Gains et pertes sont fondamentalement différents pour la minimisation du regret. Le cas sparse.

Joon Kwon
Université Pierre-et-Marie-Curie, Paris, France

Vianney Perchet
INRIA & Université Paris-Diderot, Paris, France

Mots-clefs : regret, online optimization, sparse, bandit

Nous considérons le problème de décision séquentielle appelé *minimisation de regret* [6, 7, 3, 8]. Un joueur fait face à un environnement inconnu et changeant, et choisit une *action* à chaque étape $t = 1, 2, \dots, T$. À la fin de chaque étape, la perte (ou le gain) qu'il subit dépend à la fois de son action et de l'état de l'environnement (qui varie).
Plus précisément, le joueur dispose d'un ensemble de d actions pures, et choisit, à l'instant t, une distribution de probabilités sur les actions pures, autrement dit un élément $x_t \in \Delta_d$ du simplexe. Un vecteur de gain $g_t \in [0,1]^d$ (resp. de perte $\ell_t \in [0,1]^d$) est ensuite révélé par l'environnement, et le joueur obtient le gain (resp. la perte) $\langle g_t | x_t \rangle$ (resp. $\langle \ell_t | x_t \rangle$). Le joueur souhaite maximiser (resp. minimiser) son gain (resp. sa perte) cumulé(e) $\sum_{t=1}^T \langle g_t | x_t \rangle$ (resp. $\sum_{t=1}^T \langle \ell_t | x_t \rangle$). Cependant, nous ne souhaitons faire que très peu d'hypothèses sur l'environnement. En particulier, nous n'avons pas d'a priori sur les vecteurs de gain, à la différence, par exemple, d'un cadre bayésien. Nous considérons une notion relative de performance, appelée le regret :

$$R_T = \max_i \sum_{t=1}^T g_t^{(i)} - \sum_{t=1}^T \langle g_t | x_t \rangle .$$

Cette quantité compare les gains effectivement obtenus aux pertes que le joueur aurait pu obtenir s'il avait joué la stratégie constante qui s'est trouvé être la meilleure. Le joueur souhaite minimiser cette quantité. Le but est de déterminer la meilleure garantie que peut donner une stratégie quand au regret, quelque soit la suite des vecteurs de gains choisis par l'environnement. Cette meilleure garantie, appelée minimax regret, s'écrit :

$$\min_{\text{strat.}} \max_{(g_t)_t} R_T,$$

où le minimum est pris sur l'ensemble des stratégies du joueur et le maximum sur les suites de vecteurs de gains possibles. Il est établi [5, 4] que cette quantité est de l'ordre de $\sqrt{T \log d}$. Et en l'état, le problème est le même que l'on considère des gains ou des pertes.
On ajoute à présent une hypothèse. On se donne $1 \leq s \leq d$ un entier et on suppose que les vecteurs de gains (resp. de pertes) ont au plus s composantes non nulles, et on étudie à nouveau le minimax regret. On démontre que les gains et les pertes ne sont alors plus équivalents : le minimax regret est d'ordre $\sqrt{T \log s}$ et $\sqrt{Ts \frac{\log d}{d}}$ pour les gains et les pertes respectivement.
On étudie également le cadre bandit où le joueur n'observe pas le vecteur de gains mais seulement le paiement qu'il a effectivement obtenu. Sans hypothèse particulière, il est connu [2, 1] que le minimax regret est de l'ordre de $\sqrt{Td}$. Lorsqu'on ajoute notre hypothèse, on établit que le minimax pour les pertes est de l'ordre de $\sqrt{Ts}$ à un facteur logarithmique près. Le minimax pour les gains est inconnu et reste donc un problème ouvert.

Références

[1] Jean-Yves Audibert and Sébastien Bubeck. Minimax policies for adversarial and stochastic bandits. In *COLT*, pages 217–226, 2009.

[2] Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1) :48–77, 2002.

[3] Sébastien Bubeck. Introduction to online optimization : Lecture notes. *Princeton University, New York*, 2011.

[4] Nicolo Cesa-Bianchi. Analysis of two gradient-based algorithms for on-line regression. In *Proceedings of the tenth annual conference on Computational learning theory*, pages 163–170. ACM, 1997.

[5] Nicolo Cesa-Bianchi, Yoav Freund, David Haussler, David P Helmbold, Robert E Schapire, and Manfred K Warmuth. How to use expert advice. *Journal of the ACM (JACM)*, 44(3) :427–485, 1997.

[6] Nicolo Cesa-Bianchi, Gábor Lugosi, and Gilles Stoltz. Regret minimization under partial monitoring. *Mathematics of Operations Research*, 31(3) :562–580, 2006.

[7] James Hannan. Approximation to bayes risk in repeated play. *Contributions to the Theory of Games*, 3(97-139) :2, 1957.

[8] Shai Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2) :107–194, 2011.