Problèmes de Bandits – Identification du Meilleur Bras

Le cas du budget fixé

Antoine Barrier[1,2], Aurélien Garivier[1], Gilles Stoltz[2]

1. UMPA, ÉNS de Lyon 2. LMO, Université Paris-Saclay

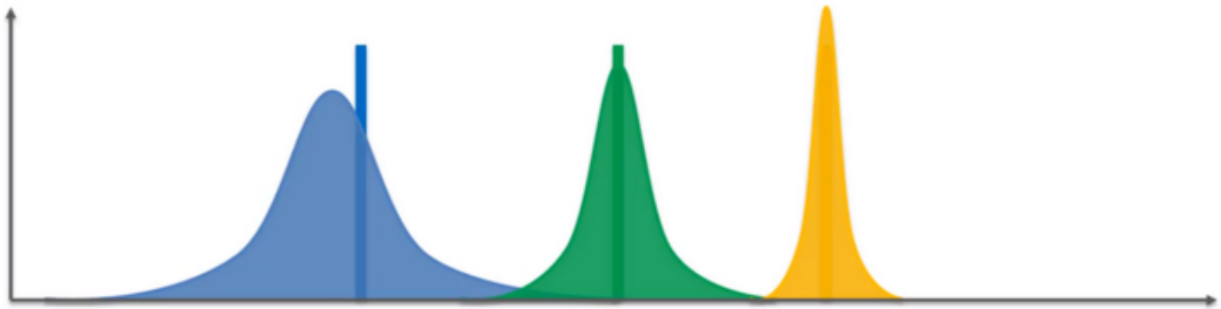
antoine.barrier@ens-lyon.fr – `http://perso.ens-lyon.fr/antoine.barrier/fr/`

16 avril 2022, JDS, Lyon

Plan

Problèmes de bandits

- K groupes (*bras*). À chaque bras $a \in \{1, \dots, K\}$ est associée une distribution de probabilité ν_a (inconnue !) d'espérance $\mu_a = \mathbb{E}[\nu_a]$. On considère des bandits ayant un **unique bras optimal** noté $a^* = a^*(\boldsymbol{\nu})$ tel que $\mu^* = \mu_{a^*} > \max_{a \neq a^*} \mu_a$



- À chaque tour $t = 1, 2, \dots,$ un joueur/agent
  - choisit un bras $A_t \in \{1, \dots, K\}$ en fonction des observations passées $(A_1, X_1, A_2, X_2, \dots, A_{t-1}, X_{t-1})$,
  - reçoit une récompense $X_t \sim \nu_{A_t}$ indépendante du passé

→ **cadre de l'apprentissage séquentiel : on reçoit les observations une par une et la stratégie du joueur influe sur celles-ci**

Applications : systèmes de recommandation, essais cliniques, A/B testing, ...

On fixe un nombre de tirages T et on veut maximiser nos gains. En cumulant sur tous les tours on a une perte relative – un **regret** – par rapport à un oracle de

$$R(T) = \sum_{t=1}^{T} \mu^* - \mathbb{E}[X_t] = T\mu^* - \sum_{t=1}^{T} \mathbb{E}[\mu_{A_t}] = \sum_{a \neq a^*} (\mu^* - \mu_a)\mathbb{E}[N_a(T)]$$

où $N_a(t) = \sum_{s=1}^{t} \mathbb{1}_{A_s=a}$ est le nombre de tirages du bras a après t tours.

PREMIER OBJECTIF **Minimiser le regret**

$\rightarrow$ **exploration vs exploitation**

- ✓ critère d'apprentissage : $R(T) = o(T)$ est sous-linéaire en T : on tire de moins en moins les mauvais bras !
- ✓ objectif très étudié dans la littérature, nombreux algorithmes vérifiant des critères d'optimalité dans différents cadres

DEUXIÈME OBJECTIF **Identifier rapidement le meilleur bras**

$\rightarrow$ **exploration pure**

- **BAI à confiance fixée** : à niveau de confiance $\delta > 0$ fixé, un algorithme doit s'arrêter après un nombre d'observations *aléatoire* τ et retourner un estimateur $\hat{a}_\tau$ du meilleur bras de manière à garantir que $\mathbb{P}_{\boldsymbol{\nu}}(\hat{a}_\tau \neq a^*) \leq \delta$ (l'algorithme est dit δ-**correct**).

  $\rightarrow$ **Objectif : minimiser le nombre moyen de tirages** $\mathbb{E}_{\boldsymbol{\nu}}[\tau]$ ✓

- **BAI à budget fixé** : à nombre total de tirages T fixé, un algorithme doit après T observations retourner un estimateur $\hat{a}_T$ du meilleur bras.

  $\rightarrow$ **Objectif : minimiser la probabilité d'erreur** $\mathbb{P}_{\boldsymbol{\nu}}(\hat{a}_T \neq a^*)$ ✗

Des résultats d'optimalité à confiance fixée

On suppose que les distributions appartiennent à un même modèle exponentiel.

On pose $\mathrm{Alt}(\boldsymbol{\nu})$ l'ensemble des problèmes de bandits ayant un bras optimal différent de celui de $\boldsymbol{\nu}$ et $\Sigma_K = \left\{ \boldsymbol{w} \in [0,1]^K \ : \ \sum_{a=1}^K w_a = 1 \right\}$.

Théorème (borne inférieure, [GK16])

Pour toute stratégie δ-correcte, on a

$$\forall \boldsymbol{\nu}, \quad \mathbb{E}_{\boldsymbol{\nu}}[\tau_\delta] \geq T(\boldsymbol{\nu}) \, \mathrm{kl}(\delta, 1-\delta) \sim_{\delta \to 0} T(\boldsymbol{\nu}) \log(1/\delta)$$

$$\text{où} \qquad T(\boldsymbol{\nu})^{-1} = \sup_{\boldsymbol{w} \in \Sigma_K} \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\nu})} \sum_{a=1}^K w_a \, \mathrm{KL}(\nu_a, \lambda_a)$$

[GK16] Garivier, A. et Kaufmann, E. (2016), **Optimal Best Arm Identification with Fixed Confidence**, In *29th Conference on Learning Theory (COLT)*

Des résultats d'optimalité à confiance fixée

Schéma de preuve.

Soit $a^* = a^*(\boldsymbol{\nu})$. Pour tout $\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\nu})$, on obtient en utilisant la chain rule pour KL, l'inégalité de data-processing et l'hypothèse de stratégie δ-correcte que

$$\sum_{a=1}^{K} \mathbb{E}_{\boldsymbol{\nu}}[N_a(\tau_\delta)]\,\mathrm{KL}(\nu_a, \lambda_a) \geq \mathrm{kl}\big(\mathbb{P}_{\boldsymbol{\nu}}(\hat{a}_{\tau_\delta} \neq a^*), \mathbb{P}_{\boldsymbol{\lambda}}(\hat{a}_{\tau_\delta} \neq a^*)\big) \geq \mathrm{kl}(\delta, 1-\delta)$$

et donc en passant à l'inf sur $\boldsymbol{\lambda}$

$$\begin{aligned}
\mathrm{kl}(\delta, 1-\delta) &\leq \mathbb{E}_{\boldsymbol{\nu}}[\tau_\delta] \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\nu})} \sum_{a=1}^{K} \underbrace{\frac{\mathbb{E}_{\boldsymbol{\nu}}[N_a(\tau_\delta)]}{\mathbb{E}_{\boldsymbol{\nu}}[\tau_\delta]}}_{= w_a(\boldsymbol{\nu})} \mathrm{KL}(\nu_a, \lambda_a) \\
&\leq \mathbb{E}_{\boldsymbol{\nu}}[\tau_\delta] \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\nu})} \sum_{a=1}^{K} w_a(\boldsymbol{\nu})\,\mathrm{KL}(\nu_a, \lambda_a) \\
&\leq \mathbb{E}_{\boldsymbol{\nu}}[\tau_\delta] \sup_{\boldsymbol{w} \in \Sigma_K} \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\nu})} \sum_{a=1}^{K} w_a\,\mathrm{KL}(\nu_a, \lambda_a) \qquad (\text{car } w(\boldsymbol{\nu}) \in \Sigma_K)
\end{aligned}$$

Une preuve similaire à budget fixé ?

Essayons de faire de même en budget fixé. Pour cela on a besoin de l'inégalité

$$\mathrm{kl}(p,q) \geq p\log(1/q) - \ln(2) \quad \Longleftrightarrow \quad q \geq \exp\left(-\frac{\mathrm{kl}(p,q)+\ln(2)}{p}\right)$$

ce qui donne

$$\begin{aligned}\mathbb{P}_{\boldsymbol{\nu}}(\hat{a}_T \neq a^*) &\geq \exp\left(-\frac{\mathrm{kl}\big(\mathbb{P}_{\boldsymbol{\lambda}}(\hat{a}_T \neq a^*), \mathbb{P}_{\boldsymbol{\nu}}(\hat{a}_T \neq a^*)\big)+\ln(2)}{\mathbb{P}_{\boldsymbol{\lambda}}(\hat{a}_T \neq a^*)}\right) \\ &\geq \exp\left(-\frac{\sum_{a=1}^K \mathbb{E}_{\boldsymbol{\lambda}}[N_a(T)]\,\mathrm{KL}(\lambda_a,\nu_a)+\ln(2)}{\mathbb{P}_{\boldsymbol{\lambda}}(\hat{a}_T \neq a^*)}\right)\end{aligned}$$

Problème : on a été contraint d'échanger les rôles de $\boldsymbol{\nu}$ et $\boldsymbol{\lambda}$ par rapport à la preuve précédente (en cause notamment la non symétrie de la kl).

$\rightarrow$ Le fait d'avoir $\mathbb{E}_{\boldsymbol{\lambda}}[N_a(T)]$ et non $\mathbb{E}_{\boldsymbol{\nu}}[N_a(T)]$ est l'une de raisons qui rend compliquée la recherche d'une borne inférieure à budget fixé.

Avec une hypothèse de monotonie

Continuons tout de même nos calculs en supposant que les bras sont rangés de sorte que $a^* = 1$ et $\mu_1 > \mu_2 \geq \cdots \geq \mu_K$.

$$\mathbb{P}_{\boldsymbol{\nu}}(\hat{a}_T \neq 1) \geq \exp\left(- \frac{\sum_{a=1}^{K} \mathbb{E}_{\boldsymbol{\lambda}}[N_a(T)]\, \mathrm{KL}(\lambda_a, \nu_a) + \ln(2)}{\mathbb{P}_{\boldsymbol{\lambda}}(\hat{a}_T \neq 1)} \right)$$

Si la stratégie est performante, elle va identifier le meilleur bras donc sous le bandit alternatif $\boldsymbol{\lambda}$, on aura $\mathbb{P}_{\boldsymbol{\lambda}}(\hat{a}_T \neq 1) \to_{T \to +\infty} 0$.

Fixons $2 \leq k \leq K$ et considérons le bandit alternatif $\boldsymbol{\lambda} = (\nu_k, \nu_2, \nu_3, \ldots, \nu_K)$. Alors on a obtenu

$$\mathbb{P}_{\boldsymbol{\nu}}(\hat{a}_T \neq 1) \geq \exp\Big(- \mathbb{E}_{\boldsymbol{\lambda}}[N_1(T)]\, \mathrm{KL}(\underbrace{\lambda_1}_{\nu_k}, \nu_1)(1 + o(1)) \Big)$$

Si de plus la stratégie est *monotone*, alors $\mathbb{E}_{\boldsymbol{\lambda}}[N_1(T)] \leq \frac{T}{k}$ et donc

$$\forall 2 \leq k \leq K, \quad \mathbb{P}_{\boldsymbol{\nu}}(\hat{a}_T \neq 1) \geq \exp\left(- \frac{T}{\frac{k}{\mathrm{KL}(\nu_k, \nu_1)}} (1 + o(1)) \right)$$

Avec une hypothèse de monotonie

Théorème (Borne inférieure, cas d'une stratégie monotone)

Soit une stratégie

- *symétrique : le choix des bras à tirer est indépendant de leur indice,*
- *consistante :* $\mathbb{P}_{\boldsymbol{\nu}}(\hat{a}_T \neq a^*) = o_{T\to+\infty}(1)$ *pour tout bandit* $\boldsymbol{\nu}$*,*
- *monotone : les espérances de tirages* $(\mathbb{E}_{\boldsymbol{\nu}}[N_a(T)])_{1\leq a\leq K}$ *sont classées dans l'ordre inverse des espérances des bras* $(\mu_a)_{1\leq a\leq K}$ *pour tout bandit* $\boldsymbol{\nu}$*.*

Alors pour tout problème de bandit $\boldsymbol{\nu}$ *on a*

$$\mathbb{P}_{\boldsymbol{\nu}}(\hat{a}_T \neq a^*) \geq C \exp\left(-\frac{2T}{\max\limits_{2\leq k\leq K} \frac{k}{\mathrm{KL}(\nu_{(k)}, \nu_{(1)})}} (1 + o_{T\to+\infty}(1)) \right)$$

pour une constante $C > 0$*, où l'on définit un ordre tel que*

$$\mu_{(1)} > \mu_{(2)} \geq \cdots \geq \mu_{(K)} \,.$$

Comparaison avec la borne de [ABM10]

Théorème (Borne inférieure, cas gaussien [ABM10])

Soit une stratégie symétrique et consistante. Alors pour tout problème de bandit ν dont les bras sont gaussiens de variance σ^2 on a

$$\mathbb{P}_\nu\left(\hat{a}_T \neq a^*\right) \geq C \exp\left(-\frac{2.5\,T}{\sigma^2 \max\limits_{2\leq k\leq K} \frac{k}{(\mu_{(1)}-\mu_{(k)})^2}}(1+o_{T\to+\infty}(1))\right)$$

pour une constante $C > 0$.

Notre borne donne quant à elle

$$\mathbb{P}_\nu\left(\hat{a}_T \neq a^*\right) \geq C \exp\left(-\frac{T}{\sigma^2 \max\limits_{2\leq k\leq K} \frac{k}{(\mu_{(1)}-\mu_{(k)})^2}}(1+o_{T\to+\infty}(1))\right)$$

[ABM10] Audibert, J.-Y., Bubeck, S. et Munos, R. (2010), **Best Arm Identification in Multi-Armed Bandits**, In *23th Conference on Learning Theory (COLT)*

Conclusion

- le cadre de BAI à budget fixé est aujourd'hui moins bien compris que le BAI à confiance fixée et semble plus difficile,

- avec des hypothèses sur les stratégies, on obtient tout de même des bornes inférieures à base de quantités informatives,

- ces hypothèses sont pour l'instant trop restrictives pour espérer des résultats avec des modèles non paramétriques.

Références

Jean-Yves Audibert, Sébastien Bubeck et Rémy Munos :
Best Arm Identification in Multi-Armed Bandits.
In Proceedings of the 23rd Annual Conference on Learning Theory, janvier 2010.

Aurélien Garivier et Emilie Kaufmann :
Optimal Best Arm Identification with Fixed Confidence.
In Conference on Learning Theory, volume 49, pages 998–1027. PMLR, juin 2016.