Problèmes de Bandits – Identification du Meilleur Bras

Antoine BARRIER

doctorant à l'UMPA (ÉNS de Lyon) et au LMO (Université Paris-Saclay)

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

04 avril 2022, Porquerolles

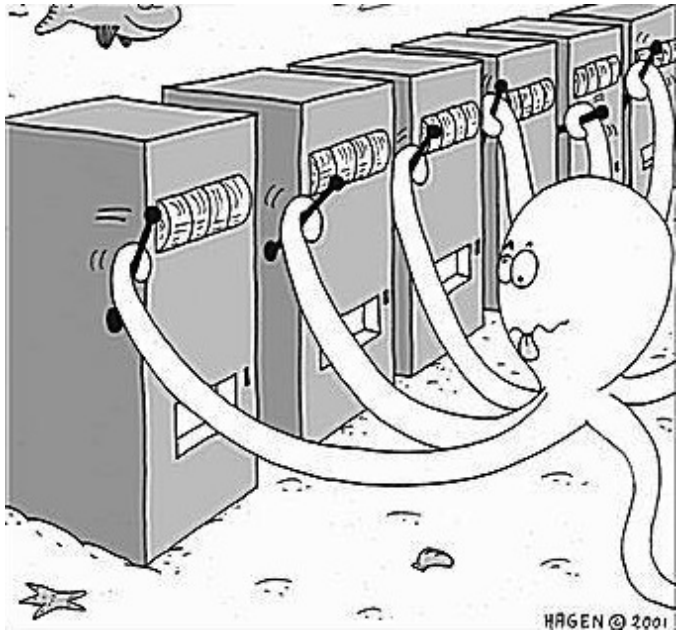
1. Introduction aux problèmes de bandits

2. BAI à confiance fixée
   - Borne inférieure
   - Track-and-Stop : un algorithme asymptotiquement optimal
   - Vers des bornes non-asymptotiques

3. BAI à budget fixé

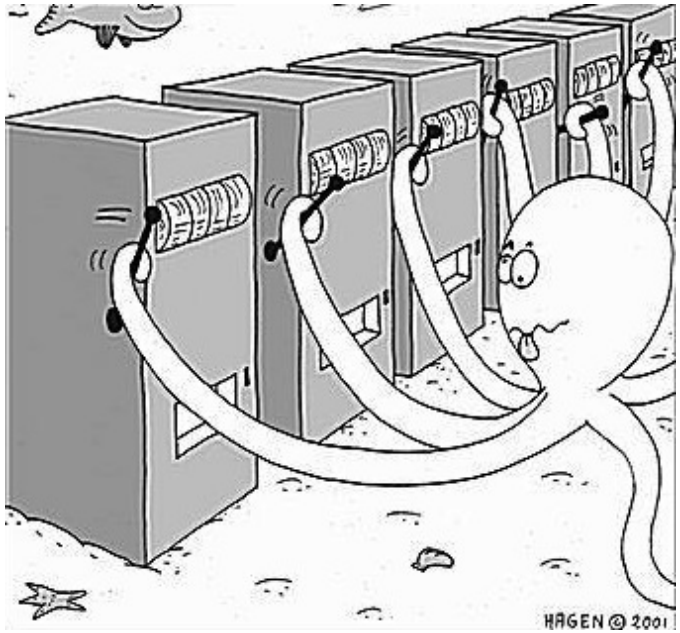
K machines à sous (*bandits manchots*) différentes

Quel bras tirer pour gagner le plus (perdre le moins ...) d'argent à long terme ?



K machines à sous (*bandits manchots*) différentes

Quel bras tirer pour gagner le plus (perdre le moins ...) d'argent à long terme ?

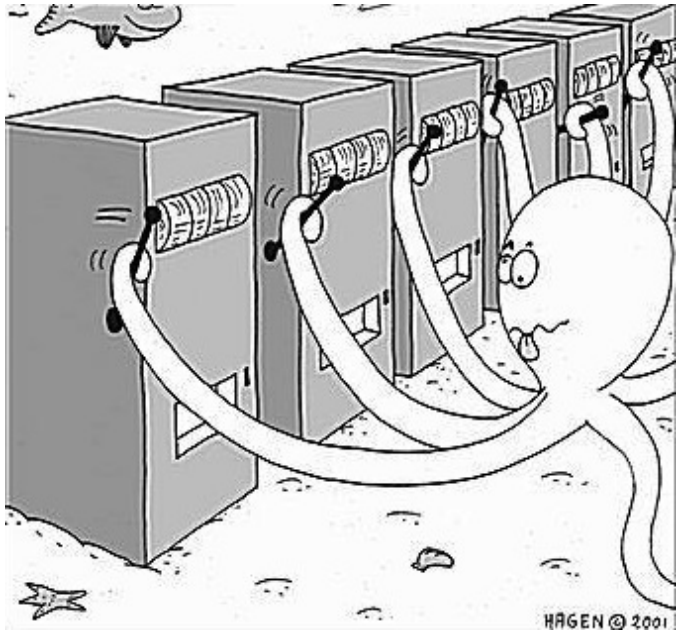


$\rightarrow$ on tire chaque machine quelques fois puis selon les gains obtenus on peut être tenté de se focaliser sur certaines machines, voire une seule

À raison ?

K machines à sous (*bandits manchots*) différentes

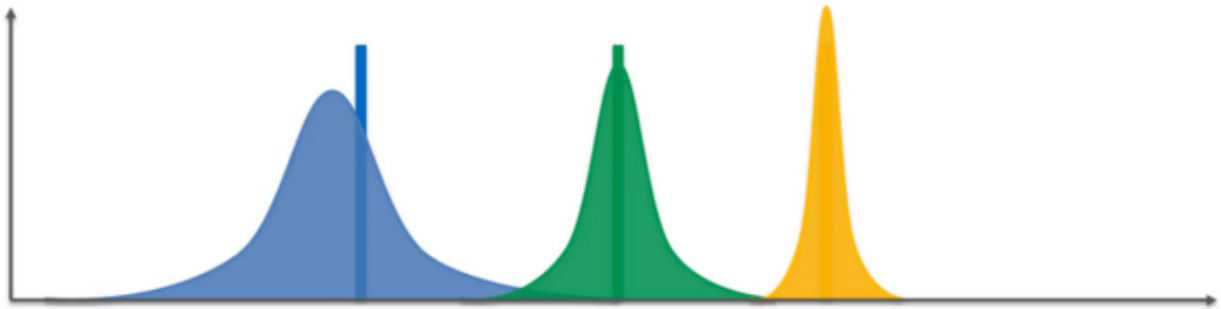
Quel bras tirer pour gagner le plus (perdre le moins ...) d'argent à long terme ?



$\rightarrow$ on tire chaque machine quelques fois puis selon les gains obtenus on peut être tenté de se focaliser sur certaines machines, voire une seule

À raison ? Dilemme **exploration** vs **exploitation**

- K groupes (*bras*). À chaque bras $a \in \{1, \dots, K\}$ on associe une distribution de probabilité ν_a (inconnue !). On notera $\mu_a = \mathbb{E}[\nu_a]$ et on supposera que le meilleur bras est le premier : $\mu_1 > \max_{a \neq 1} \mu_a$



- À tout instant $t = 1, 2, \dots,$ un joueur choisit un bras $A_t \in \{1, \dots, K\}$ en fonction du passé $(A_1, X_1, A_2, X_2, \dots, A_{t-1}, X_{t-1})$ et 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**

Si on se fixe un nombre de tirages T :

- à l'instant t, en tirant A_t on a une perte moyenne de $\mu_1 - \mu_{A_t}$ par rapport à la meilleure option qu'est le bras 1
- en cumulant sur tous les tours on a une perte relative – un **regret** – de

$$R(T) = \sum_{t=1}^{T} \mu_1 - \mathbb{E}[\mu_{A_t}] = \sum_{a \neq 1} (\mu_1 - \mu_a)\mathbb{E}[N_a(T)]$$

où $N_a(T)$ est le nombre de tirages du bras a après T étapes.

Si on se fixe un nombre de tirages T :

- à l'instant t, en tirant A_t on a une perte moyenne de $\mu_1 - \mu_{A_t}$ par rapport à la meilleure option qu'est le bras 1
- en cumulant sur tous les tours on a une perte relative – un **regret** – de

$$R(T) = \sum_{t=1}^{T} \mu_1 - \mathbb{E}[\mu_{A_t}] = \sum_{a \neq 1} (\mu_1 - \mu_a)\mathbb{E}[N_a(T)]$$

où $N_a(T)$ est le nombre de tirages du bras a après T étapes.

PREMIER OBJECTIF **Minimiser le regret**

✗ si $R(T) \simeq cT$ est linéaire en T : la proportion de tirages des mauvais bras ne diminue pas au cours du temps $\rightarrow$ on n'apprend rien

✓ si $R(T) = o(T)$ est sous-linéaire en T : on tire de moins en moins les mauvais bras ! $\rightarrow$ c'est l'objectif !

- **publicité**, **recommandation de films**, ...
  Quelle publicité parmi K différentes entraîne le plus d'achats ?
  Récompense $X_{A_t} = \mathbb{1}_{\text{le } t\text{-ième utilisateur clique}}$.
- **essais cliniques**
  Quel vaccin est le plus efficace parmi K différents ?
  Récompense $X_{A_t} = \mathbb{1}_{\text{le } t\text{-ième patient est immunisé}}$.

On peut imaginer une phase de test sur un panel avant une mise en production

- **publicité**, **recommandation de films**, ...
  Quelle publicité parmi K différentes entraîne le plus d'achats ?
  Récompense $X_{A_t} = \mathbb{1}_{\text{le } t\text{-ième utilisateur clique}}$.
- **essais cliniques**
  Quel vaccin est le plus efficace parmi K différents ?
  Récompense $X_{A_t} = \mathbb{1}_{\text{le } t\text{-ième patient est immunisé}}$.

On peut imaginer une phase de test sur un panel avant une mise en production
$\rightarrow$ changement d'objectif : **exploration pure *puis* exploitation**

DEUXIÈME OBJECTIF **Identifier rapidement le meilleur bras**

On peut travailler dans deux cadres :

- **à confiance fixée** : on fixe un niveau de confiance $\delta > 0$ et on cherche un algorithme qui retourne un bras $\hat{a}_\tau$ qu'il estime être le meilleur après un nombre *aléatoire* τ d'observations, de manière à garantir que $\mathbb{P}_{\boldsymbol{\nu}}(\hat{a}_\tau \neq 1) \leq \delta$ (l'algorithme est dit δ-**correct**).

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

- **à budget fixé** : on fixe un nombre total de tirages T, et on cherche un algorithme qui après T observations retourne un bras $\hat{a}_T$ qu'il estime être le meilleur.

  $\rightarrow$ **Objectif : minimiser la probabilité d'erreur** $\mathbb{P}_{\boldsymbol{\nu}}(\hat{a}_T \neq 1)$

1 Introduction aux problèmes de bandits

2 BAI à confiance fixée

- Borne inférieure
- Track-and-Stop : un algorithme asymptotiquement optimal
- Vers des bornes non-asymptotiques

3 BAI à budget fixé

Borne inférieure [GK16]

On suppose que les distributions appartiennent à un même modèle exponentiel (une distribution ν est donc caractérisée par sa moyenne μ) et on note $d(\mu, \mu') = KL(\nu, \nu')$.
On pose $\mathrm{Alt}(\boldsymbol{\mu})$ *l'ensemble des problèmes de bandits ayant un bras optimal différent de celui de* $\boldsymbol{\mu}$, *et* $\Delta_K = \left\{ \boldsymbol{v} \in [0,1]^K \ : \ \sum_{a=1}^K v_a = 1 \right\}$.

Théorème

Pour toute stratégie δ-correcte, on a

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

$$\text{où} \qquad T(\boldsymbol{\mu})^{-1} = \sup_{\boldsymbol{v} \in \Delta_K} \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\mu})} \sum_{a=1}^K v_a d(\mu_a, \lambda_a) \tag{1}$$

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

Borne inférieure [GK16]

$\rightarrow$ **Peut-on atteindre cette borne ? Au moins quand $\delta \to 0$, peut-on trouver une stratégie, dite asymptotiquement optimale, telle que**

$$\limsup_{\delta \to 0} \frac{\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta]}{\log(1/\delta)} = T(\boldsymbol{\mu}) \, ?$$

Si oui, elle doit approximativement observer les bras selon le **vecteur de poids optimal $\boldsymbol{w}(\mu) \in \Delta_K$** réalisant le maximum dans (1)

Idée Tracker le vecteur de poids optimal associé à la moyenne empirique $\hat{\mu}(t)$ et ajouter de l'exploration forcée pour assurer sa convergence vers $\boldsymbol{w}(\mu)$

Algorithm 1: Track-and-Stop

Input: confidence level δ, threshold function $\beta(t, \delta)$
Output: stopping time τ_δ, estimated best arm $\hat{a}_{\tau_\delta}$

Observe each arm once ; $t \leftarrow K$
while $Z(t) \leq \beta(t, \delta)$ **do**
  $\tilde{\boldsymbol{w}}(t) \leftarrow \boldsymbol{w}(\hat{\mu}(t))$
  if $U_t \triangleq \{a \in [K] \ : \ N_a(t) < \sqrt{t} - K/2\} \neq \varnothing$ **then**
    Choose $A_{t+1} \in \operatorname{argmin}_{a \in U_t} N_a(t)$; /* forced exploration */
  else
    Choose $A_{t+1} \in \operatorname{argmin}_{a \in [K]} N_a(t) - \sum_{s=K}^{t-1} \tilde{w}_a(s)$
  Observe $Y_{A_{t+1}}$ and increase t by 1
$\tau_\delta \leftarrow t$; $\hat{a}_{\tau_\delta} \leftarrow \operatorname{argmax}_{a \in [K]} \hat{\mu}_a(t)$

Théorème (Asymptotique optimalité de Track-and-Stop)

Il existe un threshold $\beta(t,\delta)$ tel que Track-and-Stop est δ-correct et satisfait :

✓ $$\limsup_{\delta \to 0} \frac{\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta]}{\log(1/\delta)} = T(\boldsymbol{\mu})$$

Théorème (Asymptotique optimalité de Track-and-Stop)

Il existe un threshold $\beta(t,\delta)$ tel que Track-and-Stop est δ-correct et satisfait :

$$\checkmark \quad \limsup_{\delta \to 0} \frac{\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta]}{\log(1/\delta)} = T(\boldsymbol{\mu})$$

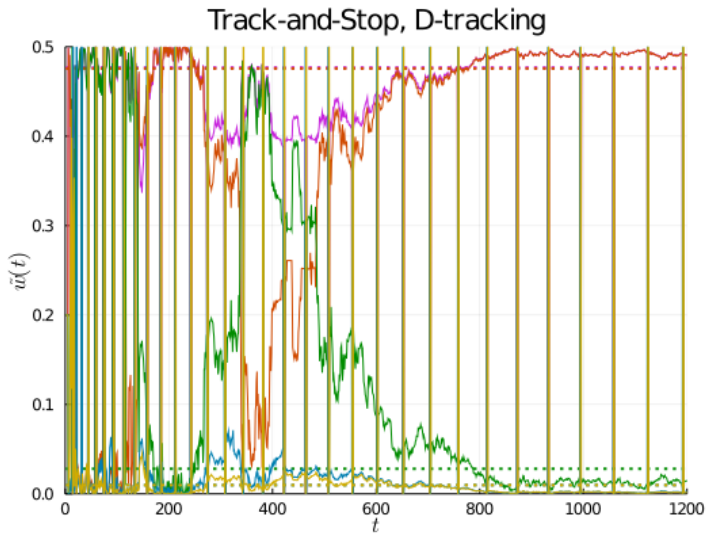


Figure – Évolution de $\tilde{\boldsymbol{w}}(t)$ sur l'exécution de Track-and-Stop avec $\delta = 0.01$ et $\boldsymbol{\mu} = (0.9, 0.8, 0.6, 0.4, 0.4)$.

✗ le tracking de $\boldsymbol{w}(\hat{\mu}(t))$ est assez hasardeux : trop instable lors des premières étapes $\rightarrow$ mauvaises estimées

✗ peut mener à un sous-échantillonnage de certains bras sans **exploration forcée** à un taux arbitraire ($\sqrt{t}$ ici)

Comment obtenir des garanties non-asymptotiques ?

Dans la suite on travaillera uniquement avec des variables gaussiennes réduites $\mathcal{N}(.,1)$. *Dans ce cas on a*

$$T(\boldsymbol{\mu})^{-1} = \sup_{\boldsymbol{v}\in\Delta_K} \inf_{\boldsymbol{\lambda}\in \mathrm{Alt}(\boldsymbol{\mu})} \sum_{a=1}^{K} v_a \frac{(\mu_a - \lambda_a)^2}{2}$$

On notera $\Delta_a = \mu_1 - \mu_a$ le gap du bras a

[BGK22] Barrier, A., Garivier, A. and Kocák, T. (2022), **A non-Asymptotic Approach to Best-Arm Identification for Gaussian Bandits**, In *25th International Conference on Artificial Intelligence and Statistics (AIStats)*

Vers des bornes non-asymptotiques

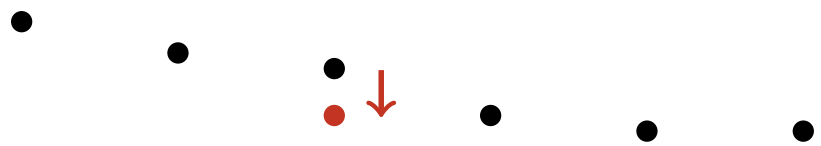
Idée Modifier le vecteur de poids tracké en le rendant plus stable : on introduit une **région de confiance** $\mathcal{CR}$ pour μ autour de $\hat{\mu}(t)$ et on track le vecteur de poids associé à un bandit $\tilde{\mu} \in \mathcal{CR}$ qui **maximises l'exploration** en satisfaisant

$$\min_{1 \leq a \leq K} w_a(\tilde{\mu}) = \max_{\nu \in \mathcal{CR}} \min_{1 \leq a \leq K} w_a(\nu) \, .$$

Vers des bornes non-asymptotiques

Idée Modifier le vecteur de poids tracké en le rendant plus stable : on introduit une **région de confiance** $\mathcal{CR}$ pour μ autour de $\hat{\mu}(t)$ et on track le vecteur de poids associé à un bandit $\tilde{\mu} \in \mathcal{CR}$ qui **maximises l'exploration** en satisfaisant

$$\min_{1 \leq a \leq K} w_a(\tilde{\mu}) = \max_{\nu \in \mathcal{CR}} \min_{1 \leq a \leq K} w_a(\nu) \, .$$



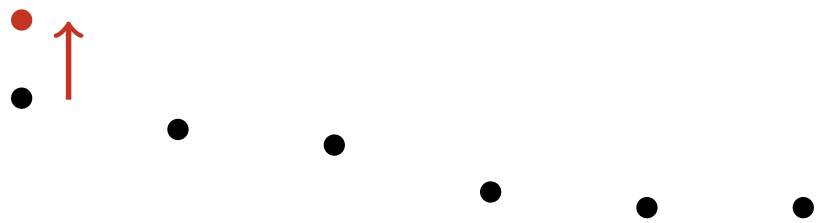
Lemme (Descente d'un bras sous-optimal)

Soient ν, ν' deux problèmes de bandits de meilleur bras 1. Supposons que $\Delta'_b > \Delta_b$ pour un $b \neq 1$ pendant que $\Delta'_a = \Delta_a$ pour $a \neq b$.
Alors $w'_b < w_b$, $w'_a > w_a$ pour $a \notin \{1, b\}$ et $T' < T$.

Vers des bornes non-asymptotiques

Idée Modifier le vecteur de poids tracké en le rendant plus stable : on introduit une **région de confiance $\mathcal{CR}$** pour μ autour de $\hat{\mu}(t)$ et on track le vecteur de poids associé à un bandit $\tilde{\mu} \in \mathcal{CR}$ qui **maximises l'exploration** en satisfaisant

$$\min_{1 \leq a \leq K} w_a(\tilde{\mu}) = \max_{\nu \in \mathcal{CR}} \min_{1 \leq a \leq K} w_a(\nu) \, .$$



Lemme (Montée du bras optimal)

Soient ν, ν' deux problèmes de bandits de meilleur bras 1. Supposons que $\Delta'_a = \Delta_a + d$ pour $a \neq 1$ et un $d > 0$.
Alors $w'_{\min} \geq w_{\min}$, avec inégalité stricte dès que $\Delta_a \neq \Delta_b$ pour un $a, b \neq 1$.

Vers des bornes non-asymptotiques

Idée Modifier le vecteur de poids tracké en le rendant plus stable : on introduit une **région de confiance** $\mathcal{CR}$ pour μ autour de $\hat{\mu}(t)$ et on track le vecteur de poids associé à un bandit $\tilde{\mu} \in \mathcal{CR}$ qui **maximises l'exploration** en satisfaisant

$$\min_{1\leq a\leq K} w_a(\tilde{\mu}) = \max_{\nu\in\mathcal{CR}} \min_{1\leq a\leq K} w_a(\nu)\,.$$

Lemme (Montée des pires bras)

Soient ν, ν' deux problèmes de bandits de meilleur bras 1. Soit $B = \mathrm{argmin}_{1\leq a\leq K} \mu_a$ (resp. $B' = \mathrm{argmin}_{1\leq a\leq K} \mu'_a$) l'ensemble des pires bras de μ (resp. μ') et supposons que $B \subset B'$ et $\Delta'_{\max} < \Delta_{\max}$, pendant que $\Delta'_a = \Delta_a$ pour $a \notin B'$.
Alors $w'_{\min} \geq w_{\min}$.

Vers des bornes non-asymptotiques

Idée Modifier le vecteur de poids tracké en le rendant plus stable : on introduit une **région de confiance $\mathcal{CR}$** pour μ autour de $\hat{\mu}(t)$ et on track le vecteur de poids associé à un bandit $\tilde{\mu} \in \mathcal{CR}$ qui **maximises l'exploration** en satisfaisant

$$\min_{1 \leq a \leq K} w_a(\tilde{\mu}) = \max_{\nu \in \mathcal{CR}} \min_{1 \leq a \leq K} w_a(\nu) \, .$$

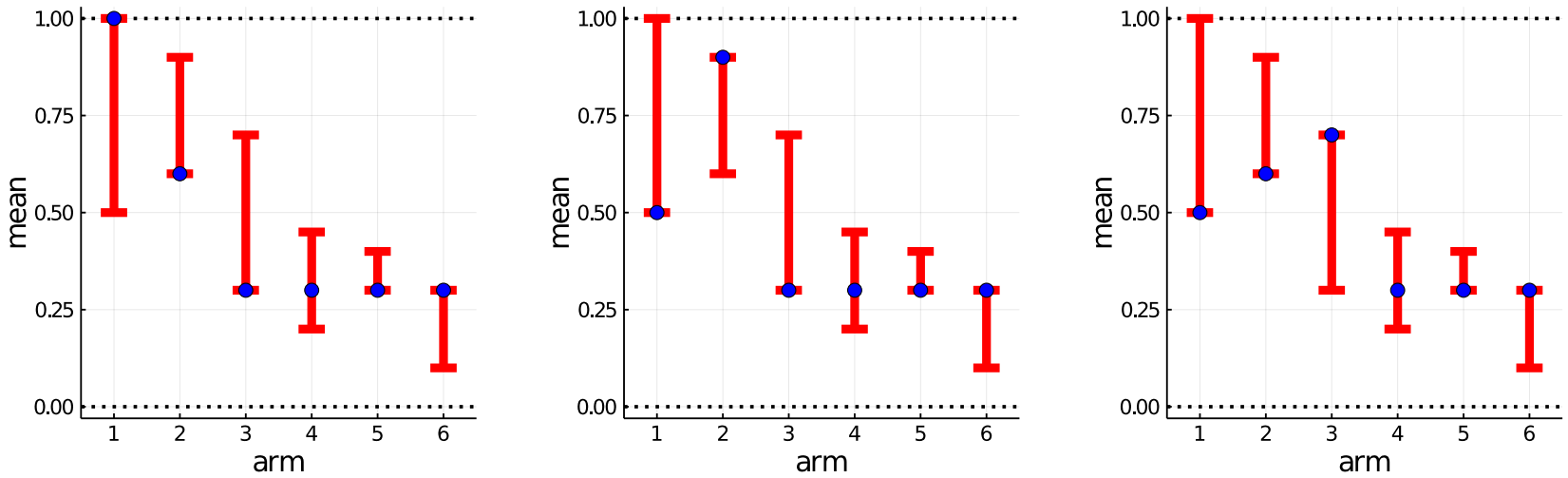


Figure – En utilisant les lemmes, on peut restreindre notre recherche à seulement quelques candidats (un par meilleur bras possible), pour calculer $\tilde{\mu}$

$\rightarrow$ le bandit $\tilde{\mu}$ est calculable

Idée **Utiliser la nouvelle stratégie d'échantillonnage** en calculant une région de confiance $\mathcal{CR}_{\boldsymbol{\mu}}(t)$ pour $\boldsymbol{\mu}$ à chaque tour

---

Algorithm 2: Exploration-Biased Sampling

Input: confidence level δ, threshold function $\beta(t,\delta)$, confidence parameter γ
Output: stopping time τ_δ, estimated best arm $\hat{a}_{\tau_\delta}$

Observe each arm once ; $t \leftarrow K$
while $Z(t) \leq \beta(t,\delta)$ **do**

$\mathcal{CR}_{\boldsymbol{\mu}}(t) \leftarrow \prod_{a\in[K]} \left[\hat{\mu}_a(t) \pm 2\sqrt{\frac{\log(4N_a(t)K/\gamma)}{N_a(t)}}\right]$

$\tilde{\boldsymbol{\mu}} \leftarrow$ optimistic bandit associated to $\mathcal{CR}_{\boldsymbol{\mu}}(t)$

$\tilde{\boldsymbol{w}}(t) \leftarrow \boldsymbol{w}(\tilde{\boldsymbol{\mu}})$

Choose $A_{t+1} \in \operatorname{argmin}_{a\in[K]} N_a(t) - \sum_{s=K}^{t-1} \tilde{w}_a(s)$

Observe $Y_{A_{t+1}}$ and increase t by 1

$\tau_\delta \leftarrow t$; $\hat{a}_{\tau_\delta} \leftarrow \operatorname{argmax}_{a\in[K]} \hat{\mu}_a(t)$

- ✓ asymptotiquement optimal
- ✓ exploration naturelle : pas besoin de forcer l'exploration !

✓ **borne non-asymptotique** avec forte probabilité :

Theorem (Borne non-asymptotique pour Exploration-Biased Sampling)

Soit $\gamma \in (0,1), \eta \in (0,1]$. Il existe un évènement $\mathcal{E}$ de probabilité au moins $1-\gamma$ et $\delta_0 > 0$ tel que pour tout $0 < \delta \leq \delta_0$, Exploration-Biased Sampling satisfait

$$\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta \mathbb{1}_{\mathcal{E}}] \leq (1+\eta)\, T(\boldsymbol{\mu}) \log(1/\delta) + o_{\delta \to 0}(1)$$

(avec une formule explicité pour δ_0 et $o_{\delta \to 0}(1)$)

✗ la convergence de $\tilde{\boldsymbol{w}}(t)$ vers $\boldsymbol{w}(\boldsymbol{\mu})$ est moins rapide que pour Track-and-Stop

Régularité du problème du problème de complexité d'échantillonnage (1)

Theorem

Soient $\boldsymbol{\mu}, \boldsymbol{\mu}'$ *de meilleur bras* 1, *et supposons que* $(1-\varepsilon)\Delta_a^2 \leq {\Delta'_a}^2 \leq (1+\varepsilon)\Delta_a^2$ *pour* $a \neq 1$ *et* $\varepsilon \in [0, 1/7]$. *Alors*

$$(1-3\varepsilon)T(\boldsymbol{\mu}) \leq T(\boldsymbol{\mu}') \leq (1+6\varepsilon)T(\boldsymbol{\mu})$$

$$\text{et} \quad \forall 1 \leq a \leq K, \quad (1-10\varepsilon)w_a(\boldsymbol{\mu}) \leq w_a(\boldsymbol{\mu}') \leq (1+10\varepsilon)w_a(\boldsymbol{\mu})\,.$$

Expériences numériques : amélioration de la stabilité

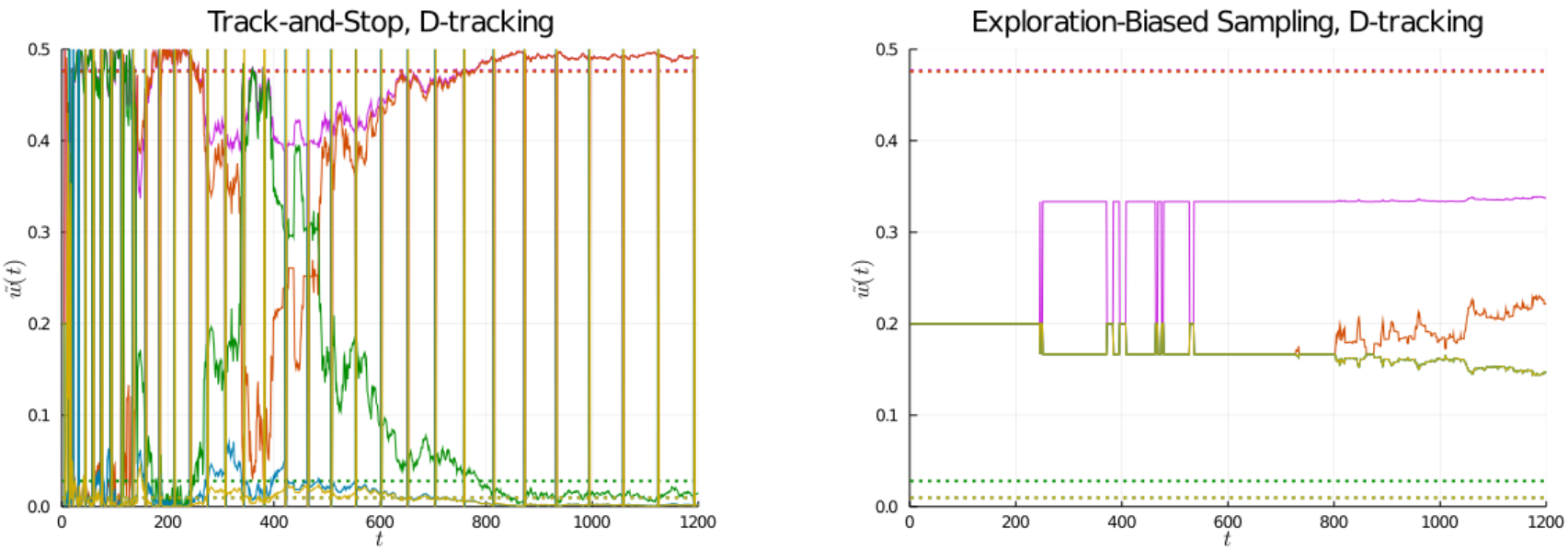


Figure – Évolution de $\tilde{\boldsymbol{w}}(t)$ sur l'exécution des deux stratégies avec $\delta = 0.01$, $\gamma = 0.2$ et $\boldsymbol{\mu} = (0.9, 0.8, 0.6, 0.4, 0.4)$.

Track-and-Stop

- ✗ instabilité des poids : les poids rouge et vert fluctuent beaucoup (mauvaises estimées initiales, alors qu'intuitivement on devrait explorer uniformément au début)
- ✗ les mauvais bras sont sous-échantilonnés sans exploration forcée (pics bleus et verts)

Expériences numériques : amélioration de la stabilité

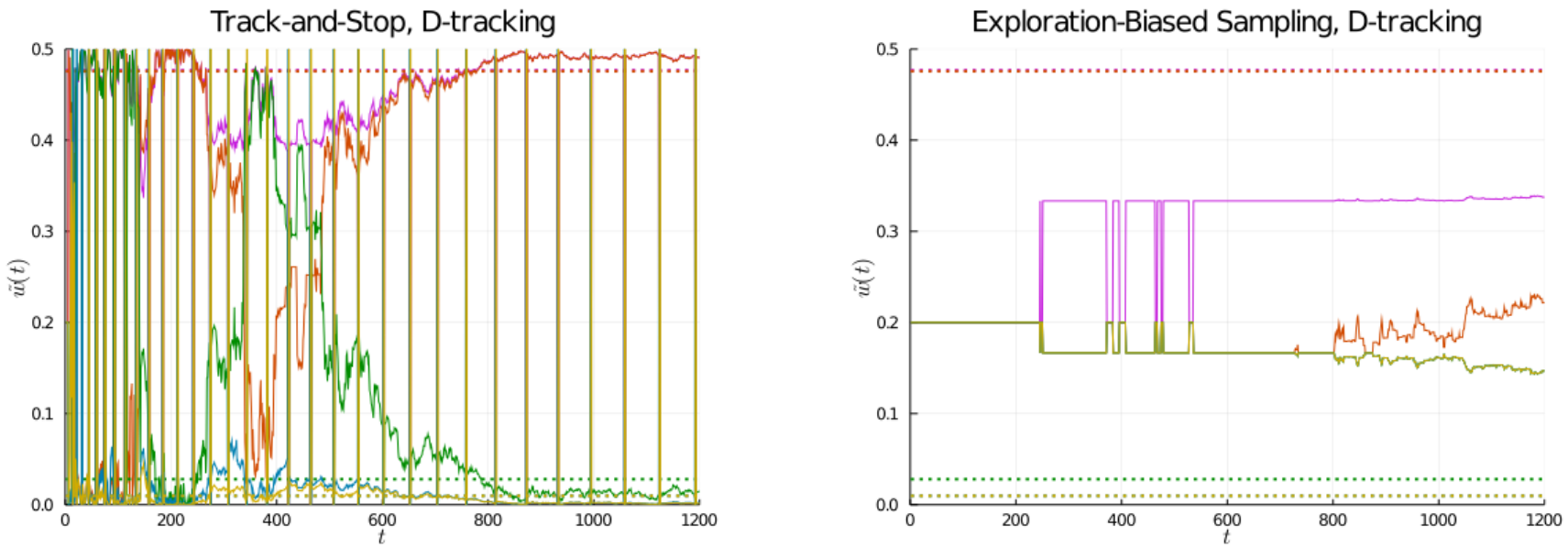


Figure – Évolution de $\tilde{\boldsymbol{w}}(t)$ sur l'exécution des deux stratégies avec $\delta = 0.01$, $\gamma = 0.2$ et $\boldsymbol{\mu} = (0.9, 0.8, 0.6, 0.4, 0.4)$.

Exploration-Biased Sampling

- ✓ exploration uniforme pendant les premiers tours
- ✓ stabilité de la stratégie de tracking
- ✓ séparation prudente des poids quand une nette distinction des estimées apparaît.

1. Introduction aux problèmes de bandits

2. BAI à confiance fixée
   - Borne inférieure
   - Track-and-Stop : un algorithme asymptotiquement optimal
   - Vers des bornes non-asymptotiques

3. BAI à budget fixé

BAI à budget fixé

Pour un bandit $\boldsymbol{\nu}$, on cherche une constante $H(\boldsymbol{\nu})$ la plus optimale possible telle que pour toute stratégie consistante on ait :

$$\mathbb{P}_{\boldsymbol{\nu}}(\hat{a}_T \neq 1) \gtrsim \exp\left(-\frac{T}{H(\boldsymbol{\nu})}\right)$$

si bien que l'on définit la complexité

$$\kappa(\boldsymbol{\nu}) = \inf_{\text{algo consistant}} \left(\limsup_{T\to+\infty} -\frac{1}{T}\log \mathbb{P}_{\boldsymbol{\nu}}(\hat{a}_T \neq 1)\right)^{-1}$$

qui assure effectivement que pour toute stratégie consistante :

$$\mathbb{P}_{\boldsymbol{\nu}}(I_T \notin a^*(\boldsymbol{\nu})) \geq \exp\left(-\frac{T}{\kappa(\boldsymbol{\nu})}(1+o(1))\right).$$

Borne inférieure

Théorème (Borne inférieure pour des Bernoulli)

Soit $\boldsymbol{\nu}$ un problème de bandits tel que $\nu_a = \mathcal{B}(\mu_a)$ pour $1 \leq a \leq K$ et $\mu_1, \dots, \mu_K \in [p, 1-p]$ pour un $p \in]0, 1/2[$ fixé. Alors pour toute stratégie, il existe une permutation $\sigma \in \mathfrak{S}_K$ telle que

$$\mathbb{P}_{\sigma(\boldsymbol{\nu})}\big(I_T \neq \sigma(1)\big) \geq \exp\left(-\frac{5T}{p(1-p)\max_{a\neq 1}\frac{a}{\Delta_{(a)}^2}}(1+o(1))\right).$$

Autrement dit $\kappa(\boldsymbol{\nu}) \geq \frac{p(1-p)}{5}\max_{a\neq 1}\frac{a}{\Delta_{(a)}^2}$.

- ✗ spécifique aux Bernoulli (adaptable aux gaussiennes)
- ✗ la preuve est très technique, on aimerait quelque chose de plus naturel

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

Borne supérieure Principalement un algorithme : l'algorithme de Rejet Successif : on divise le budget en $K-1$ phases durant lesquelles les bras sont tirés uniformément puis le moins bon bras empirique est éliminé

Théorème (Borne supérieure pour le Rejet Successif)

On regarde les distributions sur $[0,1]$. *Soit* ν *d'unique meilleur bras* 1. *L'algorithme de Rejet Successif satisfait :*

$$\mathbb{P}_\nu(\hat{a}_T \neq 1) \leq K \exp\left(-\frac{T-K}{\overline{\log}(K)} \frac{1}{\max_{2\leq k\leq K} \frac{k}{\Delta^2_{(k)}}}\right)$$

où $\overline{\log}(K) \sim_{K\to+\infty} \log(K)$.
Autrement dit, $\kappa(\nu) \leq \overline{\log}(K) \max_{2\leq a\leq K} \frac{a}{\Delta^2_{(a)}}$.

✗ bornes supérieure et inférieure diffèrent d'un facteur $\log(K)$

Peut-on obtenir des bornes plus informatives ?
Peut-on faire matcher bornes inférieure et supérieure ?

On suppose de nouveau que les distributions des bras appartiennent à un même modèle exponentiel.

On peut faire plusieurs hypothèses sur la stratégie que l'on considère :

- elle identifie asymptotiquement le meilleur bras : $\mathbb{P}_{\boldsymbol{\mu}}(\hat{a}_T \neq 1) = o(1)$ pour tout problème de bandit $\boldsymbol{\mu}$
- elle est symétrique : elle ne dépend pas de l'indice des bras
- elle est monotone : plus un bras est de moyenne élevée, plus il est tiré en moyenne

Théorème (Borne inférieure pour un algorithme monotone)

Soit $\boldsymbol{\nu}$ un problème de bandits. Alors pour toute stratégie symétrique, identifiant asymptotiquement le meilleur bras et monotone on a :

$$\mathbb{P}_{\boldsymbol{\mu}}(\hat{a}_T \neq 1) \geq \exp\left(-\frac{2T}{\frac{a}{d(\mu_{(a)},\mu_1)}}(1+o(1))\right)$$

pour une constante C.
Autrement dit, $\kappa(\boldsymbol{\nu}) \geq \frac{1}{2}\max_{2\leq a\leq K} \frac{a}{d(\mu_{(a)},\mu_1)}$.

- ✓ preuve très simple généralisant la borne de [ABM10]
- ✗ l'hypothèse de monotonicité est discutable $\rightarrow$ mais l'enlever complexifie la preuve !

Références

Jean-Yves Audibert, Sébastien Bubeck, and Rémi Munos.
Best Arm Identification in Multi-Armed Bandits.
In *Proceedings of the 23rd Annual Conference on Learning Theory*, January 2010.

Antoine Barrier, Aurélien Garivier, and Tomáš Kocák.
A Non-asymptotic Approach to Best-Arm Identification for Gaussian Bandits.
International Conference on Artificial Intelligence and Statistics, March 2022.

Aurélien Garivier and Emilie Kaufmann.
Optimal Best Arm Identification with Fixed Confidence.
In Vitaly Feldman, Alexander Rakhlin, and Ohad Shamir, editors, *Conference on Learning Theory*, volume 49, pages 998–1027. PMLR, June 2016.