Apprentissage et conflit exploration-exploitation : un essai

Yann BRAOUEZEC
ESILV- Dept Mathématiques et Ingénierie Financière.
Pole Universitaire Léonard de Vinci
92916 Paris La Défense Cedex
E-mail: yann.braouezec@devinci.fr

1er mars 2004

Mots clés : bandit à deux bras, algorithmes d'apprentissage, apprentissage bayésien, conflit exploration-exploitation.

Résumé

Nous présentons, dans ce petit essai qui n'est pas destiné à la publication, une discussion constructive de la notion d'apprentissage dans le cadre d'un problème du bandit à deux bras. Diverses règles de décisions sont étudiées ainsi que leur performance en longue période. Nous proposons divers formules de valorisation, telles que la valeur de l'adaptation ou la valeur de la mémoire. Nous terminons en présentant le conflit exploration-exploitation.

Dans ce petit essai, nous discutons deux notions que l'on retrouve désormais fréquemment dans la littérature sur la dynamique adaptative en Economie : l'apprentissage et le conflit exploration-exploitation. La première section est consacrée à la notion d'apprentissage alors que la seconde est consacrée au conflit exploration-exploitation.

1 Qu'est ce que l'apprentissage ?

Imaginons la situation de choix suivante : un individu doit effectuer une (ou plusieurs) décisions dans un environnement composé de deux machines, M_1 et M_2 dans lequel chaque machine lui procure une récompense de $R = 1$ euro avec une probabilité θ_i, $i = 1{,}2$ et une récompense nulle avec la probabilité complémentaire. Autrement dit,

$$\mathbb{P}(R = 1/\theta_i) = \theta_i \quad \text{pour } i \in \{1{,}2\} \tag{1}$$

Cet environnement est celui d'un jeu (éventuellement répété) contre la nature, par opposition à un jeu stratégique contre un autre joueur. La différence essentielle est que dans un jeu contre la nature, cette dernière ne modifie pas son comportement en fonction de celui de notre individu.

Supposons que les deux valeurs (θ_1,θ_2) aient été tirées avant toute décision dans $]0{,}1[\times]0{,}1[$ mais que l'individu n'ait pas eu accès à la réalisation de ces valeurs. Sans lui révéler ces deux valeurs, il serait toutefois possible de lui donner quelques éléments d'informations sur (θ_1,θ_2). Par exemple, on pourrait lui annoncer les deux valeurs numériques obtenues, mais ne pas lui dire quelle est celle qui correspond à chaque machine. On pourait également penser à lui annoncer l'une des deux réalisations, mais pas l'autre. Naturellement, en bon homo-economicus, s'il connaissait la réalisation (θ_1,θ_2), il choisirait uniquement la machine dont la probabilité de récompense est la plus élevée, soit $\max(\theta_1,\theta_2)$. Dans ce petit essai, nous allons présenter de manière constructive quatre conditions nécéssaires à toute forme d'apprentissage, avant d'en présenter une définition assez complète, emprunté au psychologue Paul Fraisse.

Supposons donc que notre individu ne dispose d'aucune information sur ces valeurs numériques, et qu'il n'ait qu'un seul choix à effectuer. Dans ce cas, il ne semble pas déraisonable qu'il laisse le hasard décider pour lui. Il peut ainsi lancer une pièce de monnaie et choisir, par exemple, la machine 1 s'il observe pile et la machine 2 si face est observée. Naturellement, puisqu'il n'y a qu'une seule décision, il ne peut pas y avoir d'apprentissage. Ceci nous permet d'énoncer une première condition nécéssaire à toute forme d'apprentissage.

Condition 1 *Toute forme d'apprentissage nécéssite la répétition des décisons dans le temps.*

Faisons un pas de plus en donnant la possibilité à notre individu d'effectuer une suite de n décisions, mais sans lui donner l'opportunité d'observer la conséquence de chaque décision, c'est-à-dire la récompence réalisée à chaque date. Il est alors, avant chaque décision, exactement dans la même situation que précédement du point de vue de l'information possédée. Appelons $\eta_{1,t}$ la probabilité qu'il choisisse M_1 à la date t, et $\eta_{2,t} = 1 - \eta_{1,t}$ la probabilité qu'il choisisse M_2. Sans rétroaction provenant de son environnement, il ne dispose d'aucune raison objective pour ne pas jouer avec une pièce de monnaie à chaque date t : il va donc choisir M_1 avec la probabilité $\eta_{1,t} = (1/2)$ à chaque date t. Appelons d_0 la règle de décision qui consiste à utiliser cette *stratégie mixte* [(1/2),(1/2)] . Il est alors facile de calculer son esperance de gain par essai, qui s'exprime sous la forme d'une probabilité de récompense que nous appelons $\mu(d_0)$, soit

$$\mu(d_0) = (1/2).\mathbb{P}(R = 1/\theta_1) + (1/2).\mathbb{P}(R = 1/\theta_2) = \frac{\theta_1 + \theta_2}{2} \tag{2}$$

Notons que cette probabilité $\mu(d_0)$ représente la fréquence avec laquelle notre individu obtiendra une récompense lorsque t, et donc lorsque n, tend vers l'infini. Puisqu'il ne peut pas observer la conséquence de ses choix, il ne saurait y avoir apprentissage. Ceci nous permet d'énoncer une seconde condition, nécessaire à toute forme d'apprentissage.

Condition 2 *Toute forme d'apprentissage nécéssite l'existence d'un mécanisme de rétroaction environnemental.*

Soit $i_t \in \{1,2\} \equiv \{M_1,M_2\}$ la décision de l'individu à la date t et $y_t \in \{0,1\}$ la conséquence de son choix. Désormais, l'individu observe y_t à chaque date t. Par hypothèse, notre individu n'a accès qu'à une unique source d'information; la conséquence de son choix. Il ne bénéficie en particulier pas de l'expérience d'autres individus qui auraient pu être dans la même situation de choix que lui. Nous pouvons alors énoncer une première définition élémentaire de l'apprentissage, faisant intervenir uniquement les deux conditions précédentes.

L'apprentissage est un processus d'acquisition d'informations de la part de l'individu sur son environnement.

Cette définition traduit alors bien l'idée selon laquelle il est nécéssaire, pour apprendre, que la 'quantité d'information' augmente avec l'expérience passée. En revanche, elle ne renseigne pas sur la manière avec laquelle cette information peut être utilisée, si tant est qu'elle l'est, pour améliorer ses décisions. Ceci nous permet alors d'énoncer une troisième condition, nécéssaire à toute forme d'apprentissage.

Condition 3 *Toute forme d'apprentissage nécéssite l'existence d'un mécanisme d'adaptation des décisions.*

Proposons alors une seconde définition de l'apprentissage, qui prend en compte cette troisième condition.

L'apprentissage représente la capacité d'un individu à amèliorer ses décisions sur la base de son expérience passée.

Nous allons maintenant, en suivant Herbert Robbins (1952) étudier une première règle de décision, que nous appelons d_1, et dont les prescriptions sont les suivantes.

1. Effectuer la première décision entre l'une des deux machines avec une pièce de monnaie.
2. Ensuite choisir la machine 1 si ($i_t = 1$,$y_t = 1$) ou si ($i_t = 2$,$y_t = 0$), et choisir la machine 2 si ($i_t = 2$,$y_t = 1$) ou si ($i_t = 1$,$y_t = 0$).

Cette fois, il y a bien utilisation du mécanisme de rétroaction environnemental de la part de l'individu pour améliorer ses décisons. En effet, s'il a choisit la machine 1 à la date t (i.e., si $i_t = 1$) et qu'il observe une récompense nulle (i.e., si $y_t = 0$), alors, il choisit avec probabilité un la machine 2 à la date $t + 1$. Notons que cette règle d'apprentissage prescrit à notre individu de choisir à chaque date une machine avec probabilité un [1]. Il est alors aisé de voir que l'évolution des décisions va être décrite par une chaîne de Markov, dont la matrice des probabilités de transition est $A = \begin{pmatrix} \theta_1 & 1-\theta_1 \\ 1-\theta_2 & \theta_2 \end{pmatrix}$. Par exemple, $1 - \theta_1$ est la probabilité qu'il choisisse la machine 2 en $t + 1$ sachant qu'il a choisit la machine 1 en t. Notre objectif est maintenant de determiner la fréquence limite avec laquelle il obtiendra 1 euro en longue période, c'est-à-dire lorsque t tend vers l'infini, notée $\mu(d_1)$. Toutefois, il nous faut d'abord calculer les probabilités stationaires [2] $\eta_{1,\infty}$, soit à résoudre le système de deux équations aux différences finies $A^T\eta_t = \eta_{t+1}$. Puisque $\eta_{2,t} = 1 - \eta_{1,t}$ on peut remplacer $\eta_{2,t}$ par $1 - \eta_{1,t}$ dans la premiere équation. On obtient alors que $\eta_{1,t} = (\theta_1+\theta_2-1)\eta_{1,t}+(1-\theta_2)$. Puisque $(\theta_1 + \theta_2 - 1) < 1$, il vient immédiatement que

$$\eta_{1,\infty} = \frac{1-\theta_2}{2-(\theta_1+\theta_2)} \quad et \quad \eta_{2,\infty} = \frac{1-\theta_1}{2-(\theta_1+\theta_2)} \tag{3}$$

Son espérance de gain par essai (en longue période) est $\mu(d_1) = \eta_{1,\infty}\mathbb{P}(R = 1/\theta_1) + \eta_{2,\infty}\mathbb{P}(R = 1/\theta_2)$, soit

$$\mu(d_1) = \frac{\theta_1 + \theta_2 - 2\theta_1\theta_2}{2-(\theta_1+\theta_2)} \tag{4}$$

1. On dit dans ce cas qu'il s'agit d'un algorithme déterministe. Lorsque l'individu choisit l'une des deux machines de manière aléatoire, on dit qu'il s'agit d'un algorithme stochastique.

2. La détermination de ces probabilités stationnaires n'a bien entendu de sens que si la chaîne de Markov est ergodique. Compte tenu de la matrice des probabilités de transition, l'ergodicité est évidente.

Il est facile de voir que (4) est toujours supérieur à (2), quelles que soient les valeurs de θ_1 et θ_2. Nous voyons donc que l'adaptation, c'est-à-dire la possibilité de modifier sa décision sur la base de l'expérience passée, a de la valeur puisqu'elle permet à l'individu d'améliorer ses performances sur l'ensemble du processus de décision. La possibilité de modifier ses décisions a en effet pour conséquence d'augmenter la fréquence (limite) avec laquelle il obtiendra un euro en longue période. On est assez naturellement tenté de mesurer la *valeur de l'adaptation* de la règle d_1 comme suit :

$$VA(d_1) = \mu(d_1) - \mu(d_0) \tag{5}$$

Toutefois, d_1 reste imparfaite puisque $\mu(d_1) < \max(\theta_1,\theta_2)$. Elle ne permet pas à notre individu, à long terme, de choisir la meilleure machine avec une probabilité égale à un. On peut alors également définir la valeur de l'information parfaite, avec d_1 comme référence, comme :

$$VIP(d_1) = \max(\theta_1,\theta_2) - \mu(d_1) \tag{6}$$

Les équations de valorisation (5), (6) supposent implicitement que notre individu ne dévalorise pas le futur. L'équation (5) (resp (6)) mesure alors la somme, en euros, qu'il serait prêt au maximum à payer à chaque date t (telle une redevance périodique) pour qu'on lui indique, avant sa première décision, la règle d_1 (resp $\max\{\theta_1,\theta_2\}$).

Pourquoi la règle d_1 ne permet pas à notre individu de choisir avec probabilité 1 en longue période la meilleure des deux machine? Parce que l'implèmentation de d_1 n'utilise qu'une mémoire de taille 1 : la décision i_{t+1} ne dépend en effet que du dernier couple décision-observation $(i_t,y_t$. Le fait qu'il n'utilise qu'une mémoire de taille 1 amène notre individu à sauter d'une machine à une autre selon le résultat observé. Introduisons alors de la mémoire et regardons quelles en sont les conséquences sur le processus de décision. Pour ce faire, suivons, Herbert Robbins 1956 qui propose une règle de décision dont les prescriptions sont les suivantes :

1. Effectuer la première décision entre l'une des deux machines avec une pièce de monnaie.
2. ensuite, changer de machine lorsqu'une suite consécutive de k zeros est observée avec la même machine.

Appelons d_2 cette règle de décison. Notons que maintenant, l'implèmentation de d_2 nécéssite l'existence d'une mémoire de taille k. Robbins 1956 montre alors que les probabilités stationnaires sont :

$$\eta_{1,\infty} = \frac{(1-\theta_2)^k}{(1-\theta_1)^k + (1-\theta_2)^k} \quad et \quad \eta_{2,\infty} = \frac{(1-\theta_1)^k}{(1-\theta_1)^k + (1-\theta_2)^k} \tag{7}$$

Il vient donc que :

$$\mu(d_2(k)) = \frac{\theta_1(1-\theta_2)^k + \theta_2(1-\theta_1)^k}{(1-\theta_1)^k + (1-\theta_2)^k} \tag{8}$$

Nous vérifions facilement que lorsque $k = 1$, (8)=(4). Ce qui est en revanche plus intéréssant est que lorsque $k \geq 2$, (8)>(4) . Il est maintenant simple de montrer que $\mu(d_2(k))$ est une fonction croissante de k. Ainsi, lorsque k tend vers l'infini, $\mu(d_2(k))$ tend vers max (θ_1,θ_2). L'irréversibilité du choix est donc étroitement liée à la profondeur de la mémoire. Ce cadre simple nous permet, comme précédemment d'obtenir une formule de valorisation de la profondeur de la mémoire. La règle d_2 étant donnée, on peut alors définir la valeur marginale d'une unité supplémentaire de mémoire par décision comme :

$$VM(d(k+1) = \mu(d(k+1)) - \mu(d(k)) \tag{9}$$

L'équation (9) mesure la valeur que notre individu, utilisant la règle $d_2(k)$ serait prêt à payer pour bénéficier d'une mémoire de taille $k+1$ plutot que k. Nous pouvons maintenant la dernière condition, nécéssaire à toute forme d'apprentissage.

Condition 4 *Toute forme d'apprentissage nécéssite l'existence d'un mécanisme de stockage de l'information : la mémoire.*

Nous sommes maintenant en mesure d'enoncer une définition assez complète de ce qu'est l'apprentissage, empruntée au psychologue Paul Fraisse.

L'apprentissage correspond à toutes les modifications adaptatives du comportement qui se produisent au cours d'épreuves répétées. On distingue, dans ce processus, deux phases essentielles :

- **l'acquisition de relations entre un système de signaux et de réponses**
- **la mémorisation, qui est la rétention dans le temps de ces relations.**

Cette définition ne nous dit toutefois pas si l'apprentissage doit conduire, en longue période, à la selection de la meilleure décision. Nous avons clairement vu que la règle de Robbins (1956) permet d'augmenter l'espérance de gain de notre individu en longue période, mais elle reste imparfaite. En effet, à mémoire k fixée, $\mu(d_2(k)) < \max\{\theta_1,\theta_2\}$. Pour le voir, fixons par exemple $k = 3$, $\theta_1 = 0.8$, et $\theta_2 = 0.1$. On obtient $\mu(d_2(k)) \approx 0.792$. La différence est faible mais elle

reste positive. La règle de Robbins (1956) ne permet donc pas de réaliser un apprentissage "complet" au sens ou notre individu ne finit pas par selectionner avec probabilité un, à long terme, la meilleure des deux machines.

Dans les années soixante, les deux articles de Robbins avaient engendré un véritable thème de recherche dont l'enjeu était le suivant : à mémoire donnée de taille finie[3], peut-on trouver une règle d'apprentissage qui permette de réaliser un apprentisage complet ?

C'est Cover et Hellman (1970) qui ont résolu ce problème du bandit à deux bras[4] à mémoire finie en montrant qu'il existe des règles d'apprentissage, prenant éventuellement la forme d'un algorithme stochastique, qui permettent de réaliser un apprentissage ϵ-complet, i.e., tel que la performance à long terme soit aussi près que l'on veut de $\max\{\theta_1,\theta_1\}$.

2 Apprentissage et optimisation : le conflit exploration exploitation

Nous avons jusqu'a maintenant implicitement supposé que l'objectif du décideur était *d'apprendre*, c'est-à-dire d'identifier à long terme la meilleure machine. Pourtant, en réalité, ce qui interesse le décideur n'est pas l'apprentissage *per se*, mais les gains dont il peut bénéficier de celui ci. Supposons donc que l'objectif de notre individu soit de maximiser un critère intertemporel tel que l'esperance de la somme actualisée des gains. Le fait de dévaloriser le futur au taux δ implique que le gain (espéré) maximum soit borné. Quelle règle de comportement notre individu peut-il employer pour réaliser au mieux son objectif ?

Lorsqu'il possède une certaine mémoire (finie), il peut par exemple chercher à estimer les paramètres qu'il ne connaît pas en allouant, par exemple, $(k/2)$ décisions sur chaque machine, puis estimer θ_1 et θ_2 par les fréquences observées f_1 et f_2. Il semble ensuite naturel d'utiliser cette information acquise en affectant les décisions restantes à la machine qui a obtenu la fréquence de récompense la plus élevée. Appelons $d_3(k)$ une telle règle de décision. Le programme d'optimisation de notre individu est alors :

$$\max_{k\in\mathbb{N}} \mathbb{E}\sum_{t=0}^{\infty} \delta^t y_t \tag{10}$$

Dans un tel contexte, la détermination du k optimal (qui dépend de δ), notée $k^*(\delta)$, implique comme nous allons le voir, un conflit entre exploration de l'environnement, et exploitation de l'information acquise. Voyons pourquoi.

3. Par exemple, de taille k comme dans Robbins (1956)

4. L'apprentissage dans les jeux a récemment fait l'objet d'une monographie par Fudenberg et Levine (1998), à laquelle nous renvoyons le lecteur intéressé.

Supposons que k soit faible. Dans ce cas, du fait de la petite taille de chaque échantillon, les estimations de θ_1 et θ_2 vont être peu fiables. Le risque en effet d'exploiter défintivement la mauvaise machine va être élevée, due aux importantes fluctuations d'échantillonnage. Supposons au contraire que k soit élevé. Dans ce cas, à supposer même qu'il estime correctement les deux paramètres, le gain (espéré) lié à l'exploitation de la meilleure machine va être faible, a cause du facteur d'actualisation. Clairement, la valeur de k^* qui réalise le meilleur arbitrage entre exploration et exploitation est intermédiaire. D'un point de vue économique, on peut interpreter k^* comme une demande d'information. Il est alors permis de conjecturer que l'information est un bien normal, c'est à dire dont la quantité $k^*(\delta)$ décroit lorsque son prix augmente, i.e., lorsque δ diminue.

C'est dans l'approche bayésienne que l'approche suggérée ci-dessus trouve toute sa généralité. Pour simplifier la discussion, supposons que notre individu connaîsse la réalisation de θ_2 mais pas celle de θ_1. Notre individu estime alors le paramètre inconnu $\theta_1 \in]0,1[$ via une *distribution de probabilité a priori* qu'il modifie en utilisant la règle de Bayes. Lorsque la distribution conditionnelle $\mathbb{P}(Y = y/\theta_1)$ est une loi de bernoulli, il est commode de choisir, comme distribution a priori, une loi beta de paramètre α et β. Dans ce cas, lorsque notre individu affecte k décisions sur la machine 1, on peut montrer que son estimation de θ_1, conditionnelle à $\mathcal{F}_k = \{y_1, y_2, ...y_k\}$ sera égale à $\mathbb{E}(\theta_1/\mathcal{F}_k) = \dfrac{\alpha + \sum_{t=1}^{k} y_t}{\alpha + \beta + k}$. Les valeurs de α et β, qui reprèsentent l'influence de l'a priori, sont donc en un sur k. Ainsi, pour k élevé, son estimation de θ_1 sera proche de la fréquence observée f_1, qui sera elle même proche de θ_1. Bien entendu, l'objectif de notre individu est de déterminer k tel qu'il maximise le critère donné par l'équation (10). La détermination de la valeur optimale k^*, porte le nom de *règle optimale d'arrêt*. Par exemple, si $k^* = 25$, il est alors optimal pour notre individu d'affecter les 25 premières décisions sur la machine 1. Le reste des décisions sera affecté sur la machine 1 si $\mathbb{E}(\theta_1/\mathcal{F}_{k^*}) > \theta_2$, ou sur la machine 2 si $\mathbb{E}(\theta_1/\mathcal{F}_{k^*}) < \theta_2$.

Il revient à Rothschild (1974) d'avoir explicitement considéré ce problème. Du au conflit exploration-exploitation, il montre qu'il n'est pas optimal pour notre individu d'identifier parfaitement la valeur des deux paramètres : celui ci peut donc, avec probabilité positive, opter définitivement pour la mauvaise machine. Ce résultat a été affiné par Easley et Kieffer (1988) qui ont montré qu'il existe un facteur d'actualisation critique $\delta^* < 1$, qui détermine s'il sera optimal ou non d'identifier la vraie valeur de θ_1. Selon que δ est supérieur (inférieur) à δ^*, il sera optimal d'identifier (de ne pas identifier) le paramètre inconnu.

Références

[1] Cover T (1968), "A Note on the Two-Armed Bandit Problem with Finite Memory", *Information and Control*, pp 371-377.

[2] Cover T, Hellman (1970), "The Two-Armed Bandit Problem with Time Invariant Finite Memory", *IEEE Transactions on Information Theory*, pp 185-195.

[3] Easley D, Kieffer N (1988), "Controlling a Stochastic Process with Unknown Parameters", *Econometrica*, pp 1045-1064.

[4] Fudenberg D, Levine D (1998), "Theory of Learning in Games", MIT Press.

[5] Isbell J (1959), "On a Problem of Robbins", *Annals of Mathematical Statistics*, pp 606-610.

[6] Robbins H (1952), "Some Aspect of the Sequential Design of Experiments", *Bulletin of the Americain Mathematical Society*, 527-535.

[7] Robbins H (1956), "A Sequential Decision Problem with Finite Memory", *Proceeding of the National Academy of Science*, pp 920-923.

[8] Rothschild M (1974), "A Two-Armed bandit Theory of Market Pricing", *Journal of Economic Theory*, pp 185-202.

[9] Samuels B (1968), "Randomized Rules for the Two-Armed Bandit with Finante Memory", *Annals of Mathematical Statistics*, pp 2103-2107.