MVA ENS Cachan Paris Saclay

Prediction for Individual Sequences
Notes de Cours

Cours donné par Vianney Perchet
Note prises par Adrien Lina

3 mars 2018

1 Lecture 3 : Bandits contextuels

1.1 Rappels

Cadre : pour chaque $t \geq 1$,
— le joueur choisit $p_t \in \Delta_K = \{p \in [0,1]^K : \sum p^{(k)} = 1\}$
— on observe
  — son reward $X_t^{(k)}$ "bandit"
  — $X_t^{(1)}, ..., X_t^{(K)}$ "info complête"
But : minimiser le regret $R_T^{(k)} = \sum_{t=1}^T X_t^{(k)} - \sum_{t=1}^T X_t^{(\pi_t)}$
On a vu jusqu'ici comment minimiser le pseudoregret $\max_k \mathbb{E}[R_T^{(k)}]$.
On va maintenant voir comment minimiser $\max_k R_T^{(k)}$.

1.2 EXP3.P

Définition 1.1 (EXP3.P)**.** *On choisit chaque bras avec une probabilité :*

$$p_t^{(k)} = (1-\gamma)\frac{e^{\eta \widehat{R_T^{(k)}}}}{\sum_{j=1}^K e^{\eta \widehat{R_T^{(j)}}}} + \frac{\gamma}{K}$$

où

$$\widehat{R_T^{(k)}} = \sum_{s=1}^{t-1}\left(\widehat{X_s^{(k)}} - \widehat{X_s^{(\pi_s)}}\right)$$
$$\widehat{X_t^{(k)}} = X_t^{(j)}\mathbb{1}_{\{k=\pi_t\}} + \beta$$

Théorème 1.1. *EXP3P vérifie :*

$$R_T \leq T\sqrt{TK\log K} + \sqrt{\frac{TK}{\log K}}\log\left(\frac{1}{\delta}\right)$$

pour $\eta > 0$, $\gamma \in [0,1]$ *et* $\beta \in [0,1]$ *choisis tels que*

$$\eta(1+\beta)\frac{K}{\gamma} \leq 1$$

Lemme 1.1. *pour* $\beta \in [0,1]$,

$$\sum_{t=1}^{T} \widehat{X_t^{(k)}} > \sum_{t=1}^{T} X_t^{(k)} - \frac{\log(1/\delta)}{\beta}$$

Démonstration. On pose $q_t^{(k)} = \frac{e^{\eta \widehat{R_T^{(k)}}}}{\sum_{j=1}^{K} e^{\eta \widehat{R_T^{(j)}}}}$.

D'après la preuve de EXP, on a

$$\max_k \sum_{t=1}^{T} \widehat{X_t^{(k)}} - \sum_{t=1}^{T} \mathbb{E}_{j\sim q_t}\left[\widehat{X_t^{(j)}}\right] \leq \frac{\log K}{\eta} + \eta \sum_{t=1}^{T} \mathbb{E}_{j\sim q_t}\left[\left(\widehat{X_t^{(j)}}\right)^2\right] \quad (1)$$

avec

$$\mathbb{E}_{j\sim q_t}\left[\widehat{X_t^{(j)}}\right] = \sum_{j=1}^{K} q_t^{(j)} \widehat{X_t^{(j)}}$$

On calcule :

$$\begin{aligned}\mathbb{E}_{j\sim q_t}\left[\widehat{X_t^{(j)}}\right] &= \sum_{j=1}^{K} q_t^{(j)} \widehat{X_t^{(j)}} \\ &\leq \frac{1}{1-\gamma}\sum_{j=1}^{K} p_t^{(j)} \widehat{X_t^{(j)}}\end{aligned}$$

car $p_t^{(k)} = (1-\gamma)q_t^{(k)} + \frac{\gamma}{K}$ et donc $q_t^{(k)} \leq \frac{p_t^{(k)}}{1-\gamma}$.

Donc

$$\begin{aligned}\mathbb{E}_{j\sim q_t}\left[\widehat{X_t^{(j)}}\right] &= \sum_{j=1}^{K} q_t^{(j)} \widehat{X_t^{(j)}} \\ &\leq \frac{1}{1-\gamma}\sum_{j=1}^{K} p_t^{(j)} \frac{X_t^{(j)} \mathbb{1}_{\{k=\pi_t\}} + \beta}{p_t^{(j)}} \\ &\leq X_t^{\pi_t} + K\beta\end{aligned}$$

De même,

$$\mathbb{E}_{j\sim q_t}\left[\left(\widehat{X_t^{(j)}}\right)^2\right] \leq \frac{1}{1-\gamma}\sum_{k=1}^{K} p_t^{(k)}\left(\widehat{X_t^{(j)}}\right)^2$$
$$\leq \frac{1}{1-\gamma}\sum_{k=1}^{K}\widehat{X_t^{(k)}}(1+\beta)$$

car $p_t^{(k)} = X_t^{(k)}\mathbb{1}_{\{k=\pi_t\}} + \beta \leq 1+\beta$

Donc

$$\mathbb{E}_{j\sim q_t}\left[\left(\widehat{X_t^{(j)}}\right)^2\right] \leq \frac{1+\beta}{1-\gamma}\sum_{t=1}^{T}\sum_{k=1}^{K}\widehat{X_t^{(j)}}$$
$$\leq \frac{1+\beta}{1-\gamma}\sum_{k=1}^{K}\sum_{t=1}^{T}\widehat{X_t^{(j)}}$$
$$\leq \frac{(1+\beta)K}{1-\gamma}\max_k\left(\sum_{t=1}^{T}\widehat{X_t^{(k)}}\right)$$

En remplaçant dans (1) :

$$\max_k\sum_{t=1}^{T}\widehat{X_t^{(k)}} - \frac{1}{1-\gamma}\left(\sum_{t=1}^{T}X_t^{\pi_t} - TK\beta\right) \leq \frac{\log K}{\eta} + \eta\frac{(1+\beta)K}{1-\gamma}\max_k\left(\sum_{t=1}^{T}\widehat{X_t^{(k)}}\right)$$

i.e.

$$(1-\gamma)\max_k\left(\sum_{t=1}^{T}\widehat{X_t^{(k)}}\right) \leq \sum_{t=1}^{T}X_t^{(\pi_k)} + TK\beta + \frac{1-\gamma}{\eta}\log K + \underbrace{\eta(1+\beta)K}_{\leq\gamma \text{ par hypothèse}}\max_k\sum_{t=1}^{T}\widehat{X_t^{(k)}}$$

Donc

$$(1-2\gamma)\max_k\left(\sum_{t=1}^{T}\widehat{X_t^{(k)}}\right) \leq \sum_{t=1}^{T}X_t^{(\pi_k)} + TK\beta + \frac{1-\gamma}{\eta}\log K$$

Or d'après le lemme

$$\mathbb{P}\left(\max_k\left(\sum_{t=1}^{T}\widehat{X_t^{(k)}}\right) \leq \max_k\left(\sum_{t=1}^{T}X_t^{(k)}\right) - \frac{\log\frac{K}{\delta}}{\beta}\right) = \mathbb{P}\left(\exists k : \sum_{t=1}^{T}\widehat{X_t^{(k)}} \leq \sum_{t=1}^{T}X_t^{(k)} - \frac{\log\frac{K}{\delta}}{\beta}\right)$$
$$\leq \sum_{k=1}^{K}\mathbb{P}\left(\sum_{t=1}^{T}\widehat{X_t^{(k)}} \leq \sum_{t=1}^{T}X_t^{(k)} - \frac{\log\frac{K}{\delta}}{\beta}\right)$$
$$\leq \sum_{k=1}^{K}\frac{\delta}{K}$$
$$\leq \delta$$

Donc avec proba au moins $1-\delta$:

$$(1-2\gamma)\max_k\left(\sum_{t=1}^{T}\widehat{X_t^{(k)}}\right)\leq\sum_{t=1}^{T}X_t^{(\pi_k)}+TK\beta+\frac{\log K}{\eta}$$

En reorganisant :

$$R_t=\max_k\sum_{t=1}^{T}X_t^{(k)}-\sum_{t=1}^{T}X_t^{(\pi_t)}\leq TK\beta+\frac{\log K}{\eta}+2\gamma\underbrace{\max_k\left(\sum_{t=1}^{T}X_t^{(k)}\right)}_{\leq T}+\underbrace{(1-2\gamma)}_{\leq 1}\frac{\log\frac{K}{\delta}}{\beta}$$

$$\leq TK\beta+\frac{\log\frac{K}{\delta}}{\beta}+\frac{\log K}{\eta}+2\gamma T$$

On choisit $\gamma=2\eta K$ de sorte que $\eta(1+\beta)\frac{K}{\gamma}\leq 1$ pour $\beta\in[0,1]$:

$$R_T\leq TK\beta+\frac{\log\frac{K}{\delta}}{\beta}+\frac{\log K}{\eta}+4\eta KT$$

On choisit η tel que : $\frac{\log K}{\eta}=4\eta KT$, donc $\eta=\frac{1}{2}\sqrt{\frac{\log K}{KT}}$ e
On choisit β tel que $TK\beta=\frac{\log K}{\beta}$, donc $\beta=\sqrt{\frac{\log K}{KT}}$.
En remplaçant dans la borne :

$$R_T\leq\gamma\sqrt{TK\log K}+\sqrt{\frac{TK}{\log K}}\log\frac{1}{\delta}$$

□

Remarque 1.1. Les algos précédents EXP, EXP3 et EXP3.P dépendent de paramètres η, β et γ qui ont été optimisés en fonction de T.

Les résultats qu'on a vu ne sont donc valabes que pour un horizon T connu à l'avance.

Comment faire pour tout t ?

Définition 1.2 (Doubling Trick)**.** *À chaque fois que t est une puissance de 2, on redémarre l'algorithme avec $\eta=\sqrt{\frac{\log K}{KT}}$ eb oubliant tout ce qui a été appris.*

Théorème 1.2. *Avec EXP3 et le doubling Trick, on obtient :*

$$\max_k\mathbb{E}[R_T^{(k)}]\leq 7\sqrt{TK\log K}$$

Remarque 1.2. On a juste perdu un facteur multiplicatif.

Démonstration. Si $2^M\leq T<2^{M+1}$, T inconnu.

Le pseudo-regret de EXP3 + Doubling Trick devient

$$\begin{aligned}
\max_k \mathbb{E}[R_T^{(k)}] &= \max_k \mathbb{E}\left[\sum_{t=1}^{T} X_t^{(k)} - X_t^{(\pi_t)}\right] \\
&\leq \max_k \mathbb{E}\left[\sum_{m=0}^{M}\sum_{t=2^m}^{2^{m+1}} X_t^{(k)} - X_t^{(\pi_t)}\right] \\
&\leq \sum_{m=0}^{M} \underbrace{\max_k \mathbb{E}\left[\sum_{t=2^m}^{2^{m+1}} X_t^{(k)} - X_t^{(\pi_t)}\right]}_{\text{EXP3 de paramètre } \eta = \sqrt{\frac{\log K}{K 2^m}}} \\
&\leq \sum_{m=0}^{M} 2\sqrt{2^m K \log K} \\
&\leq 2(1+\sqrt{2})\sqrt{2^{M+1} K \log K} \\
&\leq 7\sqrt{TK\log K}
\end{aligned}$$

□

Théorème 1.3. *Le doubling trick fonctionne pour n'importe quel algorithme sequentiel en regret $\mathcal{O}(T^a)$ pour le transformer en un algo de $\mathcal{O}(T^a)$ pour tout T.*

Remarque 1.3. En pratique, c'est très mauvais. On préfère utiliser des paramètres η_t où T est remplacé par t. Théoriquement, on peut prouver que cela fonctionne, mais c'est plus laborieux.

1.3 Optimalité des algos

On a vu des algos qui ont un regret
— en $\mathcal{O}(\sqrt{T\log K})$ en info complète ;
— en $\mathcal{O}(\sqrt{TK\log K})$ en bandit ;
Peut-on faire mieux ?

Théorème 1.4. *En info complète, pour tout algo $R_t \geq const \times \sqrt{T\log K}$. En bandit, pour tout algo $R_t \geq const \times \sqrt{TK}$.*

Intuition. — Pour l'info complète :
Considérons un adversaire qui choisit $X_t^{(k)} \sim \text{Ber}(\frac{1}{2})$ iid.

$$\mathbb{E}[\sum_{t=1}^{T} X_t^{(k)}] = \frac{T}{2}$$

$$\mathbb{E}[\max_k \sum_{t=1}^{T} X_t^{(k)}] = \mathbb{E}[\max\{K \text{ marches aléatoires}\}]$$

$$\simeq \frac{T}{2} + \sqrt{T \log K}$$

— Pour le bandit :
L'adversaire ne choisit que des $\text{Ber}(\frac{1}{2})$ sauf un bras k^* suivant $\text{Ber}(\frac{1}{2}+\epsilon)$. Ainsi, pour différencier les bras, il faut tirer ϵ^{-2} observations de chaque bras.
En effet, d'après le théorème limite central, après T_i observations d'une variable aléatoire, on peut estimer son espérance avec une erreur de l'ordre de $\frac{1}{\sqrt{T_i}}$.
Un des bras est tiré moins de $\frac{T}{K}$ fois. Si c'est le bras k^*, on ne pourra se rendre compte que c'est le meilleur (normalement à partir de $\frac{T}{K} \leq \epsilon^{-2}$). On aura tiré les autres bras au moins $T - \frac{T}{K} = (1 - \frac{1}{K})T$ fois. On a donc un regret pour $\epsilon = \sqrt{\frac{T}{K}}$ de $(1 - \frac{1}{K})T\epsilon \approx (1 - \frac{1}{K})\sqrt{TK}$.

□

1.4 Online learning with expertise

On a plusieurs algorithmes / experts qui nous proposent des avis à chaque temps t. On veut se rapprocher de la performance du meilleur expert.

Définition 1.3 (Online learning with expertise)**.**
***Cadre** : à chaque temps t,*
— *N expertes proposent des prévisions $\xi_t^{(i)} \in \mathcal{A}$ pour $i = 1, ..., N$;*
— *On fait notre prévision $\pi_t \in \mathcal{A}$ et l'adversaire choisit une fonction de gain $g_t : \mathcal{A} \to [0, 1]$;*
— *Le joueur observe*
  — *g_t en info complète ;*
  — *$g_t(\pi_t)$ en bandit.*
***But** : minimiser le regret par rapport au experts :*

$$R_T = \max_k \sum_{t=1}^{T} g_t(\xi_t^{(k)}) - g_t(\pi_t)$$

Remarque 1.4. Si $\mathcal{A} = \{1, ..., N\}$ et $\xi_t^{(i)} = i$ et $g_t(i) = X_t^{(i)}$, on retrouve le cardre de bandit traditionnel. Cependant, ce cadre est plus riche : on peut par exemple considérer des ensemble $\mathcal{A}$ continus.

Remarque 1.5. en appliquant les algos précédants en considérant es experts come des bras avec $X_t^{(k)} = g_t(\xi_t^{(k)})$, les algos précédants sont valables ici.

— En info complète, $R_T \leq \sqrt{T \log N}$.

— En bandit, $R_T \leq \sqrt{TN \log N}$.

Le but est d'avoir de meilleurs résultats ici.

1.4.1 Info complète avec gains exp-concaves

L'idée est la suivante : on a une suite $y_1, ..., y_T$ choisit par un adversaire à prévoir, $y_i \in \mathcal{Y} = \mathbb{R}^d$. À chaque temps t, chaque expert essaye de prévoir y_t avec $\xi_t^{(i)} \in \mathbb{R}^d$ et l'adversaire révèle les gains $g_t(\xi) = g(\xi, y_t)$ qui augmente quand ξ et y_t sont proches.

Exemple 1.1. On peut avoir $g_t(\xi, y_t) = -||\xi - y_t||^2$.

Définition 1.4 (Exp-concavité)**.** *g est η-exp-concave si $\forall y \in \mathcal{Y}, \xi \mapsto e^{\eta g(\xi, y)}$ est concave.*

Remarque 1.6. L'exp-concavité est plus forte que la concavité en le premier argument.

Exemple 1.2. Quelques expamples de fonctions exp-concaves :

— $g : (\xi, y) \in [0,1]^2 \mapsto -(\xi - y)^2$ est $(\frac{1}{2})$-exp-concave.
En effet, en notant $G_y : \xi \mapsto g(\xi, y)$, et en prennant $y \in [0,1]$, on a :

$$G_y''(\xi) = (4\eta^2(\xi - y)^2 - 2\eta)e^{-\eta(\xi - y)^2}$$

Donc

$$\begin{aligned} G_y''(\xi) \leq 0 &\Leftrightarrow 4\eta^2(\xi - y)^2 - 2\eta \leq 0 \\ &\Leftarrow 2\eta \underbrace{(\xi - y)^2}_{\leq 1} \leq 1 \\ &\Leftarrow \eta \leq \frac{1}{2} \end{aligned}$$

— De même, $g : (\xi, y) \in [0,1]^{d \times d} \mapsto -||\xi - y||^2$ est $(\frac{1}{2})$-exp-concave.

— Le gain induit par l'entropie relative est 1-exp-concave :

$$g : (\xi, y) \in [0,1]^2 \mapsto -y \log\left(\frac{y}{\xi}\right) - (1 - y) \log\left(\frac{1 - y}{1 - \xi}\right)$$

Quelques exemples de fonction qui ne sont pas η-exp-concave pour $\eta > 0$:

— $g : (\xi, y) \mapsto -|\xi - y|$

— $g : (\xi, y) \mapsto \langle \xi, y \rangle$

Définition 1.5 (EXP)**.** *On assigne le poids*

$$P_t^{(k)} = \frac{e^{\eta R_{t-1}^{(i)}}}{\sum_{j=1}^{K} e^{\eta R_{t-1}^{(j)}}}$$

à chacun des experts $i = 1, ..., N$ et on choisit

$$\pi_t = \mathbb{E}_{j\sim p_t}[\xi_t^{(j)}] = \sum_{j=1}^{N} p_t^{(j)} \xi_t^{(j)}$$

où

$$R_t^{(i)} = \sum_{s=1}^{t-1} g_s(\xi_s^{(i)}) - g_s(\pi_s)$$

Théorème 1.5. *Si*

— *$\mathcal{A}$ est convexe ;*
— *$g_t(\xi) = g(\xi, y_t)$;*
— *g est η-exp-convexe, $\eta > 0$.*

Alors EXP utilisé avec le paramètre a un regret $R_T \leq \frac{\log N}{\eta}$.

Démonstration. (C'est presque la même que celle de EXP)

On note $W_t^{(i)} = e^{\eta \sum_{s=1}^{t-1} g_s(\xi_s^{(i)})}$ et $W_t = \sum_{i=1}^{N} W_t^{(i)}$.
On majore et on minore W_t :

$$\begin{aligned}
W_t &= \sum_{i=1}^{N} W_t^{(i)} \\
&= \sum_{i=1}^{N} W_{t-1}^{(i)} e^{\eta g_t(\xi_t^{(i)})} \\
&= W_{t-1} \sum_{i=1}^{N} \underbrace{\frac{W_{t-1}^{(i)}}{W_{t-1}}}_{p_t^{(i)}} e^{\eta g_t(\xi_t^{(i)})} && = W_{t-1} \sum_{i=1}^{N} p_t^{(i)} e^{\eta g_t(\xi_t^{(i)})} \\
&\leq W_{t-1} \exp\left(\eta g\Big(\underbrace{\sum_{i=1}^{N} p_t^{(i)} \xi_t^{(i)}}_{=\pi_t}, y_t \Big) \right) \\
&= W_{t-1} \exp(\eta g_t(\pi_t))
\end{aligned}$$

Par induction, et comme $W_0 = N$, on a : $W_t \leq N \exp\left(\eta \sum_{t=1}^{T} g_t(\pi_t) \right)$
On minore W_t :

$$W_t = \sum_{t=1}^{T} W_t^{(i)} \geq \max_i e^{\eta \sum_{t=1}^{T} g_t(\xi_t^{(i)})} = e^{\eta \max_i \sum_{t=1}^{T} g_t(\xi_t^{(i)})}$$

En combinant et en prenant le log, on a : $R_t = \max_i \sum_{t=1}^{T} g_t(\xi_t^{(i)}) - g_t(\pi_t) \leq \frac{\log N}{\eta}$ □

1.4.2 Bandit avec experts

Les experts $\xi_t^{(i)} \in \{1, ..., K\}$ pour $i = 1, ..., N$ et les gains $g_t(\xi) = X_t^{(i)}$.

Appliquer maintenant EXP3 sur les experts donne un regret en $\sqrt{TN \log N}$. On peut faire mieux si $N > K$ car on n'utilise pas le fait que quand on choisit le bras $\pi_t \in \{1, ..., K\}$ on observe la performace de tous les experts tels qu $\xi_t^{(i)} = \pi_t$.

Définition 1.6 (EXP4)**.** *A l'instant t :*

- *Observer les avis d'experts* $\xi_t^{(i)} \in \{1, ..., K\}$
- *Choisir l'action* $\pi_t = \xi_t^{(i)}$ *avec proba* $q_t^{(i)}$
- *Observer le gain* $X_t^{(\pi_t)}$
- *Estimer le gain de chacun des bras :* $\widehat{X_t^{(}k)} = \frac{X_t^{(k)} - 1}{p_t^{(k)}} \mathbb{1}_{\{k=\pi_t\}}$
- *Mettre à jour les proba des experts*

$$\widehat{R_{t-1}} = \sum_{s=1}^{t-1} \widehat{X_t^{(k)}} - \widehat{X_t^{(\pi_s)}}$$

Théorème 1.6. *Le pseudoregret par rapport au meilleur expert*

$$R_t = \max_i \mathbb{E}[]$$