Jeux de bandit-manchot multijoueur : Comportement limite avec un grand nombre de joueurs et corrélation partielle

Antoine Salomon[1,2]

[1] LAGA ; Université Paris 13
[2] Doctorat HEC ; HEC Paris
salomon@math.univ-paris13.fr

Mots-Clés : *jeu de bandit-manchot, apprentissage stratégique (social learning), comportement asymptotique, bandits corrélés.*

1 Introduction

Le modèle de bandit-manchot permet d'étudier le conflit qui peut se poser entre exploitation et exploration. Chaque joueur a affaire à un bandit-manchot à un bras dont le rendement est incertain (ou à deux bras, l'un étant risqué, l'autre sûr). A chaque étape du jeu, il décide s'il parie sur ce bras risqué (c'est-à-dire s'il "explore"). Si c'est le cas, il reçoit un paiement aléatoire qui l'informe partiellement sur la rentabilité de la machine. Le but est de décider si l'on doit arrêter cette expérimentation, et le cas échéant à quelle étape : ni trop tôt pour se donner le temps de vérifier la rentabilité de la machine, ni trop tard pour éviter les paris ruineux.
On suppose couramment qu'une machine peut être soit de type Haut (si elle est profitable), soit de type Bas (si elle n'est pas rentable). Si le jeu est multijoueur, l'apprentissage devient stratégique dès lors que les types des machines sont corrélés. Un joueur a ainsi intéret à observer son voisin car cela peut lui apporter des informations sur sa propre machine. Il peut par exemple tirer avantage de l'expérience des autres joueurs sans prendre de risque lui-même.
Nous étudierons les équilibres de Nash de ce jeu suivant deux problématiques distinctes. La question principale sera l'efficacité de ces équilibres : est-ce que les joueurs acquièrent une bonne idée du type de leur machine ? Le cas échéant, combien d'entre eux ? Est-ce que cet apprentissage est rapide ?

2 Jeu de bandit-manchot à grand nombre de joueurs

Nous étudions le comportement asymptotique des joueurs lorsque leur nombre tend vers $+\infty$. Le modèle est le suivant.
A chaque étape, chacun des N joueurs décide d'abord s'il continue d'expérimenter le bras (risqué) de son bandit-manchot. S'il s'arrête, cette décision est irréversible et donne pour les étapes suivantes un paiement normalisé à zéro. S'il continue, il reçoit un paiement aléatoire X_n^i (le total des paiements étant escompté à taux δ). Enfin, il observe quels joueurs ont décidé de rester sur le bras risqué. Ainsi les paiements sont une information privée, mais les décisions sont publiques.

La distribution des paiements X_n^i (que l'on identifie au type) est la même pour toutes les machines (et pour toutes les étapes). On est donc dans un cas de corrélation parfaitement positive : toutes les machines ont le même type. Si le type est Haut, l'espérance est positive, s'il est Bas elle est négative. Ce type est inconnu des joueurs. Chacun peut l'apprendre en observant ses paiements et les décisions des autres joueurs.
Dans un premier temps l'intuition et la littérature sur le sujet [1] (notamment les modèles à continuum de joueurs [2]) suggèrent que le comportement des joueurs quand $N \to +\infty$ donne lieu à un apprentissage déterministe. Expliquons brièvement en quoi ceci consiste. Pendant les premières étapes du jeu, tous les joueurs expérimentent. Puis les plus pessimistes d'entre eux, c'est-à-dire ceux qui ont reçu les pires informations concernant le type, arrêtent l'expérimentation. Le nombre de ces départs dépend du type des machines et, comme il y a beaucoup de joueurs, les joueurs restants peuvent alors déterminer ce type avec une grande certitude. Ces derniers arrêtent donc tous l'expérimentation si le type est Bas, et expérimentent indéfiniment si le type est Haut.
En fait, ce scénario n'a pas toujours lieu. Nous donnerons les conditions d'existence de tels équilibres, dits asymptotiquement déterministes. Dans d'autres cas, nous verrons que l'apprentissage peut être aléatoire. Ainsi, si l'on considère des équilibres symétriques, le nombre limite de départs des joueurs les plus pessimistes suit une loi de Poisson dont le paramètre dépend du type. L'apprentissage n'est alors que partiel.

3 Jeu de bandit-manchot à 2 joueurs aux types partiellement corrélés

Le problème de la corrélation partielle étant original et potentiellement complexe, nous adopterons le même modèle que précédemment, mais à quelques hypothèses simplificatrices près. Tout d'abord, nous considérons le cas où il n'y a que deux joueurs. Ensuite, le jeu n'est plus séquentiel, mais en temps continu. Les machines ont deux bras : l'un sûr, l'autre risqué. Actionner le bras sûr donne un paiement constant connu. En revanche le bras risqué ne donne aucun paiement si le type de la machine est Bas ; si le type est Haut les paiements sont délivrés à des intervalles de temps aléatoires qui suivent une loi exponentielle, de sorte que dans ce cas le bras risqué est en moyenne plus rentable que le bras sûr.
L'originalité du modèle réside dans le fait que les types des machines ne sont pas parfaitement corrélés. Ceux-ci sont distribués aléatoirement suivant des probabilités arbitraires p_0^{HH}, p_0^{HB},p_0^{BH}, p_0^{HH}. Par exemple la corrélation parfaitement positive correspond au cas où $p_0^{HB} = p_0^{BH} = 0$. De plus, nous distinguerons deux cas, suivant que les paiements soient publiquement observés ou non.
Nous avons identifié un paramètre permettant de quantifier la corrélation. Si la coorélation est positive, une bonne nouvelle pour un joueur est aussi une bonne nouvelle pour l'autre joueur. Si elle est négative, une bonne nouvelle pour un joueur se transforme en mauvaise nouvelle pour l'autre. Notre travail consiste à étudier l'existence et l'unicité d'équilibres et/ou d'équilibres purs. Nous donnerons également quelques résultats qualitatifs concernant ces équilibres.

Références

[1] D. Rosenberg, E. Solan and N. Vielle (2007) : Social learning in One-Arm Bandit Problems, *Econometrica*, Vol. 75, No. 6, 1591-1611.

[2] A. Caplin and J. Leahy (1994) : Business as Usual, Market Crashes, and Wisdom After the Fact, *The American Economic Review*, Vol. 84, No. 3, 548-565.