# Explorer puis exploiter : Le véritable visage du regret dans les problèmes de bandit stochastique.

Pierre Ménard [1] & Aurélien Garivier [2] & Gilles Stoltz

[1] *Institut de Mathématiques de Toulouse; UMR5219*
*Université de Toulouse; CNRS*
*UPS IMT, F-31062 Toulouse Cedex 9, France, pierre.menard@math.univ-toulouse.fr*
[2] *Institut de Mathématiques de Toulouse; UMR5219*
*Université de Toulouse; CNRS*
*UPS IMT, F-31062 Toulouse Cedex 9, France, aurelien.garivier@math.univ-toulouse.fr*
[3] *CNRS & HEC Paris, stoltz@hec.fr*

**Résumé.** Les problèmes de bandit stochastique forment un modèle d'apprentissage stochastique d'allocation séquentielle de ressources où apparaît un dilemme entre exploitation et exploration. L'objectif du joueur dans un tel problème est de minimiser son regret accumulé lors de l'apprentissage, par rapport à un joueur omniscient. On revisite ici les bornes inférieures sur ce regret. Ces dernières sont importantes, notamment pour trancher sur l'optimalité d'une stratégie, mais aussi pour mieux comprendre le comportement typique du regret au cours de l'apprentissage. En particulier, on présente de nouvelles bornes inférieures non asymptotiques sur le regret dont les preuves reposent sur des propriétés élémentaires de la divergence de Kullback-Leibler. Ces bornes permettent de dégager deux phases dans la croissance du regret : une première où le regret croît de manière presque linéaire puis une seconde, bien connue, où ce dernier croît de manière logarithmique.

**Mots-clés.** bandit multi-bras, regret, outils de théorie de l'information, bornes inférieures non asymptotiques...

**Abstract.** Multi-armed bandit problems are stochastic sequential resource allocation problems where appears an exploration–exploitation trade-off. In such problems, the gambler have to minimize his regret for not playing like a omniscient gambler. We revisit lower bounds on the regret. These bounds are important not only to arbitrate on the optimality of an algorithm but also to provide an insight into the regret's growth. In particulary, we obtain non-asymptotic bounds and provide straightforward proofs based only on well-known properties of Kullback-Leibler divergences. These bounds show that in an initial phase the regret grows almost linearly, and that the well-known logarithmic growth of the regret only holds in a final phase.

**Keywords.** multi-armed bandit, regret, information-theoretic proof techniques, non-asymptotic lower bounds...

# 1 Introduction

On reprend ici le cadre habituel des problèmes de bandit stochastiques que l'on peut retrouver dans Bubeck and Cesa-Bianchi (2012). Soit un problème de bandit avec un nombre fini de bras indexés par $a \in \{1, .., K\}$. À chacun de ces bras est associé une loi de probabilité inconnue $\nu_a$ avec par exemple $\nu_a = \mathcal{B}(\mu_a)$. À chaque tour $t \geqslant 1$ le joueur tire un bras et reçoit une récompense distribuée selon la loi $\nu_{A_t}$. C'est la seule information à laquelle il aura accès.

**Stratégies.** Une stratégie $\psi$, adoptée par le joueur, associe un bras à l'information récoltée durant les tours précédents, et éventuellement un aléa auxiliaire, qui, sans perte de généralité, peut être donné par une suite $U_0, U_1, U_2, \ldots$ de variables aléatoires indépendantes, distribuées selon la loi uniforme sur $[0, 1]$. Ces variables sont aussi indépendantes des récompenses $Y_t$. Ainsi, une stratégie est une suite de fonctions mesurables $\psi = (\psi_t)_{t\geqslant 0}$ qui à l'information passée

$$I_t = \big(U_0,\, Y_1, U_1,\, \ldots,\, Y_t, U_t\big)\,,$$

associent un bras $\psi_t(I_t) = A_{t+1} \in \{1, \ldots, K\}$, où $t \geqslant 0$. L'information initiale se réduit à $I_1 = U_0$ et le premier bras est tirée selon $A_1 = \psi_0(U_0)$.

**Mesure de probabilité.** D'après le théorème d'extension de Kolmogorov il existe un espace de probabilité $(\Omega, \mathcal{F})$ tel que toutes les variables aléatoires définies ci-dessus, puissent être définies sur cet espace. Par la suite on aura besoin de réaliser des changements de mesures entre les différents problèmes. On pose donc le vecteur de lois de probabilité associées aux bras $\nu = (\nu_a)_{a=1,\ldots,K}$ et on définit la mesure de probabilité $\mathbb{P}_\nu$ sur $(\Omega, \mathcal{F})$ telle que pour tout $t \geqslant 0$, pour tout Boréliens $B \subseteq \mathbb{R}$ et $B' \subseteq [0, 1]$,

$$\mathbb{P}_\nu\big(Y_{t+1} \in B,\, U_{t+1} \in B' \;\big|\; I_t\big) = \nu_{\psi_t(I_t)}(B)\ \lambda(B')\,,$$

où $\lambda$ est la mesure de Lebesgue sur $[0, 1]$.

**Regret.** Avant de définir le regret donnons quelques notations habituelles. On pose $\mu_a = E\nu_a$ l'espérance du bras $a$ et $\Delta_a$ son écart avec un bras optimal :

$$\mu^\star = \max_{a=1,\ldots,K} \mu_a \qquad \text{et} \qquad \Delta_a = \mu^\star - \mu_a\,.$$

Le nombre de fois que le bras $a$ est tiré avant le temps $T$ est noté :

$$N_a(T) = \sum_{t=1}^{T} \mathbb{I}_{\{A_t = a\}}\,.$$

Le regret correspond alors à la différence entre le gain moyen obtenu en jouant uniquement un bras optimal (joueur omniscient) et celui obtenu par le joueur,

$$R_{\nu,T} = T\mu^\star - \mathbb{E}_\nu\left[\sum_{t=1}^{T} Y_t\right] = \mathbb{E}_\nu\left[\sum_{t=1}^{T}\big(\mu^\star - \mu_{A_t}\big)\right] = \sum_{a=1}^{K} \Delta_a \, \mathbb{E}_\nu\big[N_a(T)\big] \, .$$

# 2 Une borne inférieure asymptotique

Une borne inférieure fondamentale, sur le regret, pour les problèmes de bandit stochastique, est celle de Lai and Robbins (1985) généralisée par Burnetas and Katehakis (1996). Cette dernière affirme que si la stratégie est consistante i.e. $\mathbb{E}_\nu\big[N_a(T)\big] = o(T^\alpha)$ pour tout $0 < \alpha \leqslant 1$ et tout bras $a$ sous-optimal ($\Delta_a > 0$) alors on tire, en moyenne, asymptotiquement, au moins un nombre logarithmique de fois les bras sous optimaux,

$$\liminf_{T\to\infty} \ \frac{\mathbb{E}_\nu\big[N_a(T)\big]}{\ln T} \geqslant \frac{1}{\mathrm{kl}(\mu_a, \mu^\star)} \, . \tag{1}$$

Où kl est la divergence de Kullback-Leibler pour des lois de Bernoulli,

$$\forall p, q \in [0,1]^2, \qquad \mathrm{kl}(p,q) = p \ln \frac{p}{q} + (1-p) \ln \frac{1-p}{1-q} \, .$$

En sommant ces bornes pour tous les bras sous-optimaux il vient immédiatement une borne inférieure asymptotique sur le regret. Il existe des algorithmes qui atteignent cette borne, voir par exemple Cappé et al. (2013). Cependant la nature asymptotique de cette borne se révèle, de façon spectaculaire, lors des simulations numériques, comme on peut le constater avec la figure 1. En effet pour $T$ relativement faible le regret ne croît pas de manière logarithmique en $T$, il est même plutôt linéaire en $T$, au tout début du moins. Et il faut attendre des horizons $T$ très grands, trop pour que ce soit facilement visible expérimentalement, pour que le regret passe au dessus de la borne inférieure asymptotique.

# 3 Une inégalité fondamentale

On présente ici, une inégalité fondamentale qui permet de prouver efficacement les différentes bornes inférieures, en particulier la borne (1). Celle-ci est issue de Garivier et al. (2016) ainsi que les bornes inférieures non asymptotiques ci dessous. Plus précisément, cette dernière est composée d'une égalité standard, la règle du conditionnement pour l'entropie relative et d'une inégalité moins courante (dans le domaine des bandits), la contraction de l'entropie,

$$\sum_{a=1}^{K} \mathbb{E}_\nu\big[N_a(T)\big] \, \mathrm{KL}(\nu_a, \nu'_a) = \mathrm{KL}\big(\mathbb{P}_\nu^{I_T}, \, \mathbb{P}_{\nu'}^{I_T}\big) \geqslant \mathrm{kl}\big(\mathbb{E}_\nu[Z], \, \mathbb{E}_{\nu'}[Z]\big) \, , \tag{F}$$

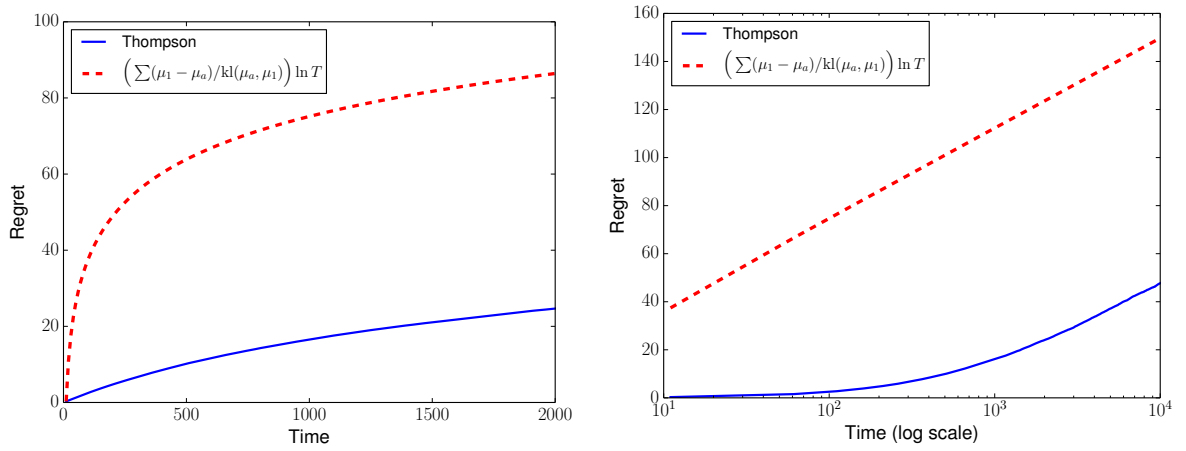


**Figure 1:** Regret pour la stratégie de Thomson Sampling, Thompson (1933) (en *bleu*) pour un problème de bandit avec des Bernoulli de paramètres $(\mu_a)_{1\leqslant a\leqslant 6} = (0.05, 0.04, 0.02, 0.015, 0.01, 0.005)$, la moyenne s'effectue sur approximativement 500 essaies.

Versus la borne inférieure de Lai and Robbins (1985) (en *rouge*).

*Gauche* : le regret n'est pas logarithmique au début, plutôt linéaire.
*Droite*: la borne asymptotique est atteinte seulement pour des horizons $T$ extrêmement grands.

où $\mathbb{P}_\nu^{I_T}$ et $\mathbb{P}_{\nu'}^{I_T}$ sont les mesures images de $\mathbb{P}_\nu$ et $\mathbb{P}_{\nu'}$ par $I_T$ et où $Z$ est une variable aléatoire $\sigma(I_T)$–mesurable à valeurs dans $[0,1]$.

Typiquement, on choisira la variable aléatoire $Z = N_a(T)/T$ pour un certain bras $a$. Contrairement aux preuves précédentes, cela permet d'éviter d'introduire la probabilité d'un événement bien choisi qu'il faudrait ensuite contrôler avec une inégalité de Markov pour revenir à des espérances. Cette technique de preuve se situe dans le prolongement de celles introduites dans Kaufmann et al. (2016).

# 4 Bornes inférieures non asymptotiques

On considère, pour simplifier, le problème de bandit suivant $\nu = (\nu_a)_{a=1,\dots,K} = \big(\mathcal{B}(\mu_a)\big)_{a=1,\dots,K}$, constitué de loi de Bernoulli. Sans perdre de généralités on peut supposer qu'il y a un unique bras optimal $\mu^\star = \mu_1 > \mu_2 \geqslant \dots \geqslant \mu_K$.

## 4.1 Régime : T petit

On présente deux bornes inférieures avec leurs défauts et qualités. On s'attend pour $T$ petit à tirer en moyenne $T/K$ chacun des bras, ce qui correspond à ce que l'on obtient avec la stratégie qui tire un bras uniformément parmi tous les bras à chaque tour. En

utilisant l'inégalité (F), on peut montrer que pour toute stratégie qui est meilleure que la stratégie uniforme (i.e. qui tire en moyenne plus que $T/K$ fois le bras optimal), pour tout problème $\nu$, pour tout bras $a$,

$$\mathbb{E}_\nu\big[N_a(T)\big] \geqslant \frac{T}{K}\Big(1 - \sqrt{2T\mathrm{kl}(\mu_a, \mu^\star)}\Big)\,.$$

En particulier,

$$\forall\, T \leqslant \frac{1}{8\mathrm{kl}(\mu_a, \mu^\star)}, \qquad \mathbb{E}_\nu\big[N_a(T)\big] \geqslant \frac{T}{2K}\,,$$

ce qui donne bien la croissance linéaire initiale du regret recherchée. Cependant ce régime est valable seulement pour $T$ au plus de l'ordre $8/\mathrm{kl}(\mu_a, \mu^\star)$, indépendamment de $K$, alors que l'on aurait pu s'attendre à une constante de l'ordre de $K/\mathrm{kl}(\mu_a, \mu^\star)$. On peut obtenir ce facteur $K$ en minorant non plus le nombre moyen de fois que l'on tire un bras donné mais seulement le nombre moyen de fois qu'un bras sous optimal est tiré,

$$\sum_{a\neq 1} \mathbb{E}_\nu\big[N_a(T)\big] \geqslant T\left(1 - \frac{1}{K} - \frac{\sqrt{2T\,\mathrm{kl}(\mu_K, \mu^\star)}}{K} - \frac{2T\,\mathrm{kl}(\mu_K, \mu^\star)}{K}\right).$$

En particulier, cela implique une borne inférieure sur le regret

$$R_{\nu,T} \geqslant \left(\min_{a\neq 1} \Delta_a\right) T\left(1 - \frac{1}{K} - \frac{1\sqrt{2T\,\mathrm{kl}(\mu_K, \mu^\star)}}{K} - \frac{2T\,\mathrm{kl}(\mu_K, \mu^\star)}{K}\right).$$

### 4.2 Régime : T grand

Grâce à une version non asymptotique de l'hypothèse utilisé pour montrer (1), il est possible d'obtenir la borne inférieure suivante,

$$\mathbb{E}_\nu\big[N_a(T)\big] \geqslant \frac{\ln T}{\mathrm{kl}(\mu_a, \mu^\star)} - \mathrm{O}\big(\ln(\ln T)\big)\,. \tag{2}$$

où le terme de second ordre peut être explicité avec les constantes intervenant dans l'hypothèse. Cette dernière implique immédiatement (1), au prix d'une hypothèse plus forte. La borne (2) n'est pas triviale, i.e. le minorant n'est pas négatif, seulement pour des horizons $T$ au moins de l'ordre de $K/\mathrm{kl}(\mu_a, \mu^\star)$.

## References

S. Bubeck and N. Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5(1):1–122, 2012.

A.N. Burnetas and M.N. Katehakis. Optimal adaptive policies for sequential allocation problems. *Advances in Applied Mathematics*, 17(2):122–142, 1996.

O. Cappé, A. Garivier, O.-A. Maillard, R. Munos, and G. Stoltz. Kullback-Leibler upper confidence bounds for optimal sequential allocation. *Annals of Statistics*, 41(3):1516–1541, 2013.

A. Garivier, P. Ménard, and G. Stoltz. Explore first, exploit next: The true shape of regret in bandit problems. *submitted*, 2016.

E. Kaufmann, O. Cappé, and A. Garivier. On the complexity of best-arm identification in multi-armed bandit models. *Journal of Machine Learning Research*, 2016. To appear.

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

W. R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25:285–294, 1933.