PROBLÈME DU BANDIT À DEUX ARMES (G7)
(03 / 08 / 2020, © Monfort, Dicostat2005, 2005-2020)

L'expression imagée de **problème du bandit à plusieurs armes** (ou **bras**), ou de **problème du bandit multi-armé**, désigne un **problème statistique** séquentiel (cf **analyse séquentielle**, **théorie séquentielle**) : l'**expérience** consiste à choisir, puis à utiliser, une « arme » à chaque instant. Cette expérience est alors répétée. Chaque fois qu'une arme est tirée (qu'un bras est actionné), une **observation** est effectuée à partir d'une **loi** dont le **paramètre** est inconnu.

Dans la **théorie bayésienne**, ce paramètre est parfois muni d'une **loi a priori**.

(i) Dans le cas de deux armes, on suppose que :

(a) il existe deux **états** de la **Nature**, ie $\Theta = \{\theta_1 , \theta_2\}$;

(b) le **temps** est discret (ie eg T = **N**) ;

(c) $(\mathcal{X}, \mathcal{B})$ désigne un **espace d'observation** donné ;

(d) à chaque instant $t \in T$, on réalise l'une des deux **expériences aléatoires** $d_1 = (\Omega_1 , \mathcal{T}_1 , (P_\theta^1)_{\theta \in \Theta})$ ou $d_2 = (\Omega_2 , \mathcal{T}_2 , (P_\theta^2)_{\theta \in \Theta})$ indexées par Θ. Autrement dit, l'ensemble des **décisions** est $D = \{d_1 , d_2\}$ (deux décisions possibles). Le choix effectué à tout instant t se fait au vu de l'**information** passée, ie à travers l'**observation** d'une **variable aléatoire** $X_t : \Omega_1$ (resp Ω_2) $\mapsto \mathcal{X}$;

(e) la **loi conditionnelle** de X_t relativement à θ et au choix de $d \in D$, est indépendante du « passé » (ie indépendante des X_s , $\forall\ s < t$). On la note $\mathcal{L}\ (X_t / \theta, d)$;

(f) si $\theta = \theta_1$ (resp $\theta = \theta_2$), c'est l'expérience d_1 (resp d_2) qui est, en principe, préférée (ie réalisée) : par suite, chaque fois que d_2 (resp d_1) est préférée (à tort), cette décision entraîne un coût $c_1 > 0$ (resp $c_2 > 0$) (cf **fonction de coût**, **théorie des tests**).

L'objectif consiste alors à trouver une **suite** de décisions (choix d'expériences $d \in D$) qui soit optimale, ie à minimiser une **fonction de risque** associée à ce problème. Plusieurs types de solutions ont été exhibées.

Ainsi, en **théorie bayésienne**, on note $\pi_1 = \Pi\ (\theta = \theta_1)$ la **probabilité a priori** de l'**état** θ_1 et $Q_t\ (. / \theta_1)$ la probabilité a posteriori de l'état θ_1 après t expériences. Qi le problème est symétrique (ie si $\mathcal{L}\ (X_t / \theta_1 , d_1) = \mathcal{L}\ (X_t / \theta_2 , d_2)$, si $\mathcal{L}\ (X_t / \theta_2 , d_1) = \mathcal{L}\ (X_t / \theta_1 , d_2)$ et si $c_2 = c_1$), on montre alors que la décision séquentielle optimale $\tilde{d} = (\tilde{d_t})_{t \in T}$ est définie selon :

(1)
$$Q_t(X_t / \theta_1) > 1/2 \Rightarrow d_{t+1}\tilde{} = d_1 ,$$
$$Q_t(X_t / \theta_1) < 1/2 \Rightarrow d_{t+1}\tilde{} = d_2 .$$

Le cas où une égalité survient peut être « randomisé » (cf **randomisation**).

(ii) Le problème du bandit à k armes, qui comporte un ensemble de k décisions ($\Theta = \{\theta_1 ,..., \theta_k\}$), étend le précédent (cf aussi **problème de classification**, **problème de décision multiple**, **problème de test**, **processus stochastique**, **test séquentiel**).