Outils et modèles statistiques pour l'allocation séquentielle de ressources

Emilie Kaufmann

Séminaire MathPark
IHP, 22 avril 2017

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?

Essais cliniques

- K traitements possibles (d'effet inconnu)

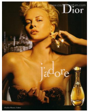


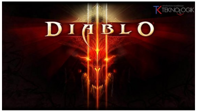



- Quel traitement allouer à chaque patient en fonction des effets observés sur les patients précédents?

Publicité en ligne

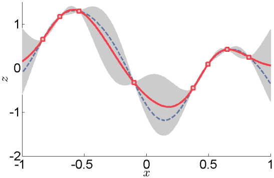
- K publicités pouvant être affichées





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

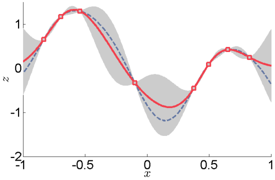
Expériences numériques:



- où effectuer la prochaine évaluation d'une fonction coûteuse ?

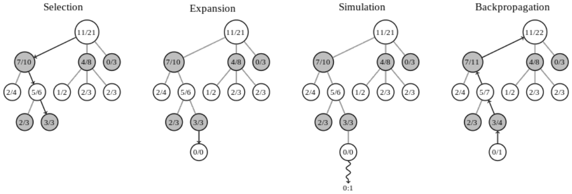
Allocation dynamique de la capacité de calcul

Expériences numériques:



- où effectuer la prochaine évaluation d'une fonction coûteuse ?

Intelligence artificielle pour les jeux:



- comment choisir la prochaine partie à simuler et à évaluer ?

Modélisation

Dans chaque exemple, un agent

- fait face à plusieurs options
- chaque option conduit à un résultat aléatoire

➜ introduction d'un **modèle probabiliste**

L'agent cherche à adopter

- une bonne stratégie de sélection des options

➜ définition mathématique d'un **objectif**

Plan

Plan

Le modèle de bandit à plusieurs bras

K **bras** (options) $= K$ lois de probabilités

$\mathcal{B}(\mu_1)$ $\mathcal{B}(\mu_2)$ $\mathcal{B}(\mu_3)$ $\mathcal{B}(\mu_4)$ $\mathcal{B}(\mu_5)$

A l'instant t, un agent

- sélectionne le bras A_t
- observe une réalisation $X_t \sim \mathcal{B}(\mu_{A_t})$

$$\mathbb{P}(X_t = 1 | A_t = a) = \mu_a \quad \text{et} \quad \mathbb{P}(X_t = 0 | A_t = a) = 1 - \mu_a.$$

La stratégie d'échantillonnage (A_t) est séquentielle :

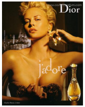
$$A_{t+1} = F_t(A_1, X_1, \ldots, A_t, X_t).$$

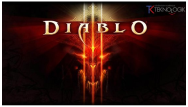
Un objectif: découvrir le meilleur bras $a^* = \mathrm{argmax}_a \ \mu_a$

...en maximisant ses récompenses!

Le modèle de bandit à plusieurs bras

K **bras** (options) $=$ K lois de probabilités





$\mathcal{B}(\mu_1)$ $\mathcal{B}(\mu_2)$ $\mathcal{B}(\mu_3)$ $\mathcal{B}(\mu_4)$ $\mathcal{B}(\mu_5)$

A l'instant t, un agent

- sélectionne le bras A_t
- observe une réalisation $X_t \sim \mathcal{B}(\mu_{A_t})$

$$\mathbb{P}(X_t = 1 | A_t = a) = \mu_a \quad \text{et} \quad \mathbb{P}(X_t = 0 | A_t = a) = 1 - \mu_a.$$

La stratégie d'échantillonnage (A_t) est séquentielle :

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

Un objectif: découvrir le meilleur bras $a^* = \text{argmax}_a \ \mu_a$
...en maximisant ses récompenses!

Le modèle de bandit à plusieurs bras

K **bras** (options) $=$ K lois de probabilités





$\mathcal{B}(\mu_1)$ $\mathcal{B}(\mu_2)$ $\mathcal{B}(\mu_3)$ $\mathcal{B}(\mu_4)$ $\mathcal{B}(\mu_5)$

A l'instant t, un agent

- sélectionne le bras A_t
- observe une réalisation $X_t \sim \mathcal{B}(\mu_{A_t})$

$$\mathbb{P}(X_t = 1 | A_t = a) = \mu_a \quad \text{et} \quad \mathbb{P}(X_t = 0 | A_t = a) = 1 - \mu_a.$$

La stratégie d'échantillonnage (A_t) est séquentielle :

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

Un objectif: découvrir le meilleur bras $a^* = \text{argmax}_a \, \mu_a$
...en maximisant ses récompenses!

Plan

Une mesure de performance: le regret

$$\mu^* = \max_a \ \mu_a \qquad a^* = \arg\max_a \ \mu_a$$

L'agent cherche une stratégie qui maximise, *en moyenne*,

$$X_1 + X_2 + \cdots + X_T.$$

Une mesure de performance: le regret

$$\mu^* = \max_a \ \mu_a \qquad a^* = \arg\max_a \ \mu_a$$

L'agent cherche une stratégie qui maximise

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

Il voudrait

$$\mathbb{E}\left[\sum_{t=1}^{T} X_t\right] \simeq \underbrace{T\mu^*}_{\substack{\text{gains moyen d'une stratégie} \\ \text{ne tirant que le bras } a^*}}$$

et cherche donc à minimiser le **regret** :

$$R_T = \underbrace{T\mu^*}_{\substack{\text{récompense cumulée d'une} \\ \text{stratégie oracle}}} - \underbrace{\mathbb{E}\left[\sum_{t=1}^{T} X_t\right]}_{\substack{\text{récompense cumulée} \\ \text{de la stratégie } (A_t)}}.$$

Une réécriture du regret

Pour simplifier les notations, on supposera dans la suite

$$\mu^* = \mu_1 > \mu_2 \geq \cdots \geq \mu_K.$$

Le regret peut se réécrire

$$R_T = \sum_{a=2}^{K} (\mu_1 - \mu_a) \times \mathbb{E}[N_a(T)]$$

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

Pour minimiser le regret:

- ➜ Tirer aussi peu que possible les bras sous-optimaux !
- ➜ Réaliser un compromis entre exploration et exploitation

Plan

Premiers exemples de stratégies

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

$\Rightarrow$ EXPLORATION

$$R(T) = \left(\frac{1}{K} \sum_{a=2}^{K} (\mu_1 - \mu_a) \right) T$$

Premiers exemples de stratégies

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

$\Rightarrow$ EXPLORATION

$$R(T) = \left(\frac{1}{K}\sum_{a=2}^{K}(\mu_1 - \mu_a)\right) T$$

- **Idée 2 :** Faire confiance au meilleur empirique jusqu'ici

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

où

$$\hat{\mu}_a(t) = \frac{\text{somme des récompenses issues du bras } a}{\text{nombre de sélections du bras } a}$$

est un estimateur de la moyenne inconnue μ_a.

$\Rightarrow$ EXPLOITATION

$$\mathbb{R}(T) \geq (1 - \mu_1) \times \mu_2 \times (\mu_1 - \mu_2) T$$

Premiers exemples de stratégies

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

$\Rightarrow$ EXPLORATION

$$R(T) = \left(\frac{1}{K}\sum_{a=2}^{K}(\mu_1 - \mu_a)\right) T$$

- **Idée 2 :** Faire confiance au meilleur empirique jusqu'ici

$$A_{t+1} = \underset{a}{\text{argmax}} \ \hat{\mu}_a(t)$$

où

$$\hat{\mu}_a(t) = \frac{Y_{a,1} + \dots + Y_{a,N_a(t)}}{N_a(t)}$$

est un estimateur de la moyenne inconnue μ_a.

$\Rightarrow$ EXPLOITATION

$$\mathbb{R}(T) \geq (1-\mu_1) \times \mu_2 \times (\mu_1 - \mu_2) T$$

Une meilleure idée: explorer puis exploiter

Etant donné $m \in \{1, \ldots, T/K\}$, on

- tire chaque bras m fois
- détermine le meilleur empirique $\hat{a} = \arg\max_a \ \hat{\mu}_a(Km)$
- on sélectionne ce bras jusqu'à la fin

$$A_{t+1} = \hat{a} \ \text{ pour } t \geq Km$$

$\Rightarrow$ EXPLORATION puis EXPLOITATION

Une meilleure idée: explorer puis exploiter

Etant donné $m \in \{1, \ldots, T/K\}$, on

- tire chaque bras m fois
- détermine le meilleur empirique $\hat{a} = \arg\max_a \ \hat{\mu}_a(Km)$
- on sélectionne ce bras jusqu'à la fin

$$A_{t+1} = \hat{a} \ \text{ pour } t \geq Km$$

$\Rightarrow$ EXPLORATION puis EXPLOITATION

$$R(T) \leq \left(\sum_{a=2}^{K} (\mu_1 - \mu_a)\right) \left[m + T \exp\left(-\frac{m\Delta^2}{2}\right)\right]$$

où

$$\Delta = \mu_1 - \mu_2$$

➜ Comment choisir m?

Une meilleure idée: explorer puis exploiter

Etant donné $m \in \{1, \dots, T/K\}$, on

- tire chaque bras m fois
- détermine le meilleur empirique $\hat{a} = \arg\max_a \ \hat{\mu}_a(Km)$
- on sélectionne ce bras jusqu'à la fin

$$A_{t+1} = \hat{a} \ \ \text{pour } t \geq Km$$

$\Rightarrow$ EXPLORATION puis EXPLOITATION

$$R(T) \leq \frac{2\left(\sum_{a=2}^{K}(\mu_1 - \mu_a)\right)}{\Delta^2}\left[\log\left(\frac{T\Delta^2}{2}\right) + 1\right]$$

où

$$\Delta = \mu_1 - \mu_2$$

➜ en prenant $m = \frac{2}{\Delta^2}\log\left(\frac{T\Delta^2}{2}\right)$

Une meilleure idée: explorer puis exploiter

Etant donné $m \in \{1, \dots, T/K\}$, on

- tire chaque bras m fois
- détermine le meilleur empirique $\hat{a} = \arg\max_a \; \hat{\mu}_a(Km)$
- on sélectionne ce bras jusqu'à la fin

$$A_{t+1} = \hat{a} \quad \text{pour } t \geq Km$$

$\Rightarrow$ EXPLORATION puis EXPLOITATION

$$\underline{\text{2 bras}} : \quad R(T) \leq \frac{2}{\Delta} \log\left(\frac{T\Delta^2}{2}\right) + \frac{2}{\Delta}$$

où

$$\Delta = \mu_1 - \mu_2$$

➜ en prenant $m = \frac{2}{\Delta^2} \log\left(\frac{T\Delta^2}{2}\right)$

Une version séquentielle de cette stratégie

Cas du bandit à deux bras

$$\mu_1 \quad > \quad \mu_2$$

Problème: $\Delta = \mu_1 - \mu_2$ inconnu

➜ impossible de fixer le “bon” m à l’avance

Une version séquentielle de cette stratégie

Cas du bandit à deux bras

$$\mu_1 \quad > \quad \mu_2$$

<u>Problème</u>: $\Delta = \mu_1 - \mu_2$ inconnu

➜ impossible de fixer le "bon" m à l'avance

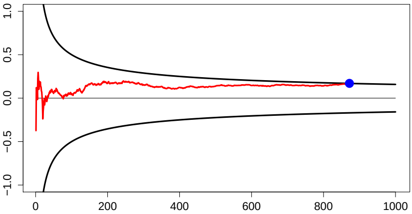
<u>Solution</u>: utiliser une phase d'exploration de durée adaptative .

➜ explorer uniformément jusqu'à l'instant

$$\tau = \inf \left\{ t \in \mathbb{N} : |\hat{\mu}_1(t) - \hat{\mu}_2(t)| > \sqrt{\frac{4\log(T/t)}{t}} \right\}$$

➜ $\hat{a} = \underset{a\in\{1,2\}}{\text{argmax}}\ \hat{\mu}_a(\tau)$ et choisir $(A_{t+1} = \hat{a})$ pour $t \in \{\tau, \dots, T\}$

Une version séquentielle de cette stratégie



$$\tau = \inf\left\{ t \in \mathbb{N} : |\hat{\mu}_1(t) - \hat{\mu}_2(t)| > \sqrt{\frac{4\log(T/t)}{t}} \right\}$$

Théorème

Il existe une constant $C > 0$ telle que

$$R_T \leq \frac{2}{\Delta}\log(T) + C\sqrt{\log(T)}.$$

➜ même de taux de croissance du regret qu'en connaissant Δ !

Peut-on faire mieux qu'un regret logarithmique?

[Lai et Robbins 1985]: tout "bon" algorithme de bandit doit tirer tous les bras une infinité de fois : pour $a \geq 2$,

$$\mathbb{E}[N_a(T)] \geq \frac{1}{d(\mu_a, \mu_1)} \log(T) \qquad \text{“pour } T \text{ grand”}$$

où $d(x, y) = x \log\left(\frac{x}{y}\right) + (1 - x) \log\left(\frac{1-x}{1-y}\right)$.

Remarque: $d(x, y) \geq 2(x - y)^2$.

$$\underline{\text{2 bras}} : \ \mathbb{E}[N_2(T)] \simeq \frac{2}{\Delta^2} \log(T) > 4 \times \frac{1}{d(\mu_2, \mu_1)} \log(T)$$

➜ Peut-on faire mieux et atteindre ce nombre minimal de tirage des bras?

Plan

Le principe d'optimisme

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

$$\mathcal{I}_a(t) = [\mathrm{LCB}_a(t), \mathrm{UCB}_a(t)]$$

$$\mathrm{LCB} = \text{Lower Confidence Bound}$$
$$\mathrm{UCB} = \text{Upper Confidence Bound}$$

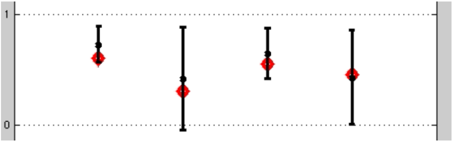


Figure: Intervalles de confiance sur les bras après t instants

Le principe d'optimisme

- On applique le principe suivant :

«agir comme si on se trouvait dans le meilleur des modèles possible »

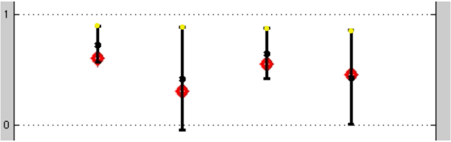


Figure: Intervalles de confiance sur les bras après t instants

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

$$A_{t+1} = \arg\max_a \ \mathrm{UCB}_a(t)$$

Comment construire des intervalles de confiance?

On cherche $\mathrm{UCB}_a(t)$ tel que

$$\mathbb{P}\left(\mu_a \leq \mathrm{UCB}_a(t)\right) \simeq 1 - \frac{1}{t}.$$

- Une première idée:

Théorème Centrale Limite

Z_i i.i.d. $\sim \mathcal{B}(\mu)$ et $X \sim \mathcal{N}(0,1)$.

$$\mathbb{P}\left(\sqrt{\frac{s}{\mu(1-\mu)}}\left(\mu - \frac{Z_1 + \dots + Z_s}{s}\right) \geq x\right) \underset{s\to\infty}{\rightarrow} \mathbb{P}(X \geq x)$$

Comment construire des intervalles de confiance?

On cherche $\mathrm{UCB}_a(t)$ tel que

$$\mathbb{P}\left(\mu_a \leq \mathrm{UCB}_a(t)\right) \simeq 1 - \frac{1}{t}.$$

- Une première idée:

Théorème Centrale Limite

Z_i i.i.d. $\sim \mathcal{B}(\mu)$ et $X \sim \mathcal{N}(0,1)$.

$$\mathbb{P}\left(\sqrt{\frac{s}{\mu(1-\mu)}}\left(\mu - \frac{Z_1 + \dots + Z_s}{s}\right) \geq x\right) \underset{s\to\infty}{\to} \mathbb{P}(X \geq x)$$

Pour s assez grand,

$$\mathbb{P}\left(\sqrt{\frac{s}{\mu_a(1-\mu_a)}}\left(\mu_a - \frac{Y_{a,1} + \dots + Y_{a,s}}{s}\right) \geq \sqrt{2\log\left(1/\delta\right)}\right) \lesssim \delta.$$

Comment construire des intervalles de confiance?

On cherche $\mathrm{UCB}_a(t)$ tel que

$$\mathbb{P}\left(\mu_a \leq \mathrm{UCB}_a(t)\right) \simeq 1 - \frac{1}{t}.$$

- Une première idée:

Théorème Centrale Limite

Z_i i.i.d. $\sim \mathcal{B}(\mu)$ et $X \sim \mathcal{N}(0,1)$.

$$\mathbb{P}\left(\sqrt{\frac{s}{\mu(1-\mu)}}\left(\mu - \frac{Z_1 + \cdots + Z_s}{s}\right) \geq x\right) \underset{s\to\infty}{\to} \mathbb{P}(X \geq x)$$

Pour s assez grand,

$$\mathbb{P}\left(\mu_a \leq \frac{Y_{a,1} + \cdots + Y_{a,s}}{s} + \sqrt{\frac{\log(1/\delta)}{2s}}\right) \gtrsim 1 - \delta.$$

➜ intervalle de confiance asymptotique

➜ le nombre d'observation de a à l'instant t est aléatoire !

Comment construire des intervalles de confiance?

- Une deuxième idée: Utiliser une inégalité de déviation

Inégalité de Hoeffding

Z_i i.i.d. de moyenne μ avec $Z_i \in [0,1]$. Pour tout entier $s \geq 1$

$$\mathbb{P}\left(\frac{Z_1 + \dots + Z_s}{s} \leq \mu - x\right) \leq e^{-2x^2 s}.$$

On a ainsi, pour tout $s \geq 1$,

$$\mathbb{P}\left(\frac{Y_{a,1} + \dots + Y_{1,s}}{s} \leq \mu_a - \sqrt{\frac{\log(1/\delta)}{2s}}\right) \leq \delta$$

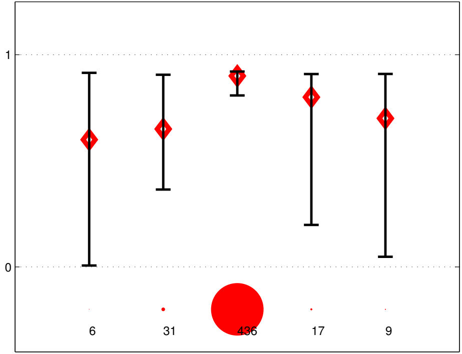
et en utilisant une borne de l'union

$$\mathbb{P}\left(\mu_a \leq \hat{\mu}_a(t) + \sqrt{\frac{\log(t)}{N_a(t)}}\right) \geq 1 - \frac{1}{t}.$$

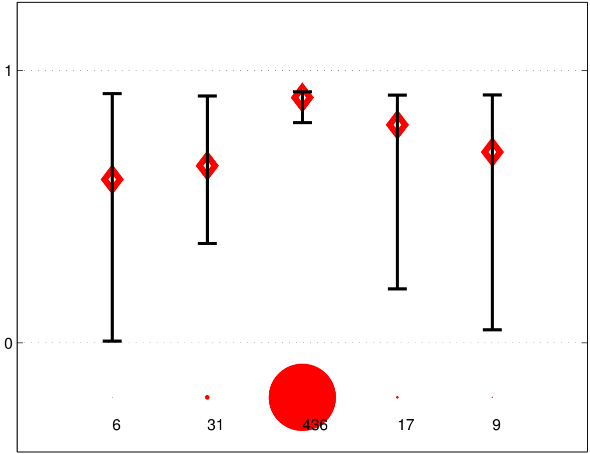
L'algorithme UCB1

UCB1 [Auer et al. 02] choisit $A_{t+1} = \text{argmax}_a \ \text{UCB}_a(t)$ avec

$$\text{UCB}_a(t) = \underbrace{\hat{\mu}_a(t)}_{\text{terme d'exploitation}} + \underbrace{\sqrt{\frac{2\log(t)}{N_a(t)}}}_{\text{bonus d'exploration}}.$$



UCB en action !



Cet algorithme est-il optimal?

Théorème [Auer et al. 02]

Pour l'algorithme UCB1

$$\mathbb{E}[N_a(T)] \leq \frac{8}{(\mu_1 - \mu_a)^2} \log(T) + \left(1 + \frac{\pi^2}{6}\right).$$

$$\underline{\text{2 bras}} : \quad R_T \leq \frac{8}{\Delta} \log(T) + C.$$

- L'analyse peut être rafinée pour une variante de l'algorithme:

$$\underline{\text{2 bras}} : \quad R_T \leq \frac{1}{2\Delta} \log(T) + C\sqrt{\log(T)}.$$

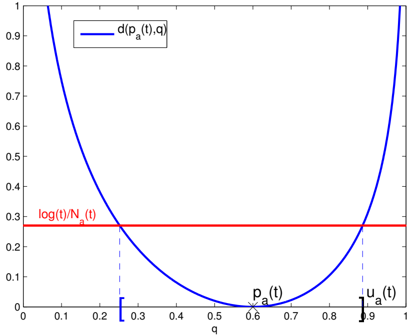
➜ mieux que la stratégie "explore puis exploite"!

➜ sous-optimal par rapport à la borne inférieure

L'algorithme KL-UCB : un algorithme optimal

- KL-UCB [Cappé et al. 13] utilise

$$\mathrm{UCB}_a(t) = \max\left\{q : d\left(\hat{\mu}_a(t), q\right) \leq \frac{\log(t)}{N_a(t)}\right\}$$



On peut montrer que

$$\mathbb{E}[N_a(T)] \leq \frac{1}{d(\mu_a, \mu_1)} \times \log T + C\sqrt{\log(T)}.$$

Plan

Le modèle bayésien

Deux visions du modèle de bandit:

- **fréquentiste**: $\mu_1, \ldots, \mu_K$ sont des paramètres inconnus

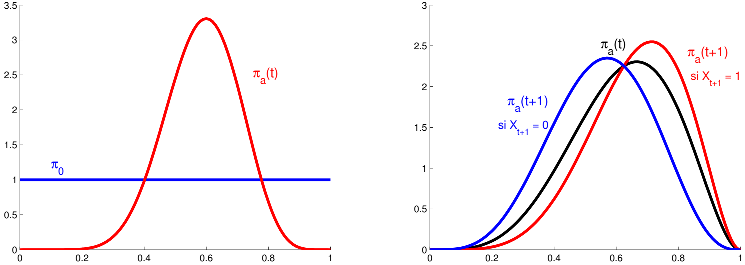
➜ estimation, construction d'intervalles de confiance

- **bayésienne**: $\mu_1, \ldots, \mu_K$ sont des variables aléatoires

$$\text{loi a priori}: \quad \mu_a \sim \mathcal{U}([0,1])$$

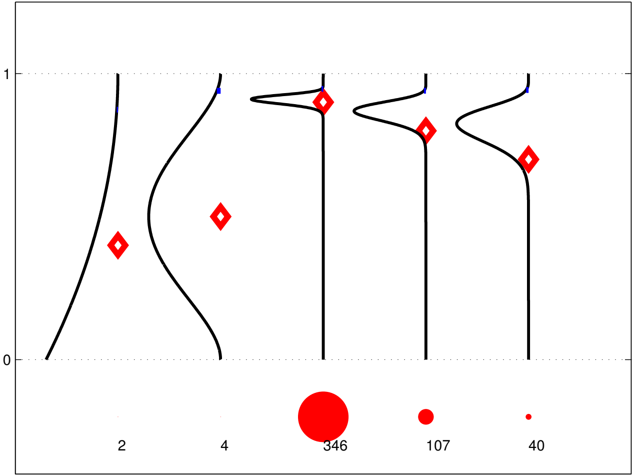
➜ étant données des observations, on peut calculer la

$$\begin{aligned} \text{loi a posteriori} : \quad \pi_a(t) &= \mathcal{L}\left(\mu_a \middle| Y_{a,1}, \ldots, Y_{a,N_a(t)}\right) \\ &= \mathrm{Beta}(S_a(t)+1, N_a(t)-S_a(t)+1) \end{aligned}$$



Algorithme bayésien

Un algorithme de bandit bayésien est un algorithme qui exploite les lois a posteriori des μ_a pour décider quel bras sélectionner.



Un example: l'échantillonnage de Thompson

Thompson Sampling:

$$\begin{cases} \forall a \in \{1..K\}, \quad \theta_a(t) \sim \pi_a(t) \\ A_{t+1} = \underset{a=1...K}{\text{argmax}}\ \theta_a(t). \end{cases}$$

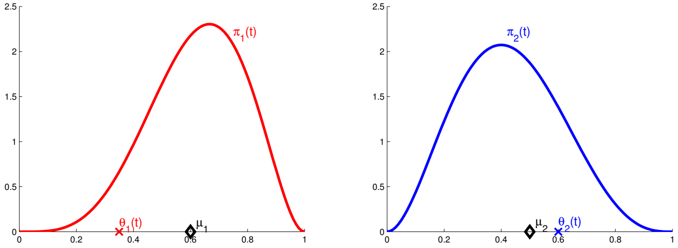


Figure: L'algorithme sélectionne le bras 2 car $\theta_2(t) \geq \theta_1(t)$

- ➜ Le premier algorithme de bandit ! [Thompson 33]
- ➜ Très efficace en pratique, même dans des modèles complexes
- ➜ Atteint la borne inférieure de Lai et Robbins [K. et al. 12]

Plan

Faut-il minimiser le regret?





$\mathcal{B}(\mu_1)$ $\mathcal{B}(\mu_2)$ $\mathcal{B}(\mu_3)$ $\mathcal{B}(\mu_4)$ $\mathcal{B}(\mu_5)$

Pour le t-ème patient d'une étude clinique, un docteur

- choisit un traitement A_t
- observe une réponse $X_t \in \{0,1\}$: $\mathbb{P}(X_t = 1) = \mu_{A_t}$

Minimiser le regret:
maximiser le nombre de patients guéris durant l'étude

$\mathcal{B}(\mu_1)$ $\mathcal{B}(\mu_2)$ $\mathcal{B}(\mu_3)$ $\mathcal{B}(\mu_4)$ $\mathcal{B}(\mu_5)$

Pour le t-ème patient d'une étude clinique, un docteur

- choisit un traitement A_t
- observe une réponse $X_t \in \{0, 1\}$: $\mathbb{P}(X_t = 1) = \mu_{A_t}$

Minimiser le regret:
maximiser le nombre de patients guéris durant l'étude

Objectif alternatif: allouer les traitement de sorte à identifier le plus rapidement possible le meilleur traitement
(sans objectif thérapeutique immédiat)

Un nouveau problème de bandit

Paramètre:

- $\delta \in]0,1[$ un paramètre de risque

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 de tirer des bras
- une **règle de recommendation**, indiquant le bras choisi

$$\hat{a}_\tau = \underset{a=1\dots K}{\text{argmax}} \ \hat{\mu}_a(\tau)$$

Son objectif :

- $\mathbb{P}(\hat{a}_\tau = a^*) \geq 1 - \delta$
- Le nombre total de tirages des bras utilisés $\mathbb{E}[\tau]$ est faible

L'algorithme LUCB

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

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

par exemple

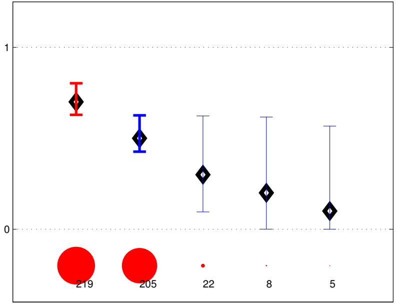
$$\begin{aligned} U_a(t) &= \hat{\mu}_a(t) + \sqrt{\frac{\log(2Kt^2/\delta)}{2N_a(t)}}, \\ L_a(t) &= \hat{\mu}_a(t) - \sqrt{\frac{\log(2Kt^2/\delta)}{2N_a(t)}}. \end{aligned}$$

➜ On se sert aussi des bornes de confiance inférieures.

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

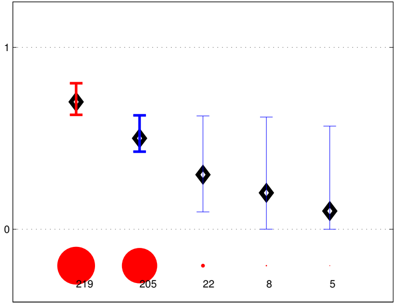


meilleur bras empirique, l_t bras sous-optimaux, u_t en gras

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]

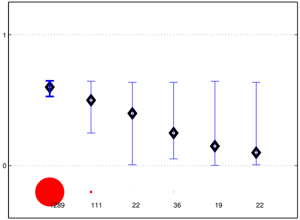
L'algorithme LUCB vérifie

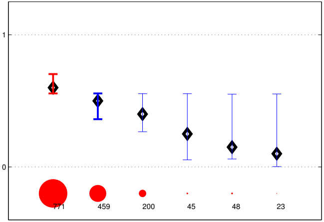
$$\mathbb{P}\left(\hat{a}^* = a^*\right) \geq 1 - \delta \quad \text{et} \quad \mathbb{E}[\tau] = C \times H \log \frac{1}{\delta},$$

où

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

UCB versus LUCB





Conclusion

On a vu:

- ➜ des algorithmes efficaces basés sur des intervalles de confiance
  ... pour plusieurs objectifs !
- ➜ un algorithme basé sur des outils de statistique bayésienne

Pour maximiser ses récompenses dans un modèle de bandit:

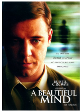
- ➜ il faut mêler exploration et exploitation

En fonction de l'application, plusieurs objectifs peuvent être considérés dans un modèle de bandit.

modèle de bandit $\neq$ problème de bandit

Plan

Example: recommendation de film



Quel film Netflix va recommander à un utilisateur, en se basant sur les notes données par les utilisateurs précédents?

➜ On a besoin d'un modèle prenant en compte les caractéristiques des films présentés

Exemple: le modèle logistique

$$\mathbb{P}(X_t = +|A_t = a) = \frac{1}{1 + e^{-x_a^T\theta}}$$

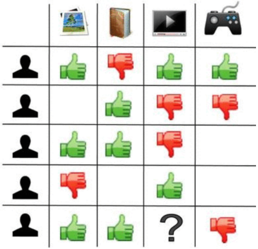
où

- $x_a \in \mathbb{R}^d$ est un vecteur (connu) de caractéristique du film
- $\theta \in \mathbb{R}^d$ (inconnu) indique les préférences de l'utilisateur

Example: recommendation de film

Quel film Netflix va recommander à un utilisateur, en se basant sur les notes données par les utilisateurs précédents?

➜ Comment combiner bandits et filtrage collaboratif ?

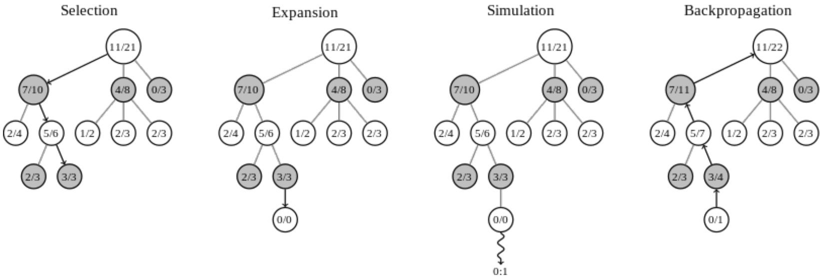


source: Wikipédia

Pour décider du prochain coup à jouer:

- choisir successivement des trajectoires dans l'arbre de jeu
- effectuer des évaluations (aléatoires) de certains positions

➜ Comment sélectionner séquentiellement les trajectoires ?



source: Wikipédia

Algorithme UCT [Kocsis & Szepesvari 06]: **UC**B for **T**rees !

Des questions?