Introduction to Linear Bandits

Yoan Russac

CNRS, Inria, ENS, Université PSL

Roadmap

Stochastic Bandit Model

ν_1

ν_2

ν_3

ν_4

Setting:

- K arms. Each arm associated with an unknown distribution ν_a with mean μ_a
- action $A_t \in \{1, ..., K\}$ is chosen at time t based on previous observations and rewards
- reward X_t observed

$$X_t = \mu_{A_t} + \epsilon_t \quad (\epsilon_t \text{ centered noise})$$

- $a^\star = \arg\max_{a \in \{1,...,K\}} \mu_a$

Specificity of Bandit Models

- Sequential Learning: learning on the fly
- Incomplete information: at time t we don't know the rewards we would have obtained by selecting a different arm
- Difference with General Reinforcement Learning: choosing an action does not impact the state of the environment

Stochastic Bandit Model: Mesure of performance

Objective: maximize the expected sum of the rewards or equivalently minimizing the regret

$N_a(t)$: number of times the arm a has been pulled up to time t
$\Delta_a = \mu_{a^\star} - \mu_a$: sub-optimality gap of arm a

Regret of an algorithm $\mathcal{A}$ on a bandit instance ν:

$$
\begin{aligned}
R(T) &= T\mu_{a^\star} - \mathbb{E}\left[\sum_{t=1}^{T} X_t\right] \\
&= \sum_{a=1}^{K} \Delta_a \mathbb{E}[N_a(T)]
\end{aligned}
$$

Strategy with small regrets

How to design a strategy with a small regret ?

$$R(T) = \sum_{a=1}^{K} \Delta_a \mathbb{E}[N_a(T)]$$

$\hookrightarrow$ Not selecting too frequently the arms where $\Delta_a > 0$

Problem: The μ_a are unknown, so Δ_a is unknown ! Need to try all the arms to estimate Δ_a's

$\hookrightarrow$ Exploration - Exploitation trade-off

Exploration and Exploitation

- Naive idea for exploration: Select each arm T/K times
- Naive idea for exploitation: Select the arm with the best empirical mean: $A_t = \arg\max_{a \in \{1,\dots K\}} \hat{\mu}_a(t)$, where

$$\hat{\mu}_a(t) = \frac{1}{N_a(t-1)} \sum_{s=1}^{t-1} X_s \mathbb{1}(A_s = a)$$

$\hookrightarrow$ Linear regret !

Optimism in the face of uncertainty

- For each arm build a confidence interval on the mean μ_a

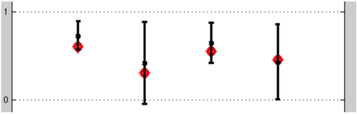
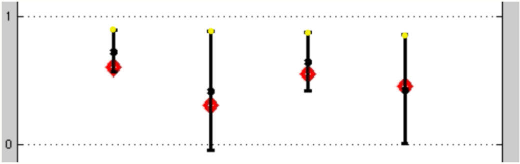


Figure: Confidence interval for the different arms at time t

- Act as if the best possible model is the true model



$\hookrightarrow$ Select the arm

$$A_t = \arg\max_{a=\{1,\dots,K\}} \mathsf{UCB}_{t-1}(a)$$

UCB(α) algorithm

Under the assumption of Gaussian rewards,

$$\mathrm{UCB}_t(a) = \hat{\mu}_a(t) + \sqrt{\frac{\alpha \log(t)}{N_a(t-1)}}$$

Problem dependent Bound [Auer et al. 2002]

UCB(α) with $\alpha = 2$ and gaussian rewards with variance 1, satisfies

$$R(T) \leq 8 \left(\sum_{a \neq a^\star} \frac{1}{\Delta_a} \right) \log(T) + (1 + \pi^2/3) \sum_{a=1}^{K} \Delta_a$$

UCB(α) algorithm

Sometimes we prefer problem independent bounds.

$\varepsilon(K, G) = \{\nu = (\nu_1, ..., \nu_K), \text{where} \, \forall i \in \{1, ..., K\}, \, \nu_i = \mathcal{N}(\mu_i, 1), \text{with} \, \mu_i \in [0, 1]\}$

Problem independent Bound

If $\delta = \frac{1}{n^2}$, the regret of UCB(α) with $\alpha = 2$ on any bandit instance in $\varepsilon(K, G)$ is bounded by

$$R(T) \leq 4\sqrt{KT \log(T)} + (1 + \pi^2/3)$$

Roadmap

Contextual bandits

Use case: Recommender system

- At time t a user arrives on a website with some characteristics
- Several items with some characteristics could be recommended to the user
- For each item a context $A \in \mathbb{R}^d$ is build based on the user features + item features. Those contexts form a set $\mathcal{A}_t$
- By choosing a context A the associated product is displayed to the user
- A reward X_t depending on A_t is then observed

$$X_t = f(A_t) + \epsilon_t$$

Contextual bandits

How to specify f?

- Linear Models: $\exists \theta^\star$, $X_t = A_t^\top \theta^\star + \epsilon_t$
- Generalized Linear Models $\exists \theta^\star$, $X_t = \mu(A_t^\top \theta^\star) + \epsilon_t$
  $\hookrightarrow$ μ is called inverse link function

In this talk we focus on Linear Models

Linear Bandits Setting

- In round t a set of K actions $\mathcal{A}_t = \{A_{t,1}, ..., A_{t,K}\}$ is available
- By selecting the context A_t, one observes the reward

$$X_t = A_t^\top \theta^\star + \epsilon_t$$

- Assumption on the noise: ϵ_t are supposed to be i.i.d and normally distributed $\epsilon_t \sim \mathcal{N}(0,1)$
- Bounded Actions
- Bounded $\theta^\star$

Best action at time t:

$$A_t^\star = \arg\max_{a \in \mathcal{A}_t} a^\top \theta^\star$$

Difference with the Stochastic Bandit Model

- In the Stochastic Bandit Model the arms are independent
- The Linear Bandit model is a structured bandit problem: The rewards of each arm are connected by a common unknown parameter $\theta^\star$

  $\hookrightarrow$ Learning transfer from one context to another

Goal

Regret Minimization

$$\max \mathbb{E}\left(\sum_{t=1}^{T} X_t\right) \iff \min \ \mathbb{E}\left[\sum_{s=1}^{T} \max_{a \in \mathcal{A}_t} \langle a, \theta^\star \rangle - \sum_{t=1}^{T} X_t\right]$$

$$\iff \min \ \mathbb{E}\left(\sum_{t=1}^{T} \max_{a \in \mathcal{A}_t} \langle a - A_t, \theta^\star \rangle\right)$$

How to choose an action A_t at time t to minimize the regret ?

Estimating the unknown parameter $\theta^\star$

- Say we already played $t-1$ rounds where the actions $A_1, ..., A_{t-1}$ have been selected and the rewards $X_1, ..., X_{t-1}$ have been collected
- How to estimate $\theta^\star$ based on those observations?
  $\hookrightarrow$ Regularized Least-Squares Estimator

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

- Closed form solution: $\hat{\theta}_t = V_{t-1}^{-1} \sum_{s=1}^{t-1} A_s X_s$, where

$$V_{t-1} = \sum_{s=1}^{t-1} A_s A_s^\top + \lambda I_d$$

Link with the Linear Regression

- Closed form solution $\hat{\theta}_t = (\sum_{s=1}^{t-1} A_s A_s^\top + \lambda I_d)^{-1} \sum_{s=1}^{t-1} A_s X_s$
- For $\lambda = 0$ we find the usual estimator for the Linear Regression $(X^\top X)^{-1} X^\top Y$, where X is the matrix containing the data of up time $t-1$ and Y is the associated reward vector

Optimism in the face of uncertainty

- Acting as if the environment is as nice as plausibly possible
- In the stochastic bandit model it means selecting the action with the largest Upper Confidence Bound
- In the Linear Model, the form of the confidence bound is more complicated because rewards received give information about more than just the arm played.

  $\hookrightarrow$ Constructing a confidence set $\mathcal{C}_t \in \mathbb{R}^d$ that contains the unknown parameter $\theta^\star$ with high probability given the observations available up to time $t-1$

Exploration/Exploitation dilemna and Linear Bandits

- Greedy Policy: Chooses the action A_t that maximizes

$$A_t = \arg\max_{a \in \mathcal{A}_t} a^\top \hat{\theta}_t$$

$\hookrightarrow$ not enough exploration

- Linear Upper Confidence Bound algorithm (LinUCB): Chooses the action A_t that maximizes

$$A_t = \arg\max_{a \in \mathcal{A}_t} \max_{\theta \in \mathcal{C}_t} a^\top \theta$$

with a particular $\mathcal{C}_t$

How to choose the confidence ellipsoid ?

Let $\beta_t(\delta) = \lambda + \sqrt{2\log(1/\delta) + d\log\left(1 + \frac{t}{\lambda d}\right)}$. The confidence ellipsoid is defined as:

$$\mathcal{C}_t(\delta) = \{\theta \in \mathbb{R}^d : \|\theta - \hat{\theta}_t\|_{V_{t-1}} \leq \beta_{t-1}(\delta)\}$$

Theorem

$\mathcal{C}_t(\delta)$ is a confidence set for $\theta^\star$ at level $1-\delta$,

$$\forall \delta > 0, \mathbb{P}\left(\forall t \geq 1,\ \theta^\star \in \mathcal{C}_t(\delta)\right) \geq 1 - \delta$$

- With this choice of confidence ellipsoid the previous optimization program is equivalent to maximizing

$$A_t = \arg\max_{a \in \mathcal{A}_t} \left(a^\top \hat{\theta}_t + \beta_{t-1}(\delta)\|a\|_{V_{t-1}^{-1}}\right)$$

LinUCB

Algorithm 1: LinUCB

Input: Probability δ, dimension d, regularization λ.

Initialization: $b = 0_{\mathbb{R}^d}$, $V = \lambda I_d$, $\hat{\theta} = 0_{\mathbb{R}^d}$

for $t \geq 1$ **do**

Receive $\mathcal{A}_t$, compute

$$\beta_{t-1} = \sqrt{\lambda} + \sqrt{2\log\left(\frac{1}{\delta}\right) + d\log\left(1 + \frac{t-1}{\lambda d}\right)}$$

for $a \in \mathcal{A}_t$ **do**

Compute $\text{UCB}(a) = a^\top \hat{\theta} + \beta_{t-1}\sqrt{a^\top V^{-1} a}$

$A_t = \arg\max_a(\text{UCB}(a))$

Play action A_t and **receive reward** X_t

Updating phase: $V = V + A_t A_t^\top$

$b = b + X_t A_t$

$\hat{\theta} = V^{-1} b$

LinUCB

Regret of LinUCB

Under the previous assumptions, with probability $1-\delta$ the regret of LinUCB satisfies

$$R_T \leq \sqrt{dT}\sqrt{8\beta_T(\delta)\log\left(1+\frac{TL^2}{\lambda d}\right)} = \tilde{O}(d\sqrt{T})$$

$\hookrightarrow$ Independent of the number of actions K

Roadmap

Linear Bandits Setting

- In round t a set of K actions $\mathcal{A}_t = \{A_{t,1}, ..., A_{t,K}\}$ is available
- By selecting the context A_t, one observes the reward

$$X_t = A_t^\top \theta_t^\star + \epsilon_t$$

- Assumption on the noise: ϵ_t are supposed to be i.i.d and normally distributed $\epsilon_t \sim \mathcal{N}(0, 1)$
- Bounded Actions
- Bounded $\theta_t^\star$

Best action at time t:

$$A_t^\star = \arg\max_{a \in \mathcal{A}_t} a^\top \theta_t^\star$$

Optimality Criteria

Dynamic Regret Minimization

$$\max\ \mathbb{E}\left(\sum_{t=1}^{T} X_t\right) \iff \min\ \mathbb{E}\left[\sum_{s=1}^{T} \max_{a\in\mathcal{A}_t}\langle a, \theta_t^\star\rangle - \sum_{t=1}^{T} X_t\right]$$

$$\iff \min\ \underbrace{\mathbb{E}\left(\sum_{t=1}^{T} \max_{a\in\mathcal{A}_t}\langle a - A_t, \theta_t^\star\rangle\right)}_{\text{dynamic regret}}$$

Our Approach

We only focus on robust policies

With that in mind, the non-stationarity in the $\theta^\star_t$ parameter is measured with the variation budget

$$\sum_{s=1}^{T-1} \|\theta^\star_s - \theta^\star_{s+1}\|_2 \leq B_T$$

$\hookrightarrow$ A large variation budget can be either due to large scarce changes of $\theta^\star_t$ or frequent but small deviations

Weighted Least Squares Estimator

Least Squares Estimator

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

Weighted Least Squares Estimator

$$\hat{\theta}_t = \arg\min_{\theta \in \mathbb{R}^d} \sum_{s=1}^{t} w_s (X_s - A_s^\top \theta)^2 + \frac{\lambda_t}{2} \|\theta\|_2^2$$

The Case of Exponential weights

Exponential Discount (Time-Dependent Weights)

$$\hat{\theta}_t = \arg\min_{\theta \in \mathbb{R}^d} \sum_{s=1}^{t} \underbrace{\gamma^{t-s}}_{w_{t,s}} (X_s - A_s^\top \theta)^2 + \frac{\lambda}{2} \|\theta\|_2^2$$

D-LinUCB Algorithm (1)

Algorithm 2: D-LinUCB

Input: Probability δ, dimension d, regularization λ, discount factor γ.

Initialization: $b = 0_{\mathbb{R}^d}$, $V = \lambda I_d$, $\widetilde{V} = \lambda I_d$, $\hat{\theta} = 0_{\mathbb{R}^d}$

for $t \geq 1$ **do**

Receive $\mathcal{A}_t$, compute

$$\beta_{t-1} = \sqrt{\lambda} + \sqrt{2\log\left(\frac{1}{\delta}\right) + d\log\left(1 + \frac{1-\gamma^{2(t-1)}}{\lambda d(1-\gamma^2)}\right)}$$

for $a \in \mathcal{A}_t$ **do**

Compute $\text{UCB}(a) = a^\top\hat{\theta} + \beta_{t-1}\sqrt{a^\top V^{-1}\widetilde{V}V^{-1}a}$

$A_t = \arg\max_a(\text{UCB}(a))$

Play action A_t and **receive reward** X_t

Updating phase: $V = \gamma V + A_t A_t^\top + (1-\gamma)\lambda I_d$,

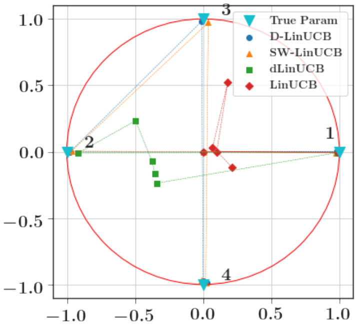
$\widetilde{V} = \gamma^2\widetilde{V} + A_t A_t^\top + (1-\gamma^2)\lambda I_d$

$b = \gamma b + X_t A_t$,

$\hat{\theta} = V^{-1}b$

Roadmap

Performance in Abruptly-Changing Environment



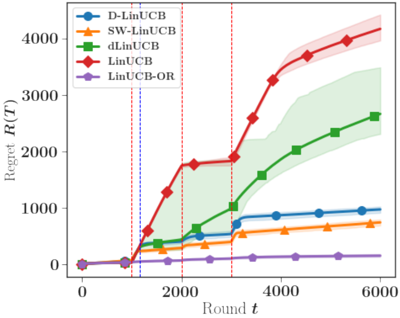


Figure: Performances of the algorithms in the abruptly-changing environment. The plot on the left correspond to the estimated parameter and the one on the right to the accumulated regret, averaged on $N = 100$ independent experiments

Performance in Slowly-Changing Environment

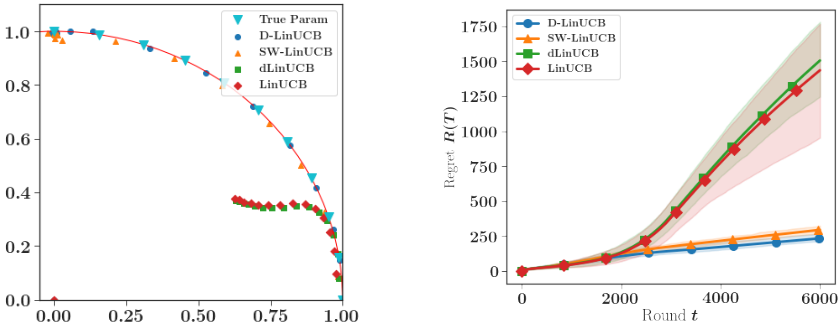


Figure: Performances of the algorithms in the slowly-varying environment. The plot on the left correspond to the estimated parameter and the one on the right to the accumulated regret, averaged on $N = 100$ independent experiments

Thank you !

Concentration Result in Stationary Environments

Theorem 1

Assuming that $\theta_t^\star = \theta^\star$, for any $\mathcal{F}_t$-predictable sequences of actions $(A_t)_{t\geq 1}$ and positive weights $(w_t)_{t\geq 1}$ and for all $\delta > 0$, with probability higher than $1-\delta$,

$$\mathbb{P}\left(\forall t, \|\hat{\theta}_t - \theta^\star\|_{V_t \widetilde{V}_t^{-1} V_t} \leq \frac{\lambda_t}{\sqrt{\mu_t}} S + \sigma \sqrt{2\log(1/\delta) + d \log\left(1 + \frac{L^2 \sum_{s=1}^t w_s^2}{d\mu_t}\right)}\right)$$

where

$$V_t = \sum_{s=1}^{t} w_s A_s A_s^\top + \lambda_t I_d,$$

$$\widetilde{V}_t = \sum_{s=1}^{t} w_s^2 A_s A_s^\top + \mu_t I_d$$

Concentration in the Non-Stationary Case

Moving back to the non-stationary environment $X_s = A_s^\top \theta_s^\star + \eta_s$ and assuming that $w_s = \gamma^{-s}$, $\lambda_s = \lambda\gamma^{-s}$

Let $\bar{\theta}_t = V_{t-1}^{-1}\left(\sum_{s=1}^{t-1} \gamma^{-s} A_s A_s^\top \theta_s^\star + \gamma^{t-1}\theta_t^\star\right)$ denote a "noiseless" proxy value for $\theta_t^\star$

Concentration in the Non-Stationary Case

Moving back to the non-stationary environment $X_s = A_s^\top \theta_s^\star + \eta_s$ and assuming that $w_s = \gamma^{-s}$, $\lambda_s = \lambda\gamma^{-s}$

Let $\bar{\theta}_t = V_{t-1}^{-1} \left(\sum_{s=1}^{t-1} \gamma^{-s} A_s A_s^\top \theta_s^\star + \gamma^{t-1} \theta_t^\star \right)$ denote a "noiseless" proxy value for $\theta_t^\star$

Theorem 2

Let $\mathcal{C}_t = \{\theta \in \mathbb{R}^d : \|\theta - \hat{\theta}_{t-1}\|_{V_{t-1}\widetilde{V}_{t-1}^{-1}V_{t-1}} \leq \beta_{t-1}\}$ *denote the confidence ellipsoid with*

$$\beta_t = \lambda\sqrt{S} + \sigma\sqrt{2\log(1/\delta) + d\log\left(1 + \frac{L^2(1-\gamma^{2t})}{\lambda d(1-\gamma^2)}\right)}$$

Then, $\forall \delta > 0$,

$$\mathbb{P}\left(\forall t \geq 1, \bar{\theta}_t \in \mathcal{C}_t\right) \geq 1 - \delta$$