Exposé de maîtrise
Un problème de casino : les bandits à N bras

Sylvain Carré et Joseph Picot - sujet proposé par Gilles Stoltz

26 juin 2008

Table des matières

1 Le problème des bandits

1.1 Cadre et formalisation

Dans ce travail, on s'intéresse à la mise en place de bonnes stratégies pour des jeux de casino appelés bandits à N bras : on dispose de N bras qu'on peut actionner, chacun permettant de gagner avec une certaine probabilité inconnue au départ. Le problème est donc, sur un grand nombre d'itérations, d'obtenir un gain approximativement égal à celui qu'aurait procuré, en espérance, le meilleur des N bras.

Bien sûr, si on connaissait les lois des bras au préalable, en actionnant systématiquement celui de plus grande espérance, on aurait par loi des grands nombres le gain souhaité. Ici, le seul moyen d'estimer les performances d'un bras est a priori de l'actionner un nombre grand de fois. Il faut donc arriver à déterminer le meilleur bras (exploration) tout en maximisant le gain courant (exploitation).

La première étape est de définir rigoureusement le problème et notamment ce qu'est une stratégie. Formellement, on se donne une suite $(E_{t,1}, \ldots, E_{t,N})_{t\geq 1}$ de v.a. positives iid de loi $P_1 \otimes P_2 \ldots \otimes P_N$, les P_i étant des lois de variance σ_i^2 avec $0 < \sigma_i < \infty$, et d'espérance $\mu_i < \infty$. On désignera par des astérisques les variables relatives au bras

d'espérance maximale (et de plus petit indice), en particulier $\mu^* = \max_i \mu_i$. On suppose bien sûr l'existence d'un indice j avec $\mu_j < \mu^*$. Les indices $i = 1, \dots, N$ correspondent aux N bras, qu'on va actionner à chaque temps $t \geq 1$, *ie* on choisit $C_t \in \{1, \dots, N\}$ et on obtient un gain $X_t = E_{t,C_t}$. On notera G_t la moyenne des gains :

$$G_t = \frac{1}{t} \sum_{k=1}^{t} X_k.$$

On souhaite donc essentiellement trouver une stratégie qui permette d'assurer que, p.s. $G_t \to \mu^*$. Dans ce cadre, la stratégie correspond au processus (C_t), qui, du fait qu'on dispose seulement de l'information donnée par les actions antérieures et leur rétribution, est prévisible pour la filtration $(\mathcal{F}_{t-1})$ où $\mathcal{F}_t = \sigma(C_1, X_1, \dots, C_t, X_t)$ (et par convention $\mathcal{F}_0$ est la tribu triviale). Remarquons qu'on est en train de chercher une stratégie *déterministe*, *ie* fondée uniquement sur les observations antérieures (on pourrait également se servir d'une randomisation, c'est-à-dire se donner d'autres variables aléatoires intervenant dans le choix des bras, ce qui revient formellement à prendre une tribu plus grosse). Bien sûr, pour une telle stratégie déterministe, C_t étant fonction de $X_1, \dots, X_{t-1}$ on a en fait $\mathcal{F}_t = \sigma(X_1, \dots, X_t)$.

1.2 Consistance des estimateurs

Une première idée pourrait être de se donner un certain $n \geq 1$ et de tester les N bras, pendant les Nn premiers coups afin d'estimer leur loi, c'est-à-dire choisir pour $k < n-1$, $j < N$, $C_{kN+j} = j$ puis pour $t > Nn$, $C_t = j$ où j est l'indice du bras de meilleur moyenne empirique après Nn coups, soit l'indice maximisant $\hat{\theta}_{Nn,k}$ où pour $k = 1, \dots, N$:

$$\hat{\theta}_{Nn,k} = \frac{1}{n} \sum_{j=0}^{n-1} X_{k+jN}.$$

Mais si j est tel que $\mu_j = \min_k \mu_k < \mu^*$, on a par indépendance

$$\mathbf{P}(\hat{\theta}_{Nn,j} > \hat{\theta}_{Nn,j^*}) \geq \left(\prod_{k \neq j}^{N} \mathbf{P}(E_{1,j} > E_{1,k}) \right)^n.$$

Dans certains cas triviaux (par exemple $N = 2$, $E_{t,1}$ de loi à support dans $[0,1]$ et $E_{t,2}$ de loi à support dans $[2,3]$) la stratégie fonctionne (dans cet exemple, on identifie en un seul coup le meilleur bras). Mais, bien sûr, on ne connaît pas les lois des v.a. $E_{t,i}$ en jeu, et par exemple pour des lois de Bernoulli, ou des lois dont l'intersection des supports contient un intervalle non trivial, on a pour $k \neq j$, $\mathbf{P}(E_{1,j} > E_{1,k}) > 0$. Donc avec probabilité positive on aura $\hat{\theta}_{Nn,j} > \hat{\theta}_{Nn,j^*}$ puis pour tout $t > Nn$, $C_t = j$ et avec probabilité positive (par loi des grands nombres)

$$G_t \to \mu_j < \mu^*.$$

Ainsi dans le cas général cette stratégie ne permet pas un gain asymptotique optimal p.s.

Pour s'assurer qu'on actionne majoritairement le bon bras, une possibilité consiste donc à trouver des *estimateurs (fortement, ie avec convergence presque sûre) consistants* de μ_i pour $1 \leq i \leq N$, c'est-à-dire des processus $(\hat{\theta}_{t,i})$ adaptés à la filtration $(\mathcal{F}_t)$, convergeant

p.s. vers μ_i. Or de tels estimateurs existent, pour peu qu'on ne néglige pas les étapes d'exploration. Les processus

$$\hat{\theta}_{t,i} = \frac{1}{N_{t,i}} \sum_{k=1}^{t} X_k \mathbf{1}_{\{C_k=i\}}$$

si $N_{t,i} > 0$ (et $\hat{\theta}_{t,i} = 0$ si $N_{t,i} = 0$) conviennent en effet dès lors que pour tout i, $N_{t,i} = \sum_{k=1}^{t} \mathbf{1}_{\{C_k=i\}}$ (nombre d'activations du bras i après t coups) tend vers $+\infty$. (Bien sûr dans le cas particulier précédent la définition des $\hat{\theta}_{t,i}$ coïncidait avec celle-ci). Cela découle d'une loi des grands nombres pour les martingales, dont nous avons besoin car les choix des bras ne sont *pas indépendants* ici et la loi usuelle ne s'applique donc pas.

Avant d'exposer cette loi, introduisons une définition pratique : étant donnée une filtration $(\mathcal{G}_n)_{n\geq 0}$ et une martingale $(M_n)_{n\geq 0}$ adaptée à $(\mathcal{G}_n)_{n\geq 0}$, on appelle accroissements de martingale correspondants à (M_n) les variables aléatoires $Y_n = M_n - M_{n-1}$ pour $n \geq 1$ (et $Y_0 = E[M_n]$). Le processus $(Y_n)_{n\geq 1}$ est alors une $(\mathcal{G}_n)_{n\geq 1}$ martingale, vérifiant $E[Y_n] = 0$ pour $n \geq 1$. Réciproquement, partant d'un processus $(Y_n)_{n\geq 1}$ vérifiant ces propriétés et d'une constante y_0 on peut reconstruire une suite $(M_n)_{n\geq 0}$, martingale adaptée à $(\mathcal{G}_n)_{n\geq 0}$ dont les éléments ont pour espérance commune y_0, en posant $Y_0 = y_0$ puis

$$M_n = Y_0 + Y_1 + \ldots + Y_n.$$

Dans la suite, il pourra être commode de définir ainsi les martingales (M_n) utilisées.

Théorème 1 (loi des grands nombres pour les martingales dans L^2) *Soit $(\mathcal{G}_n)_{n\geq 0}$ une filtration et $(M_n)_{n\geq 0}$ une martingale dans L^2, adaptée à $(\mathcal{G}_n)$. Si (Y_n) désigne la suite des accroissements de martingale correspondants et V_n le compensateur prévisible de (M_n^2),*

$$V_n = E[M_0]^2 + \sum_{t=1}^{n} \mathbf{E}\left[Y_t^2 \mid \mathcal{G}_{t-1}\right],$$

et que p.s. $V_n \to +\infty$, alors

$$M_n = o(V_n).$$

On peut maintenant justifier la consistance des estimateurs annoncée. Fixons $i \in \{1, \ldots, N\}$. Prenons

$$Y_{t,i} = \mathbf{1}_{\{C_t=i\}}(E_{t,i} - \mu_i).$$

D'après une remarque précédente on peut se donner une martingale $(M_{t,i})_{t\geq 1}$ (et $M_0 = 0$) correspondant à ces accroissements ; elle vaut ici

$$M_{t,i} = \sum_{k=1}^{t} \mathbf{1}_{\{C_k=i\}}(E_{k,i} - \mu_i) = N_{t,i}(\hat{\theta}_{t,i} - \mu_i).$$

On a $M_{t,i} \in L^2(\mathcal{F}_t)$ et $\mathbf{E}[M_{t,i} \mid \mathcal{F}_{t-1}] = M_{t-1,i}$ car

$$\mathbf{E}[Y_{t,i} \mid \mathcal{F}_{t-1}] = \mathbf{E}\left[(E_{t,i} - \mu_i)\mathbf{1}_{\{C_t=i\}} \mid \mathcal{F}_{t-1}\right] = \mathbf{E}[E_{t,i} - \mu_i]\mathbf{1}_{\{C_t=i\}} = 0.$$

$(M_{t,i})$ est donc une $(\mathcal{F}_t)$-martingale dans L^2 de compensateur prévisible

$$V_{t,i} = \sum_{k=1}^{t} \mathbf{E}\left[Y_k^2 \mid \mathcal{F}_{k-1}\right] = \sum_{k=1}^{t} \mathbf{E}\left[(E_{k,i} - \mu_i)^2 \mathbf{1}_{\{C_k=i\}} \mid \mathcal{F}_{k-1}\right] = N_{t,i}\ \sigma_i^2.$$

Si $N_{t,i}$ tend p.s. vers $+\infty$ alors $V_{t,i}$ tend vers $+\infty$ et le théorème donne alors bien $\hat{\theta}_{t,i} \to \mu_i$ p.s.

Il nous faut donc activer un grand nombre de fois les N bras, tout en évitant d'affecter le gain asymptotique, ce qui impose des instants d'exploration de plus en plus éloignés (sinon, comme on le verra plus précisemment en 1.3, dans une proportion de temps asymptotiquement non nulle, on se sert du bras le moins performant, ce qui rend impossible la convergence de G_t vers μ^*).

En respectant cette condition, on va construire une stratégie répondant au problème initial.

1.3 Une stratégie assurant p.s. un gain asymptotique optimal

Soit (c_k) une suite d'entiers strictement croissante telle que $k \ll c_k$, au temps t on active un bras selon la règle suivante : si $t = c_{Nk+j}$ on joue de façon déterministe le bras j et sinon on active celui de plus grand estimateur.

Par construction de (c_k) on a pour tout i, $N_{t,i} \to +\infty$ p.s. Ainsi nos estimateurs $\hat{\theta}_{t,i}$ sont consistants. Le nombre de bras étant fini, on a l'existence d'un temps aléatoire T fini p.s. tel que pour $t \geq T$:

$$\max_{\mu_j<\mu^*} \hat{\theta}_{t,j} < \hat{\theta}_{t,j^*}.$$

Ainsi, on jouera toujours après T le bras de meilleure espérance sauf lors des tours d'exploration. Leur nombre étant négligeable devant t, par choix des (c_k) on a pour tout j tel que $\mu_j < \mu^*$, $\frac{N_{t,j}}{t} \to 0$. En conséquence

$$\frac{1}{t}\sum_{j=1}^{N} N_{j,t}\mu_j \to \mu^*.$$

Pour obtenir le résultat souhaité, à savoir la convergence p.s. de (G_t) vers μ^*, il faut s'assurer que la convergence (intuitive) de $G_t - \frac{1}{t}\sum_{j=1}^{N} N_{t,j}\mu_j$ vers 0 a effectivement lieu. Or cela est encore une conséquence de la loi des grands nombres pour les martingales.

En effet, posons $M_t = tG_t - \sum_{j=1}^{N} N_{t,j}\mu_j = M_{t,1} + \ldots + M_{t,N}$. Avec des arguments similaires à ceux utlisés pour les suites $(M_{t,i})$ on voit que (M_t) est une $(\mathcal{F}_t)$-martingale, dans L^2 et que si $\delta = \min_j \sigma_j^2 > 0$, le compensateur prévisible (V_t) associé à (M_t) s'écrit alors :

$$V_t = N_{t,1}\sigma_1^2 + \ldots + N_{t,N}\sigma_N^2 \geq t\delta$$

car $N_{t,1} + \ldots + N_{t,N} = t$. De là $V_t \to +\infty$ et le théorème s'applique bien. Ainsi, p.s., $G_t \to \mu^*$ et la stratégie atteint l'objectif souhaité.

Remarque : la quantité $\sum_{j=1}^{N} \frac{N_{t,j}}{t}\mu_j$ est un barycentre des paramètres μ_j. On constate,

comme annoncé en 1.2, que s'il existe j tel que $\mu_j < \mu^*$ avec $\frac{N_{t,j}}{t}$ qui ne converge pas p.s. vers 0, alors sur un ensemble de probabilité strictement positive la convergence $\frac{1}{t}\sum_{j=1}^{N} N_{j,t}\mu_j \to \mu^*$ n'a pas lieu. Donc G_t ne peut pas converger p.s. vers μ^*.

2 Inégalités de concentration et vitesses de convergence

2.1 L'inégalité d'Hoeffding

La stratégie vue en première partie nous assure p.s. un gain asymptotique optimal, mais en pratique il est surtout intéressant de savoir à quelle vitesse a lieu cette convergence, ou plus précisement d'estimer avec quelle probabilité on dévie du résultat attendu d'une quantité préalablement choisie. Pour cela, on a besoin d'inégalités dites de concentration.

Lemme 1 (Hoeffding) *Soit $\mathcal{F}$ une tribu et Y une v.a. telle que $\mathbf{E}[Y \mid \mathcal{F}] = 0$, avec $a \leq Y \leq b$. Alors*

$$\mathbf{E}[e^{tY} \mid \mathcal{F}] \leq e^{\frac{t^2(b-a)^2}{8}}.$$

Preuve : soit la v.a.

$$\alpha = \frac{b-Y}{b-a},$$

de sorte que $Y = \alpha a + (1-\alpha)b$. Par convexité :

$$e^{tY} \leq \alpha e^{ta} + (1-\alpha)e^{tb}.$$

En prenant l'espérance conditionnelle par rapport à $\mathcal{F}$, utilisant que $\mathbf{E}[Y \mid \mathcal{F}] = 0$ donc $\mathbf{E}[\alpha \mid \mathcal{F}] = \frac{b}{b-a}$ et remplaçant α par sa valeur on voit que

$$\mathbf{E}[e^{tY} \mid \mathcal{F}] \leq \frac{a}{a-b}e^{tb} + \frac{b}{b-a}e^{ta}.$$

On veut majorer cette quantité par $e^{\frac{t^2(b-a)^2}{8}}$, il est donc naturel d'essayer de mettre le majorant sous la forme de l'exponentielle d'une certaine fonction g puis de majorer g par $\frac{t^2(b-a)^2}{8}$. Or

$$\frac{a}{a-b}e^{tb} + \frac{b}{b-a}e^{ta} = e^{ta}\left(\frac{b}{b-a} + \frac{a}{a-b}e^{t(b-a)}\right)$$

$$= \exp\left(ta + \ln\left(1 + \frac{a}{b-a} + \frac{a}{a-b}e^{t(b-a)}\right)\right) = \exp(g(t)).$$

$$g'(t) = a - \frac{a}{\frac{b}{b-a}e^{-t(b-a)} + \frac{a}{a-b}}, \qquad g''(t) = (b-a)^2\frac{(-a)be^{-t(b-a)}}{\left(be^{-t(b-a)} - a\right)^2}.$$

Or si $x, y > 0$, $\frac{xy}{x^2+y^2} \leq \frac{1}{2}$, donc $g''(t) \leq \frac{(b-a)^2}{4}$. De plus $g(0) = g'(0) = 0$ donc par inégalité de Taylor le résultat est acquis.

On va utiliser ce résultat pour démontrer l'inégalité d'Hoeffding, ainsi qu'en 2.3 sous la forme du corollaire suivant.

Corollaire : si U est une v.a. à valeurs dans $[0,1]$ et si Φ désigne la transformée de Cramer de U :

$$\Phi(\lambda) = \ln \mathbf{E}[\exp(\lambda U)],$$

alors pour tout $\lambda \geq 0$ on a :

$$\Phi(\lambda) \leq \lambda \mathbf{E}[U] + \frac{\lambda^2}{8}.$$

(Prendre pour $\mathcal{F}$ la tribu triviale, $a = -\mathbf{E}[U]$, $b = 1 - \mathbf{E}[U]$ et $Y = U - \mathbf{E}[U]$).

Théorème 2 (Inégalité d'Hoeffding) *Si (M_n) est une martingale par rapport à $(\mathcal{F}_n)$ d'accroissements Y_n, avec, p.s., $a \leq Y_n \leq b$, alors*

$$\mathbf{P}\left(\sum_{t=1}^{n} Y_t \geq m\right) \leq e^{\frac{-2m^2}{n(b-a)^2}}.$$

En d'autres termes, si $0 < \delta < 1$ alors avec probabilité au moins $1-\delta$ on a

$$\sum_{t=1}^{n} Y_t \leq (b-a)\sqrt{\frac{n}{2}\ln\frac{1}{\delta}}.$$

Preuve : soit $m > 0$, par inégalité de Chernoff on a pour tout $t > 0$,

$$\mathbf{P}\left(\sum_{i=1}^{n} Y_i \geq m\right) = \mathbf{P}\left(e^{t\sum_{i=1}^{n} Y_i} \geq e^{tm}\right) \leq e^{-tm}\mathbf{E}\left[e^{t\sum_{i=1}^{n} Y_i}\right] = e^{-tm}\mathbf{E}\left[e^{t\sum_{i=1}^{n-1} Y_i}\mathbf{E}[e^{tY_n} \mid \mathcal{F}_{n-1}]\right].$$

En utilisant le lemme d'Hoeffding et par une récurrence immédiate il vient

$$\mathbf{P}\left(\sum_{i=1}^{n} Y_i \geq m\right) \leq e^{-tm+n\frac{t^2(b-a)^2}{8}}.$$

Si $t = \frac{4m}{n(b-a)^2}$ le majorant vaut $e^{\frac{-2m^2}{n(b-a)^2}}$ d'où la première inégalité. La deuxième résulte de l'inversion de la fonction de m ainsi obtenue.

Remarque : on peut améliorer ce résultat en utilisant l'inégalité maximale de Doob. On a vu qu'on pouvait écrire

$$\mathbf{P}\left(\sum_{i=1}^{n} Y_i \geq m\right) = \mathbf{P}\left(e^{t\sum_{i=1}^{n} Y_i} \geq e^{tm}\right) \leq e^{-tm}\mathbf{E}\left[e^{t\sum_{i=1}^{n-1} Y_i}\mathbf{E}[e^{tY_n} \mid \mathcal{F}_{n-1}]\right].$$

Mais comme par convexité de l'exponentielle, $\left(e^{t\sum_{i=1}^{n} Y_i - tm}\right)$ est une sous-martingale positive, l'inégalité maximale de Doob permet d'obtenir en fait

$$\mathbf{P}\left(\max_{1\leq s\leq n}\sum_{i=1}^{s} Y_i \geq m\right) \leq e^{-tm}\mathbf{E}\left[e^{t\sum_{i=1}^{n} Y_i - tm}\right].$$

Cela ne change rien à la suite de la preuve, et on obtient en définitive qu'avec probabilité au moins $1-\delta$, l'inégalité

$$\max_{1\leq s\leq n}\sum_{i=1}^{s} Y_i \leq (b-a)\sqrt{\frac{n}{2}\ln\frac{1}{\delta}}$$

a lieu.

2.2 Contrôle après un nombre aléatoire d'activations des bras

Dans la troisième section, nous aurons besoin d'un raffinement de l'inégalité d'Hoeffding, adapté au fait que pour un bras à un instant donné, son nombre d'activations dans le passé est aléatoire. Ici on voit comment s'affranchir de cette difficulté (cf. [3]).

Théorème 3 (Inégalité d'Hoeffding pour un nombre aléatoire de termes) *Soit $(\mathcal{G}_n)$ une filtration, (U_n) une suite $(\mathcal{G}_n)$ adaptée de v.a. positives indépendantes uniformément bornées par 1 avec aussi si $m > n$, U_m indépendante de $\mathcal{G}_n$. On notera $u_n = \mathbf{E}[U_n]$. Soit (ε_n) une suite de v.a. prévisibles (ie ε_n est $\mathcal{G}_{n-1}$-mesurable) valant 0 ou 1. Notons $\gamma = 2^{1/4} + 2^{-1/4}$,*

$$S_n = \sum_{k=1}^{n} U_k \varepsilon_k, \qquad M_n = \sum_{k=1}^{n} u_k \varepsilon_k, \qquad N_n = \sum_{k=1}^{n} \varepsilon_k.$$

Avec ces notations et pour tous $n \geq 1$, $K > 0$, on a :

$$\mathbf{P}\left(\frac{S_n - M_n}{\sqrt{N_n}} > K\right) \leq \lceil \log_2 n \rceil \exp\left(-\frac{8K^2}{\gamma^2}\right).$$

Remarques : ce théorème concerne bien une inégalité de type Hoeffding puisque le terme $S_n - M_n$ renormalisé par $\sqrt{N_n}$ est à comparer au terme $\frac{\sum_{i=1}^{n} Y_i}{\sqrt{n}}$ qu'on pouvait majorer avec probabilité au moins $1-\delta$ par $(b-a)\sqrt{\frac{1}{2}\ln\frac{1}{\delta}}$. On avait (pour $a = 0$ et $b = 1$, ce qui correspond au cas présent)

$$\mathbf{P}\left(\frac{S_n - \mathbf{E}[S_n]}{\sqrt{n}} > K\right) \leq e^{-2K^2}.$$

On voit que l'inégalité a la même forme, à un facteur constant (lorsque n est fixé) logarithmique près (conséquence du caractère aléatoire des sommations), et avec une constante $\frac{8}{\gamma^2}$, soit environ 1,94, très légèrement moins bonne que la constante 2.

Preuve : introduisons la transformée de Cramer de U_n :

$$\Phi_n(\lambda) = \ln \mathbf{E}[\exp(\lambda U_n)].$$

Nous aurons besoin d'un lemme facile.

Lemme 2 *Soient $x > 0$ et $\lambda_k = \sqrt{\frac{8x}{2^{k-\frac{1}{2}}}}$. Si k est tel que $2^{k-1} \leq N_n < 2^k$ alors*

$$\frac{x}{\lambda_k \sqrt{N_n}} + \frac{\lambda_k \sqrt{N_n}}{8} \leq \sqrt{\frac{x}{8}}\gamma.$$

Preuve : en élevant les deux membres au carré, l'inégalité est équivalente à

$$\frac{x}{8}\left(\frac{2^{k-\frac{1}{2}}}{N_n} + 2 + \frac{N_n}{8 \times 2^{k-\frac{1}{2}}}\right) \leq \frac{x}{8}\left(2^{\frac{1}{2}} + 2 + 2^{\frac{-1}{2}}\right),$$

ce qui est clair vu que $2^{k-1} \leq N_n < 2^k$.

Revenant à la preuve, on observe que

$$\mathbf{E}[\exp(\lambda U_n \varepsilon_n) \mid \mathcal{G}_{n-1}] = \mathbf{1}_{\{\varepsilon_n = 0\}} + \mathbf{1}_{\{\varepsilon_n = 1\}} \mathbf{E}[e^{\lambda U_n}] = \exp(\Phi_n(\lambda)\varepsilon_n).$$

Une récurrence immédiate donne alors

$$\mathbf{E}[\exp(\lambda S_n - \sum_{k=0}^{n-1} \Phi_k(\lambda)\varepsilon_k] = 1.$$

Par le lemme 1 $\Phi_n(\lambda) \leq \lambda u_n + \frac{\lambda^2}{8}$ donc

$$\mathbf{E}\left[\exp(\lambda(S_n - M_n) - \frac{\lambda^2}{8}N_n)\right] \leq 1.$$

De là, si $x > 0$,

$$\mathbf{P}\left(\frac{S_n - M_n}{\sqrt{N_n}} > \frac{x}{\lambda\sqrt{N_n}} + \frac{\lambda\sqrt{N_n}}{8}\right) = \mathbf{P}\left(\exp(\lambda(S_n - M_n) - \frac{\lambda^2}{8}N_n) \geq e^x\right) \leq e^{-x},$$

par inégalité de Markov.

Pour exploiter cette inégalité, il suffit maintenant d'éliminer la dépendance en $\sqrt{N_n}$ dans le terme minorant $\frac{S_n - M_n}{\sqrt{N_n}}$ (on veut le remplacer par une constante K), et c'est ici que le lemme 2 intervient. En effet, il montre que si $\frac{S_n - M_n}{\sqrt{N_n}} > K$ et que x est tel que $K = \sqrt{\frac{x}{8}}\gamma$, alors $\frac{S_n - M_n}{\sqrt{N_n}} > \frac{x}{\lambda_k\sqrt{N_n}} + \frac{\lambda_k\sqrt{N_n}}{8}$ pour un certain $k \in \{1, \ldots, D\}$ où $D = \lceil \log_2 n \rceil$. (En effet $0 \leq N_n \leq n$, et donc si $k = \log_2 N_n + 1$ on a bien $k \leq D$ et $2^{k-1} \leq N_n < 2^k$ et le lemme s'applique bien). Ainsi,

$$\mathbf{P}\left(\frac{S_n - M_n}{\sqrt{N_n}} > K\right) \leq \sum_{k=1}^{D} \mathbf{P}\left(\frac{S_n - M_n}{\sqrt{N_n}} > \frac{x}{\lambda_k\sqrt{N_n}} + \frac{\lambda_k\sqrt{N_n}}{8}\right) \leq De^{-x}.$$

En réécrivant x comme fonction de K ($x = \frac{8K^2}{\gamma^2}$) on trouve le résultat souhaité.

2.3 Applications

Revenons à l'estimation $M_n = o(n)$ où $M_n = nG_n - \sum_{j=1}^{N} N_{j,n}\mu_j$, démontrée dans la première partie. Nous voulons maintenant la préciser. Notant Y_n les accroissements de martingales correspondants, on a $-1 \leq Y_n \leq 1$ et donc avec probabilité au moins $1 - \frac{\delta}{2}$ on a

$$M_n \leq \sqrt{2n \ln \frac{2}{\delta}}$$

par inégalité d'Hoeffding. On peut faire de même avec les $-Y_n$ pour obtenir, par une union d'événements, qu'avec probabilité au moins $1 - \delta$,

$$|M_n| \leq \sqrt{2n \ln \frac{2}{\delta}}.$$

Soit $\delta_n = \frac{1}{n^2}$. Appliquons ce qui précède pour $\delta = \delta_n$. Comme $\sum \delta_n$ converge, on peut appliquer le lemme de Borel-Cantelli. On sait alors que, p.s., il existe n_0 tel que si $n \geq n_0$ alors $|M_n| \leq \sqrt{2n \ln(2n^2)}$. C'est l'amélioration souhaitée, *ie*, presque sûrement :

$$\limsup_{n\to\infty} \frac{|M_n|}{\sqrt{4n \ln n}} \leq 1.$$

Grâce à la remarque de la section 2.1. on peut obtenir un résultat plus précis. L'inégalité maximale s'écrit ici

$$\max_{1\leq s\leq n} \sum_{i=1}^{s} Y_i \leq \sqrt{2n \ln \frac{2}{\delta}}$$

avec probabilité au moins $1-\frac{\delta}{2}$. Toujours en considérant les $-Y_n$ on voit qu'en fait avec probabilité au moins $1-\delta$:

$$\max_{1\leq s\leq n} |M_s| \leq \sqrt{2n \ln \frac{2}{\delta}}.$$

On prend maintenant $n = 2^r$ et $\delta = \frac{1}{r^2}$ pour trouver, par le même argument que précédemment : p.s.,

$$\limsup_{r\ \to+\infty} \max_{1\leq s\leq 2^r} \frac{|M_s|}{\sqrt{(2^r)\ln\ln(2^r)}} \leq 1. \tag{1}$$

Pour un n quelconque, on prend r tel que $2^r \leq n < 2^{r+1}$ (par souci de clarté on notera $2^{r+1} = a(n)$). Comme $\frac{a(n)}{2} \leq n < a(n)$, on a

$$\limsup_{n\to+\infty} \frac{a(n)\ln\ln a(n)}{n\ln\ln n} = 2.$$

En divisant l'inégalité évidente $|M_n| \leq \max_{1\leq k\leq a(n)} |M_k|$ par $\sqrt{a(n)\ln\ln a(n)}$ et en utilisant (1) et cette égalité il vient en définitive, p.s. :

$$\limsup_{n\to+\infty} \frac{|M_n|}{\sqrt{2n\ln\ln n}} \leq 1,$$

soit une majoration analogue à celle de la loi du logarithme itéré.

3 La stratégie UCB

3.1 L'algorithme UCB

Les phases d'exploration induisent un manque à gagner qui croît moins vite que linéairement pour de bonnes stratégies comme on l'a vu en première partie. Cependant, il a été montré (cf. [4]) que ce manque, en espérance, est nécessairement au moins logarithmique et l'objectif de cette section est de mettre en place une stratégie qui atteint cette borne.

Les machines à sous distribuant des lots compris dans un intervalle borné, il est raisonnable de considérer des lois P_j à support borné, mettons dans $[0,1]$. Notons $\Delta_i = \mu^* - \mu_i$. Le *regret (en espérance)* après n coups est la quantité

$$R_t = t\mathbf{E}[\mu^* - G_t] = t\mu^* - \sum_{j=1}^{N} \mu_j \mathbf{E}[N_{t,j}] = \sum_{\mu_j<\mu^*} \Delta_j \mathbf{E}[N_{t,j}].$$

R_t mesure, en espérance, l'écart entre ce qu'on aurait pu gagner dans le meilleur des cas et ce qui s'est effectivement passé. On souhaite trouver une borne explicite de R_t

pour tout t. La stratégie UCB (*Upper Confidence Bound*) permet d'avoir une inégalité du type

$$R_t \leq A \ln t + B$$

avec $A, B > 0$ (dépendant des lois P_j en jeu). On procède comme suit. Lors des N premiers coups on actionne chaque bras une fois. Par la suite, au $(t+1)$-ème coup, on joue le bras maximisant la quantité $\hat{\theta}_{t,j} + c_{t,N_{t,j}}$ où

$$c_{t,s} = \sqrt{\frac{2 \ln t}{s}}.$$

Cette quantité est reliée à la notion d'intervalle de confiance (d'où le nom de la stratégie UCB), c'est-à-dire qu'on va pouvoir contrôler la probabilité que les estimateurs soient "loin" de la variable estimée, en obtenant des majorations des probabilités des événements suivants :

$$\hat{\theta}_{N^*_{s-1},j^*} + c_{s-1,N^*_{s-1}} \leq \mu^*, \tag{2}$$

$$\hat{\theta}_{N_{s-1,j},j} - c_{s-1,N_{s-1,j}} \geq \mu_j. \tag{3}$$

Plus précisemment, on se donne ici comme intervalles de confiance : pour μ_j, $I_{s,j} =]\hat{\theta}_{N_{s-1,j},j} - c_{s-1,N_{s-1,j}}, +\infty[$ et pour μ^*, $J^*_s =]-\infty, \hat{\theta}_{N_{s-1,j^*},j} + c_{s-1,N^*_{s-1}}[$. On veut obtenir des minorations des événements $\{\mu_j \in I_{s,j}\}$ et $\{\mu^* \in J^*_s\}$, dont on notera $A_{s,2}$ et $A_{s,3}$ les négations, qui correspondent respectivement aux événements (2) et (3).

Cela va être possible grâce au théorème 3. Par exemple pour (3), posons $U_n = E_{n,j}$ et $\mathcal{G}_n = \sigma(E_{t,i})_{1 \leq t \leq n, 1 \leq i \leq N}$, ainsi que $\varepsilon_n = \mathbf{1}_{\{C_n = j\}}$. (ε_n) est bien une suite prévisible *car la stratégie est déterministe* et on est donc dans les conditions d'application du théorème. Ici $N_n = N_{n,j}$, $S_n = N_{n,j}\hat{\theta}_{n,j}$ et $M_n = \mu_j N_{n,j}$. On peut alors obtenir la probabilité de trouver la variable μ_j dans l'intervalle de confiance $I_{s,j}$. Cela est réalisé avec probabilité $1-\delta$ ($1-\delta$ est le *niveau* de l'intervalle de confiance) où d'après le théorème 3 :

$$\delta = \lceil \log_2(s-1) \rceil (s-1)^{-k}.$$

En effet,

$$\mathbf{P}(\mu_j \in I_{s,j}) = \mathbf{P}\left(N_{s-1,j}(\hat{\theta}_{N_{s-1,j}} - \mu_j) < \sqrt{2 \ln(s-1) N_{s-1,j}} \right)$$

$$= \mathbf{P}\left(\frac{S_{s-1} - M_{s-1}}{\sqrt{N_{s-1}}} < \sqrt{2 \ln(s-1)} \right) \geq 1 - \lceil \log_2(s-1) \rceil (s-1)^{-k},$$

où $k = \frac{16}{\gamma^2}$ par application du théorème 3. $I_{s,j}$ est un intervalle de confiance pertinent : δ est "petit" (en un sens qui sera précisé dans la deuxième section), ce qui signifie qu'on a peu de chance de trouver μ_j hors de $I_{s,j}$. $A_{s,3}$ n'étant rien d'autre que la négation de l'évènement $\{\mu_j \in I_{s,j}\}$ on a :

$$P(A_{s,3}) \leq \lceil \log_2(s-1) \rceil \, e^{-k \ln(s-1)} = \lceil \log_2(s-1) \rceil \, (s-1)^{-k}.$$

En procédant de même avec le bras d'indice $*$ on obtient la même majoration pour

$\mathbf{P}(A_{s,2})$. En effet, si une suite (U_n) vérifie les hypothèses du théorème 3, alors il en est de même de la suite $(1-U_n)$. L'inégalité du théorème 3 vaut donc également pour la quantité $\frac{M_n-S_n}{\sqrt{N_n}}$, ce qu'on utilise pour majorer $\mathbf{P}(A_{s,2})$.

Ainsi apparaît la logique du choix de $c_{t,s}$. Dans les deux cas précédents ((2) et (3)) il s'agit comme on l'a dit d'estimer la probabilité que les estimateurs soient hors d'un intervalle de confiance. Un intervalle de confiance mesurant la probabilité de déviation par rapport à la moyenne, si on se rappelle du théorème central limite, il n'est pas surprenant de voir apparaître des facteurs $\sqrt{s}$ (ou $\sqrt{N_n}$ dans le théorème 3). Aussi, on comprend mieux pourquoi, grâce au choix de $c_{t,s}$ on retombe exactement sur les probabilités estimées par le théorème 3.

3.2 Borne sur le regret

Théorème 4 *On dispose d'une majoration logarithmique du regret. Plus précisement, il existe une constante $B>0$ telle que pour tout $t \geq 1$:*

$$R_t \leq \left(8 \sum_{\mu_i<\mu^*} \frac{1}{\Delta_i} \right) \ln t + B.$$

Preuve : par souci de lisibilité on notera ici $\{A\}$ la fonction indicatrice de l'événement A (et non $\mathbf{1}_A$). Vue l'expression de R_t il est naturel de majorer $\mathbf{E}[N_{t,j}]$. Or pour tout entier $\ell \geq 1$ on a

$$N_{t,j} \leq \ell + \sum_{s=N+1}^{t} \{C_s = j,\ N_{s-1,j} \geq l\}$$

$$\leq \ell + \sum_{s=N+1}^{t} \{\hat{\theta}_{N^*_{s-1},j*} + c_{s-1,N^*_{s-1}} \leq \hat{\theta}_{N_{s-1,j},j} + c_{s-1,N_{s-1,j}}\}\ \{N_{s-1,j} \geq \ell\}.$$

Or si le terme d'indice s de la somme vaut 1 on est nécessairement dans l'un des trois cas (2), (3) ou (4) où (4) est l'événement

$$\mu^* \quad < \quad \mu_j + 2c_{s-1,N_{s-1,j}}. \tag{4}$$

On a vu que $\mathbf{P}(A_{s,3}) \leq \lceil \log_2(s-1) \rceil (s-1)^k$ où $1 < k = \frac{16}{\gamma^2}$, soit environ 3,88, avec la même majoration pour $\mathbf{P}(A_{s,2})$.

Or $\sum_{s\geq 1} (\ln s) s^{-k}$ converge (c'est en ce sens que dans l'étude faite dans la section précédente, on pouvait dire que les δ étaient petits, où $1-\delta$ était le niveau de l'intervalle de confiance), et on obtient donc que la somme sur s des probabilités de (2) et (3) est une constante finie majorée par

$$B_0 = 2 \sum_{s=N}^{+\infty} \lceil \log_2 s \rceil s^{-k}.$$

Enfin, si $\ell = \left\lceil \frac{8}{\Delta_j^2} \ln t \right\rceil$, (4) ne peut pas se produire puisque

$$c_{s-1,N_{s-1,j}} \leq \sqrt{2\frac{s-1}{\ell}} \leq \Delta_j \sqrt{\frac{2\ln(s-1)}{8\ln t}} \leq \frac{1}{2}\Delta_j.$$

Rappelons qu'on avait établi que :

$$N_{t,j} \leq \ell + \sum_{s=N+1}^{t} \{\hat{\theta}_{N^*_{s-1},j*} + c_{s-1,N^*_{s-1}} \leq \hat{\theta}_{N_{s-1,j},j} + c_{s-1,N_{s-1,j}}\} \ \{N_{s-1,j} \geq \ell\}.$$

En prenant l'espérance dans l'inégalité ci-dessus on trouve :

$$\mathbf{E}[N_{t,j}] \leq \frac{8 \ln t}{\Delta_j^2} + B_0,$$

et ainsi, si $B = B_0 \sum_{j=1}^{N} \Delta_j$:

$$R_t \leq \left(8 \sum_{\mu_i < \mu^*} \frac{1}{\Delta_i} \right) \ln t + B.$$

Références

[1] G. Stoltz, V. Rivoirard. *Onze thèmes de statistique* (en cours de rédaction).

[2] P. Auer, N. Cesa-Bianchi, P. Fischer, 2002. Finite-time analysis of the multiarmed bandit problem, *Machine Learning*, 47, 235–256.

[3] A. Garivier, E. Moulines, 2008. On upper-confidence bound policies for non-stationary bandit problem (prépublication).

[4] T. Lai, H. Robbins, 1985. Asymptotically efficient adaptive allocation rules, *Advances in Applied Mathematics*, 6, 4–22.

[5] P. Massart, 2006. *Concentration inequalities and Model Selection*, Springer.