MVA ENS Cachan Paris Saclay

Prediction for Individual Sequences
Notes de Cours

Cours donné par Pierre Gaillard

9 mars 2018

Introduction, notations

On ne fait plus d'hypothèse d'indépendance sur les variables $X_t^{(i)}$ (hypothèse trop forte).

2 cadres différents :

— **Adversaire oblivious**
L'adversaire choisit la suite $(X_1^{(1)}...X_T^{(K)})$. A chaque instant le joueur choisit une action π_t et il observe $X_t^{(\pi_t)}$.

— **Adversaire adaptatif**
A chaque instant l'adversaire choisit $(X_t^{(1)}, ..., X_t^{(K)}$. Le joueur choisit une action π_T et observe $X_t^{(\pi_t)}$.

Définition 0.1. *On définit le regret par rapport à une action j comme*

$$R_T^{(j)} = \sum_{t=1}^{T} \left(X_t^{(j)} - X_t^{(\pi_t)} \right)$$

Le regret est alors $R_T = \max_j R_T^{(j)}$.

Remarque 0.1. Le joueur doit avoir une stratégie aléatoire. En effet toute stratégie déterministe peut avoir un regret linéaire, i.e. il existe une séquence $(X_t^{(k)})_{t,k}$ qui donne un regret linéaire. Il suffit de prendre $X_t^{(\pi_t)} = 0$ et $X_t^{(j)} = 1$ si $j \neq \pi_t$. Entre autres, UCB et ε-greedy ne peuvent pas fonctionner.

1 Stratégie par poids exponentiels (EXP) en information complète

Plut un bras j a été bon dans le passé, plus $R_T^{(j)}$ est grand. On veut donc choisir les actions avec des probabilités qui augmentent en $R_T^{(j)}$. Les poids exponentiels choisissent le bras j avec probabilité

$$p_t^{(j)} = \frac{e^{\eta R_{t-1}^{(j)}}}{\sum_i e^{\eta R_{t-1}^{(i)}}}$$

Théorème 1.1. *La stratégie EXP pour η bien choisi vérifie $E[R_T] \leq 2\sqrt{T \log K}$.*

Démonstration. On définit $W_t(j) = e^{\eta \sum_{s=1}^{t} X_s(j)}$ et $W_t = \sum_i W_t^{(i)}$, de sorte que $p_t^{(j)} = \frac{W_{t-1}^{(j)}}{W_{t-1}}$. On utilisera une notation vectorielle pour p_t et X_t.

On va majorer et minorer W_t. On commence par majorer :

$$\begin{aligned}
W_t &= \sum_j W_{t-1}^{(j)} e^{\eta X_t^{(j)}} \\
&= W_{t-1} \sum_j p_t(j) e^{\eta X_t^{(j)}} \\
&\leq W_{t-1} \sum_j p_t(j)(1 + \eta X_t^{(j)} + \eta^2 (X_t^{(j)})^2) \\
&= W_{t-1}(1 + \eta p_t \cdot X_t + \eta p_t \cdot X_t^2) \\
&\leq W_{t-1} e^{\eta p_t \cdot X_t + \eta p_t \cdot X_t^2}
\end{aligned}$$

Finalement $W_T \leq W_0 e^{\eta \sum_t p_t \cdot X_t + \eta^2 \sum_t p_t \cdot X_t^2}$.

On minore : $W_T = \sum_j e^{\eta \sum_t X_t^{(j)}} \geq e^{\eta \max_j \sum_t X_t^{(j)}}$. En prenant le log :

$$\eta \max_j \sum_t X_t^{(j)} \leq \log K + \eta \sum_t p_t \cdot X_t + \eta^2 \sum_t p_t \cdot X_t^2$$

ce qui donne

$$\max_j \sum_t X_t^{(j)} - \sum_t p_t \cdot X_t \leq \frac{\log K}{\eta} + \eta \sum_t p_t \cdot X_t^2$$

Pour tout t, $p_t \cdot X_t^2 \leq 1$, donc finalement

$$E[R_T] \leq \frac{\log K}{\eta} + \eta T$$

On obtient la borne souhaitée en choisissant η optimal : $\eta = \sqrt{\log K / T}$. □

2 Bandit

Le choix de p_t dans EXP nécessite l'information complète. Ce n'est pas possible dans un cadre bandit. Il va falloir remplacer $R_{t-1}^{(j)}$ par un estimateur $\hat{R}_{t-1}^{(j)}$.

Objectifs possibles :

1. Majorer $E[R_T] = E[\max_j R_T^{(j)}]$
2. Majorer R_T avec grande probabilité : $R_T \leq \varepsilon$ avec probabilité $\geq 1 - \delta$.
3. Majorer le pseudo-regret $\max_j E[R_T^{(j)}] \leq E[R_T]$

Remarque 2.1. (2) $\implies$ (1) $\implies$ (3)

On veut trouver un estimateur de $X_T^{(j)}$. On pourrait prendre $\hat{X}_t^{(j)} = X_t^{(j)} \mathbb{1}_{j=\pi_t}$, mais cet estimateur serait biaisé. On prend donc $\hat{X}_t^{(j)} = (X_t^{(j)} - 1)\mathbb{1}_{j=\pi_t}/p_t^{(j)}$.

L'algorithme EXP3 choisit le bras j avec probabilité

$$p_t^{(j)} = \frac{e^{\eta \hat{R}_{t-1}^{(j)}}}{\sum\limits_i e^{\eta \hat{R}_{t-1}^{(i)}}}$$

2.1 Borne sur le pseudo-regret

Théorème 2.1. *Le pseudo-regret de EXP3 est majoré :* $\max_j E[R_T^{(j)}] \leq 2\sqrt{TK \log K}$ *pour* $\eta = \sqrt{\frac{\log K}{TK}}$.

Démonstration. La preuve est la même que la précédent en remplaçant les X par des $\hat{X}$. On obtient l'inégalité

$$E[\max_j \hat{R}_T^{(j)}] \leq \frac{\log K}{\eta} + \eta \sum_t E[p_t \cdot \hat{X}_t^{\,2}]$$

Or on a $E[\max_j \hat{R}_T^{(j)}] \geq \max_j E[\hat{R}_T^{(j)}]$.

De plus,

$$\begin{aligned} E[p_t \cdot \hat{X}_t^{\,2}] &= E[\sum_j p_t^{(j)} (\hat{X}_t^{(j)})^2] \\ &= E[\sum_j p_t^{(j)} \left((X_t^{(j)} - 1)\mathbb{1}_{j=\pi_t}/p_t^{(j)} \right)^2] \\ &= E[\sum_k p_t^{(k)} \sum_j p_t^{(j)} \left((X_t^{(j)} - 1)\mathbb{1}_{j=k}/p_t^{(j)} \right)^2] \leq K \end{aligned}$$

Donc finalement

$$\max_j E[\hat{R}_T^{(j)}] \leq \frac{\log K}{\eta} + \eta TK \leq 2\sqrt{TK \log K}$$

□

2.2 Borne en grande probabilité sur le regret

Le problème de EXP3 est que c'est très instable sur les $\hat{X}_t^{(j)}$, qui peuvent être très proches de $-\infty$. Il faut donc s'assurer que les $p_t^p(j)$ sont assez loin de 0. De plus, il faut une borne du type $E[\max_j \hat{R}_T^{(j)}] \geq E[\max_j R_T^{(j)}]$.

On peut l'avoir avec $\hat{X}_t^{(j)} = (X_t^{(j)} \mathbb{1}_{j=\pi_t} + \beta)/p_t^{(j)}$.

Lemme 2.1. *Avec probabilité au moins* $1-\delta$*, on a*

$$\sum_t \hat{X}_t^{(j)} \geq \sum_t X_t^{(j)} - \frac{\log 1/\delta}{\beta}$$

Démonstration. On utilise l'inégalité de Markov $P(X \geq \varepsilon) \leq \frac{E[X]}{\varepsilon}$ pour $X \geq 0$, qui donne également $p(X \geq \log \varepsilon) \leq \frac{E[e^X]}{\varepsilon}$.

Ainsi $p(\beta \sum_t (X_t^{(j)} - \hat{X}_t^{(j)}) \geq \log 1/\delta) \leq \delta E[e^{\beta \sum_t X_t^{(j)} - \hat{X}_t^{(j)}}]$. Il suffit de montrer que $E[e^{\beta \sum_t X_t^{(j)} - \hat{X}_t^{(j)}}] \leq 1$.

Or

$$\begin{aligned} E[e^{\beta(X_t^{(j)} - \hat{X}_t^{(j)})}] &= E[e^{\beta(X_t^{(j)} - \frac{X_t^{(j)} \mathbb{1}_{j=\pi_t})}{p_t^{(j)}} - \frac{\beta}{p_t^{(j)}}}] = E[e^{\beta(X_t^{(j)} - \frac{X_t^{(j)} \mathbb{1}_{j=\pi_t})}{p_t^{(j)}}} e^{-\frac{\beta^2}{p_t^{(j)}}}] \\ &\leq \ldots \\ &\leq E[(1 + \frac{\beta^2}{p_t^{(j)}}) e^{-\frac{\beta^2}{p_t^{(j)}}}] \leq 1 \end{aligned}$$

$\square$

L'algorithme EXP3.P choisit le bras j avec probabilité

$$p_t^{(j)} = (1-\gamma) \frac{e^{\eta \hat{R}_{t-1}^{(j)}}}{\sum_i e^{\eta \hat{R}_{t-1}^{(i)}}} + \frac{\gamma}{K}$$

avec les estimateurs définis précédemment.

Théorème 2.2. *EXP3.P vérifie, pour* γ, η, β *bien choisis, pour tout* $\delta > 0$ *:*

$$R_T \leq 6\sqrt{TK \log K} + \sqrt{\frac{TK}{\log K}} \log 1/\delta$$

avec probabilité $1 - \delta$*.*