GdT Pbs de décision séquentielle

Introduction aux bandits antagonistes

07/02/2013

Bandits = à la fin de chaque tour, on observe seulement le gain (ou la perte) associé à l'action qu'on a choisie.

Dans un premier temps, on va relâcher cette hypothèse en supposant qu'on a accès à tous les gains qu'on aurait pu encourir si on avait procédé différemment ("information parfaite").

I/ Apprentissage séquentiel robuste : combinaison d'actions en information parfaite

1) Formulation du problème

A chaque tour $t \in \mathbb{N}^*$, le statisticien choisit une action $I_t \in \{1, \dots, K\}$ selon le protocole suivant.

Jeu de prédiction : à chaque tour $t \in \mathbb{N}^*$,

1) Le statisticien choisit $p_t \in \Delta(K) = \mathcal{M}_1^+(\{1, \dots, K\})$ en fonction des données disponibles $(\underline{\ell}_s, I_s)$, $s = 1, \dots, t-1$. $\rightsquigarrow p_t$ révélé à l'environnement.

2) Simultanément :
- Le statisticien tire $I_t \sim p_t$ (conditionnellement au passé)
- L'environnement choisit $\underline{\ell}_t = (\ell_{i,t})_{1 \le i \le K} \in [0, M]^K$ en fonction des données disponibles (I_s, p_s), $s = 1, \dots, t-1$ et même p_t (mais pas I_t).

3) Le statisticien encourt la perte $\ell_{I_t, t}$ et observe le vecteur $\underline{\ell}_t$; l'environnement observe I_t.

Objectif de prédiction : minimiser le "regret"

$$R_T = \sum_{t=1}^{T} \ell_{I_t, t} - \min_{1 \le i \le K} \sum_{t=1}^{T} \ell_{i,t}$$

Plus précisément, on cherche des stratégies/politiques (= choix séquentiel des p_t) qui assurent que pour toute suite de fonctions $\ell_t(\cdot)$ (POUR TOUT ADVERSAIRE) :

$$\begin{cases} \bullet\ \mathbb{E}[R_T] \le o(T) \quad \left(\text{ex : } \mathbb{E}[R_T] \le \square \sqrt{T \ln K}\right) \\ \bullet\ \text{p.s., } \limsup_{T \to +\infty} \dfrac{R_T}{T} \le 0. \end{cases}$$

2) Le prédicteur par pondération exponentielle

Rappel : pour espérer vérifier $\mathbb{E}[R_T] \le o(T)$, il faut randomiser ! Autrement dit : $p_t = \text{dirac}$ est proscrit.

Algorithme : EWA(η) ["Exponentially Weighted Average"]

- Paramètre : $\eta > 0$
- à chaque date $t \in \mathbb{N}^*$,

$$p_{i,t} = \frac{\exp\left(-\eta \sum_{s=1}^{t-1} \ell_{i,s}\right)}{\sum_{j=1}^{K} \exp\left(-\eta \sum_{s=1}^{t-1} \ell_{j,s}\right)} \quad , \quad 1 \le i \le K.$$

Rem : $p_1 = \left(\frac{1}{K}, \dots, \frac{1}{K}\right)$

Algo ne dépend que des pertes passées $\underline{\ell}_s$, $s = 1, \dots, t-1$.

η à calibrer

Proposition 1 : borne de regret de EWA(η)

Supposons que les pertes $\ell_{i,t}$ sont à valeurs $[0, M]$.

Alors, p.s.,

$$\sum_{t=1}^{T} \sum_{i=1}^{K} p_{i,t} \ell_{i,t} - \min_{1 \le i \le K} \sum_{t=1}^{T} \ell_{i,t} \le \frac{\ln K}{\eta} + \frac{\eta T M^2}{8}$$

$$\le \sqrt{\frac{T}{2} \ln K} \quad \text{pour le choix de } \eta = \frac{1}{M}\sqrt{\frac{8 \ln K}{T}}.$$

Preuve : on notera $p_t \cdot \underline{\ell}_t = \sum_{i=1}^{K} p_{i,t} \ell_{i,t}$

Pour tout $t \in \{1, \dots, T\}$, on a :

$$p_t \cdot \underline{\ell}_t = -\frac{1}{\eta} \ln\left(\sum_{i=1}^{K} p_{i,t} e^{-\eta \ell_{i,t}}\right) + \underbrace{\frac{1}{\eta} \ln\left(\sum_{i=1}^{K} p_{i,t} e^{-\eta\left(\ell_{i,t} - p_t \cdot \underline{\ell}_t\right)}\right)}$$

$$\overset{\text{déf de } p_t}{\le} -\frac{1}{\eta} \ln\left(\frac{\sum_{i=1}^{K} \exp\left(-\eta \sum_{s=1}^{t} \ell_{i,s}\right)}{\sum_{i=1}^{K} \exp\left(-\eta \sum_{s=1}^{t-1} \ell_{i,s}\right)}\right) + \frac{\eta M^2}{8} \qquad \le \frac{\eta^2 M^2}{8} \text{ d'après le lemme de Hoeffding}$$

En sommant sur $t = 1, \dots, T$ et en posant $W_t = \frac{1}{K} \sum_{i=1}^{K} \exp\left(-\eta \sum_{s=1}^{t-1} \ell_{i,s}\right)$ ($W_1 = 1$)

il vient :

$$\sum_{t=1}^{T} p_t \cdot \underline{\ell}_t \le -\frac{1}{\eta} \sum_{t=1}^{T} \ln\left(\frac{W_{t+1}}{W_t}\right) + \frac{\eta T M^2}{8}$$

$$= -\frac{1}{\eta} \ln \frac{W_{T+1}}{W_1} + \frac{\eta T M^2}{8}$$

$$= -\frac{1}{\eta} \ln\left(\frac{1}{K} \sum_{i=1}^{K} \exp\left(-\eta \sum_{t=1}^{T} \ell_{i,t}\right)\right) + \frac{\eta T M^2}{8}$$

$$\leq -\frac{1}{\eta}\ln\left(\max_{1\leq i\leq K}\exp\left(-\eta\sum_{t=1}^{T}\ell_{i,t}\right)\right) + \frac{\ln K}{\eta} + \frac{\eta T M^2}{8}$$

$$= \min_{1\leq i\leq K}\sum_{t=1}^{T}\ell_{i,t} + \frac{\ln K}{\eta} + \frac{\eta T M^2}{8}$$ ■ cf explication bayésienne + formule de dualité sur la Kullback pour davantage d'intuition.

Rappel Hoeffding :
Soit Z une v.a. centrée t.q. $a \leq Z \leq b$ p.s.
Alors : $\forall \lambda \in \mathbb{R}$, $\ln \mathbb{E}[e^{\lambda Z}] \leq \frac{\lambda^2}{2}\frac{(b-a)^2}{4}$. ("$Z$ sous-gaussien de facteur de variance $\frac{(b-a)^2}{4}$")

Conséquences de la proposition :

- Majoration du regret en espérance : comme $\mathbb{E}[\ell_{I_t,t}] = \mathbb{E}\left[\mathbb{E}[\ell_{I_t,t} \mid \mathcal{F}_{t-1}]\right]$
  on a : $\mathbb{E}\left[\sum_{t=1}^{T}\ell_{I_t,t}\right] = \mathbb{E}\left[\sum_{t=1}^{T} p_t \cdot \ell_t\right]$ $= \mathbb{E}[p_t \cdot \ell_t]$ ↑ engendrée par $(I_1,\dots,I_{t-1})$

  d'où $\mathbb{E}[R_T] \leq \sqrt{\frac{T}{2}\ln K} = o(T)$.

- Majoration p.s. — concentration de martingales (Hoeffding-Azuma)
  $\forall T \geq 1$, avec proba $\geq 1-\delta$, $\sum_{t=1}^{T}\ell_{I_t,t} \leq \sum_{t=1}^{T} p_t\cdot\ell_t + M\sqrt{\frac{T}{2}\ln\frac{1}{\delta}}$
  En prenant $\delta_T = \frac{6}{\pi^2 T^2}$ et en appliquant Borel-Cantelli, il vient :

  p.s., pour T suffisamment grand, $R_T \leq \sqrt{\frac{T}{2}\ln K} + M\sqrt{\frac{T}{2}\ln\frac{\pi^2 T^2}{6}}$

  d'où p.s. $\overline{\lim_{T\to+\infty}}\ \frac{R_T}{T} \leq 0$.

  N.B. On peut faire mieux : $\sqrt{T \ln\ln T}$ (lg-itéré)

- Prédiction avec avis d'experts et perte convexe
  Si $\ell_{i,t} = \ell(a_{i,t}, y_t)$ avec $\ell(\cdot, y)$ convexe $\forall y \in \mathcal{Y}$, alors

  $$\sum_{t=1}^{T}\ell\left(\sum_{i=1}^{K} p_{i,t}\, a_{i,t}, y_t\right) \leq \sum_{t=1}^{T} p_t\cdot\ell_t \leq \min_{1\leq i\leq K}\sum_{t=1}^{T}\ell_{i,t} + \sqrt{\frac{T}{2}\ln K} \quad \text{p.s.}$$

  ↑ autre moyen de convexifier ! (plus besoin de randomiser)

3) Calibration séquentielle du paramètre de température

Pb : le choix de $\eta = \frac{1}{M}\sqrt{\frac{8\ln K}{T}}$ dépend de T (et de M).
On va choisir η_t (éventuellement en fonction des données observées).

Algo : $p_{i,t} = \dfrac{\exp\left(-\eta_t \sum_{s=1}^{t-1} \ell_{i,s}\right)}{\sum_{j=1}^{K} \exp\left(-\eta_t \sum_{s=1}^{t-1} \ell_{j,s}\right)}$, $1 \le i \le K$.

Proposition : p.s., si la suite $(\eta_t)_{t \ge 1}$ est décroissante, alors :

$$\sum_{t=1}^{T} p_t \cdot \ell_t - \min_{1 \le i \le K} \sum_{t=1}^{T} \ell_{i,t} \le \frac{\ln K}{\eta_{T+1}} + \sum_{t=1}^{T} \underbrace{\frac{1}{\eta_t} \ln\left(\sum_{i=1}^{K} p_{i,t} e^{-\eta_t (\ell_{i,t} - p_t \cdot \ell_t)}\right)}_{=: \delta_t}$$

Preuve : adaptation de la preuve ↑

$$p_t \cdot \ell_t \overset{\text{def } p_t}{=} -\frac{1}{\eta_t} \ln\left(\frac{\sum_{i=1}^{K} \exp\left(-\eta_t \sum_{s=1}^{t} \ell_{i,s}\right)}{\sum_{i=1}^{K} \exp\left(-\eta_t \sum_{s=1}^{t-1} \ell_{i,s}\right)}\right) + \delta_t$$

$$= -\frac{1}{\eta_t} \ln\left(\frac{W'_{t+1}}{W_t}\right) + \delta_t \quad \text{où} \quad W'_{t+1} = \frac{1}{K} \sum_{i=1}^{K} \exp\left(-\eta_t \sum_{s=1}^{t} \ell_{i,s}\right)$$

↑ au lieu de η_{t+1}

$$= \frac{\ln W_t}{\eta_t} - \frac{\ln W_{t+1}}{\eta_{t+1}} + \underbrace{\left(\frac{\ln W_{t+1}}{\eta_{t+1}} - \frac{\ln W'_{t+1}}{\eta_t}\right)}_{\le 0 \text{ d'après Jensen et car } \eta_{t+1} \le \eta_t}$$

En sommant sur $t = 1, \dots, T$, il vient :

$$\sum_{t=1}^{T} p_t \cdot \ell_t \le -\frac{1}{\eta_{T+1}} \ln W_{T+1} + \sum_{t=1}^{T} \delta_t \qquad (\text{car } \ln W_1 = \ln 1 = 0)$$

$$\le \min_{1 \le i \le K} \sum_{t=1}^{T} \ell_{i,t} + \frac{\ln K}{\eta_{T+1}} + \sum_{t=1}^{T} \delta_t \quad \blacksquare$$

Conséquences :

- $\eta_t = \dfrac{c}{M} \sqrt{\dfrac{\ln K}{t}}$ et majoration Hoeffding $\Rightarrow$ regret conditionnel $\le \square \sqrt{T \ln K}$ $\forall T \ge 1$
- Autres choix plus fins de η_t, cf par ex Cesa-Bianchi, Mansour et Stoltz (2007) pour une borne en variance : regret conditionnel $\le \square \sqrt{\sum_{t=1}^{T} V_t \ln K} + \dots$
- Ex. de majoration plus fine que Hoeffding : **majoration de type Bennett**

  Rappel : Si Z est une v.a. tq $Z \le b$ p.s. ($b > 0$) et $\mathbb{E}[Z^2] = v$, alors : $\forall \lambda > 0$, $\ln \mathbb{E}\left[e^{\lambda(Z - \mathbb{E}Z)}\right] \le \dfrac{v}{b^2} \phi(\lambda b)$ où $\phi(x) = e^x - x - 1$.

Le contrôle sous-gaussien sera utile ensuite. Mentionnons une conséquence :

Si $Z \leq 0$ p.s. et $\mathbb{E}[Z^2] = v$, alors

(*) $\forall \lambda > 0$, $\ln \mathbb{E}\left[e^{\lambda(Z - \mathbb{E}Z)}\right] \leq \frac{\lambda^2 v}{2}$ (passage à la limite $b \to 0$).

Application : majoration de δ_t de type Bennett

$$\delta_t = \frac{1}{\eta_t} \ln\left(\sum_{i=1}^K p_{i,t}\, e^{-\eta_t(\ell_{i,t} - p_t \cdot \ell_t)}\right) = \frac{1}{\eta_t} \ln \mathop{\mathbb{E}}_{i \sim p_t}\left[e^{-\eta_t\left(\ell_{i,t} - \mathbb{E}_{j \sim p_t} \ell_{j,t}\right)}\right]$$

d'après (*) avec $\begin{cases} Z = \ell_{i,t} \\ i \sim p_t \end{cases}$ $\leq \frac{\eta_t}{2} \sum_{i=1}^K p_{i,t} \ell_{i,t}^2$ car $\ell_{i,t} \geq 0$.

II / Information imparfaite : les bandits antagonistes

1) Formulation du problème

⚠ Dorénavant, le statisticien n'observe plus $\ell_{i,t}$ pour $i \neq I_t$.

Protocole de décision (bandits antagonistes à K bras) : pour chaque $t \in \mathbb{N}^*$,

1) Le statisticien choisit et révèle $p_t \in \mathcal{M}_1^+(\{1, \dots, K\})$ en fonction des données disponibles $(\ell_{I_s,s}, I_s)_{s \leq t-1}$.

2) Simultanément : • Le statisticien tire $I_t \sim p_t$ (conditionnellement au passé)
• L'environnement choisit $\ell_t = (\ell_{i,t})_{1 \leq i \leq K} \in [0,M]^K$ en fonction des données disponibles $(I_s, p_s)_{s \leq t-1}$ et même p_t (mais pas I_t).

3) Le statisticien encourt et observe la perte $\ell_{I_t,t}$ (les $\ell_{j,t}$, $j \neq I_t$, restent cachés) ; l'environnement observe I_t.

Objectifs : minimiser le regret $R_T = \sum_{t=1}^T \ell_{I_t,t} - \min_{1 \leq i \leq K} \sum_{t=1}^T \ell_{i,t}$

Plus précisément, on cherche des stratégies (choix séquentiels des p_t) telles que, pour toute suite de fonctions $\ell_t(\cdot)$ (i.e. pour tout adversaire),

- $\mathbb{E}[R_T] \leq o(T)$ (on obtiendra $\mathbb{E}[R_T] \leq \square \sqrt{TK \ln K}$)
- p.s., $\overline{\lim_{T \to +\infty}} \frac{R_T}{T} \leq 0$ (on obtiendra $\overline{\lim_{T \to +\infty}} \frac{R_T}{\sqrt{T \ln \ln T}} \leq 0$)

Dans un premier temps, on va s'intéresser à une quantité plus faible :

$$\mathbb{E}\left[\sum_{t=1}^T \ell_{I_t,t}\right] - \min_{1 \leq i \leq K} \mathbb{E}\left[\sum_{t=1}^T \ell_{i,t}\right] \leftarrow \text{le "pseudo-regret".}$$