MVA ENS Cachan Paris Saclay

Prediction for Individual Sequences
Notes de Cours

Cours donné par Vianney Perchet
Note prises par Adrien Lina

2 février 2018

1 Lecture 2 : Modèle en full monitoring

Rappel de ce qui a été vu durant la Lecture 1 :
— K bras iid, σ^2 sous-gaussien dans $[0,1]$ X_t^k ;
— $\mathbb{E}[X_t^k] = \mu^k$; Par notation, $\mu^1 > \mu^2 > ... > \mu^K$;
— On les échantillone séquentiellement $\pi_t \in [K]$
— $R_T = T\mu^* - \sum_{t=1}^{T} \mu^{\pi_k}$
— Full-info : tous les X_t^k sont observés
— $\pi_{t+1} = \text{argmax}\, \overline{X_t^k}$ ce qui donne un regret

$$\mathbb{E}[R_T] \lesssim \frac{\sigma^2}{\Delta}$$

— Tout algo a un regret pour tout bandit borné inférieurement

$$\mathbb{E}[R_T] \gtrsim \frac{\sigma^2}{\Delta}$$

On regarde aujourd'hui le cas bandit.

1.1 Modification du problème

On observe uniquement le résultat de la décision. Les observations disponibles sont :

$$X_1^{\pi_1}, ..., X_T^{\pi_T}$$
$$\pi_1, ..., \pi_T$$

1.2 Approche naïve

Exemple 1.1. On peut essayer :

$$\pi_{t+1} = \operatorname{argmax} \widehat{X_t^k} \qquad \text{où } \widehat{X_t^k} = \frac{\sum\limits_{s:\pi_s=k} X_s^{\pi_s}}{\sum\limits_{s:\pi_s=k} 1}$$

avec $\frac{0}{0} = +\infty$.

On peut penser que cela fonctionnera : si les $X_s^{\pi_s}$ sont IID, avec la loi des grands nombres, $\widehat{X_t^k} \approx \overline{X_t^k}$. Dans les faits, par la manière dont on choisit les π_k, les échantillons ne sont pas IID, et donc ça ne fonctionne pas.

C'est en fait le pire algo possible, le regret est linéaire.

$$V^1 \sim \mathcal{B}(1/2)$$
$$V^2 = \delta_{\frac{1}{4}}$$

L'algo va prendre $\pi_1 = 1$ puis $\pi_2 = 2$ (ou vice versa).

$$\widehat{X_2^2} = \frac{1}{4}$$
$$\widehat{X_2^1} = \begin{cases} 1 & \text{avec proba } 1/2 \\ 0 & \text{avec proba } 1/2 \end{cases}$$

Donc, avec proba 1/2, l'algo sample le bras 2 chaque étape.

Donc

$$\mathbb{E}[R_T] \geq \frac{1}{2}(T-1)\frac{1}{4} \geq$$
$$\geq \frac{T}{8} - \frac{1}{2}$$

Le problème vient du fait que $\mathbb{E}[\widehat{X_t^1}] < \frac{1}{2}$, et donc $\mathbb{E}[\widehat{X_t^1}] \neq \mu^1$: **la moyenne empirique est biasée**. L'algorithme va renforcer ses propres erreurs.

Remarque 1.1. On peut effectuer des tests pour vérifier qu'un algo ne drifte pas. Par exemple, si on a déjà un algo, et qu'on met en place un nouveau.

En comparant les performances au cours du temps des deux algos (neuf vs ancien), on peut voir le drift : le nouvel algo, qui a priori parait mieux (ex : +5% de gain), peut finir par drifter par rapport à l'ancien (ex : -10% de gain au bout de 3 mois).

Cela permet de vérifier que le nouvel algo ne corrompe pas les données. C'est un garde-fou.

1.3 UCB

Il y a deux solutions :

1. "Forcer l'exploration" (ϵ-greedy) : en choisissant à l'étape t un bras complètement au hasard avec proba ϵ et avec proba $1-\epsilon$ en exploitant l'argmax $\widehat{X_t^k}$.
   Par exemple, $\epsilon = 5\%$ ou $\epsilon_t \lesssim \frac{1}{t}$.
2. "Essayer de réduire le biais", ou avoir un biais positif [1] : en rajoutant un terme d'erreur à $\widehat{X_t^k}$ "donné" par Hoeffding.

Le premier cas est moins intéressant car moins fin et dispose de moins de résultats théoriques. L'autre est une catégorie d'algorithme que nous allons étudier.

Définition 1.1 (Upper Confidence Bound (UCB))**.** *On prend*

$$\pi_{t+1} = \operatorname{argmax} \widehat{X_t^k} + \sqrt{\frac{8\,\sigma^2 \log(t)}{N_t^k}}$$

avec $\frac{0}{0} = +\infty$.

Théorème 1.1 (Regret espéré UCB)**.**

$$\mathbb{E}[R_T] \leq \sum_{k=2}^{K} \frac{32\,\sigma^2 \log(T)}{\Delta_k} + 5 \sum_{k=2}^{K} \Delta_k$$

avec $\Delta_k = \mu^* - \mu^k$.

Démonstration. L'idée : on sample k au lieu de k^* si :

$$\underbrace{\widehat{X_t^k}}_{\approx \mu_k} + \sqrt{\frac{8\,\sigma^2 \log(t)}{N_t^k}} \geq \underbrace{\widehat{X_t^*}}_{\approx \mu^*} + \underbrace{\sqrt{\frac{8\,\sigma^2 \log(t)}{N_t^*}}}_{\approx 0}$$

i.e. si $\sqrt{\frac{8\,\sigma^2 \log(t)}{N_t^k}} \geq \Delta_k$ i.e. si $N_t^k \leq \frac{8\sigma^2 \log(t)}{\Delta_k^2}$.

La preuve formelle : on a

$$\begin{aligned}
\{\pi_t = k\} &\subset \{t = k\} \cup \left\{ \widehat{X_t^k} + \sqrt{\frac{8\,\sigma^2 \log(t)}{N_t^k}} \geq \widehat{X_t^*} + \sqrt{\frac{8\,\sigma^2 \log(t)}{N_t^*}} \right\} \\
&\subset \{t = k\} \cup \left\{ \widehat{X_t^k} + \sqrt{\frac{8\,\sigma^2 \log(t)}{N_t^k}} \geq \mu_k \right\} \cup \left\{ \widehat{X_t^*} + \sqrt{\frac{8\,\sigma^2 \log(t)}{N_t^*}} \leq \mu^* \right\} \\
&\qquad \cup \left\{ \mu^k + 2\sqrt{\frac{8\sigma^2 \log(t)}{N_t^k}} \geq \mu^*; \pi_{t+1} = k \right\}
\end{aligned}$$

1. Un biais positif va faire qu'on échantillone trop le bras en question. Mais si le biais positif tend vers 0 avec le nombre d'échantillon, ce n'est pas gênant.

On utilise $\{A \geq B\} \subset \{C \geq B\} \cup \{A \geq D\} \cup \{D \geq C\}$ qui vient des événements contraires $\{A < B\} \subset \{C < B\} \cup \{A < D\} \cup \{D < C\}$ i.e. $A < D < C < B \Rightarrow A < B$.

Donc

$$\begin{aligned}\mathbb{E}[N_t^k] &= \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{\pi_t = k\}\right] \\ &= 1 + \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left\{\widehat{X_t^k} + \sqrt{\frac{8\,\sigma^2 \log(t)}{N_t^k}} \geq \mu_k\right\}\right] + \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left\{\widehat{X_t^*} + \sqrt{\frac{8\,\sigma^2 \log(t)}{N_t^*}} \leq \mu^*\right\}\right] \\ &\qquad + \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left\{\mu^k + 2\sqrt{\frac{8\sigma^2 \log(t)}{N_t^k}} \geq \mu^*; \pi_{t+1} = k\right\}\right]\end{aligned}$$

Or Hoeffding ne peut s'appliquer à $\mathbb{P}\left(\widehat{X_t^k} - \mu^k \geq \sqrt{\frac{2\sigma^2 \log(t^4)}{N_t^k}}\right)$, mais

$$\begin{aligned}\mathbb{P}\left(\widehat{X_t^k} - \mu^k \geq \sqrt{\frac{2\sigma^2 \log(t^4)}{N_t^k}}\right) &\leq \mathbb{P}\left(\exists s \in [1,t], \overline{X_s^k} - \mu^k \geq \sqrt{\frac{2\sigma^2 \log(t^4)}{s}}\right) \\ &\leq \sum_{s=1}^{t} \mathbb{P}\left(\overline{X_s^k} - \mu^k \geq \sqrt{\frac{2\sigma^2 \log(t^4)}{s}}\right) \\ &\leq t\frac{1}{t^4} \qquad \text{(Hoeffding)} \\ &\leq \frac{1}{t^3}\end{aligned}$$

Donc le deuxième terme est bornée par $\sum_{t=1}^{T} \frac{1}{t^3} \leq \frac{\pi^2}{6}$. Idem pour le troisième terme.

Enfin

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left\{\widehat{X_t^k} + \sqrt{\frac{8\,\sigma^2 \log(t)}{N_t^k}} \geq \mu_k\right\}\right] \leq 32\frac{\sigma^2 \log(T)}{\Delta_k^2}$$

□

Remarque 1.2. Donc le regret de UCB est de l'ordre de $\min\left\{\frac{\sigma^2 \log(T)}{\Delta}, \Delta T\right\}$ si $k = 2$. En particulier, dans le pire es cas, son regret est de l'ordre de $6\sqrt{T \log(T)}$.

Le coût de passer en bandit par rapport au contexte full info est de multiplier par $\log(T)$ le regret : c'est peu par rapport à toute l'info qu'on a perdu.

De plus, le pire des cas n'est pas plus dur en bandit qu'en full info.

On peut donc se demander si UCB est optimal.

1.4 Bornes inférieures

Théorème 1.2. *Soit $V_1 = (\delta_0; \mathcal{N}(-\Delta, 1))$ et $V_2 = (\delta_0; \mathcal{N}(\Delta, 1))$ deux bandits à 2 bras.*

Alors, pour tout algo,

$$\max\{\mathbb{E}_1[R_T]; \mathbb{E}_2[R_T]\} \quad \gtrsim \quad \frac{\log(T\Delta^2)}{\Delta}$$

pour tout $T \gtrsim \frac{1}{\sqrt{\Delta}}$.

Démonstration. Le bras 1 est non-informatif. Donc tout repose sur N_t^2 car c'est ce qui porte de l'information.

On remarque que $\mathbb{E}_1[R_T] = \Delta\mathbb{E}_1[N_T^2]$.

On sait que, lemme prouvé dans la lecture 1,

$$\mathbb{E}_1[R_T] + \mathbb{E}_2[R_T] \geq \frac{\Delta}{2}\sum_{t=1}^{T} e^{-KL(V_1,V_2)}$$

Dans notre cas, et comme le bras 1 est non informatif,

$$KL(V_1^{\otimes t}, V_2^{\otimes t}) = 2\Delta^2\mathbb{E}_1[N_t^2]$$

Donc

$$\mathbb{E}_1[R_T] + \mathbb{E}_2[R_T] \geq \frac{\Delta}{2}Te^{-2\Delta^2\mathbb{E}_1[N_T^2]}$$

Donc

$$\begin{aligned}
\max\{\mathbb{E}_1[R_T]; \mathbb{E}_2[R_T]\} &\geq \max\left\{\mathbb{E}_1[R_T]; \frac{\mathbb{E}_1[R_T] + \mathbb{E}_2[R_T]}{2}\right\} \\
&\geq \max\left\{\Delta\mathbb{E}_1[N_T^2]; \frac{\Delta}{4}Te^{-2\Delta^2\mathbb{E}_1[N_T^2]}\right\} \\
&\geq \min_{x\in[0,T]} \max\left\{\Delta x; \frac{\Delta}{4}Te^{-2\Delta^2 x}\right\} \\
&\gtrsim \frac{\log(T\Delta^2)}{\Delta} \qquad \text{si} \quad T \geq \frac{1}{\sqrt{\Delta}}
\end{aligned}$$

□

Remarque 1.3. Ainsi UCB est presque optimal, parce qu'aucun autre algorithme ne peut faire toujours mieux qu'UCB (à $\log(\Delta^2)/\Delta$ près).

Proposition 1.1 (Borne asymptotique)**.** *Soit π un algo dont le regret est toujours $o(T^\alpha)$ $\forall\alpha > 0$. Alors, nécessairement*

$$\liminf_{T\to+\infty} \frac{\mathbb{E}[R_T]}{\log(T)} \quad \gtrsim \quad \frac{1}{\Delta}$$

1.5 Un autre algo optimal

C'est à dire optimal en $\frac{\log(T\Delta^2)}{\Delta}$.
On prend pour simplifier $k = 2$.

Définition 1.2 (Explore Then Comit (ETC) *aka* Successive Elimination (SC))**.** *Etant donné un horizon T, on alterne entre 1 et 2 jusqu'à ce que*

$$\widehat{X_t^i} - \sqrt{\frac{16\sigma^2 \log(\frac{2T}{t})}{t}} \geq \widehat{X_t^j} + \sqrt{\frac{16\sigma^2 \log(\frac{2T}{t})}{t}}$$

On arrête alors et on tire i jusqu'à T.

Remarque 1.4. On cherche en fait dès le début quel est le meilleur bras, et après on y reste une fois qu'on sait que celui là est le meilleur.

Remarque 1.5. Pour l'appliquer à un horizon non fixé, on prend $T = 2^p$, et si on n'a pas atteint la condition, on double T, et on continue, etc.

Théorème 1.3.

$$\mathbb{E}[R_T] \lesssim \frac{\log(T\Delta^2)}{\Delta}$$

Démonstration. Si tout se passe bien on arrête d'alterner quand

$$\begin{aligned} & \mu^1 - \sqrt{\frac{16\sigma^2 \log(\frac{2T}{t})}{t}} \geq \mu^2 + \sqrt{\frac{16\sigma^2 \log(\frac{2T}{t})}{t}} \\ \Leftrightarrow \quad & \Delta^2 \geq \frac{64\sigma^2 \log(\frac{T}{t})}{t} \end{aligned}$$

"En gros", $t \approx 2\frac{64\sigma^2 \log(\frac{T\Delta^2}{64\sigma^2})}{\Delta^2}$ car $t \geq \frac{64\sigma^2}{\Delta^2}\log(\frac{T}{t})$.
Donc $t \approx \frac{64\sigma^2}{\Delta^2}$ et donc $t \approx \frac{64\sigma^2}{\Delta^2}\log(\frac{T}{64\sigma^2})$.
Pour la preuve, il suffit de prendre t^* tel que

$$\sqrt{\frac{16\sigma^2 \log(\frac{2T}{t^*})}{t^*}} \leq \frac{\Delta}{4}$$

et démontrer grace à Hoeffding "maximal" que

$$\mathbb{P}\left(\widehat{X_{t^*}^i} - \sqrt{\frac{16\sigma^2 \log(\frac{2T}{t^*})}{t^*}} \geq \widehat{X_{t^*}^j} + \sqrt{\frac{16\sigma^2 \log(\frac{2T}{t^*})}{t^*}}\right) \leq \frac{t^*}{T}$$

□