Introduction to Linear Bandits

Yoan Russac

CNRS, Inria, ENS, Université PSL
Roadmap

1. Stochastic Multi Armed Bandits
2. Linear Bandits
3. Non-Stationary Bandits
4. Empirical Performances
Stochastic Bandit Model

Setting:

- $K$ arms. Each arm associated with an unknown distribution $\nu_a$ with mean $\mu_a$
- action $A_t \in \{1, \ldots, K\}$ is chosen at time $t$ based on previous observations and rewards
- reward $X_t$ observed
  \[ X_t = \mu_{A_t} + \epsilon_t \quad (\epsilon_t \text{ centered noise}) \]
- $a^* = \arg \max_{a \in \{1, \ldots, K\}} \mu_a$
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
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 \( t \)

\( \Delta_a = \mu_{a^*} - \mu_a: \) sub-optimality gap of arm \( a \)

Regret of an algorithm \( \mathcal{A} \) on a bandit instance \( \nu \):

\[
R(T) = T\mu_{a^*} - \mathbb{E}\left[ \sum_{t=1}^{T} X_t \right] \\
= \sum_{a=1}^{K} \Delta_a \mathbb{E}[N_a(T)]
\]
Strategy with small regrets

How to design a strategy with a small regret?

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

\[ \hookrightarrow \text{Not selecting too frequently the arms where } \Delta_a > 0 \]

Problem: The \( \mu_a \) are unknown, so \( \Delta_a \) is unknown! Need to try all the arms to estimate \( \Delta_a \)'s

\[ \hookrightarrow \text{Exploration - Exploitation trade-off} \]
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: $A_t = \arg \max_{a \in \{1, \ldots, K\}} \hat{\mu}_a(t)$, where

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

$\rightarrow$ Linear regret!
Optimism in the face of uncertainty

- For each arm build a confidence interval on the mean $\mu_a$

![Figure: Confidence interval for the different arms at time $t$](image)

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

$\mapsto$ Select the arm

$$A_t = \arg\max_{a=\{1,\ldots,K\}} \text{UCB}_{t-1}(a)$$
UCB(α) algorithm

Under the assumption of Gaussian rewards,

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

Problem dependent Bound [Auer et al. 2002]

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

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

UCB($\alpha$) algorithm

Sometimes we prefer problem independent bounds.

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

**Problem independent Bound**

If $\delta = \frac{1}{n^2}$, the regret of UCB($\alpha$) 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

1. Stochastic Multi Armed Bandits
2. Linear Bandits
3. Non-Stationary Bandits
4. Empirical Performances
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 $A_t$
- By choosing a context $A$ the associated product is displayed to the user
- A reward $X_t$ depending on $A_t$ is then observed

$$X_t = f(A_t) + \epsilon_t$$
How to specify $f$?

- **Linear Models**: $\exists \theta^*, \ X_t = A_t^\top \theta^* + \epsilon_t$
- **Generalized Linear Models**: $\exists \theta^*, \ X_t = \mu(A_t^\top \theta^*) + \epsilon_t$

$\mu$ is called inverse link function

In this talk we focus on **Linear Models**
Linear Bandits Setting

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

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

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

Best action at time $t$:

$$A_t^* = \arg\max_{a \in A_t} a^\top \theta^*$$
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 $\theta^*$.

$\Rightarrow$ Learning transfer from one context to another.
Goal

Regret Minimization

\[
\begin{aligned}
\max E\left( \sum_{t=1}^{T} X_t \right) & \iff \min E \left[ \sum_{s=1}^{T} \max_{a \in \mathcal{A}_t} \langle a, \theta^* \rangle - \sum_{t=1}^{T} X_t \right] \\
& \iff \min E \left( \sum_{t=1}^{T} \max_{a \in \mathcal{A}_t} \langle a - A_t, \theta^* \rangle \right)
\end{aligned}
\]
How to choose an action $A_t$ at time $t$ to minimize the regret?
Estimating the unknown parameter $\theta^*$

- 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 $\theta^*$ based on those observations?
  - 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

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

Constructing a confidence set $C_t \in \mathbb{R}^d$ that contains the unknown parameter $\theta^*$ with high probability given the observations available up to time $t - 1$
Greedy Policy: Chooses the action $A_t$ that maximizes

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

→ not enough exploration

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

$$A_t = \arg \max_{a \in A_t} \max_{\theta \in C_t} a^\top \theta$$

with a particular $C_t$
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:

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

Theorem

$C_t(\delta)$ is a confidence set for $\theta^*$ at level $1 - \delta$,

$$\forall \delta > 0, \mathbb{P} (\forall t \geq 1, \theta^* \in C_t(\delta)) \geq 1 - \delta$$

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

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

**Input:** Probability $\delta$, dimension $d$, regularization $\lambda$.

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

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

Receive $A_t$, 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** $a \in A_t$ **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:**

$V = V + A_t A_t^\top$

$b = b + X_t A_t$

$\hat{\theta} = V^{-1} b$
### Regret of LinUCB

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

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

→ Independent of the number of actions $K$
Roadmap

1. Stochastic Multi Armed Bandits
2. Linear Bandits
3. Non-Stationary Bandits
4. Empirical Performances
Linear Bandits Setting

- In round $t$ a set of $K$ actions $A_t = \{A_{t,1}, \ldots, A_{t,K}\}$ is available.
- By selecting the context $A_t$, one observes the reward
  \[ X_t = A_t^\top \theta_t^* + \epsilon_t \]

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

Best action at time $t$:
\[ A_t^* = \arg\max_{a \in A_t} a^\top \theta_t^* \]
Non-Stationary Bandits

Optimality Criteria

Dynamic Regret Minimization

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

\[
\iff \min \mathbb{E} \left( \sum_{t=1}^{T} \max_{a \in A_t} \langle a - A_t, \theta^*_t \rangle \right)
\]

dynamic regret
Our Approach

We only focus on robust policies

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

$$\sum_{s=1}^{T-1} \| \theta^*_s - \theta^*_{s+1} \|_2 \leq B_T$$

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

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

**Algorithm 2: D-LinUCB**

**Input:** Probability $\delta$, dimension $d$, regularization $\lambda$, discount factor $\gamma$.

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

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

- Receive $A_t$, 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** $a \in A_t$ **do**
  - Compute $\text{UCB}(a) = a^\top \hat{\theta} + \beta_{t-1} \sqrt{a^\top V^{-1} \tilde{V} V^{-1} a}$
  - $A_t = \arg \max_a (\text{UCB}(a))$

- **Play action** $A_t$ and **receive reward** $X_t$

- **Updating phase:**
  - $V = \gamma V + A_t A_t^\top + (1 - \gamma) \lambda I_d$
  - $\tilde{V} = \gamma^2 \tilde{V} + A_t A_t^\top + (1 - \gamma^2) \lambda I_d$
  - $b = \gamma b + X_t A_t$
  - $\hat{\theta} = V^{-1} b$
Roadmap

1. Stochastic Multi Armed Bandits
2. Linear Bandits
3. Non-Stationary Bandits
4. Empirical Performances
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 $\theta_t^* = \theta^*$, 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 $1 - \delta$,

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

where

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

$$\tilde{V}_t = \sum_{s=1}^{t} w_s^2 A_s A_s^\top + \mu_t I_d$$
Concentration in the Non-Stationary Case

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

Let $\bar{\theta}_t = V_{t-1}^{-1} \left( \sum_{s=1}^{t-1} \gamma^{-s} A_s A_s^T \theta_s^* + \gamma^{t-1} \theta_t^* \right)$ denote a “noiseless” proxy value for $\theta_t^*$
Concentration in the Non-Stationary Case

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

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

**Theorem 2**

Let $C_t = \{ \theta \in \mathbb{R}^d : \|\theta - \hat{\theta}_{t-1}\|_{V_{t-1}\tilde{V}_{t-1}V_{t-1}} \leq \beta_{t-1} \}$ 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, $\forall \delta > 0$,

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