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

Lecture 12. Stochastic Bandits II

Advanced Optimization (Fall 2025)

Peng Zhao

zhaop@lamda.nju.edu.cn

Nanjing University

Outline

- Linear Bandits
- Advanced Topics

Part 1. Linear Bandits

- Formulation
- Estimator and UCB Construction
- LinUCB and Regret Analysis

Bandits: Interactive Learning

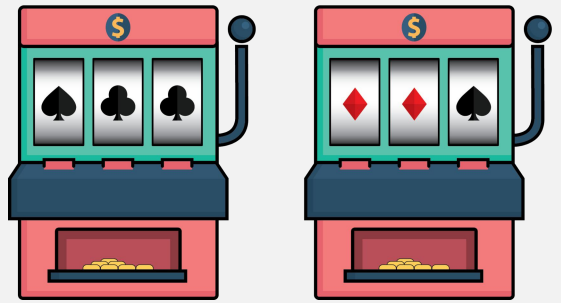
- Multi-armed bandits: a simplest formulation for bandit problems

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

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

(2) environment reveals a reward $r_t(a_t) \sim$ distribution $\mathcal{D}_{a_t}$;

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



The goal is to minimize the ***regret*** :

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

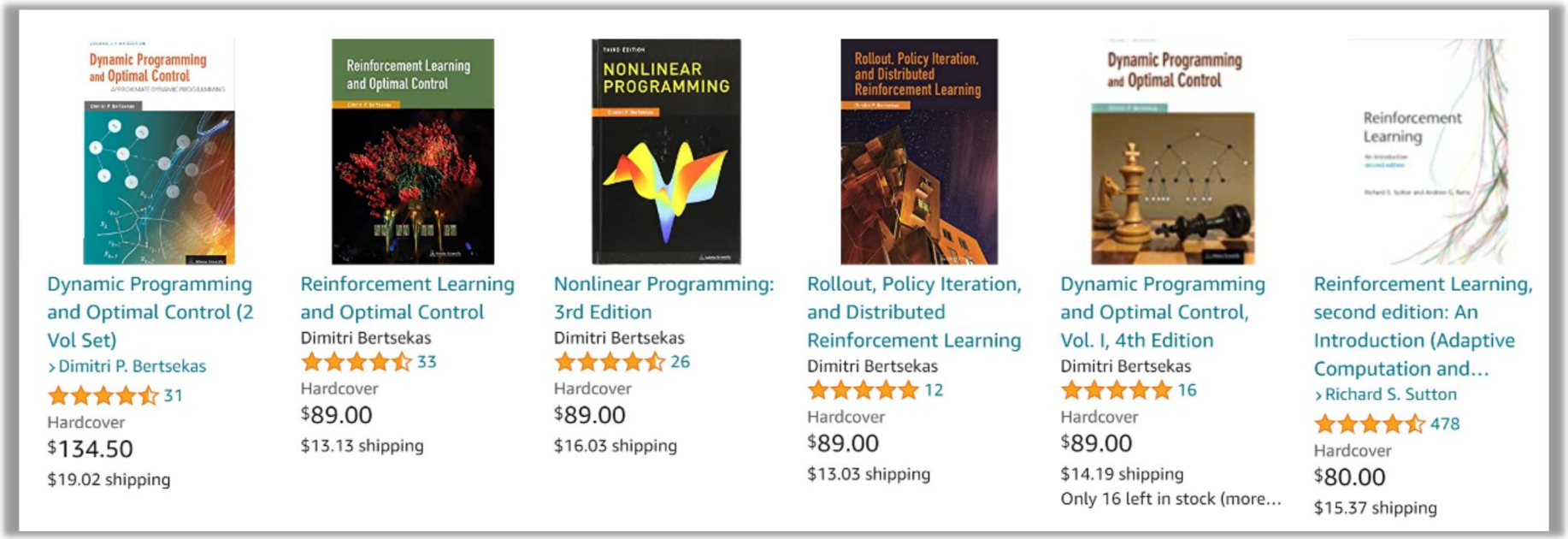
Exploration-Exploitation tradeoff

- **Exploitation**: pull the best arm so far
- **Exploration**: try other arms that may be better

i.e., difference between the cumulative reward of the best arm and that obtained by the bandit algorithm

Stochastic Linear Bandits

- A ubiquitous problem in real life: ***feature information***



- 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: $r_t(\mathbf{x}) = \mathbf{x}^\top \theta_* + \eta_t$
  - 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;

Stochastic LB: Formulation

Stochastic Linear Bandits

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

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 $r_t \in \mathbb{R}$.

- Linear modeling assumption: $r_t(\mathbf{x}) = \mathbf{x}^\top \theta_* + \eta_t$
- Regret measure: $\bar{R}_T \triangleq T \max_{\mathbf{x} \in \mathcal{X}} \mathbf{x}^\top \theta_* - \sum_{t=1}^{T} X_t^\top \theta_*$

For simplicity, we use a fixed arm set $\mathcal{X}$, and results in this lecture can be extended to changing set, i.e., $\mathcal{X}_t$.

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	$\mathbb{E}[r(a)] = \mu(a)$ $\forall t \in [T], a \in [K], r_t(a) \in [0,1]$	$r_t = X_t^\top \theta_* + \eta_t \qquad \mu(\mathbf{x}) = \mathbf{x}^\top \theta_*$ 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: Linear Bandits with UCB

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 $\widehat{\theta}_t$

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

- ***Estimator***: construct an estimation of the reward (linearly parameterized)
- ***Arm selection***: upper confidence bound selection

$$X_{t+1} = \arg\max_{\mathbf{x} \in \mathcal{X}} \left\{ \mathbf{x}^\top \widehat{\theta}_t + \beta_{t-1} \|\mathbf{x}\|_{V_{t-1}^{-1}} \right\}$$

exploit *explore*

LinUCB: Estimator

- Input: historical feature-reward pairs

$$\{(X_1, r_1), (X_2, r_2), \ldots, (X_{t-1}, r_{t-1})\}$$

- 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: $\widehat{\theta}_t = V_{t-1}^{-1} b_{t-1}$

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

- This LS estimator can be updated incrementally.

"one-pass" incremental update

online data item is processed only once, don't need to store it along the time

$$\widehat{\theta}_{t+1} = V_t^{-1} b_t, \text{ where}$$
$$V_t = V_{t-1} + X_t X_t^\top$$
$$b_t = b_{t-1} + r_t X_t^\top$$

LinUCB: Estimator

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: UCB construction and selection

Key question: how to construct a proper UCB?

Construct $\mathbf{UCB}_t(\mathbf{x})$ to ensure that

$$\mu(\mathbf{x}) = \mathbf{x}^\top \theta_* \le \mathbf{UCB}_t(\mathbf{x})$$

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

exploit *explore*

$\Longrightarrow$ $X_{t+1} = \arg\max_{\mathbf{x}\in\mathcal{X}} \mathbf{UCB}_t(\mathbf{x})$

$\Downarrow$

Submit X_{t+1}, observe $r_{t+1} \in \mathbb{R}$

$\Downarrow$

Learning History $(X_1, r_1), ..., (X_t, r_t)$

$\Longleftarrow$ $r_t = X_t^\top \theta_* + \eta_t$

Regularized Least Square Estimator

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

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

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 $t \in [T]$*

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

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

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

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

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

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

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

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

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

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

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

□

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

$$\bar{R}_T = \sum_{t=1}^{T} \left(X_*^\top \theta_* - X_t^\top \theta_*\right) \le 2\beta_T \sum_{t=1}^{T} \|X_t\|_{V_{t-1}^{-1}} \le 2\beta_T \sqrt{T \sum_{t=1}^{T} \|X_t\|^2_{V_{t-1}^{-1}}} \qquad V_t = \lambda I + \textstyle\sum_{s=1}^{t} X_s X_s^\top$$

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*

$$\sum_{t=1}^{T} \|X_t\|^2_{V_t^{-1}} \le d \log\left(1 + \frac{L^2 T}{\lambda d}\right) \qquad \text{proved in Lecture 6}$$

$$\bar{R}_T \le 2\beta_T \sqrt{T \sum_{t=1}^{T} \|X_t\|^2_{V_{t-1}^{-1}}} \le 2\beta_T \sqrt{Td \log\left(1 + \frac{L^2 T}{\lambda d}\right)} \quad \text{(actually requires slight twists on Lemma 4)}$$

LinUCB: Regret Bound

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

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

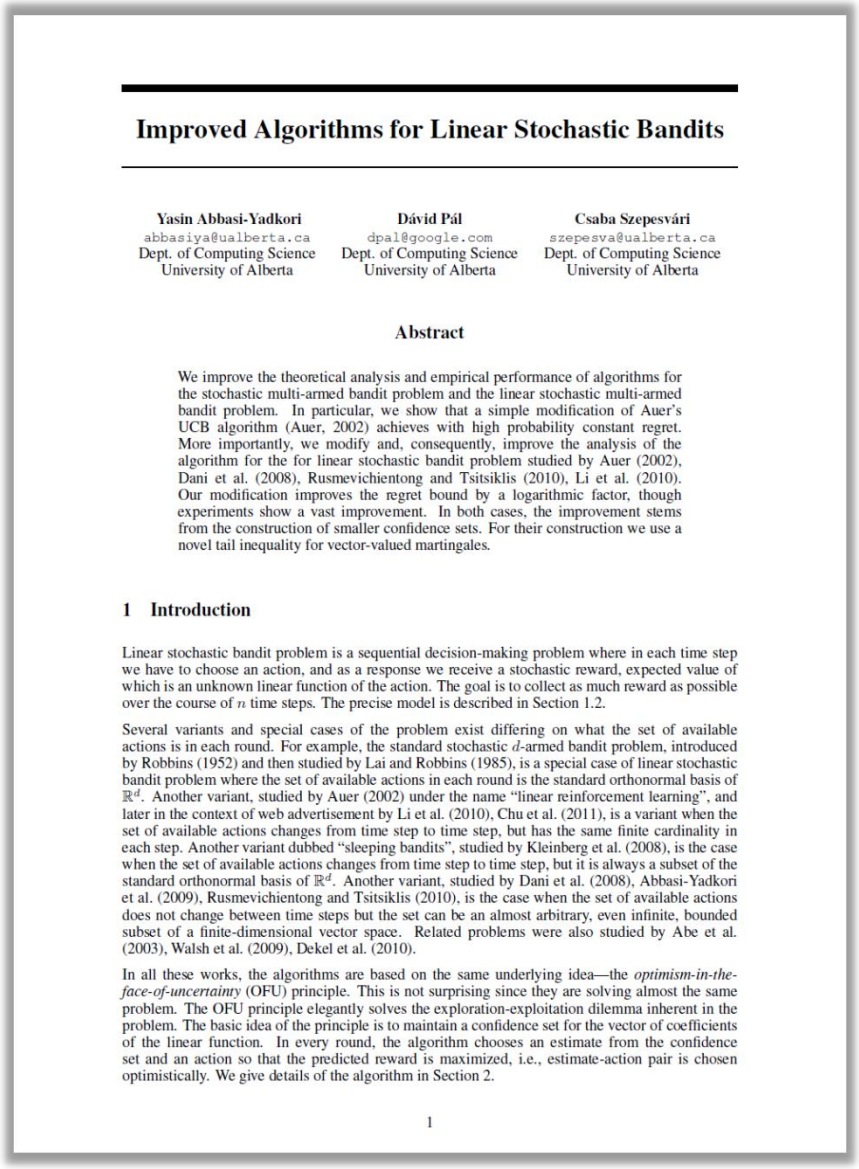
$$\le 2\left(R\sqrt{2\log\frac{1}{\delta}+d\log\left(1+\frac{TL^2}{\lambda d}\right)}+\sqrt{\lambda}S\right)\sqrt{Td\log\left(1+\frac{L^2T}{\lambda d}\right)}$$

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.

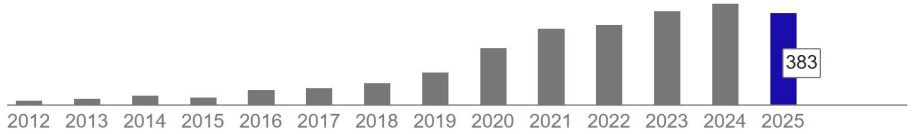
Several variants and special cases of the problem exist differing on what the set of available actions is in each round. For example, the standard stochastic d-armed bandit problem, introduced by Robbins (1952) and then studied by Lai and Robbins (1985), is a special case of linear stochastic bandit 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, Dávid Pál
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：2510



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.

Part 2. Advanced Topics

- Self-Normalized Concentration
- Connection of Linear bandits to RL theory
- More generalized model

Part 2. Advanced Topics

- Self-Normalized Concentration
- Connection of Linear bandits to RL theory
- More Generalized Model

Proof of Estimation Error Bound

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

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

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

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.,*

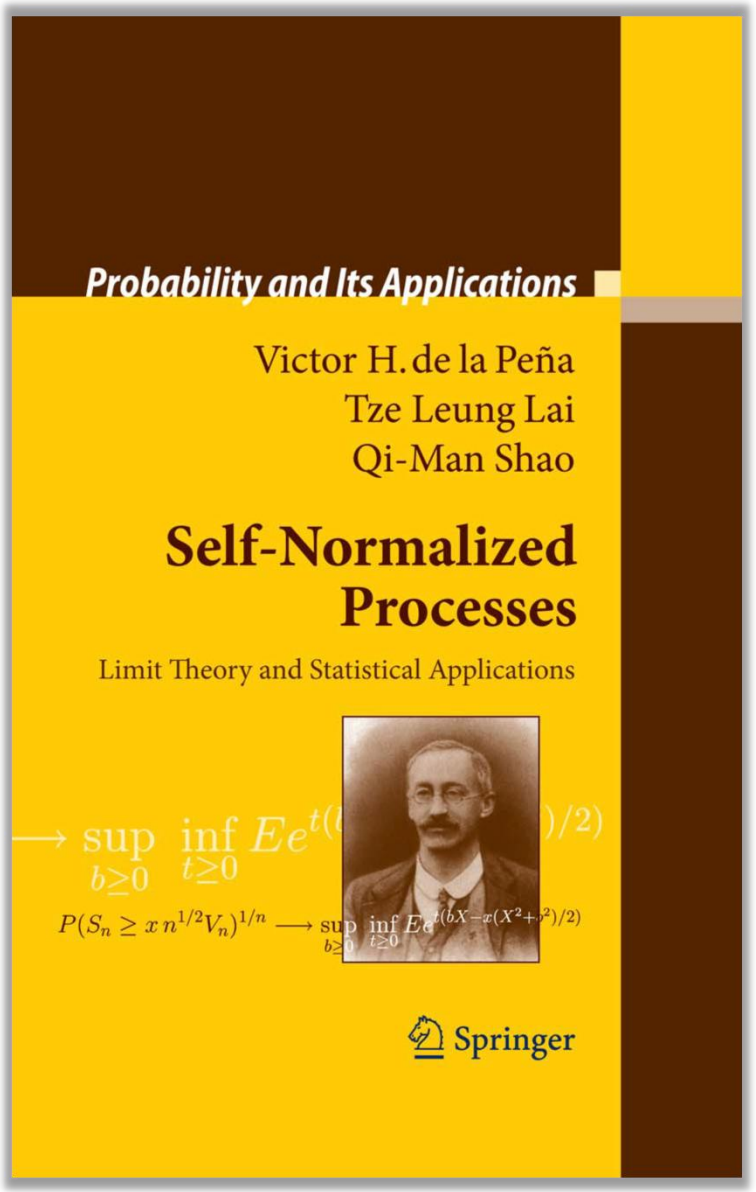
$$\forall \lambda \in \mathbb{R}, \;\; \mathbb{E}\left[\exp(\lambda \eta_t) \mid X_{1:t}, \eta_{1:t-1}\right] \leq \exp\left(\frac{\lambda^2 R^2}{2}\right).$$

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 $t \geq 0$,

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



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

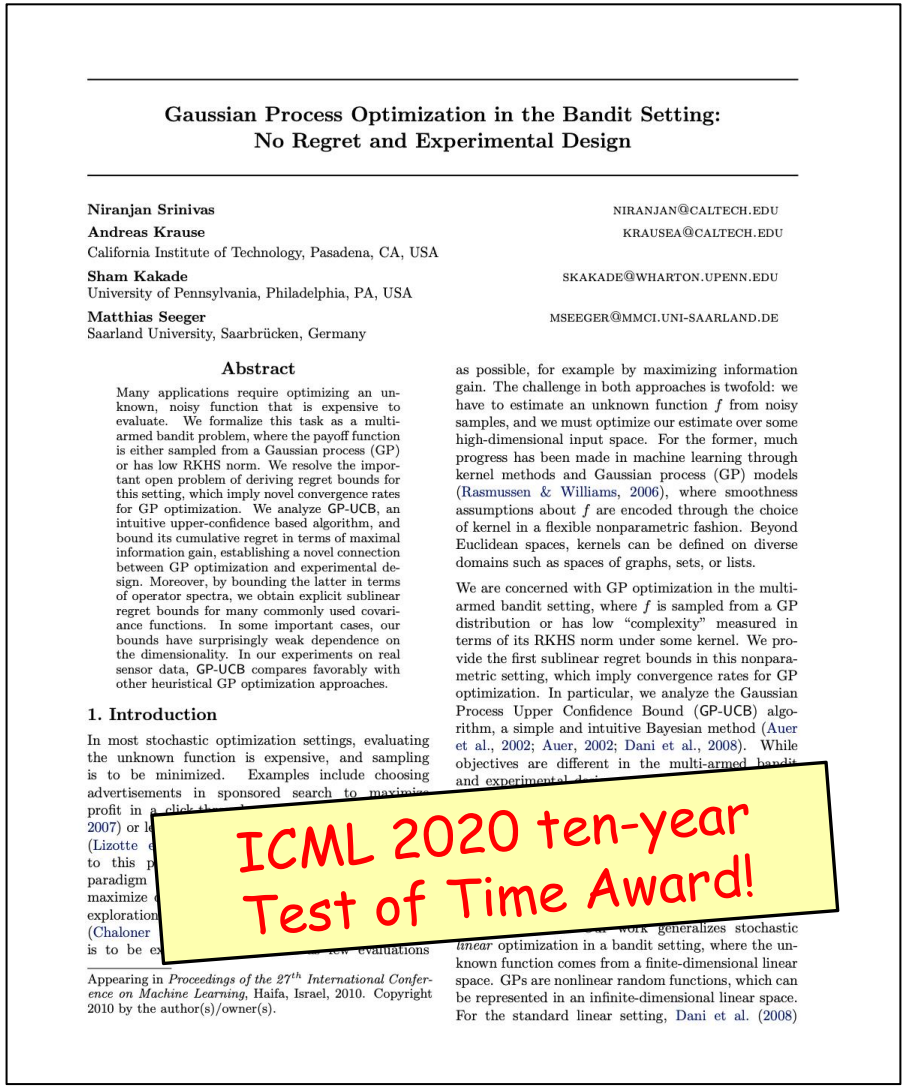
Let Π_j $(j = 1, \ldots, k)$ denote statistical populations (treatments, manufacturing processes, etc.) specified respectively by univariate density functions $f(x; \theta_j)$ with respect to some measure ν, where $f(\cdot;\cdot)$ is known and the θ_j are unknown parameters belonging to some set Θ. Assume 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
Copyright © 1985 by Academic Press, Inc.
All rights of reproduction in any form reserved.

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.

Copyright 2010 by the author(s)/owner(s).

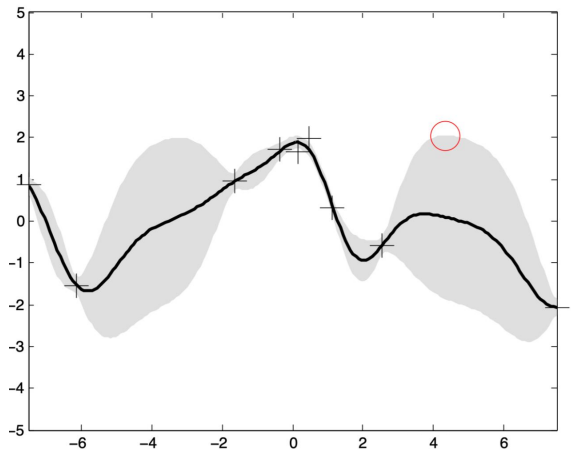
ICML 2020 ten-year Test of Time Award!

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

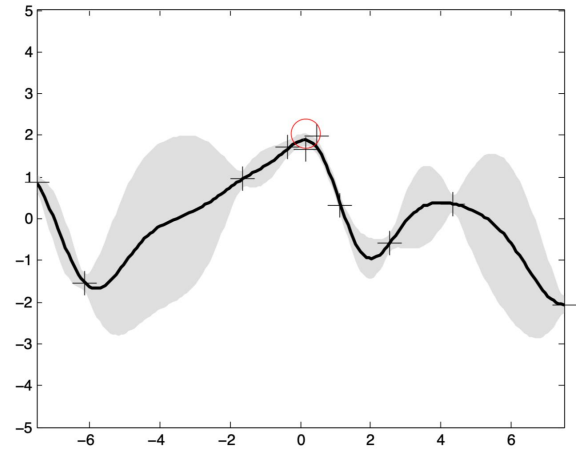
$f(\mathbf{x})$ belongs to RKHS with $k(\mathbf{x}, \mathbf{x}') = \sum_{m=1}^{|\mathcal{H}|} \varphi_m(\mathbf{x})\varphi_m(\mathbf{x}')$

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 $t+1$

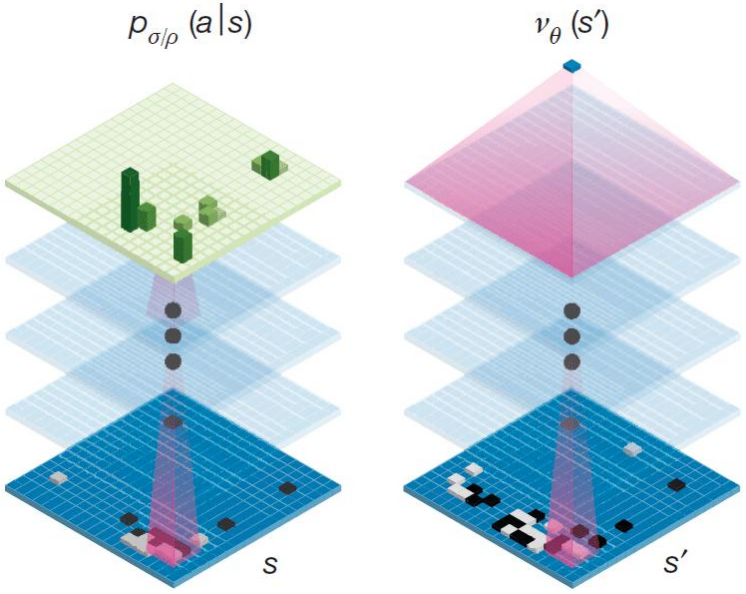
Reference: Gaussian Process Optimization in the Bandit Setting: No Regret and Experimental Design. ICML 2010.

Part 2. Advanced Topics

- Self-Normalized Concentration
- Connection of Linear bandits to RL theory
- More Generalized Model

Linear bandits for RL Theory

Function Approximation



a technique with huge success (especially by involving DNN) , crucially useful for the AlphaGo's success

Provably Efficient Reinforcement Learning with Linear Function Approximation

Chi Jin
University of California, Berkeley
chijin@cs.berkeley.edu

Zhuoran Yang
Princeton University
zy6@princeton.edu

Zhaoran Wang
Northwestern University
zhaoranwang@gmail.com

Michael I. Jordan
University of California, Berkeley
jordan@cs.berkeley.edu

COLT 2020

Reinforcement Learning in Feature Space: Matrix Bandit, Kernels, and Regret Bound

Lin F. Yang
Princeton University
lin.yang@princeton.edu

Mengdi Wang
Princeton University
mengdiw@princeton.edu

June 14, 2019

ICML 2020

Function Approximation

Tabular MDPs: usually maintain a table to store values for all states (or state-action pairs), which scales with state number S and action number A.

Figure 1

We discover through experience that this state is bad

Figure 2

In tabular methods, we know nothing about this state.

Figure 3

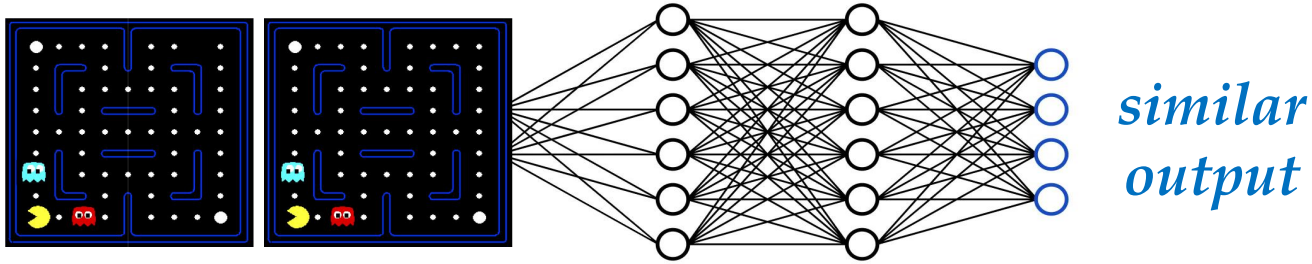
*We know **nothing** about this state either!*

But this has a poor scalability in practical scenarios; and many structures yet to exploit…

Function Approximation

RL Function approximation: approximate using a parameterized function.

- To avoid bad dependence on #states S, #action A in tabular MDPs
- Describe states (or state-actions) using feature representations in $\mathbb{R}^d$.
- A modern choice: DNN as a feature representer



parameterize MDP model with a low-dimensional representation

⇨ regret bound should not dependent on S or A, but rather the intrinsic dimension d

Deploying bandit techniques

- Linear Mixture MDPs

$$r_h(x,a) = \langle \boldsymbol{\phi}(x,a), \boldsymbol{\theta}_h^* \rangle$$

$$\mathbb{P}_h\left(s' \mid s,a\right) = \langle \psi\left(s' \mid s,a\right), \mathbf{w}_h^* \rangle$$

- $\boldsymbol{\phi} : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}^d$ is known feature map
- $\psi : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}^d$ is known feature map
- $\{\boldsymbol{\theta}_h^*\}_{h=1}^H$ is the unknown reward parameter
- $\{\mathbf{w}_h^*\}_{h=1}^H$ is the unknown transition parameter

- Linear Bandits

(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: $r_t(x) = x^\top \theta_* + \eta_t$

Linear bandits serve as a foundational tool for understanding linear mixture MDPs

Linear Mixture MDPs

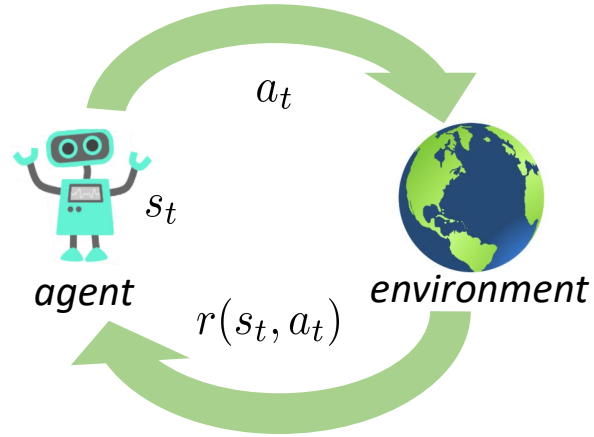
- Least square for parameter estimation

Reward estimation

$$\widehat{\boldsymbol{\theta}}_h = \arg\min_{\theta \in \mathbb{R}^d} \left\{ \frac{\lambda_\theta}{2} \|\theta\|_2^2 + \sum_{j=1}^{k-1} \left(r_h(s_h, a_h) - \phi(s_h, a_h)^\top \theta \right)^2 \right\}$$

Transition estimation

$$\widehat{\mathbf{w}}_h = \arg\min_{\mathbf{w} \in \mathbb{R}^d} \left\{ \frac{\lambda_{\mathbf{w}}}{2} \|\mathbf{w}\|_2^2 + \sum_{j=1}^{k-1} \left(\langle \psi_{h+1}(s_h, a_h), \mathbf{w} \rangle - V_{h+1}(s_{h+1}) \right)^2 \right\}$$



$$V_h^\pi(s) = \mathbb{E}_\pi \left[\sum_{h'=h}^{H} r_{h'} \left(s_{h'}, a_{h'} \right) \mid s_h = s \right]$$

Estimation error

$$\|\widehat{\mathbf{w}}_h - \mathbf{w}_h\|_{\Sigma_h} \leq \mathcal{O}\left(\sqrt{d} H (\log(t/\delta))^2 \right)$$

Regret bound

$$\mathrm{REG}_T \leq \widetilde{\mathcal{O}}\left(d\sqrt{H^3 K} \right)$$

K: the number of epsiodes

H: the length of each epsiode

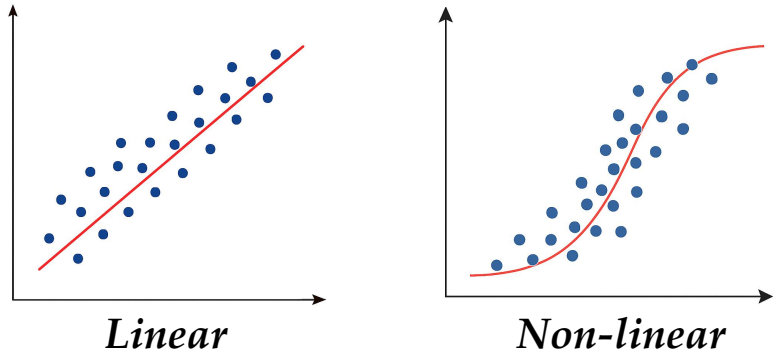
Part 2. Advanced Topics

- Self-Normalized Concentration
- Connection of Linear bandits to RL theory
- More Generalized Model

Beyond: More Expressivity

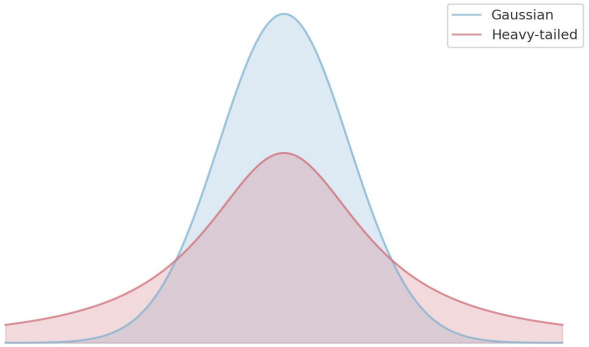
(i) Generalized linear bandits

$$r_t = \mu(X_t^\top \theta_*) + \eta_t$$



(ii) Heavy-tailed linear bandits

$$r_t = X_t^\top \theta_* + \eta_t$$



Goal: computationally efficient (better "one-pass") algorithm with optimal regret

- [Wang-Zhang-Z-Zhou, ICML'25] Heavy-Tailed Linear Bandits: Huber Regression with One-Pass Update.
- [Zhang-Xu-Z-Sugiyama, NeurIPS'25] Generalized Linear Bandits: Almost Optimal Regret with One-Pass Update.

① GLB: Problem Formulation

Generalized Linear Bandits

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

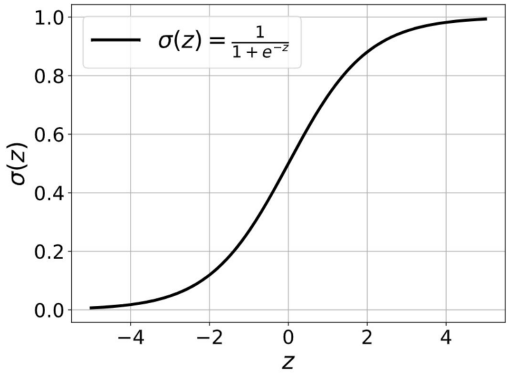
(1) the player first chooses an arm X_t from arm set $\mathcal{X}$;

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

- Generalized linear reward function: $r_t = \mu(X_t^\top \theta_*) + \eta_t$

Examples: logistic bandit

$$r_t = \begin{cases} 0 \text{ (“not click”)} & \text{w.p. } \mu(X_t^\top \theta_*) \\ 1 \text{ (“click”)} & \text{otherwise} \end{cases} \quad \Rightarrow \quad \mu(z) = \frac{1}{1+\exp(-z)}$$



① GLB: Existing Algorithm

- GLM-UCB Algorithm [Filippi et al., NIPS 2010]
  - ***Estimator***: maximum likelihood estimator

$$\widehat{\theta}_t = \arg\min_{\theta\in\Theta} \frac{\lambda}{2}\|\theta\|_2^2 + \sum_{s=1}^{t-1} \ell_s^{\mathrm{GLB}}(\theta), \text{ with } \ell_s^{\mathrm{GLB}}(\theta) = -\log \mathbb{P}_\theta\left(r_{s+1} \mid X_s\right)$$

Estimation error: $\left|\mu(\mathbf{x}^\top\widehat{\theta}_t) - \mu(\mathbf{x}^\top\theta_*)\right| \le \frac{k_\mu}{c_\mu}\beta_{t-1}\|\mathbf{x}\|_{V_{t-1}^{-1}}$

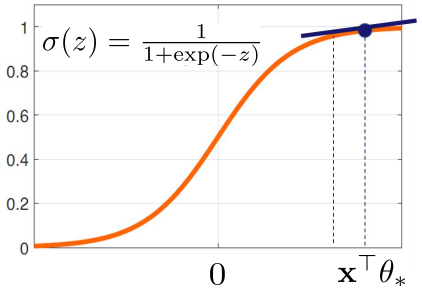
  - ***Arm selection***: upper confidence bound

$$X_t = \arg\max_{\mathbf{x}\in\mathcal{X}} \left\{\mu(\mathbf{x}^\top\widehat{\theta}_t) + \beta_{t-1}\|\mathbf{x}\|_{V_{t-1}^{-1}}\right\}$$

Regret bound: $\mathrm{REG}_T \le \widetilde{\mathcal{O}}\left(\frac{k_\mu}{c_\mu} d\sqrt{T}\right)$

* Note: $c_\mu \le \mu'(z) \le k_\mu, \forall z \in [-S, S]$

The non-linear term k_μ/c_μ can be as large as $\mathcal{O}(e^S)$!



There are recent works using "warm-up" to remove κ, but is still not one-pass

② Hvt-LB: Problem Formulation

- Linear reward with sub-Gaussian noise $r_t = X_t^\top \theta_* + \eta_t$

Assumption 1 (sub-Gaussian noise). The noise η_t is conditionally R-sub-Gaussian for some $R \geq 0$ i.e.

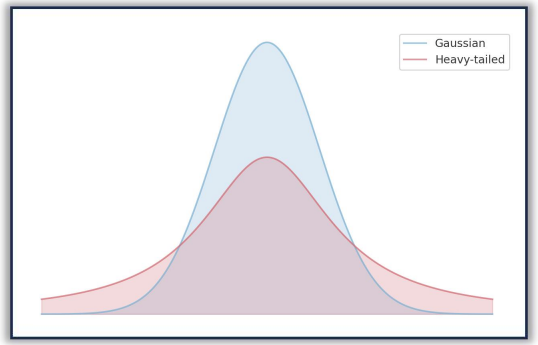
$$\forall \lambda \in \mathbb{R}, \mathbb{E}\left[\exp\left(\lambda \eta_t\right) \mid X_{1:t}, \eta_{1:t-1}\right] \leq \exp\left(\frac{\lambda^2 R^2}{2}\right).$$

In many scenarios, the noise can be heavy-tailed !

- Linear bandits with heavy-tailed noise

Assumption 2 (heavy-tailed noise). The noise $\{\eta_t, \mathcal{F}_t\}$ is is martingale difference ($\mathbb{E}\left[\eta_t \mid \mathcal{F}_{t-1}\right] = 0$), and satisfies that for some $\varepsilon \in (0, 1], \nu_t > 0$,

$$\mathbb{E}\left[\left|\eta_t\right|^{1+\varepsilon} \,\middle|\, \mathcal{F}_{t-1}\right] \leq \nu_t^{1+\varepsilon}.$$



② Hvt-LB: Existing Algorithm

- HEAVY-OFUL Algorithm [Huang et al., NeurIPS 2023]

  - ***Estimator***: adaptive Huber regression

$$\widehat{\theta}_t = \arg\min_{\theta\in\Theta} \frac{\lambda}{2}\|\theta\|_2^2 + \sum_{s=1}^{t-1} \ell_s^{\mathrm{Hvt}}(\theta)$$

Estimation error: $\left\|\hat{\theta}_{t+1} - \theta_*\right\|_{V_t} \le \widetilde{\mathcal{O}}\left(t^{\frac{1-\varepsilon}{2(1+\varepsilon)}}\right)$

Huber loss is defined using a threshold $\tau_s > 0$,

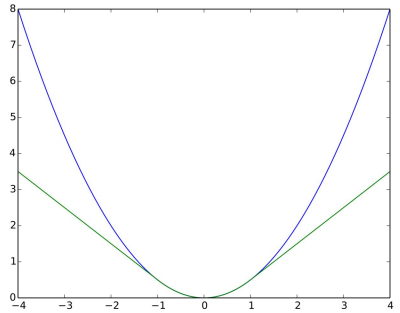
$$\ell_s^{\mathrm{Hvt}}(\theta) = \begin{cases} \frac{z_s(\theta)^2}{2} & \text{if } |z_s(\theta)| \le \tau_s, \\ \tau_s|z_s(\theta)| - \frac{\tau_s^2}{2} & \text{if } |z_s(\theta)| > \tau_s, \end{cases}$$

with $z_s(\theta) = \frac{r_s - X_s^\top \theta}{\sigma_s}$.

  - ***Arm selection***: upper confidence bound

$$X_t = \arg\max_{\mathbf{x}\in\mathcal{X}} \left\{\mathbf{x}^\top \widehat{\theta}_t + \beta_{t-1}\|\mathbf{x}\|_{V_{t-1}^{-1}}\right\}$$

Regret bound: $\mathrm{REG}_T \le \widetilde{\mathcal{O}}\left(dT^{\frac{1}{1+\varepsilon}}\right)$



Squared loss

Huber loss

reduce penalty for large deviation

Efficiency Concerns

- **Stochastic LB**: least squares (closed-form solution)

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

$\Longrightarrow$ *one-pass update*

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

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

- **Generalized LB**: maximum likelihood estimator

$$\widehat{\theta}_t = \arg\min_{\theta \in \Theta} \frac{\lambda}{2}\|\theta\|_2^2 + \sum_{s=1}^{t-1} \ell_s^{\mathrm{GLB}}(\theta)$$

- **Heavy-tailed LB**: adaptive Huber regression

$$\widehat{\theta}_t = \arg\min_{\theta \in \Theta} \frac{\lambda}{2}\|\theta\|_2^2 + \sum_{s=1}^{t} \ell_s^{\mathrm{Hvt}}(\theta)$$

inefficiency due to non-quadratic loss

$\Longrightarrow$ **The cost at round** t

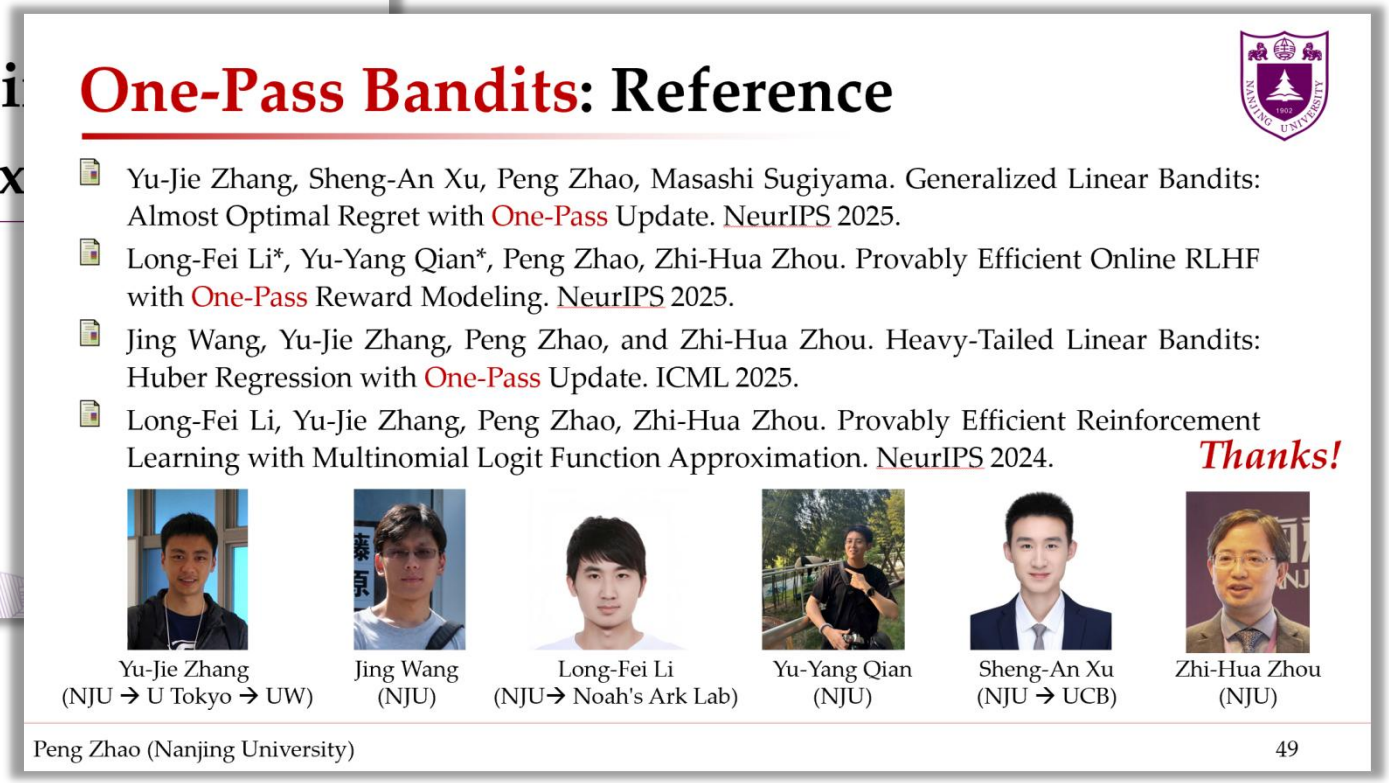
Computational cost: $\mathcal{O}(t \log T)$

Storage cost: $\mathcal{O}(t)$

infeasible!

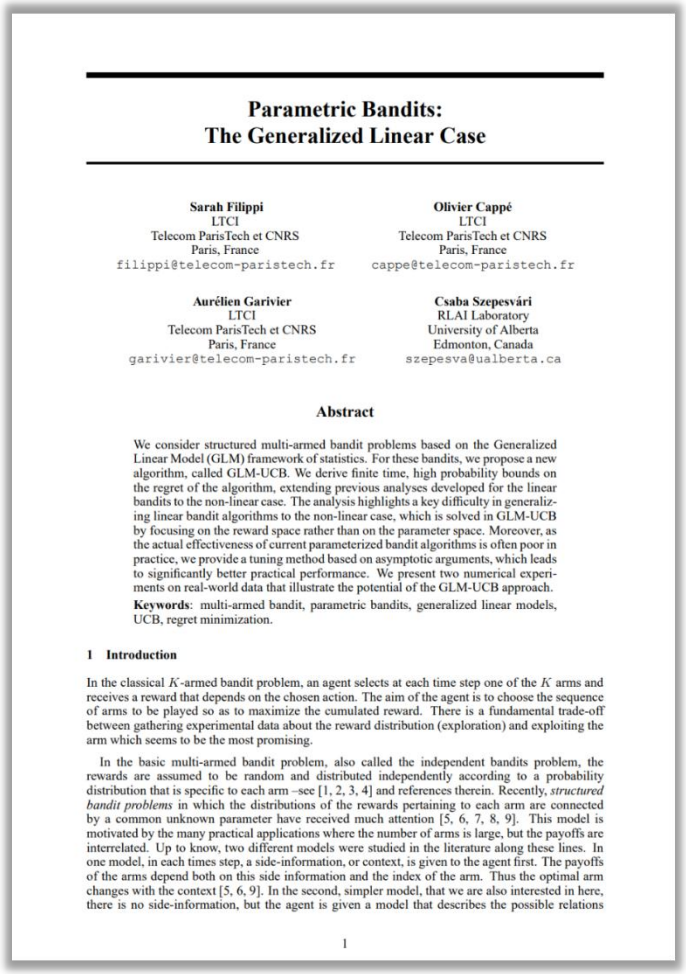
Question: Can Generalized/Heavy-tailed LB enjoy one-pass algorithms?

More Recent Progress



- One-Pass Bandit Learning for RLHF and Function Approximation
  2025.11.23, 第二十届中国人工智能基础年会 (CFAI 2025)·强化学习论坛, 湖南长沙.

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]

Generalized Linear Bandits:
Almost Optimal Regret with One-Pass Update

Yu-Jie Zhang[1], **Sheng-An Xu**[2,3], **Peng Zhao**[2,3], **Masashi Sugiyama**[1,4]
[1] RIKEN AIP, Tokyo, Japan
[2] National Key Laboratory for Novel Software Technology, Nanjing University, China
[3] School of Artificial Intelligence, Nanjing University, China
[4] The University of Tokyo, Chiba, Japan

Abstract

We study the generalized linear bandit (GLB) problem, a contextual multi-armed bandit framework that extends the classical linear model by incorporating a non-linear link function, thereby modeling a broad class of reward distributions such as Bernoulli and Poisson. While GLBs are widely applicable to real-world scenarios, their non-linear nature introduces significant challenges in achieving both computational and statistical efficiency. Existing methods typically trade off between two objectives, either incurring high per-round costs for optimal regret guarantees or compromising statistical efficiency to enable constant-time updates. In this paper, we propose a jointly efficient algorithm that attains a nearly optimal regret bound with $\mathcal{O}(1)$ time and space complexities per round. The core of our method is a tight confidence set for the online mirror descent (OMD) estimator, which is derived through a novel analysis that leverages the notion of mix loss from online prediction. The analysis shows that our OMD estimator, even with its one-pass updates, achieves statistical efficiency comparable to maximum likelihood estimation, thereby leading to a jointly efficient optimistic method.

1 Introduction

Stochastic multi-armed bandits [Robbins, 1952] represent a fundamental class of sequential decision-making problems where a learner interacts with environments by selecting actions (or arms) and receiving feedback in the form of rewards. In this paper, we study the contextual multi-armed bandit problem under the framework of generalized linear models (GLMs). In this setting, each action is characterized by a contextual feature vector $\mathbf{x} \in \mathcal{X}_t \subset \mathbb{R}^d$, where the arm set $\mathcal{X}_t$ may vary over time. More specifically, the learning process can be seen as a T round game between the learner and environments: at each round t, the learner selects an action $X_t \in \mathcal{X}_t$ and then observes a stochastic reward $r_t \in \mathbb{R}$ generated according to a GLM (see Definition 2.1). The goal of the learner is to maximize the cumulative expected reward obtained over the time horizon T. Under the GLM model, the expectation of the reward satisfies $\mathbb{E}[r_t \mid X_t] = \mu(X_t^\top \theta_*)$, where $\mu : \mathbb{R} \to \mathbb{R}$ is a non-linear link determined by the GLM model and is known to the learner. The unknown part is the underlying parameter $\theta_* \in \mathbb{R}^d$, which needs to be estimated from the observed action-reward pairs.

Compared with the classical linear case [Abbasi-Yadkori et al., 2011], the generalized linear bandit (GLB) framework allows for a richer class of reward distributions, including Gaussian, Bernoulli, and Poisson distributions. This flexibility enables the modeling of various real-world tasks, such as recommendation systems [Li et al., 2010] and personalized medicine [Tewari and Murphy, 2017], where the feedback is binary (Bernoulli) or count-based (Poisson) and inherently non-linear. Besides

*Correspondence: Peng Zhao <zhaop@lamda.nju.edu.cn>

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

[NeurIPS'25 Generalized Linear Bandits: Almost Optimal Regret with One-Pass Update]

Generalized Linear Bandits (GLB)

Table 1: Comparison of regret guarantees and computational complexity per round for GLBs. Here, $\kappa_* = 1/\left(\frac{1}{T}\sum_{t=1}^{T}\mu'(\mathbf{x}_{t,*}^\top\theta_*)\right)$ is the slope at the optimal action $\mathbf{x}_{t,*} = \arg\max_{\mathbf{x}\in\mathcal{X}_t}\mu(\mathbf{x}^\top\theta_*)$, with $\kappa_* \le \kappa$ (see Section 2 for details). † indicates the amortized time complexity, i.e., average per-round cost over T rounds.

Method	Regret	Time per Round	Memory	Jointly Efficient
GLM-UCB [Filippi et al., 2010]	$\mathcal{O}(\kappa(\log T)^{\frac{3}{2}}\sqrt{T})$	$\mathcal{O}(t)$	$\mathcal{O}(t)$	✗
GLOC [Jun et al., 2017]	$\mathcal{O}(\kappa\log T\sqrt{T})$	$\mathcal{O}(1)$	$\mathcal{O}(1)$	✗
OFUGLB [Lee et al., 2024, Liu et al., 2024]	$\mathcal{O}(\log T\sqrt{T/\kappa_*})$	$\mathcal{O}(t)$	$\mathcal{O}(t)$	✗
RS-GLinCB [Sawarni et al., 2024]	$\mathcal{O}(\log T\sqrt{T/\kappa_*})$	$\mathcal{O}\big((\log t)^2\big)^\dagger$	$\mathcal{O}(t)$	✗
GLB-OMD (Theorem 2 of this paper)	$\mathcal{O}(\log T\sqrt{T/\kappa_*})$	$\mathcal{O}(1)$	$\mathcal{O}(1)$	✓

Generalized Linear Bandits (GLB)

A variety of usage, especially for the logistic link function…

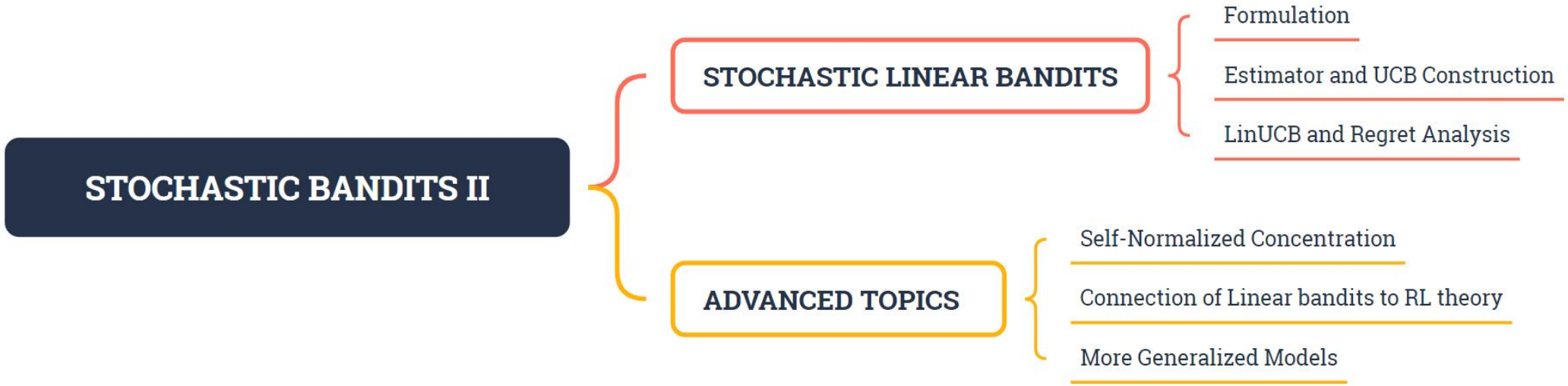
❑ ***RL with function approximation***: MNL mixture MDPs (related to GLB)

Long-Fei Li*, Yu-Yang Qian*, Peng Zhao, Zhi-Hua Zhou. Provably Efficient Online RLHF with One-Pass Reward Modeling. NeurIPS 2025.

❑ ***RLHF***: BT model naturally related to logistic bandits, etc.

Long-Fei Li, Yu-Jie Zhang, Peng Zhao, Zhi-Hua Zhou. Provably Efficient Reinforcement Learning with Multinomial Logit Function Approximation. NeurIPS 2024.

Summary



Q & A

Thanks!