LECTURE 15 - Stochastic Bandits

Contents

I - Introduction

We call stochastic (multi-armed) bandit problem the following sequential decision problem.

(In American slang, a "one-armed bandit" refers to a slot machine. A multi-armed bandit therefore refers to a collection of many slot machines)

First let $N \geqslant 2$ be known to the player. Consider probability distibutions $\nu_1, \dots, \nu_N$ over $\mathbb{R}$, unknown to the player.

→ At each round $t \geqslant 1$, the player chooses an action $I_t \in \{1, \dots, N\}$ (in the bandit jargon, actions $1, \dots, N$ are also called arms an when the player chooses action I_t we say he "pulls" arm I_t).

→ At the same time, the environment draws a random vector $\ell_t = (\ell_t(1), \dots, \ell_t(N))$ where
- $\ell_t(i) \sim \nu_i$
- Variables $\{\ell_t(i)\}_{i=1}^{N}$ are independent
- Variable $\ell_t(i)$ is independent from $\{I_s\}_{s=1}^{t}$ and $\{\ell_s(j)\}_{1 \leqslant s \leqslant t-1, 1 \leqslant j \leqslant N}$

→ The environment only reveals $\ell_{I_t, t}$ to the player. He incurs this loss and moves on to the next round.

General Goal:

Minimize the regret which is defined here as

$$R_T = \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(I_t)\right] - \min_{1 \leq i \leq N} \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(i)\right].$$

Alternative representation of the regret:

Notation

m_i : mean value of probability measure ν_i

$m^* = \min_{1 \leq i \leq N} m_i$ and $i^* \in \operatorname{argmin}_{1 \leq i \leq N} m_i$

(in particular $m^* = m_{i^*}$)

$\Delta_i = m_i - m^*$

$n_i(t) = \sum_{s=1}^{t} \mathbb{1}\{I_s = i\}$: Number of times the player took action i up to time t.

Lemma 1

$$R_T = \sum_{i : \Delta_i > 0} \Delta_i \, \mathbb{E}[n_i(T)]$$

Proof. Since $\ell_t(I_t) = \sum_{i=1}^{N} \ell_t(i)\, \mathbb{1}\{I_t = i\}$ we obtain

$$\mathbb{E}\left[\sum_{t=1}^{T} \ell_t(I_t)\right] = \sum_{t=1}^{T}\sum_{i=1}^{N} \mathbb{E}\left[\ell_t(i)\, \mathbb{1}\{I_t = i\}\right]$$

$$= \sum_{t=1}^{T}\sum_{i=1}^{N} \mathbb{E}\left[\ell_t(i)\right] \mathbb{E}\left[\mathbb{1}\{I_t = i\}\right]$$

(Since $\ell_t(i)$ is independent from I_t)

$$= \sum_{t=1}^{T}\sum_{i=1}^{N} m_i\, \mathbb{E}\left[\mathbb{1}\{I_t = i\}\right]$$

(Since $\ell_t(i) \sim \nu_i$ and definition of m_i)

$$= \sum_{i=1}^{N} m_i\, \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{I_t = i\}\right]$$

$$= \sum_{i=1}^{N} m_i\, \mathbb{E}\left[n_i(T)\right].$$

Similarly, we easily observe that

$$\min_{1 \leq i \leq N} \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(i)\right] = T m^*$$

$$= \sum_{i=1}^{N} m^*\, \mathbb{E}\left[n_i(T)\right],$$

Since clearly $\sum_{i=1}^{N} n_i(T) = T$. □

II - Reminder : Sub-gaussian distributions

In the sequel, we'll focus on the case where the loss distributions $\{\nu_i\}_{i=1}^N$ are subgaussian. This restriction is made for simplicity and can be easily removed (at the price of worse regret bounds).

Definition

A probability distribution ν over $\mathbb{R}$ is said sub-gaussian with parameter $\sigma^2 > 0$ if, given a random variable $X \sim \nu$, we have

$$\forall \lambda \in \mathbb{R} : \log \mathbb{E}\, e^{\lambda (X - \mathbb{E}X)} \leq \frac{\lambda^2 \sigma^2}{2} .$$

Remarks.

- The terminology sub-gaussian comes from the fact that, if ν is the gaussian distribution $\mathcal{N}(m, \sigma^2)$ then, $\forall \lambda \in \mathbb{R}$,
$$\log \mathbb{E}\, e^{\lambda (X - m)} = \frac{\lambda^2 \sigma^2}{2} .$$
- When ν is a general sub-gaussian distribution with parameter $\sigma^2 > 0$, we'll always implicitly assume that σ^2 is the smallest constant for which the property of the definition holds, i.e. that

$$\sigma^2 = \sup_{\lambda \neq 0} \frac{2}{\lambda^2} \log \mathbb{E} e^{\lambda (X - \mathbb{E}X)} .$$

σ^2 is also called the "variance proxy" of ν and it may be shown that inequality

$$\mathrm{Var}(\nu) \leqslant \sigma^2 ,$$

always holds.

As shown by the next result, the set of sub-gaussian distributions includes all distributions with bounded support.

Lemma 2 (Hoeffding Lemma)

Let X be a real-valued random variable such that $\mathbb{P}(a \leqslant X \leqslant b) = 1$ for some $a < b \in \mathbb{R}$. Then, $\forall \lambda \in \mathbb{R}$,

$$\log \mathbb{E} e^{\lambda (X - \mathbb{E}X)} \leqslant \frac{\lambda^2 (b-a)^2}{8}$$

(admitted)

In otherwords, the previous lemma states that any probability distribution supported on a finite interval $[a, b]$ is subgaussian with parameter

$$\sigma^2 = \frac{(b-a)^2}{4} .$$

Next, we review the concentration properties of sub-gaussian random variables. From now on, we say that a random variable X is sub-gaussian if its distribution is subgaussian.

Lemma 3

Suppose X is sub-gaussian with parameter $\sigma^2 > 0$. Then, $\forall\, t > 0$,

$$\mathbb{P}(X - \mathbb{E}X \geq t) \leq e^{-t^2/2\sigma^2}, \qquad \mathbb{P}(X - \mathbb{E}X \leq -t) \leq e^{-\frac{t^2}{2\sigma^2}}$$

and therefore

$$\mathbb{P}(|X - \mathbb{E}X| \geq t) \leq 2e^{-\frac{t^2}{2\sigma^2}}.$$

Proof: It is enough to prove the first inequality. Indeed, the 2nd follows immediatly by observing that "X sub-gaussian with parameter $\sigma^2 > 0$" $\iff$ "$-X$ sub-gaussian with parameter $\sigma^2 > 0$". The 3rd inequality follows also directly by combining the 1st and 2nd inequalities.

Now, observe that $\forall\, \lambda > 0$,

$$\begin{aligned}
\mathbb{P}(X - \mathbb{E}X \geq t) &= \mathbb{P}\left(e^{\lambda(X - \mathbb{E}X)} \geq e^{\lambda t}\right) \\
&\leq e^{-\lambda t}\, \mathbb{E}\, e^{\lambda(X - \mathbb{E}X)} && \text{(Markov's ineq.)} \\
&\leq e^{-\lambda t}\, e^{\frac{\lambda^2 \sigma^2}{2}} && \text{(X sub-gaussian).}
\end{aligned}$$

Since this holds for all $\lambda > 0$, we deduce that

$$\mathbb{P}(X - \mathbb{E}X \geq t) \leq \inf_{\lambda > 0} \left\{ e^{-\lambda t + \frac{\lambda^2 \sigma^2}{2}} \right\}$$

$$= \exp\left(- \sup_{\lambda > 0} \left\{ \lambda t - \frac{\lambda^2 \sigma^2}{2} \right\} \right)$$

$$= \exp\left(- \frac{t^2}{2\sigma^2} \right),$$

which concludes the proof □.

Lemma 4

Suppose $\{X_i\}_{i=1}^n$ are independent sub-gaussian random variables with parameters $\{\sigma_i^2\}_{i=1}^n$ respectively. Then

$$\sum_{i=1}^n X_i$$

is sub-gaussian with parameter $\sigma^2 \leq \sum_{i=1}^n \sigma_i^2$.

Proof: Denote $Y_i = X_i - \mathbb{E}X_i$. Then, by independence, we get

$$\log \mathbb{E}\, e^{\lambda \sum_i Y_i} = \log \mathbb{E} \prod_i e^{\lambda Y_i}$$

$$= \log \prod_i \mathbb{E}\, e^{\lambda Y_i}$$

$$= \sum_i \log \mathbb{E}\, e^{\lambda Y_i}.$$

By assumption, we then deduce that

$$\log \mathbb{E}\, e^{\lambda \sum_i Y_i} \leqslant \sum_i \frac{\lambda^2 \sigma_i^2}{2} = \frac{\lambda^2}{2}\left(\sum_i \sigma_i^2\right),$$

which concludes the proof. □

We arrive at the most important fact needed in the sequel, which easily follows by combining Lemmas 3 and 4.

Corollary 5

Suppose $X_1, \dots, X_n$ iid and sub-gaussians with parameter $\sigma^2 > 0$. Then $\forall\, \alpha, t > 0$ such that $t^\alpha \geqslant 1$,

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n X_i - \mathbb{E}X_1 \geqslant \sigma\sqrt{\frac{2\alpha \ln(t)}{n}}\right)$$

and

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n X_i - \mathbb{E}X_1 \leqslant -\sigma\sqrt{\frac{2\alpha \ln(t)}{n}}\right) \leqslant \frac{1}{t^\alpha}.$$

Proof: According to Lemma 4, $\sum_{i=1}^n X_i$ is sub-gaussian with parameter $\leqslant n\sigma^2$. Hence, Lemma 3 implies that

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n X_i - \mathbb{E}X_1 \geqslant \varepsilon\right) = \mathbb{P}\left(\sum_{i=1}^n X_i - \mathbb{E}\left[\sum_{i=1}^n X_i\right] \geqslant n\varepsilon\right)$$

$$\leqslant \exp\left(-\frac{(n\varepsilon)^2}{2n\sigma^2}\right)$$

$$= \exp\left(-\frac{n\varepsilon^2}{2\sigma^2}\right).$$

Setting $\frac{1}{t^\alpha} = \exp\left(-\frac{n\varepsilon^2}{2\sigma^2}\right)$, the previous inequality reads ($\forall\ \alpha, t > 0$ such that $t^\alpha \geqslant 1$)

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}[X_1] \geqslant \sigma\sqrt{\frac{2\alpha \log t}{n}}\right) \leqslant \frac{1}{t^\alpha}.$$

Changing X_i to $-X_i$ implies the second bound.

□

III Upper Confidence Bound (UCB) Strategy.

In this section, we are back to the stochastic bandit problem intoduced in Section I. We'll work under the following assumption:

Assumption (Sub-gaussian losses)

The unknown loss distibutions $\nu_1, \dots, \nu_N$ are all sub-gaussian with parameter $\sigma^2 > 0$ for a known $\sigma^2 > 0$.

The UCB algorithm we'll describe next combines exploration (of the behavior of the different actions) with exploitation (of good actions already identified).

UCB algorithm

Parameters: Subgaussian parameter $\sigma^2 > 0$ and some parameter $\alpha > 2$.

Initialisation: $n_i(0) = 0$, $\hat{m}_i(0) = 0$, $\forall i \in [N]$.

For $t \geqslant 1$:

→ Select $I_t \in \underset{1 \leqslant i \leqslant N}{\operatorname{argmin}} B_t(i)$

where

$$B_t(i) = \begin{cases} \hat{m}_i(t-1) - \sigma \sqrt{\dfrac{2\alpha \ln(t)}{n_i(t-1)}} \,, & \text{if } n_i(t-1) > 0 \\ -\infty \,, & \text{if } n_i(t-1) = 0 \end{cases}$$

and where

$$\hat{m}_i(t-1) = \frac{1}{n_i(t-1)} \sum_{s=1}^{t-1} l_s(I_s) \mathbb{1}\{I_s = i\} .$$

is the sample average of all losses obtained from action i at time $t-1$.

→ Receive $l_t(I_t)$ and update:

$$n_i(t) := n_i(t-1) + \mathbb{1}\{I_t = i\}$$

and

$$\hat{m}_i(t) = \frac{n_i(t-1)\hat{m}_i(t-1) + l_t(I_t)\mathbb{1}\{I_t = i\}}{n_i(t)}$$

Remark: We have phrased the bandit pbl in terms of "losses", instead of "gains" as is usually the case in the literature, to stay coherent with the previous chapter. In this framework, the term $B_t(i)$ has the flavour of a "lower" bound for a confidence interval of m_i instead of an "upper" bound suggested by the name of the algorithm. In the UCB algorithm, the regime switches from a more exploratory phase in the beginning (the confidence intervals are wide in the beginning) to exploitation as time goes on (we come to identify m^* with large prob).

Next we study the performance of the UCB algorithm.

Theorem 6

Supposing the distributions of losses are all subgaussian with parameter $\sigma^2 > 0$, the UCB algorithm with parameter $\alpha > 2$ satisfies

$$R_T \leq \sum_{i:\Delta_i>0} \Delta_i \left(8\sigma^2\alpha \frac{\ln T}{\Delta_i^2} + \frac{\alpha}{\alpha-2} \right)$$

Proof: Recall from Lemma 1 that

$$R_T = \sum_{i:\Delta_i>0} \Delta_i \, E[n_i(T)]$$

To prove the result, we therefore need to show that, $\forall i \in [N]$ such that $\Delta_i > 0$, we have the bound

$$\mathbb{E}\left[n_i(T)\right] \leqslant 8\sigma^2\alpha \frac{\ln T}{\Delta_i^2} + \frac{\alpha}{\alpha - 2}. \quad *_0$$

Hence, from now on, we fix $i \in [N]$ such that $\Delta_i > 0$. Denote

$$t^* = \max\left\{ t \geqslant 1 \ : \ n_i(t) \leqslant \left\lceil 8\sigma^2\alpha \frac{\ln T}{\Delta_i^2} \right\rceil \right\},$$

where $\lceil x \rceil$ denotes the smallest integer larger or equal to x.

Useful observations about t^*:

a) t^* is clearly a random variable, due to the randomness of $n_i(t)$, $t \geqslant 1$.

b) By constuction, the first N actions of the UCB algorithm takes each of the N possible actions once. In particular, this implies that, $\forall j \in [N]$, $\forall t \leqslant N$, $n_j(t) \leqslant 1$. Since $\lceil x \rceil \geqslant 1$ for $x \geqslant 0$, this implies that $t^* \geqslant N$. Also, $\forall j \in [N]$, $\forall t \geqslant N$, $n_j(t) \geqslant 1$.

c) Finally, note that we clearly have that

$$n_i(t^*) = \left\lceil 8\sigma^2\alpha \frac{\ln T}{\Delta_i^2} \right\rceil \geqslant 8\sigma^2\alpha \frac{\ln T}{\Delta_i^2}.$$

In the rest of the proof, we'll look separately at the behaviour of $n_i(T)$ on the event $\{t^* \geqslant T\}$ and $\{t^* < T\}$.

On the event $\{t^* \geqslant T\}$:

Clearly, we have that

$$\begin{aligned}
\mathbb{E}\left[\mathbb{1}\{t^* \geqslant T\}\, n_i(T)\right] &\leqslant \mathbb{E}\left[\mathbb{1}\{t^* \geqslant T\} \left\lceil \frac{8\sigma^2 \alpha \ln T}{\Delta_i^2} \right\rceil\right] \\
&= \left\lceil \frac{8\sigma^2 \alpha \ln T}{\Delta_i^2} \right\rceil \mathbb{P}(t^* \geqslant T) \\
&\leqslant \left(\frac{8\sigma^2 \alpha \ln T}{\Delta_i^2} + 1 \right) \mathbb{P}(t^* \geqslant T). \qquad \ast_1
\end{aligned}$$

On the event $\{t^* < T\}$:

We can write that

$$\begin{aligned}
\mathbb{E}\left[\mathbb{1}\{t^* < T\}\, n_i(T)\right] &= \mathbb{E}\left[\mathbb{1}\{t^* < T\} \sum_{t=1}^{T} \mathbb{1}\{I_t = i\}\right] \\
&= \mathbb{E}\left[\mathbb{1}\{t^* < T\} \left(n_i(t^*) + \sum_{t=t^*+1}^{T} \mathbb{1}\{I_t = i\} \right)\right]
\end{aligned}$$

Since $n_i(t^*) = \left\lceil \dfrac{8\sigma^2 \alpha \ln T}{\Delta_i^2} \right\rceil \leqslant \dfrac{8\sigma^2 \alpha \ln T}{\Delta_i^2} + 1$,

we deduce that

$$\mathbb{E}\left[\mathbb{1}\{t^* < T\}\, n_i(T)\right]$$
$$\leq \left(\frac{8\sigma^2\alpha \ln T}{\Delta_i^2} + 1\right)\mathbb{P}(t^* < T) + \mathbb{E}\left[\mathbb{1}\{t^* < T\}\sum_{t=t^*+1}^{T} \mathbb{1}\{I_t = i\}\right] \tag{$*_2$}$$

Combining $*_1$ and $*_2$, we deduce that

$$\mathbb{E}\left[n_i(T)\right] = \mathbb{E}\left[\mathbb{1}\{t^* \geq T\}\, n_i(T)\right] + \mathbb{E}\left[\mathbb{1}\{t^* < T\}\, n_i(T)\right]$$
$$\leq 8\,\frac{\sigma^2\alpha \ln T}{\Delta_i^2} + 1 + \mathbb{E}\left[\mathbb{1}\{t^* < T\}\sum_{t=t^*+1}^{T} \mathbb{1}\{I_t = i\}\right]$$

Since $1 + \frac{2}{\alpha - 2} = \frac{\alpha}{\alpha - 2}$, the proof is complete provided we show that

$$\mathbb{E}\left[\mathbb{1}\{t^* < T\}\sum_{t=t^*+1}^{T} \mathbb{1}\{I_t = i\}\right] \leq \frac{2}{\alpha - 2}.$$

In the rest of the proof, we therefore focus on this fact. Let us rewrite

$$\mathbb{E}\left[\mathbb{1}\{t^* < T\}\sum_{t=t^*+1}^{T} \mathbb{1}\{I_t = i\}\right] = \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{t^* < t \leq T,\ I_t = i\}\right]$$
$$= \sum_{t=1}^{T} \mathbb{P}\left(t^* < t \leq T,\ I_t = i\right)$$

Then, observe that the following fact holds:

Observation

We have the inclusion of events

$$\{ t^* < t \leqslant T \, , \, I_t = i \} \subset A_t \cup B_t$$

where $A_t := \left\{ \hat{m}_{i^*}(t-1) - \sigma \sqrt{\frac{2\alpha \ln(t)}{n_{i^*}(t-1)}} \geqslant m_{i^*} \right\}$,

and $B_t := \left\{ \hat{m}_i(t-1) - \sigma \sqrt{\frac{2\alpha \ln(t)}{n_i(t-1)}} < m_i \right\}$.

Proof of this fact: By contradiction, suppose that the above inclusion does not hold, i.e., that we may have: $t^* < t \leqslant T$, $I_t = i$ as well as

$$\hat{m}_{i^*}(t-1) - \sigma \sqrt{\frac{2\alpha \ln(t)}{n_{i^*}(t-1)}} < m_i^* \qquad *_3$$

and

$$\hat{m}_i(t-1) - \sigma \sqrt{\frac{2\alpha \ln(t)}{n_i(t-1)}} \geqslant m_i . \qquad *_4$$

Then, we get (next page)

$$\hat{m}_{i^*}(t-1) - \sigma\sqrt{\frac{2\alpha\ln(t)}{n_{i^*}(t-1)}} < m_{i^*} \qquad (\text{this is } *_3)$$

$$= m_i - \Delta_i \qquad (\text{definition of } \Delta_i)$$

But since $t-1 \geqslant t^*$ we get by definition of t^* that

$$n_i(t-1) \geqslant \left\lceil \frac{8\sigma^2\alpha\ln(T)}{\Delta_i^2} \right\rceil \geqslant \frac{8\sigma^2\alpha\ln(T)}{\Delta_i^2},$$

which implies that

$$\Delta_i \geqslant 2\sigma\sqrt{\frac{2\alpha\ln(T)}{n_i(t-1)}} \overset{\overset{t\leqslant T}{\downarrow}}{\geqslant} 2\sigma\sqrt{\frac{2\alpha\ln(t)}{n_i(t-1)}}.$$

In particular, we get that

$$\hat{m}_{i^*} - \sigma\sqrt{\frac{2\alpha\ln(t)}{n_{i^*}(t-1)}} < m_i - 2\sigma\sqrt{\frac{2\alpha\ln(t)}{n_i(t-1)}}$$

$$\leqslant \hat{m}_i - \sigma\sqrt{\frac{2\alpha\ln(t)}{n_i(t-1)}} \qquad (\text{because of } *_4).$$

Now this contradicts the fact that $I_t = i$ since $\Delta_i > 0$ imposes that $i^* \neq i$ □.

Given this observation, we can now write that

$$\sum_{t=1}^{T} \mathbb{P}(t^* < t \leqslant T,\ I_t = i) \overset{\substack{t^* \geqslant N \\ \downarrow}}{=} \sum_{t=N+1}^{T} \mathbb{P}(t^* < t \leqslant T,\ I_t = i)$$

$$\leqslant \sum_{t=N+1}^{T} \mathbb{P}(A_t) + \sum_{t=N+1}^{T} \mathbb{P}(B_t).$$

Finally we'll show that both of these sums are at most $\frac{1}{\alpha - 2}$. We prove this fact for the 1st sum only since the 2nd may be treated similarly.

For every $t \geqslant N+1$, $\overset{\forall j}{n_j}(t-1) \geqslant 1$. Hence, for $t \geqslant N+1$,

$$\mathbb{P}(A_t) = \mathbb{P}\left(\bigcup_{s=1}^{t-1} \{ n_{i^*}(t-1) = s,\ A_t \} \right)$$

$$\leqslant \sum_{s=1}^{t-1} \mathbb{P}\left(n_{i^*}(t-1) = s,\ A_t \right) \quad \text{(union bound)}$$

$$\leqslant \sum_{s=1}^{t-1} \mathbb{P}\left(\frac{1}{s} \sum_{j=1}^{s} l_j(i^*) - \sigma \sqrt{\frac{2\alpha \ln(t)}{s}} \geqslant m_{i^*} \right)$$

$$\text{(since } \{ l_j(i^*) \}_{j \geqslant 1} \text{ are i.i.d.)}$$

$$\leqslant \sum_{s=1}^{t-1} \frac{1}{t^{\alpha}} \quad \text{(Corollary 5)}$$

$$= (t-1)\, t^{-\alpha}.$$

As a result

$$\sum_{t=N+1}^{T} \mathbb{P}(A_t) \leqslant \sum_{t=N+1}^{T} (t-1)\, t^{-\alpha} \leqslant \sum_{t=2}^{+\infty} t^{1-\alpha}$$

$$\leqslant \int_{1}^{+\infty} t^{1-\alpha}\, dt = \frac{1}{\alpha - 2}.$$

This concludes the proof.

□