Une introduction à l'allocation séquentielle de ressources

Emilie Kaufmann

3ème journées YSP, IHP, 30 janvier 2015

Plan

Plan

Allocation séquentielle de ressources : des exemples

Essais cliniques

- K traitements possibles (d'effet inconnu)






- Quel traitement allouer à chaque patient en fonction des effets observés sur les patients précédents ?

Allocation séquentielle de ressources : des exemples

Essais cliniques

- K traitements possibles (d'effet inconnu)

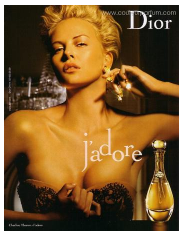


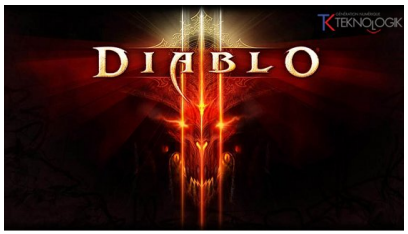



- Quel traitement allouer à chaque patient en fonction des effets observés sur les patients précédents ?

Publicité en ligne

- K publicités pouvant être affichées





- Quelle publicité montrer à chaque utilisateur en fonction des clics des utilisateurs précédents ?

Un cadre général : le modèle de bandit à plusieurs bras

- K options possibles
- option a : loi de probabilité ν_a de moyenne μ_a

A l'instant t, un agent

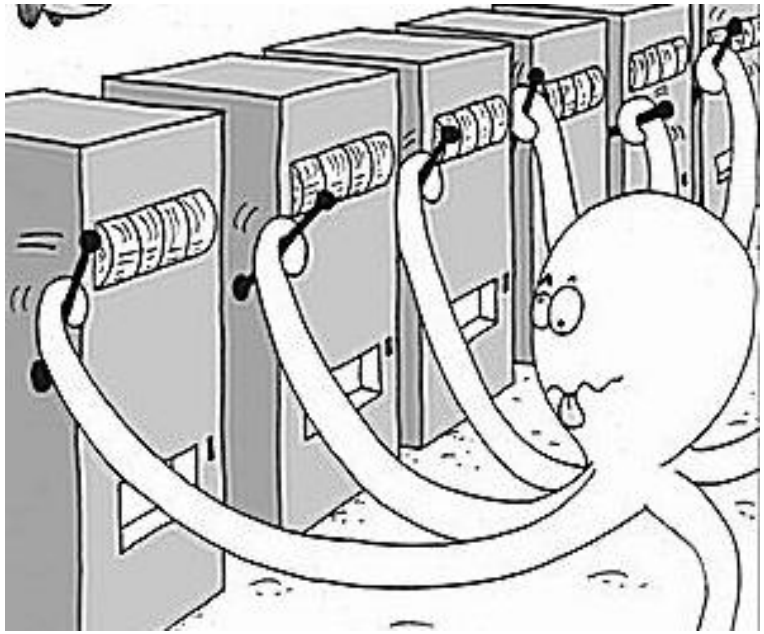
- choisit une option A_t
- observe un réalisation X_t de la loi associée ν_{A_t}

L'agent adopte une **stratégie séquentielle** (A_t), telle que

$$A_{t+1} = "F"(A_1, \dots, A_t, X_1, \dots, X_t)$$

Pourquoi "bandit" ?

Bandit manchot = machine à sous.



Si chaque bras a donne des récompenses tirées sous une loi ν_a,
quelle stratégie de tirage des bras faut-il adopter ?

Un cadre général : le modèle de bandit à plusieurs bras

- K **bras**
- **bras** a : loi de probabilité ν_a de moyenne μ_a

A l'instant t, un agent

- **tire un bras** A_t
- observe un réalisation X_t de la loi associée ν_{A_t}

L'agent adopte une stratégie séquentielle (A_t) (ou **algorithme de bandit**), telle que

$$A_{t+1} = "F"(A_1, \dots, A_t, X_1, \dots, X_t)$$

Une stratégie séquentielle : pour quel objectif ?

Objectif global : apprendre quels sont les meilleurs bras

$$\mu^* = \max_a \ \mu_a \qquad a^* = \text{argmax}_a \ \ \mu_a.$$

Maximiser ses récompenses	Identifier le meilleur bras
maximiser $\sum_{t=1}^{T} X_t$	proposer $\hat{a}^*$ tel que $\mathbb{P}(\hat{a}^* = a^*) \geq 1 - \delta$
Compromis exploration - exploitation	Exploration optimale

Plan

Des stratégies optimales

L'agent cherche à trouver une stratégie qui maximise

$$\mathbb{E}\left[\sum_{t=1}^{T} X_t\right]$$

ou de manière équivalente minimise le **regret** :

$$R_T = \mathbb{E}\left[T\mu^* - \sum_{t=1}^{T} X_t\right].$$

Des stratégies optimales

L'agent cherche à trouver une stratégie qui maximise

$$\mathbb{E}\left[\sum_{t=1}^{T} X_t\right]$$

ou de manière équivalente minimise le **regret** :

$$R_T = \mathbb{E}\left[T\mu^* - \sum_{t=1}^{T} X_t\right].$$

Une réécriture :

$$R_T = \sum_{a=1}^{K} (\mu^* - \mu_a)\mathbb{E}[N_a(T)]$$

où $N_a(T)$ est le nombre de tirages du bras a jusqu'à l'instant T.

Des stratégies optimales

$$R_T = \sum_{a=1}^{K} (\mu^* - \mu_a)\mathbb{E}[N_a(T)].$$

[Lai et Robbins 1985] : tout algorithme de bandit 'uniformément bon' doit tirer tous les bras une infinité de fois :

$$\mu_a < \mu^* \Rightarrow \liminf_{T\to\infty} \frac{\mathbb{E}[N_a(T)]}{\log T} \geq \frac{1}{\mathrm{KL}(\nu_a, \nu_{a*})}$$

Definition

Un algorithme de bandit est **asymptotiquement optimal** si

$$\mu_a < \mu^* \Rightarrow \limsup_{T\to\infty} \frac{\mathbb{E}[N_a(T)]}{\log T} \leq \frac{1}{\mathrm{KL}(\nu_a, \nu_{a^*})}$$

Des exemples de stratégies

- **Idée 1 :** Tirer chaque bras T/K fois

$\Rightarrow$ EXPLORATION

Des exemples de stratégies

- **Idée 1 :** Tirer chaque bras T/K fois

$\Rightarrow$ EXPLORATION

- **Idée 2 :**Toujours choisir le bras qui a donné les meilleures récompenses jusqu'ici

$$A_{t+1} = \operatorname*{argmax}_a \hat{\mu}_a(t)$$

$\Rightarrow$ EXPLOITATION

Des exemples de stratégies

- **Idée 1 :** Tirer chaque bras T/K fois

$\Rightarrow$ EXPLORATION

- **Idée 2 :**Toujours choisir le bras qui a donné les meilleures récompenses jusqu'ici

$$A_{t+1} = \operatorname*{argmax}_a \hat{\mu}_a(t)$$

$\Rightarrow$ EXPLOITATION

- **Idée 3 :** Tirer les bras uniformément pendant $T/2$ instants (EXPLORATION)
  Puis tirer le meilleur bras empirique jusqu'à la fin (EXPLOITATION)

$\Rightarrow$ EXPLORATION puis EXPLOITATION

Des algorithmes optimistes

- Pour chaque bras a, construire un intervalle de confiance sur la moyenne inconnue μ_a :

$$\mu_a \leq UCB_a(t) \quad w.h.p$$

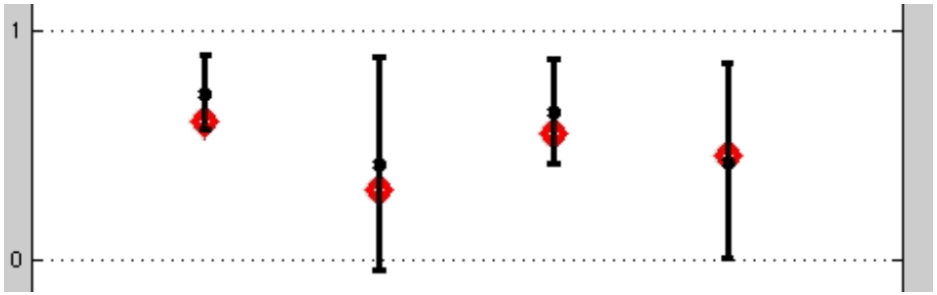


FIGURE: Intervalles de confiance sur les bras après t instants

Des algorithmes optimistes

- Utiliser le principe d'optimisme :

"agir comme si le meilleur des modèles possible était le vrai modèle"

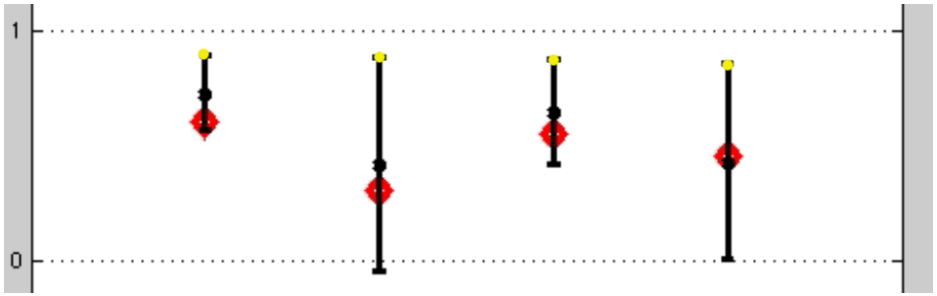


FIGURE: Intervalles de confiance sur les bras après t instants

- Ceci revient à choisir à l'instant $t+1$

$$A_{t+1} = \arg\max_a \ UCB_a(t)$$

L'algorithme UCB1

Hypothèse : pour tout a, ν_a est à support dans $[0,1]$.

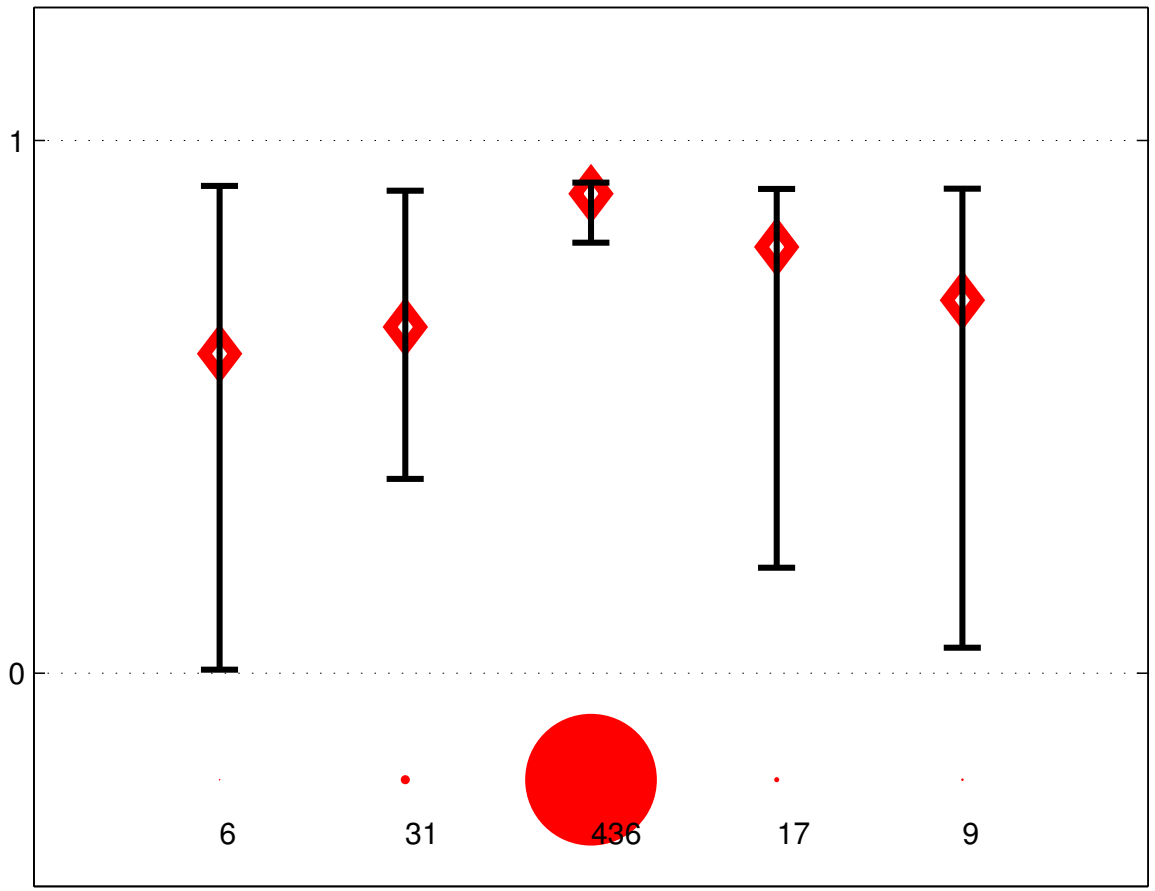
L'algorithme UCB1 inspiré par [Auer et al. 02] utilise l'indice :

$$\mathrm{UCB}_a(t) = \hat{\mu}_{a,N_a(t)} + \sqrt{\frac{\alpha \log(t)}{2N_a(t)}}$$

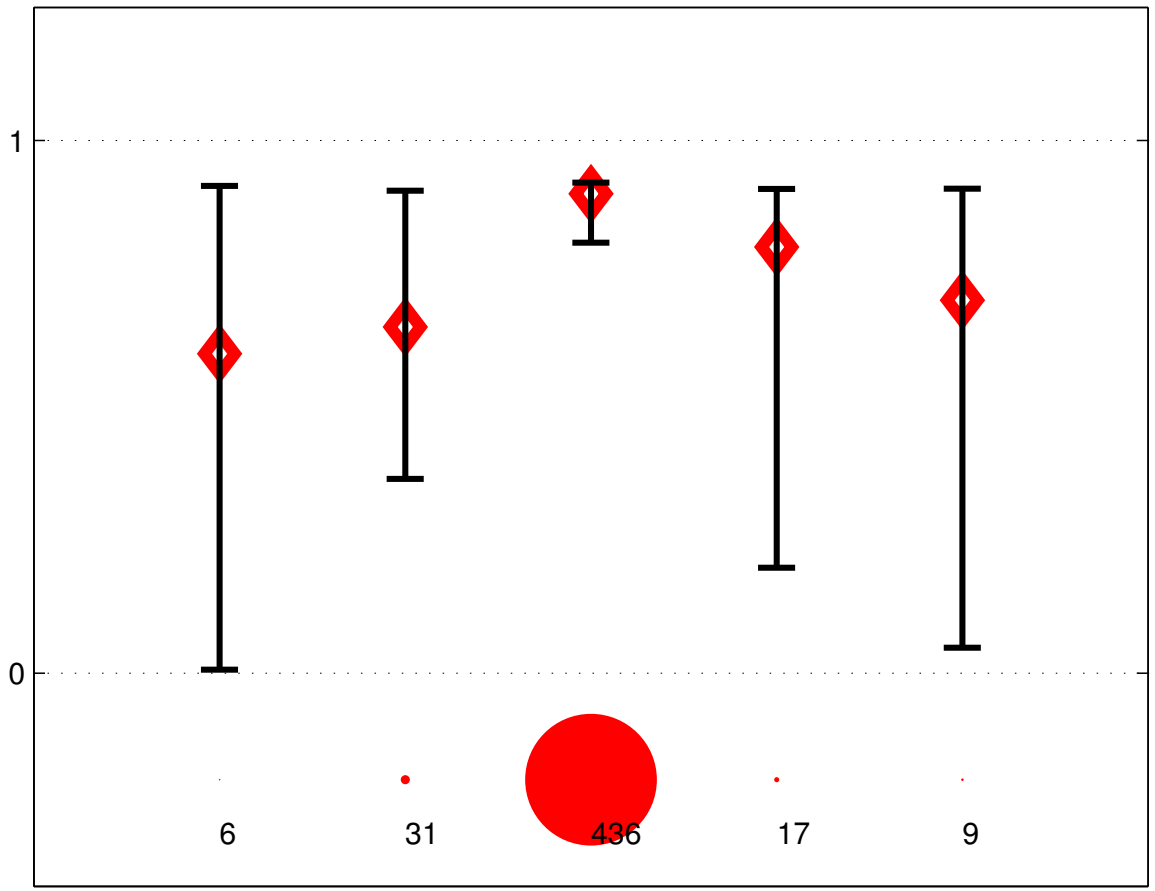
avec

- $N_a(t)$: nombre de tirages du bras a entre les instants 1 et t
- $\hat{\mu}_{a,s} = \frac{1}{s}\sum_{i=1}^{s} Y_{a,i}$ moyenne empirique des s premières observations du bras a

UCB1 en action



UCB1 en action



Construction des intervalles de confiance

$$\mathrm{UCB}_a(t) = \hat{\mu}_{a,N_a(t)} + \sqrt{\frac{\alpha \log(t)}{2N_a(t)}}$$

- L'inégalité de Hoeffding donne

$$\mathbb{P}\left(\hat{\mu}_{a,s} + \sqrt{\frac{\alpha \log(t)}{2s}} \leq \mu_a\right) \leq \exp\left(-2s\left(\frac{\alpha \log(t)}{2s}\right)\right) = \frac{1}{t^\alpha}.$$

- Il reste à gérer le *nombre aléatoire d'observations*

Construction des intervalles de confiance

$$\mathrm{UCB}_a(t) = \hat{\mu}_{a,N_a(t)} + \sqrt{\frac{\alpha \log(t)}{2N_a(t)}}$$

- L'inégalité de Hoeffding donne

$$\mathbb{P}\left(\hat{\mu}_{a,s} + \sqrt{\frac{\alpha \log(t)}{2s}} \leq \mu_a\right) \leq \exp\left(-2s\left(\frac{\alpha \log(t)}{2s}\right)\right) = \frac{1}{t^\alpha}.$$

- Il reste à gérer le *nombre aléatoire d'observations*

$\Rightarrow$ Borne de l'union ?

$$\begin{aligned}\mathbb{P}(\mathrm{UCB}_a(t) \leq \mu_a) &\leq \mathbb{P}\left(\exists s \in \{1, t\} : \hat{\mu}_{a,s} + \sqrt{\frac{\alpha \log(t)}{2s}} \leq \mu_a\right) \\ &\leq \sum_{s=1}^{t} \frac{1}{t^\alpha} \leq \frac{1}{t^{\alpha-1}}\end{aligned}$$

Construction des intervalles de confiance

$$\mathrm{UCB}_a(t) = \hat{\mu}_{a,N_a(t)} + \sqrt{\frac{\alpha \log(t)}{2N_a(t)}}$$

- L'inégalité de Hoeffding donne

$$\mathbb{P}\left(\hat{\mu}_{a,s} + \sqrt{\frac{\alpha \log(t)}{2s}} \leq \mu_a\right) \leq \exp\left(-2s\left(\frac{\alpha \log(t)}{2s}\right)\right) = \frac{1}{t^\alpha}.$$

- Il reste à gérer le *nombre aléatoire d'observations*

$\Rightarrow$ On peut faire mieux : argument de 'peeling'

$$\begin{aligned}\mathbb{P}(\mathrm{UCB}_a(t) \leq \mu_a) &\leq \mathbb{P}\left(\exists s \in \{1,t\} : \hat{\mu}_{a,s} + \sqrt{\frac{\alpha \log(t)}{2s}} \leq \mu_a\right) \\ &\leq e\alpha \frac{\log(t)^2}{t^\alpha}\end{aligned}$$

Résultat théorique

Théorème

Pour tout $\alpha > 1$ et tout bras sous-optimal a, il existe une constante $C_\alpha > 0$ telle que

$$\mathbb{E}[N_a(T)] \leq \frac{2\alpha}{(\mu^* - \mu_a)^2} \log(T) + C_\alpha.$$

Preuve : 1/3

Notation : $a^* = 1$ et contrôlons $N_2(T)$ pour $\mu_2 < \mu_1$.

$$
\begin{aligned}
N_2(T) &= \sum_{t=0}^{T-1} \mathbb{1}_{(A_{t+1}=2)} \\
&= \sum_{t=0}^{T-1} \mathbb{1}_{(A_{t+1}=2)\cap(\mathrm{UCB}_1(t)\leq\mu_1)} + \sum_{t=0}^{T-1} \mathbb{1}_{(A_{t+1}=2)\cap(\mathrm{UCB}_1(t)>\mu_1)} \\
&\leq \sum_{t=0}^{T-1} \mathbb{1}_{(\mathrm{UCB}_1(t)\leq\mu_1)} + \sum_{t=0}^{T-1} \mathbb{1}_{(A_{t+1}=2)\cap(\mathrm{UCB}_2(t)>\mu_1)}
\end{aligned}
$$

Preuve : 1/3

Notation : $a^* = 1$ et contrôlons $N_2(T)$ pour $\mu_2 < \mu_1$.

$$
\begin{aligned}
N_2(T) &= \sum_{t=0}^{T-1} \mathbb{1}_{(A_{t+1}=2)} \\
&= \sum_{t=0}^{T-1} \mathbb{1}_{(A_{t+1}=2)\cap(\mathrm{UCB}_1(t)\leq\mu_1)} + \sum_{t=0}^{T-1} \mathbb{1}_{(A_{t+1}=2)\cap(\mathrm{UCB}_1(t)>\mu_1)} \\
&\leq \sum_{t=0}^{T-1} \mathbb{1}_{(\mathrm{UCB}_1(t)\leq\mu_1)} + \sum_{t=0}^{T-1} \mathbb{1}_{(A_{t+1}=2)\cap(\mathrm{UCB}_2(t)>\mu_1)}
\end{aligned}
$$

$$
\mathbb{E}[N_2(T)] \leq \underbrace{\sum_{t=0}^{T-1} \mathbb{P}(\mathrm{UCB}_1(t) \leq \mu_1)}_{A} + \underbrace{\sum_{t=0}^{T-1} \mathbb{P}(A_{t+1} = 2, \mathrm{UCB}_2(t) > \mu_1)}_{B}
$$

Preuve : 2/3

$$\mathbb{E}[N_2(T)] \leq \underbrace{\sum_{t=0}^{T-1} \mathbb{P}(\mathrm{UCB}_1(t) \leq \mu_1)}_{A} + \underbrace{\sum_{t=0}^{T-1} \mathbb{P}(A_{t+1} = 2, \mathrm{UCB}_2(t) > \mu_1)}_{B}$$

- Contrôle du terme A

$$\begin{aligned}
\sum_{t=0}^{T-1} \mathbb{P}(\mathrm{UCB}_1(t) \leq \mu_1) &\leq 1 + e\alpha \sum_{t=1}^{T-1} \frac{\log(t)^2}{t^\alpha} \\
&\leq 1 + e\alpha \sum_{t=1}^{\infty} \frac{\log(t)^2}{t^\alpha} := C_\alpha/2.
\end{aligned}$$

Preuve : 3/3

- Contrôle du terme B

$$(B) \quad \leq \quad \sum_{t=0}^{T-1} \mathbb{P}(A_{t+1} = 2, \mathrm{UCB}_2(t) > \mu_1, \mathrm{LCB}_2(t) \leq \mu_2) + C_\alpha/2$$

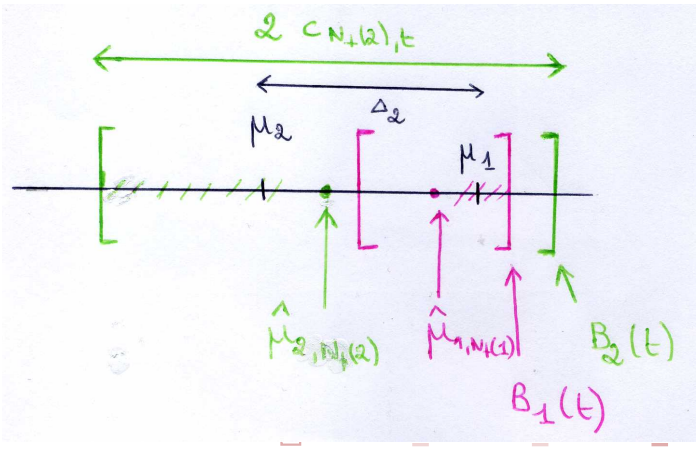
avec

$$\mathrm{LCB}_2(t) = \hat{\mu}_{2,N_2(t)} - \sqrt{\frac{\alpha \log t}{2N_2(t)}}.$$

$$(\mathrm{LCB}_2(t) < \mu_2 < \mu_1 \leq \mathrm{UCB}_2(t))$$

$$\Rightarrow \quad (\mu_1 - \mu_2) \leq 2\sqrt{\frac{\alpha \log(T)}{2N_2(t)}}$$

$$\Rightarrow \quad N_2(t) \leq \frac{2\alpha}{(\mu_1 - \mu_2)^2} \log(T)$$



Preuve : 3/3

- Contrôle du terme B

$$
\begin{aligned}
(B) &\leq \sum_{t=0}^{T-1} \mathbb{P}(A_{t+1} = 2, \mathrm{UCB}_2(t) > \mu_1, \mathrm{LCB}_2(t) \leq \mu_2) + C_\alpha/2 \\
&\leq \sum_{t=0}^{T-1} \mathbb{P}\left(A_{t+1} = 2, N_2(t) \leq \frac{2\alpha}{(\mu_1 - \mu_2)^2} \log(T)\right) + C_\alpha/2 \\
&\leq \frac{2\alpha}{(\mu_1 - \mu_2)^2} \log(T) + C_\alpha/2
\end{aligned}
$$

- Conclusion

$$
\mathbb{E}[N_2(T)] \leq \frac{2\alpha}{(\mu_1 - \mu_2)^2} \log(T) + C_\alpha.
$$

UCB1 est-il optimal ?

Théorème

Pour tout $\alpha > 1$ et tout bras sous-optimal a, il existe une constante $C_\alpha > 0$ telle que

$$\mathbb{E}[N_a(T)] \leq \frac{2\alpha}{(\mu^* - \mu_a)^2} \log(T) + C_\alpha.$$

UCB1 est-il optimal ?

Théorème

Pour tout $\alpha > 1$ et tout bras sous-optimal a, il existe une constante $C_\alpha > 0$ telle que

$$\mathbb{E}[N_a(T)] \leq \frac{2\alpha}{(\mu^* - \mu_a)^2} \log(T) + C_\alpha.$$

Pour des modèles où les récompenses sont binaires (Bernoulli)

$$\frac{2\alpha}{(\mu^* - \mu_a)^2} > \frac{4\alpha}{\mathrm{KL}(\nu_a, \nu^*)}$$

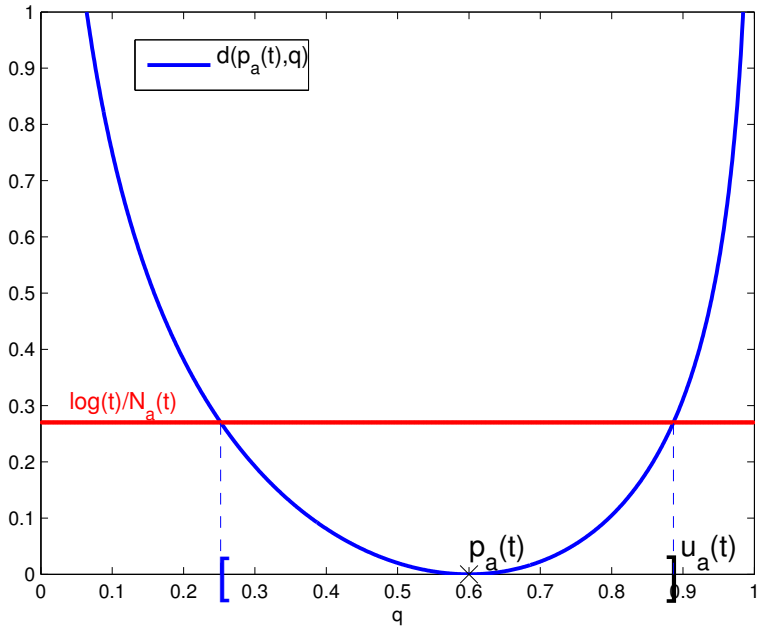
donc l'algorithme n'est pas asymptotiquement optimal...

L'algorithme KL-UCB : un algorithme optimal

- KL-UCB [Cappé et al. 13] utilise

$$\text{UCB}_a(t) = \underset{q > \hat{\mu}_{a,N_a(t)}}{\text{argmax}} \left\{ d\left(\hat{\mu}_{a,N_a(t)}, q\right) \leq \frac{\alpha \log(t)}{N_a(t)} \right\}$$

avec $d(p,q) = \text{KL}\left(\mathcal{B}(p), \mathcal{B}(q)\right)$.



On peut montrer (inégalité de Chernoff) que

$$\mathbb{P}(\text{UCB}_a(t) \leq \mu_a) \leq e\alpha \frac{\log(t)^2}{t^\alpha}$$

Pour les bandits binaires, $\mathbb{E}[N_a(T)] \leq \frac{\alpha}{\text{KL}(\nu_a, \nu^*)} \times \log T + C_\alpha$

Plan

Identification du meilleur bras : le cadre

Bras ordonnés tels que $\mu_1 > \mu_2 \geq \cdots \geq \mu_K$

Paramètres :

- $\delta \in]0,1[$ un paramètre de risque
- $a^* = 1$ le bras optimal

La stratégie de l'agent consiste en :

- une **règle d'échantillonnage** : bras A_t choisi à l'instant t
- une **règle d'arrêt** : à l'instant τ, il arrête l'échantillonnage
- une **règle de recommendation**, indiquant le bras choisi

$$\hat{a}^* = \underset{a=1\ldots K}{\text{argmax}}\ \hat{\mu}_{a,N_a(\tau)}$$

Son objectif :

- $\mathbb{P}(\hat{a}^* = a^*) \geq 1-\delta$ (algorithme δ-PAC)
- La nombre moyen d'échantillons $\mathbb{E}[\tau]$ est faible

L'algorithme LUCB

L'algorithme utilise un intervalle de confiance $\mathcal{I}_a(t)$ sur μ_a :

$$\mathcal{I}_a(t) = [L_a(t), U_a(t)]$$

$L_a(t)$= **L**ower Confidence Bound
$U_a(t)$= **U**pper **C**onfidence **B**ound

- On utilisera

$$U_a(t) = \hat{\mu}_a(t) + \sqrt{\frac{\beta(t,\delta)}{2N_a(t)}}$$

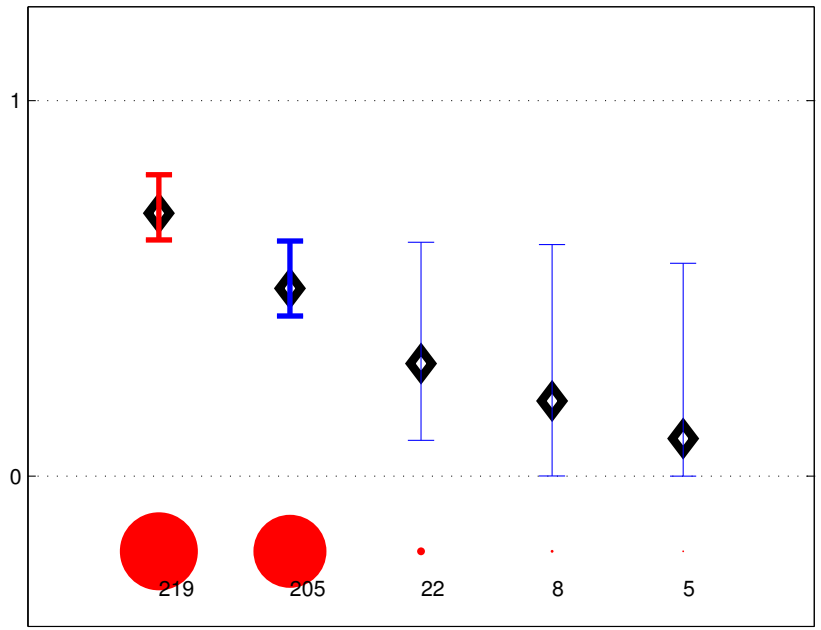
$$L_a(t) = \hat{\mu}_a(t) - \sqrt{\frac{\beta(t,\delta)}{2N_a(t)}}$$

où $\beta(t,\delta)$ est un taux d'exploration.

L'algorithme LUCB

A l'instant t, l'algorithme :

- tire deux bras bien choisis, u_t et l_t (en gras)
- s'arrête quand l'IC du bras optimal et ceux des bras sous-optimaux sont séparés

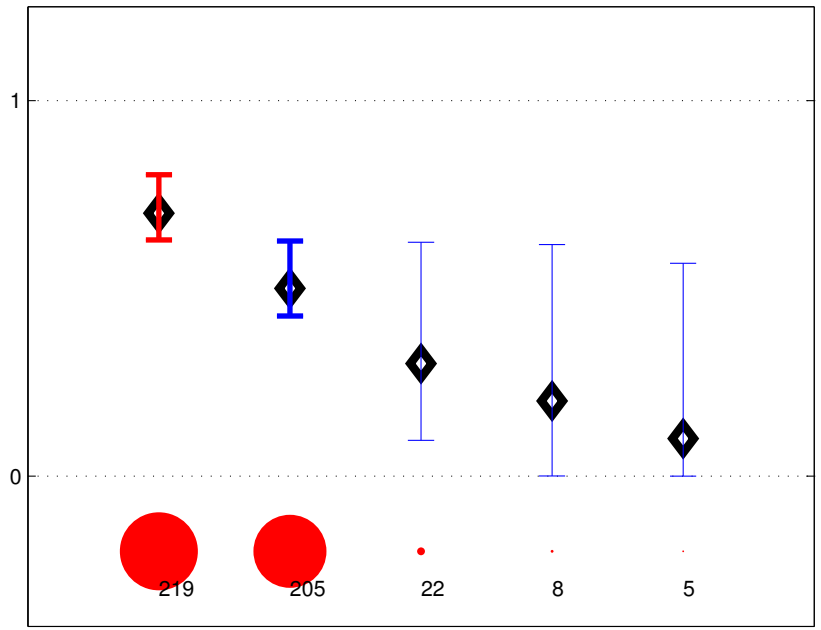


meilleur bras empirique, l_t bras sous-optimaux, u_t en gras

L'algorithme LUCB

A l'instant t, l'algorithme :

- tire deux bras bien choisis, u_t et l_t (en gras)
- s'arrête quand l'IC du bras optimal et ceux des bras sous-optimaux sont séparés



meilleur bras empirique, l_t bras sous-optimaux, u_t en gras

Propriétés théoriques de LUCB

Théorème [Kalyanakrishnan et al. 2012]

Avec le taux d'exploration

$$\beta(t,\delta) = \log\left(\frac{2Kt^2}{\delta}\right),$$

l'algorithme LUCB vérifie

$$\mathbb{P}\left(\hat{a}^* = a^*\right) \geq 1-\delta \quad \text{et} \quad \mathbb{E}[\tau] = O\left(H \log \frac{1}{\delta}\right),$$

où

$$H = \frac{1}{(\mu_1-\mu_2)^2} + \sum_{a=2}^{K} \frac{1}{(\mu_1-\mu_a)^2}.$$

Plan

Conclusion et perspectives

Bilan

L'utilisation d'intervalles de confiance est cruciale pour l'obtention de bons algorithmes de bandits pour la minimisation du regret ou l'identification du meilleur bras, mais ceux-ci ne sont pas utilisés de la même façon dans les deux cas

Pour aller plus loin :

- La notion de meilleur bras peut être modifiée : la moyenne est-elle toujours un bon critère ?
  [exposé d'Odalric]
- Peut-on incorporer de l'information sur les bras pour identifier le meilleur plus rapidement ?
  [exposé de Marta]