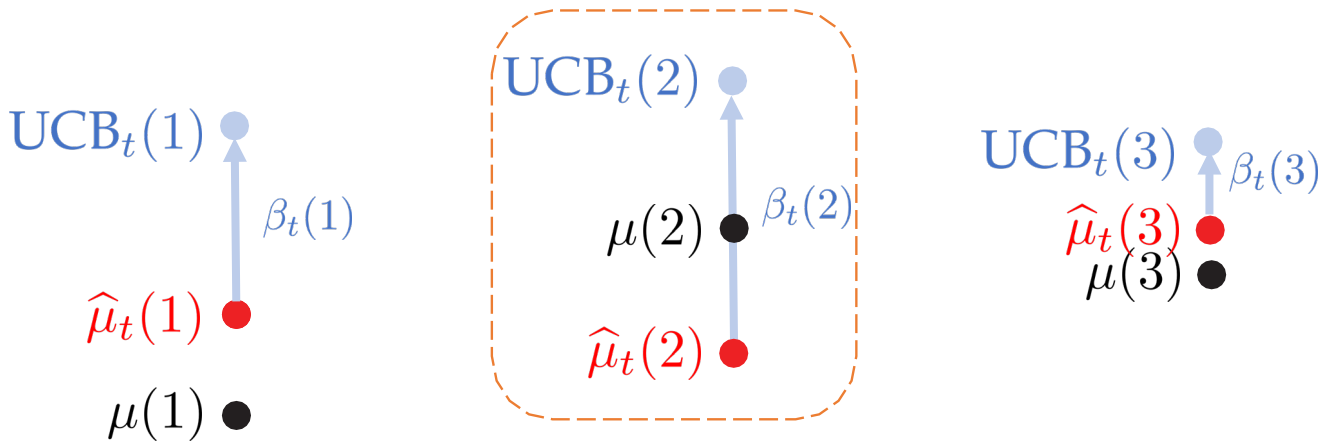


Optimism in Face of Uncertainty: $a_t = \arg\max_{a \in [K]} \mathrm{UCB}_t(a)$

Upper Confidence Bound

- UCB

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

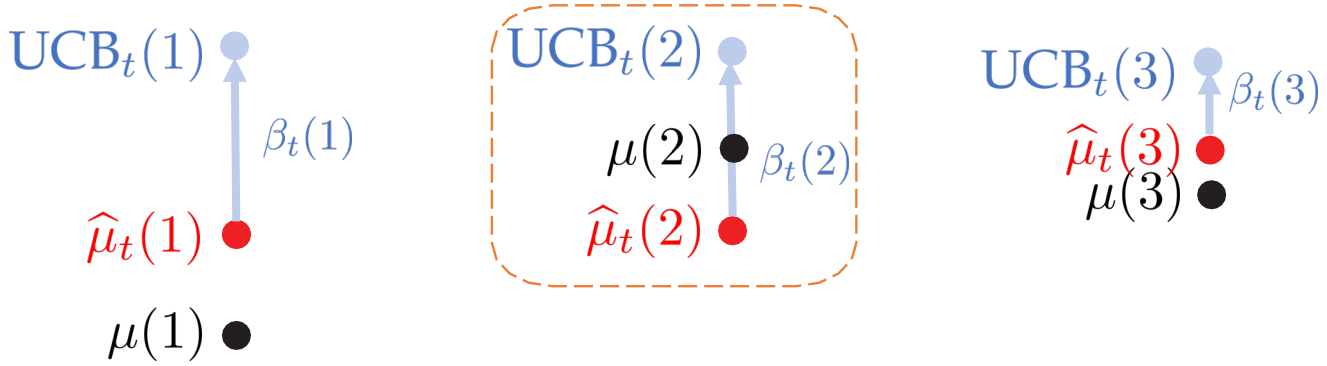


Optimism in Face of Uncertainty: $a_t = \arg\max_{a \in [K]} \mathrm{UCB}_t(a)$

Upper Confidence Bound

- UCB

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

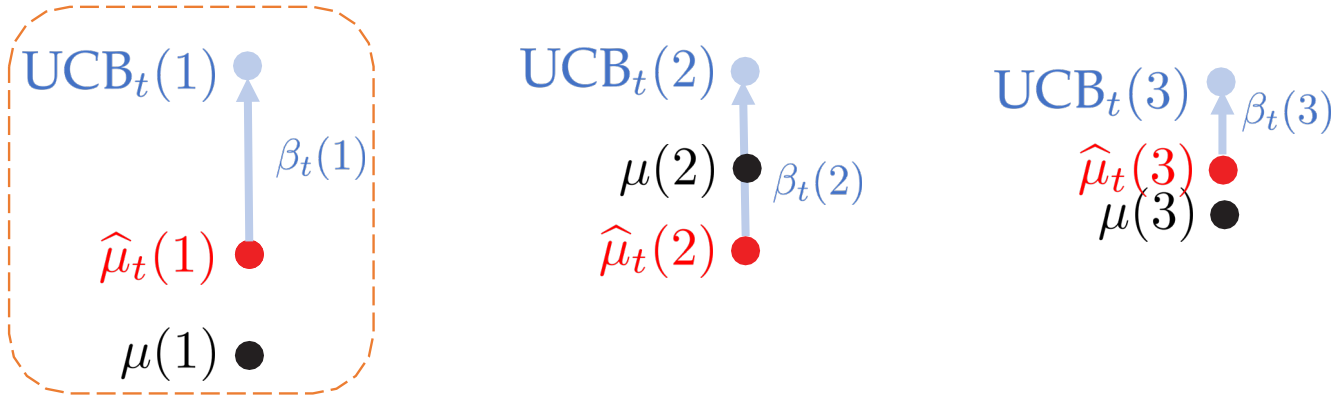


Optimism in Face of Uncertainty: $a_t = \arg\max_{a \in [K]} \mathrm{UCB}_t(a)$

Upper Confidence Bound

- UCB

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

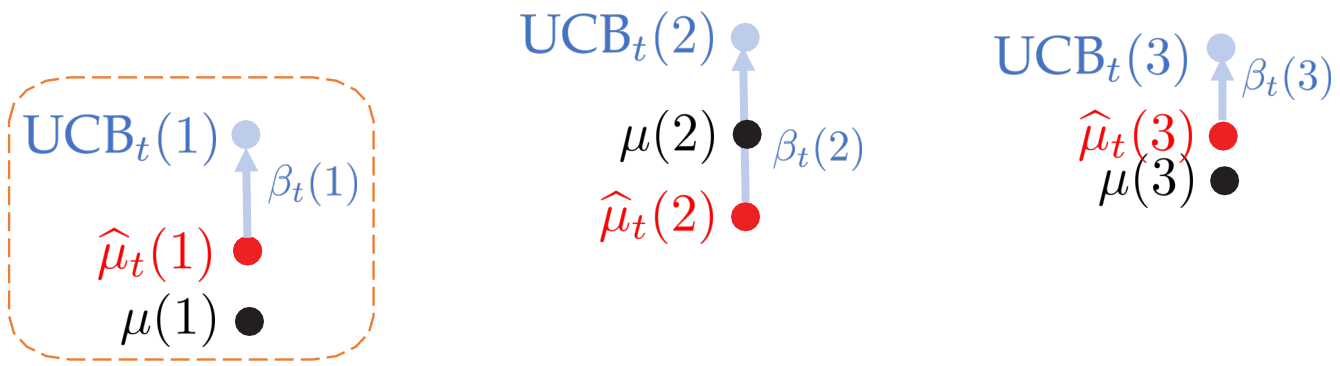


Optimism in Face of Uncertainty: $a_t = \arg\max_{a \in [K]} \mathrm{UCB}_t(a)$

Upper Confidence Bound

- UCB

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

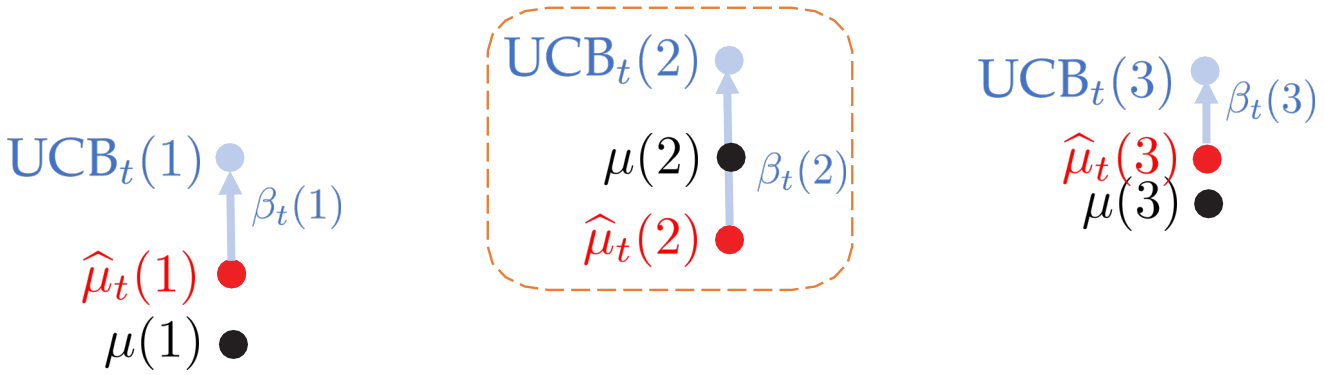


Optimism in Face of Uncertainty: $a_t = \arg\max_{a \in [K]} \mathrm{UCB}_t(a)$

Upper Confidence Bound

- UCB

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



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

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

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

(3) Update upper confidence bounds $\mathbf{UCB}_t(a)$ by new estimation

- Estimation: empirical average

$$\widehat{\mu}_t(a) = \frac{1}{n_t(a)} \sum_{s=1}^{t} \mathbf{1}\{a_s = a\} r_s(a), \quad \text{where } n_t(a) \text{ is the pulled times of arm } a$$

- UCB construction: Hoeffding's inequality

Construct UCB

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

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

Therefore, it suggests $\mathbf{UCB}_t(a) \triangleq \widehat{\mu}_t(a) + \sqrt{\frac{\ln T}{n_t(a)}}$*, ensuring* $\mu(a) \leq \mathbf{UCB}_t(a)$.

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

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

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

Furthermore, by the union bound over all arms and all rounds and letting $\delta = 1/T^2$,

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

UCB: Distribution-Dependent Bound

Theorem 2 (Distribution-dependent). *Suppose that 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*

$$\bar{R}_T \le \sum_{a:\Delta_a>0} \frac{4 \ln T}{\Delta_a} + \Delta_a = \mathcal{O}\left(\sum_{a:\Delta_a>0} \frac{\log T}{\Delta_a} \right).$$

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

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

Proof of UCB Regret Bound

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

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

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

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

$$\Rightarrow \quad \bar{R}_T = \sum_{a \in [K]} \Delta_a n_T(a) \le \sum_{a:\Delta_a > 0} \Delta_a \left(4\frac{\ln T}{\Delta_a^2} + 1\right) = \sum_{a:\Delta_a > 0} 4\frac{\ln T}{\Delta_a} + \Delta_a.$$

□

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*

$$\bar{R}_T \le \sum_{a:\Delta_a>0} \frac{4 \ln T}{\Delta_a} + \Delta_a = \mathcal{O}\left(\sum_{a:\Delta_a>0} \frac{\log T}{\Delta_a} \right).$$

- 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 $\bar{R}_T = \Delta_a T$.

$$\Rightarrow \quad \bar{R}_T \le \min \left\{ \max_{a \in [K]} \Delta_a T, \sum_{a:\Delta_a>0} \frac{4 \ln T}{\Delta_a} + \Delta_a \right\}$$

distribution-dependent
also called gap/instance-dependent

Upper Bound and Lower 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*

$$\bar{R}_T \le \sum_{a:\Delta_a>0} \frac{4 \ln T}{\Delta_a} + \Delta_a = \mathcal{O}\left(\sum_{a:\Delta_a>0} \frac{\log T}{\Delta_a} \right).$$

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

$$\inf_{\mathcal{A}} \sup_{\ell_1,\ldots,\ell_T} \mathbb{E}\left[\text{Regret}_T\right] = \Omega(\sqrt{TK})$$

Is there any contradiction between the upper bound and lower bound?

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*

$$\bar{R}_T \le 2\sqrt{TK \ln T} + \sum_{a \in [K]} \Delta_a = \mathcal{O}\left(\sqrt{TK \log T}\right).$$

Proof.

$$\bar{R}_T = \sum_{a \in [K]} \Delta_a n_T(a) = \sum_{a : \Delta_a < \Delta} \Delta_a n_T(a) + \sum_{a : \Delta_a \ge \Delta} \Delta_a n_T(a)$$

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

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

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

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

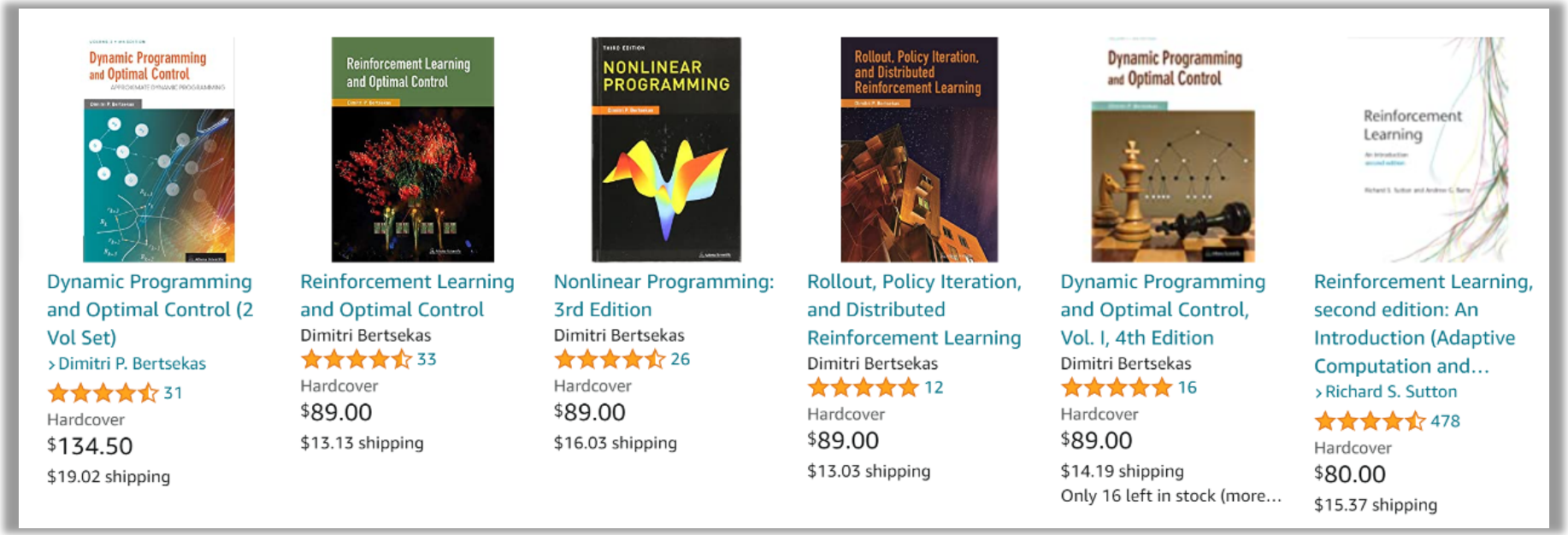
□

Part 2. Linear Bandits

- Formulation
- Estimator and UCB construction
- LinUCB and Extensions

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

Stochastic Linear Bandits

Each arm is associated with a feature vector $\mathbf{x} \in \mathcal{X} = \{\mathbf{x} \in \mathbb{R}^d \mid \|\mathbf{x}\|_2 \leq L\}$

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 $r_t \in \mathbb{R}$.

- Linear modeling assumption: $r_t(x) = x^\top \theta_* + \eta_t$
  - for some unknown parameter $\theta_* \in \Theta = \{\theta \mid \|\theta\|_2 \leq S\}$;
  - for some unknown noise: η_t is R-sub-Gaussian random noise;
- Regret measure: $\bar{R}_T \triangleq T \max_{\mathbf{x} \in \mathcal{X}} \mathbf{x}^\top \theta_* - \sum_{t=1}^{T} X_t^\top \theta_*$

Stochastic LB: Formulation

Each arm is associated with a feature vector $\mathbf{x} \in \mathcal{X} = \{\mathbf{x} \in \mathbb{R}^d \mid \|\mathbf{x}\|_2 \le L\}$

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 $\mathcal{X} = \{\|\mathbf{x}\|_2 \le L\}$
Model	$\mathbb{E}[r(a)] = \mu(a)$ $\forall t \in [T], a \in [K], r_t(a) \in [0,1]$	$r_t = X_t^\top \theta_* + \eta_t$ $\quad \mu(\mathbf{x}) = \mathbf{x}^\top \theta_*$ η_t: sub-Gaussian noise
Regret	$\bar{R}_T = T \max_{a \in [K]} \mu(a) - \sum_{t=1}^{T} \mu(a_t)$	$\bar{R}_T = T \max_{\mathbf{x} \in \mathcal{X}} \mathbf{x}^\top \theta_* - \sum_{t=1}^{T} X_t^\top \theta_*$

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 for all* $t \in [T]$ *and* $a \in [K]$, $0 \leq r_t(a) \leq 1$, *then UCB satisfies*

$$\bar{R}_T \leq 2\sqrt{TK \ln T} + \sum_{a \in [K]} \Delta_a = \mathcal{O}\left(\sqrt{TK \log T}\right).$$

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

⇨ Haven't exploited the additional ***contextual feature information*** !

LinUCB Algorithm: Formulation

LinUCB Algorithm

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

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

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

(3) update upper confidence bounds $\mathbf{UCB}_t(\mathbf{x})$ by new estimation

- Estimation: regularized least square (ridge 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

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

- This LS estimator can be updated incrementally.
- Even accelerated by using rank-1 update (Sherman-Morrison-Woodbury formula), which reduces the computational complexity from $\mathcal{O}(d^3)$ to $\mathcal{O}(d^2)$

$$K_t = \frac{P_{t-1}X_t}{1 + X_t^\top P_{t-1} X_t}$$

$$\widehat{\theta}_t = \widehat{\theta}_{t-1} + K_t\left[r_t - X_t^\top \widehat{\theta}_{t-1}\right]$$

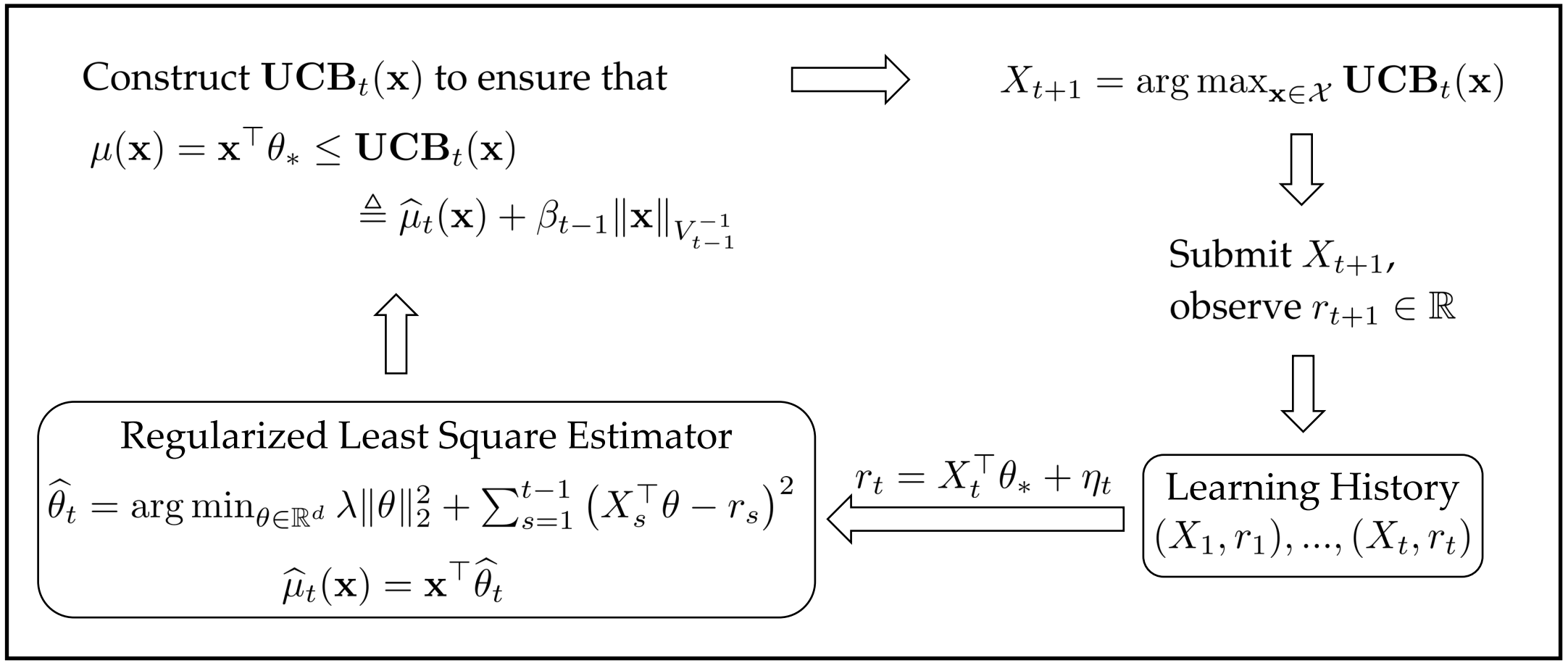
$$P_t = P_{t-1} - K_t X_t^\top P_{t-1}.$$

known as the Recursive Least Square (RLS) estimator

provably equivalent to the standard LS estimator

LinUCB Algorithm

Key question: how to construct a proper UCB?



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

$$\widehat{\mu}_t(a) = \frac{1}{n_t(a)} \sum_{s=1}^{t} \mathbf{1}\{a_s = a\} r_s(a)$$

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 $t \in [T]$*

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

Therefore, it suggests $\mathbf{UCB}_t(\mathbf{x}) \triangleq \mathbf{x}^\top\widehat{\theta}_t + \beta_{t-1}\|\mathbf{x}\|_{V_{t-1}^{-1}}$, ensuring $\mu(\mathbf{x}) \le \mathbf{UCB}_t(\mathbf{x})$.

Proof.

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

$$= V_{t-1}^{-1}\left(\sum_{s=1}^{t-1}\left(X_s^\top\theta_* + \eta_s\right)X_s\right) - V_{t-1}^{-1}\left(\lambda I_d + \sum_{s=1}^{t-1} X_s X_s^\top\right)\theta_*$$

$$= V_{t-1}^{-1}\left(\sum_{s=1}^{t-1}\eta_s X_s - \lambda\theta_*\right) \qquad V_{t-1} = \lambda I + \textstyle\sum_{s=1}^{t-1} X_s X_s^\top$$

Proof of Estimation Error Bound

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

$$\left|\mathbf{x}^\top\left(\widehat{\theta}_t - \theta_*\right)\right| \le \|\mathbf{x}\|_{V_{t-1}^{-1}} \left\|\widehat{\theta}_t - \theta_*\right\|_{V_{t-1}} \qquad \text{Cauchy-Schwarz inequality: } |a^\top b| \le \|a\|\|b\|_*$$

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

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.,*

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

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 $t\ge 0$,*

$$\left\|\sum_{s=1}^{t}\eta_s X_s\right\|_{V_t^{-1}}^2 \le 2R^2\log\left(\frac{\det(V_t)^{\frac{1}{2}}\det(V_0)^{-\frac{1}{2}}}{\delta}\right).$$

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

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

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

Proof of Estimation Error Bound

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

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

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

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

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

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

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

□

LinUCB: Regret Bound

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

$$\bar{R}_T \leq 2\left(R\sqrt{2\log T + d\log\left(1+\frac{TL^2}{\lambda d}\right)} + \sqrt{\lambda}S\right)\sqrt{Td\log\left(1+\frac{L^2T}{\lambda d}\right)} = \widetilde{\mathcal{O}}\left(d\sqrt{T}\right).$$

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 $1-2\delta$,

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

LinUCB: Regret Bound

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

$$\bar{R}_T = \sum_{t=1}^{T} \left(X_*^\top \theta_* - X_t^\top \theta_*\right) \le 2\beta_T \sum_{t=1}^{T} \|X_t\|_{V_{t-1}^{-1}} \le 2\beta_T \sqrt{T \sum_{t=1}^{T} \|X_t\|_{V_{t-1}^{-1}}^2}$$

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*

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

proved in Lecture 5

$$\bar{R}_T \le 2\beta_T \sqrt{T \sum_{t=1}^{T} \|X_t\|_{V_{t-1}^{-1}}^2} \le 2\beta_T \sqrt{Td \log\left(1 + \frac{L^2 T}{\lambda d}\right)}$$

LinUCB: Regret Bound

Proof. With probability at least $1-2\delta$, $\bar{R}_T \le 2\beta_T \sqrt{Td\log\left(1+\frac{L^2T}{\lambda d}\right)}$

$$\bar{R}_T \le 2\beta_T \sqrt{Td\log\left(1+\frac{L^2T}{\lambda d}\right)} \qquad \beta_t = R\sqrt{2\log\left(\frac{1}{\delta}\right)+d\log\left(1+\frac{tL^2}{\lambda d}\right)}+\sqrt{\lambda}S$$

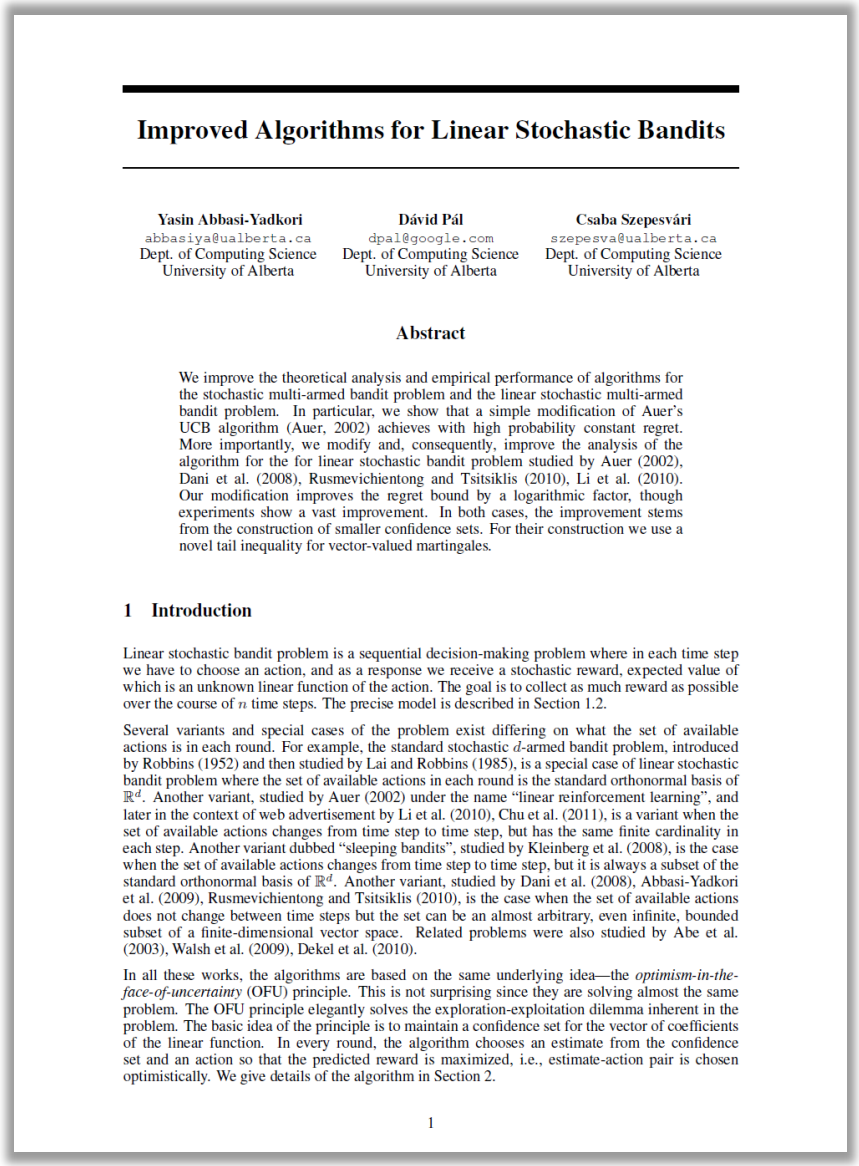
$$\le 2\left(R\sqrt{2\log\left(\frac{1}{\delta}\right)+d\log\left(1+\frac{TL^2}{\lambda d}\right)}+\sqrt{\lambda}S\right)\sqrt{Td\log\left(1+\frac{L^2T}{\lambda d}\right)}$$

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

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

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

□



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.

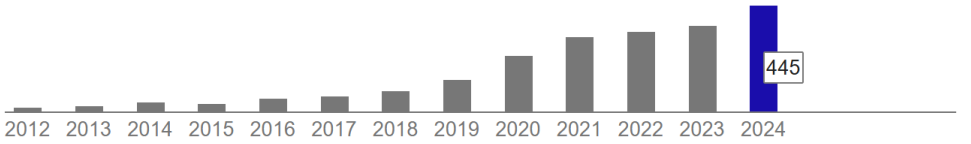
Several variants and special cases of the problem exist differing on what the set of available actions is in each round. For example, the standard stochastic d-armed bandit problem, introduced by Robbins (1952) and then studied by Lai and Robbins (1985), is a special case of linear stochastic bandit 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

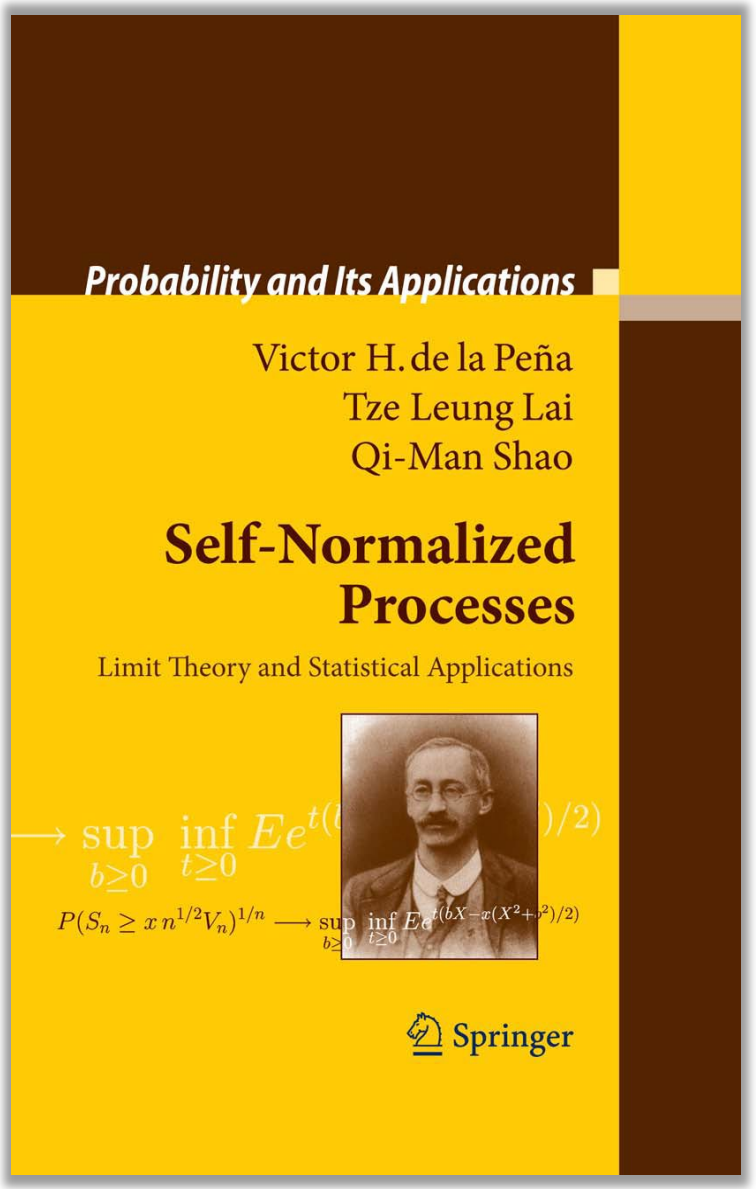
作者	Yasin Abbasi-Yadkori, Csaba Szepesvári, David Pal
发表日期	2011
研讨会论文	Advances in Neural Information Processing Systems
页码范围	2312-2320
简介	We improve the theoretical analysis and empirical performance of algorithms for the stochastic multi-armed bandit problem and the linear stochastic multi-armed bandit 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.
引用总数	被引用次数：2133



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.

Tze Leung Lai (黎子良)

1945 – 2023

斯坦福大学统计系前任系主任

第一位华人COPSS总统奖获得者

Statistical Science
1986, Vol. 1, No. 2, 276–284

The Contributions of Herbert Robbins to Mathematical Statistics

Tze Leung Lai and David Siegmund

Herbert Robbins was born on January 12, 1915, in New Castle, Pennsylvania. In 1931 he entered Harvard College at the age of 16. Although his interests until then had been predominantly literary, he found himself increasingly attracted to mathematics under the influence of Marston Morse, who during many long conversations conveyed a vivid sense of the intellectual challenge of creative work in that field (cf. Page, 1984, p. 7). He received the A.B. summa cum laude in 1935, and the Ph.D. in 1938, also from Harvard. His thesis, in the field of combinatorial topology and written under the supervision of Hassler Whitney, was published in 1941 [3]. (Numbers in brackets refer to Robbins' bibliography at the end of this article.)

After graduation, Robbins worked for a year at the Institute for Advanced Study at Princeton as Marston Morse's assistant. He then spent the next three years at New York University as instructor in mathematics. He became nationally known in 1941 as the coauthor, with Richard Courant, of the classic *What Is Mathematics?* [4]. This important book has influenced generations of mathematics students here and abroad in many editions and translations. To date more than 100,000 copies have been sold.

North Carolina at Chapel Hill. Having read [7] and [10], and greatly impressed by Robbins' mathematical skills, Hotelling offered him the position of associate professor to teach measure theory and probability to the graduate students in the new department. Robbins accepted the position and spent the next six years at Chapel Hill. During this relatively short period Robbins not only studied and developed an increasingly deep interest in statistics, but he also made a number of profound contributions to his new field: complete convergence [12], compound decision theory [25], stochastic approximation [26], and the sequential design of experiments [28], to name a few.

After a Guggenheim Fellowship at the Institute for Advanced Study during 1952–1953, Robbins moved from Chapel Hill to Columbia University as professor and chairman of the Department of Mathematical Statistics. Since 1953, with the exception of the three years 1965–1968 spent at Minnesota, Purdue, Berkeley, and Michigan, he has been at Columbia, where he is Higgins Professor Emeritus of Mathematical Statistics. During this period he has published over 100 papers on a variety of topics in probability and statistics. His most notable contributions include the creation of the empirical Bayes methodology, the theory

Bandit strategies [edit]

A major breakthrough was the construction of optimal population selection strategies, or policies (that possess uniformly maximum convergence rate to the population with highest mean) in the work described below.

Optimal solutions [edit]

Further information: Gittins index

In the paper "Asymptotically efficient adaptive allocation rules", Lai and Robbins[21] (following papers of Robbins and his co-workers going back to Robbins in the year 1952) constructed convergent population selection policies that possess the fastest rate of convergence (to the population with highest mean) for the case that the population reward distributions are the one-parameter exponential family. Then, in Katehakis and Robbins[22] simplifications of the policy and the main proof were given for the case of normal populations with known variances. The next notable progress was obtained by Burnetas and Katehakis in the paper "Optimal adaptive policies for sequential allocation problems",[23] where index based policies with uniformly maximum convergence rate were constructed, under more general conditions that include the case in which

https://en.wikipedia.org/wiki/Multi-armed_bandit

ADVANCES IN APPLIED MATHEMATICS 6, 4–22 (1985)

Asymptotically Efficient Adaptive Allocation Rules*

T. L. LAI AND HERBERT ROBBINS

Department of Statistics, Columbia University, New York, New York 10027

1. INTRODUCTION

Let Π_j $(j = 1, \ldots, k)$ denote statistical populations (treatments, manufacturing processes, etc.) specified respectively by univariate density functions $f(x; \theta_j)$ with respect to some measure ν, where $f(\cdot\,;\cdot)$ is known and the θ_j are unknown parameters belonging to some set Θ. Assume that $\int_{-\infty}^{\infty} |x| f(x; \theta)\, d\nu(x) < \infty$ for all $\theta \in \Theta$. How should we sample $x_1, x_2, \ldots$ sequentially from the k populations in order to achieve the greatest possible expected value of the sum $S_n = x_1 + \cdots + x_n$ as $n \to \infty$? Starting with [3] there has been a considerable literature on this subject, which is often called the multi-armed bandit problem. The name derives from an imagined slot machine with $k \geq 2$ arms. (Ordinary slot machines with one arm are one-armed bandits, since in the long run they are as effective as human bandits in separating the victim from his money.) When an arm is pulled, the player wins a random reward. For each arm j there is an unknown probability distribution Π_j of the reward. The player wants to choose at each stage one of the k arms, the choice depending in some way on the record of previous trials, so as to maximize the long-run total expected reward. A more worthy setting for this problem is in the context of sequential clinical trials, where there are k treatments of unknown efficacy to be used in treating a long sequence of patients.

An *adaptive allocation rule* φ is a sequence of random variables $\varphi_1, \varphi_2, \ldots$ taking values in the set $\{1, \ldots, k\}$ and such that the event $\{\varphi_n = j\}$ ("sample from Π_j at stage n") belongs to the σ-field $\mathcal{F}_{n-1}$ generated by the previous values $\varphi_1, x_1, \ldots, \varphi_{n-1}, x_{n-1}$. Let $\mu(\theta) = \int_{-\infty}^{\infty} x f(x; \theta)\, d\nu(x)$.

*Research supported by the National Science Foundation and the National Institutes of Health. This paper was delivered at the Statistical Research Conference at Cornell University, July 6–9, 1983, in memory of Jack Kiefer and Jacob Wolfowitz.

0196-8858/85 $7.50

Generalized Linear Bandits (GLB)

Extension: want to model ***non-linear*** reward functions.

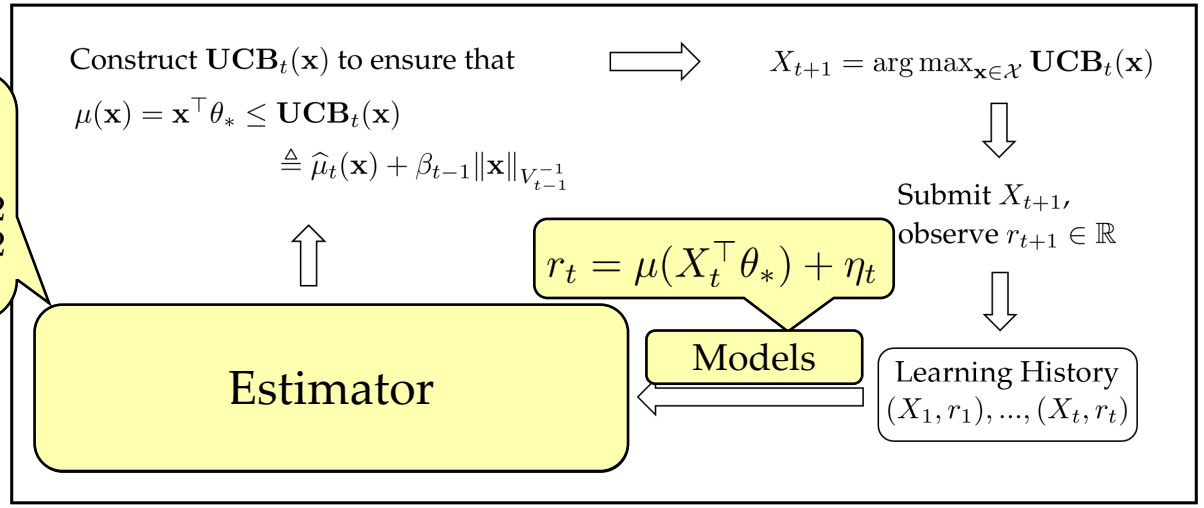
- Generalized linear model: $r_t = \mu(X_t^\top \theta_*) + \eta_t$
  - Link function $\mu : \mathbb{R} \mapsto \mathbb{R}$, which is supposed to be k_μ-Lipschitz

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

    $c_\mu = \inf_{\{\|\theta\|_2 \le S, \mathbf{x} \in \mathcal{X}\}} \dot{\mu}\left(\theta^\top \mathbf{x}\right) > 0$ is an important constant
- GLM-UCB

Maximum quasi-likelihood estimator

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



Generalized Linear Bandits (GLB)

Parametric Bandits:
The Generalized Linear Case

Sarah Filippi
LTCI
Telecom ParisTech et CNRS
Paris, France
filippi@telecom-paristech.fr

Olivier Cappé
LTCI
Telecom ParisTech et CNRS
Paris, France
cappe@telecom-paristech.fr

Aurélien Garivier
LTCI
Telecom ParisTech et CNRS
Paris, France
garivier@telecom-paristech.fr

Csaba Szepesvári
RLAI Laboratory
University of Alberta
Edmonton, Canada
szepesva@ualberta.ca

Abstract

We consider structured multi-armed bandit problems based on the Generalized Linear Model (GLM) framework of statistics. For these bandits, we propose a new algorithm, called GLM-UCB. We derive finite time, high probability bounds on the regret of the algorithm, extending previous analyses developed for the linear bandits to the non-linear case. The analysis highlights a key difficulty in generalizing linear bandit algorithms to the non-linear case, which is solved in GLM-UCB by focusing on the reward space rather than on the parameter space. Moreover, as the actual effectiveness of current parameterized bandit algorithms is often poor in practice, we provide a tuning method based on asymptotic arguments, which leads to significantly better practical performance. We present two numerical experiments on real-world data that illustrate the potential of the GLM-UCB approach.

Keywords: multi-armed bandit, parametric bandits, generalized linear models, UCB, regret minimization.

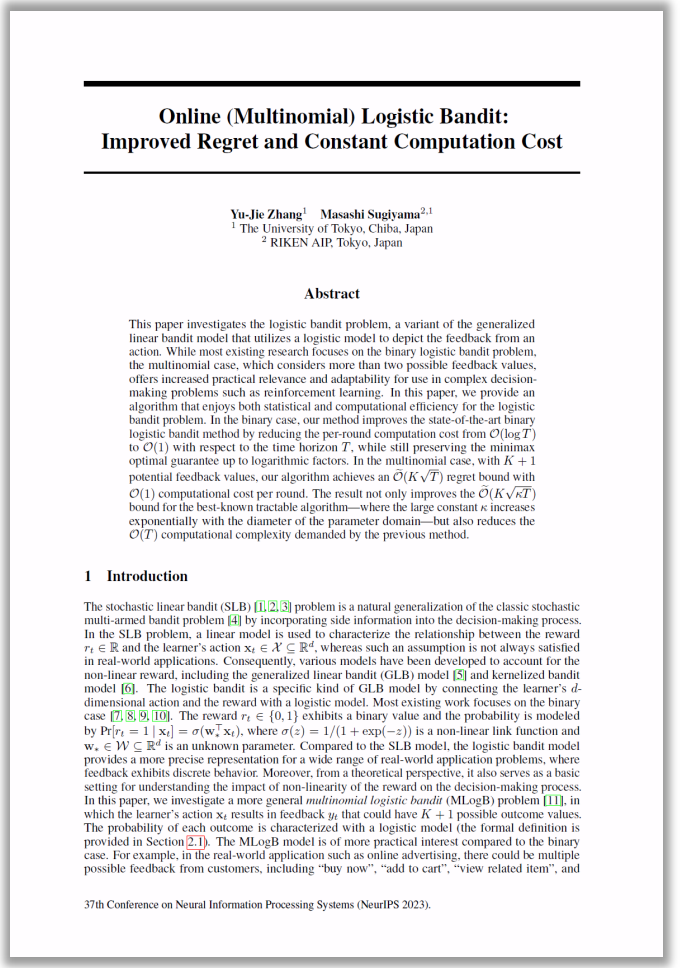
1 Introduction

In the classical K-armed bandit problem, an agent selects at each time step one of the K arms and receives a reward that depends on the chosen action. The aim of the agent is to choose the sequence of arms to be played so as to maximize the cumulated reward. There is a fundamental trade-off between gathering experimental data about the reward distribution (exploration) and exploiting the arm which seems to be the most promising.

In the basic multi-armed bandit problem, also called the independent bandits problem, the rewards are assumed to be random and distributed independently according to a probability distribution that is specific to each arm –see [1, 2, 3, 4] and references therein. Recently, *structured bandit problems* in which the distributions of the rewards pertaining to each arm are connected by a common unknown parameter have received much attention [5, 6, 7, 8, 9]. This model is motivated by the many practical applications where the number of arms is large, but the payoffs are interrelated. Up to know, two different models were studied in the literature along these lines. In one model, in each times step, a side-information, or context, is given to the agent first. The payoffs of the arms depend both on this side information and the index of the arm. Thus the optimal arm changes with the context [5, 6, 9]. In the second, simpler model, that we are also interested in here, there is no side-information, but the agent is given a model that describes the possible relations

1

[NIPS'10 Parametric Bandits: The Generalized Linear Case]



Online (Multinomial) Logistic Bandit:
Improved Regret and Constant Computation Cost

Yu-Jie Zhang[1] **Masashi Sugiyama**[2,1]
[1] The University of Tokyo, Chiba, Japan
[2] RIKEN AIP, Tokyo, Japan

Abstract

This paper investigates the logistic bandit problem, a variant of the generalized linear bandit model that utilizes a logistic model to depict the feedback from an action. While most existing research focuses on the binary logistic bandit problem, the multinomial case, which considers more than two possible feedback values, offers increased practical relevance and adaptability for use in complex decision-making problems such as reinforcement learning. In this paper, we provide an algorithm that enjoys both statistical and computational efficiency for the logistic bandit problem. In the binary case, our method improves the state-of-the-art binary logistic bandit method by reducing the per-round computation cost from $\mathcal{O}(\log T)$ to $\mathcal{O}(1)$ with respect to the time horizon T, while still preserving the minimax optimal guarantee up to logarithmic factors. In the multinomial case, with $K+1$ potential feedback values, our algorithm achieves an $\widetilde{\mathcal{O}}(K\sqrt{T})$ regret bound with $\mathcal{O}(1)$ computational cost per round. The result not only improves the $\widetilde{\mathcal{O}}(K\sqrt{\kappa T})$ bound for the best-known tractable algorithm—where the large constant κ increases exponentially with the diameter of the parameter domain—but also reduces the $\mathcal{O}(T)$ computational complexity demanded by the previous method.

1 Introduction

The stochastic linear bandit (SLB) [1, 2, 3] problem is a natural generalization of the classic stochastic multi-armed bandit problem [4] by incorporating side information into the decision-making process. In the SLB problem, a linear model is used to characterize the relationship between the reward $r_t \in \mathbb{R}$ and the learner's action $\mathbf{x}_t \in \mathcal{X} \subseteq \mathbb{R}^d$, whereas such an assumption is not always satisfied in real-world applications. Consequently, various models have been developed to account for the non-linear reward, including the generalized linear bandit (GLB) model [5] and kernelized bandit model [6]. The logistic bandit is a specific kind of GLB model by connecting the learner's d-dimensional action and the reward with a logistic model. Most existing work focuses on the binary case [7, 8, 9, 10]. The reward $r_t \in \{0, 1\}$ exhibits a binary value and the probability is modeled by $\Pr[r_t = 1 \mid \mathbf{x}_t] = \sigma(\mathbf{w}_*^\top \mathbf{x}_t)$, where $\sigma(z) = 1/(1+\exp(-z))$ is a non-linear link function and $\mathbf{w}_* \in \mathcal{W} \subseteq \mathbb{R}^d$ is an unknown parameter. Compared to the SLB model, the logistic bandit model provides a more precise representation for a wide range of real-world application problems, where feedback exhibits discrete behavior. Moreover, from a theoretical perspective, it also serves as a basic setting for understanding the impact of non-linearity of the reward on the decision-making process. In this paper, we investigate a more general *multinomial logistic bandit* (MLogB) problem [11], in which the learner's action $\mathbf{x}_t$ results in feedback y_t that could have $K+1$ possible outcome values. The probability of each outcome is characterized with a logistic model (the formal definition is provided in Section 2.1). The MLogB model is of more practical interest compared to the binary case. For example, in the real-world application such as online advertising, there could be multiple possible feedback from customers, including "buy now", "add to cart", "view related item", and

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

[NeurIPS'23 Online (Multinomial) Logistic Bandit: Improved Regret and Constant Computation Cost]

Part 3. Advanced Topics

- Best of Both Worlds
- Bayesian optimization
- Linear (Mixture) MDPs

Advanced Topic: Best of Both Worlds

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

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

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 the adversarial setting.
- Exp3: can't have adaptive regret bound in the stochastic case.

⇨ Surprisingly, 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 heuristical 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-th... 2007) or l... (Lizotte ... to this p... paradigm ... maximize ... exploration ... (Chaloner ... is to be ex... ...ew evaluations 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 d... ...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)

Appearing in *Proceedings of the 27th International Conference on Machine Learning*, Haifa, Israel, 2010.

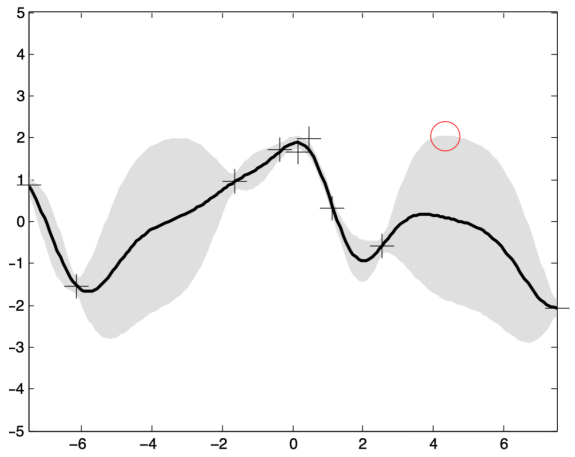
ICML 2020 ten-year Test of Time Award!

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

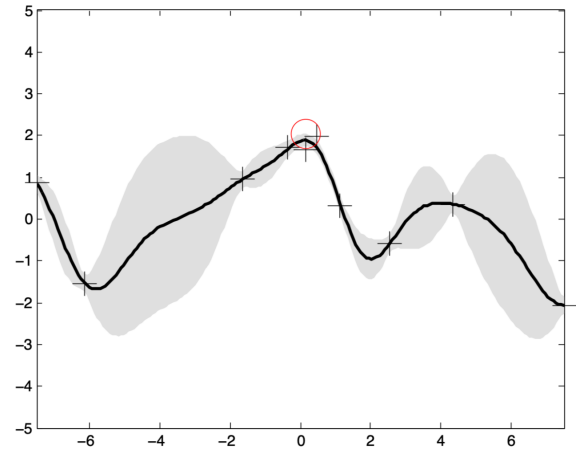
$f(\mathbf{x})$ belongs to RKHS with $k(\mathbf{x}, \mathbf{x}') = \sum_{m=1}^{|\mathcal{H}|} \varphi_m(\mathbf{x})\varphi_m(\mathbf{x}')$

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

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



Iteration t



Iteration $t+1$

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

Advanced Topic: Linear (Mixture) 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

UCB-VI for Linear MDPs

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

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

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

  2. Construct the exploration bonuses

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

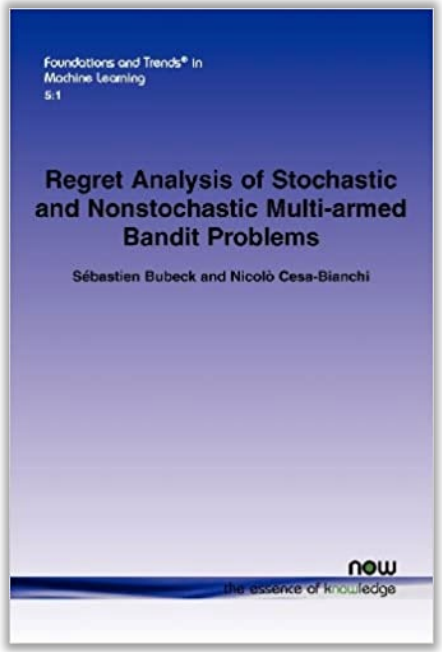
  3. Run optimistic value iterations, and update greedy policy

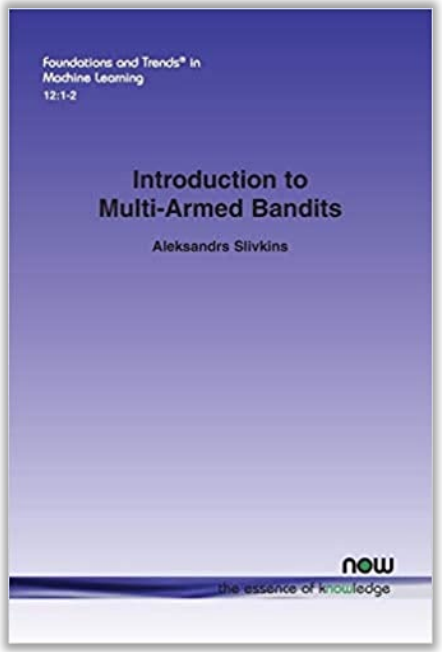
11

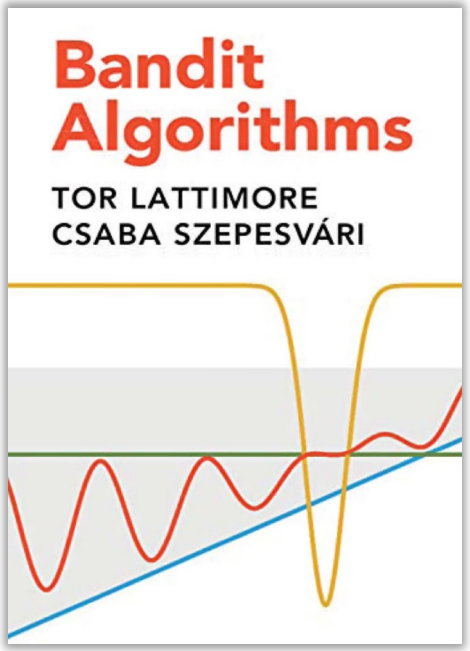
Reference: Yu-Xiang Wang's course CS292F Lecture 10 Exploration IV: Linear MDP

Many more results

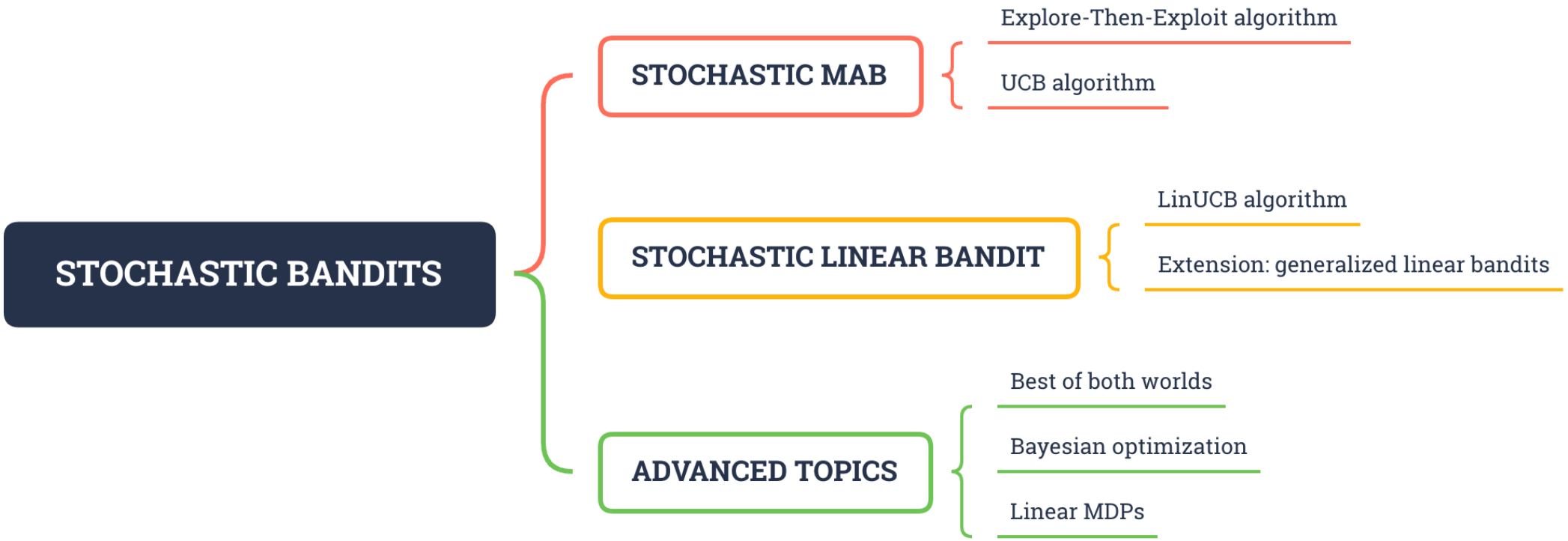
- Techniques developed in bandit problems have been applied in many areas, including machine learning, statistics, operational research, and information theory [Bubeck and Cesa-Bianchi, 2012; Slivkins, 2019; Lattimore and Szepesvári, 2020].







Summary



Q & A

Thanks!