L'algorithme KL-UCB pour les bandits bornés, et au delà [arXiv:1102.2490]

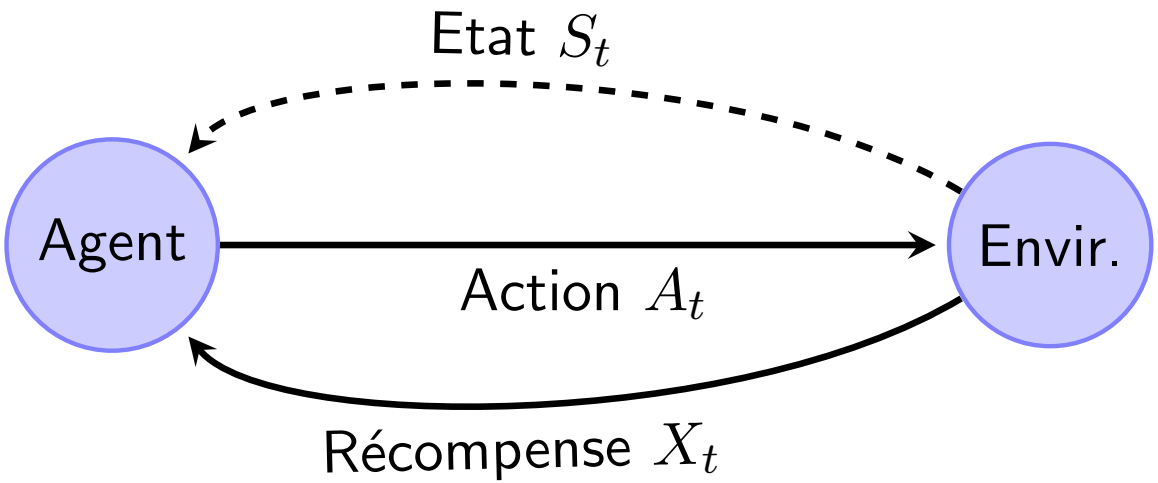
Aurélien Garivier et Olivier Cappé

CNRS & Telecom ParisTech

10 juin 2011

Plan de l'exposé

Apprentissage par renforcement



dilemme
exploration
|
exploitation

RL $\neq$ apprentissage classique (notion de récompense)
RL $\neq$ théorie des jeux (environnement indifférent)

Exemple : essais cliniques séquentiels

Pour fixer les idées, on considère le cas de figure suivant :

problème : des patients atteints d'une certaine maladies sont diagnostiqués au fil du temps

outils : on dispose de plusieurs traitements mal dont l'efficacité est a priori inconnue

déroulement : on traite chaque patient avec un traitement, et on observe le résultat (binaire)

objectif : soigner un maximum de patients (et pas connaître précisément l'efficacité de chaque traitement)

Le problème des bandits stochastiques

Environment K bras, paramètre $\theta = (\theta_1, \dots, \theta_K) \in [0,1]^K$
L'allocation de bras $a_t \in \{1, \dots, K\}$ conduit à récompense

$$Y_t = X_{a_t,t}$$

où $X_{i,s} = \mathbb{1}\{U_s \leq \theta_i\}$, pour $1 \leq i \leq K$, $s \geq 1$, et $(U_s)_s \overset{iid}{\sim} \mathcal{U}[0,1]$.

Stratégie règle d'allocation dynamique : $\pi = (\pi_1, \pi_2, \dots)$ tq

$$A_t = \pi_t(Y_1, \dots, Y_{t-1})$$

Nombre de tirages du bras $b \in \{1, \dots, K\}$:

$$N_t^\pi(b) = \sum_{s \leq t} \mathbb{1}\{A_s = b\}$$

Performance, regret

Récompense cumulée : $S_n = Y_1 + \cdots + Y_n, \quad n \geq 1$

Notre objectif : choisir π de manière à maximiser

$$E[S_n] = \sum_{t=1}^{n} \sum_{b=1}^{K} \mathbb{E}\big[\mathbb{E}[Y_t \mathbb{1}\{A_t = b\} | Y_1, \ldots, Y_{t-1}]\big]$$
$$= \sum_{b=1}^{K} \theta_b \mathbb{E}[N_n^\pi(b)]$$

Objectif équivalent : minimiser le regret

$$R_n(\theta) = n\theta^* - E[S_n] = \sum_{b:\theta_b < \theta^*} (\theta^* - \theta_b)\mathbb{E}[N_n^\pi(b)]$$

où $\theta^* = \max\{\theta_b : 1 \leq b \leq K\}$.

Plan de l'exposé

Stratégie consistante

- une stratégie π est dite consistante si, pour tout $\theta \in [0,1]^K$,

$$\frac{1}{n}\mathbb{E}[S_n] \to \theta^*$$

  c'est-à-dire si elle finit par se concentrer sur le meilleur traitement
- elle est efficace si pour tout $\theta \in [0,1]^K$ et pour tout $a > 0$,

$$R_n(\theta) = o(n^a)$$

  c'est-à-dire si le nombre de mauvais traitements adminsitrés est sous-polynômial
- on construit assez aisément des stratégies efficaces, mais moins facilement des stratégies *optimales*

La borne de Lai et Robbins

Théorème [Lai&Robbins, '85]

Si π est une stratégie efficace, alors pour tout $\theta \in [0,1]^K$

$$\liminf_{n\to\infty} \frac{R_n(\theta)}{\log(n)} \geq \sum_{b:\theta_b<\theta^*} \frac{\theta^* - \theta_b}{\mathrm{kl}(\theta_b, \theta^*)}$$

où

$$kl(p,q) = p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q}$$

désigne la divergence de Kullback-Leibler entre la loi $\mathcal{B}(p)$ et la loi $\mathcal{B}(q)$, $0 \leq p, q \leq 1$.

Plan de l'exposé

Principe d'optimisme

Algorithmes **optimistes** : [Lai&Robins '85 ; Agrawal '95]

Fais comme si tu te trouvais dans l'environnement qui t'est le plus favorable parmi tous ceux qui rendent les observations suffisamment vraisemblables

De façon plutôt inattendue, les méthodes optimistes se révèlent pertinentes dans des cadres très différents, efficaces, robustes et simples à mettre en oeuvre

Stratégies "Upper Confidence Bound"

UCB [Lai&Robins '85 ; Auer&al '02 ; Audibert&al '07]

- Construit une UCB pour chaque bras :

$$\underbrace{\frac{S_t(a)}{N_t(a)}}_{\text{récompense moyenne estimée}} + \underbrace{\sqrt{\frac{\log(t)}{2N_t(a)}}}_{\text{bonus d'exploration}}$$

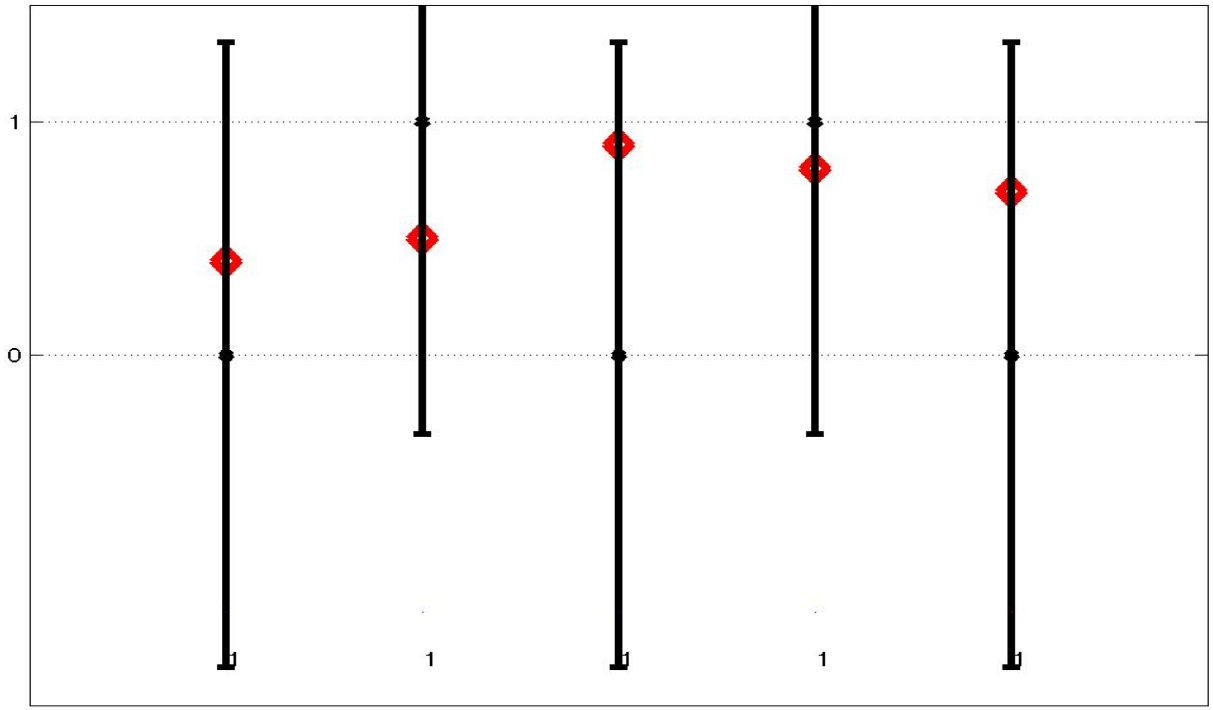
- Choisis le bras qui la plus grande UCB

Avantage : comportement facilement interprétable et "acceptable"
Borne sur le regret :

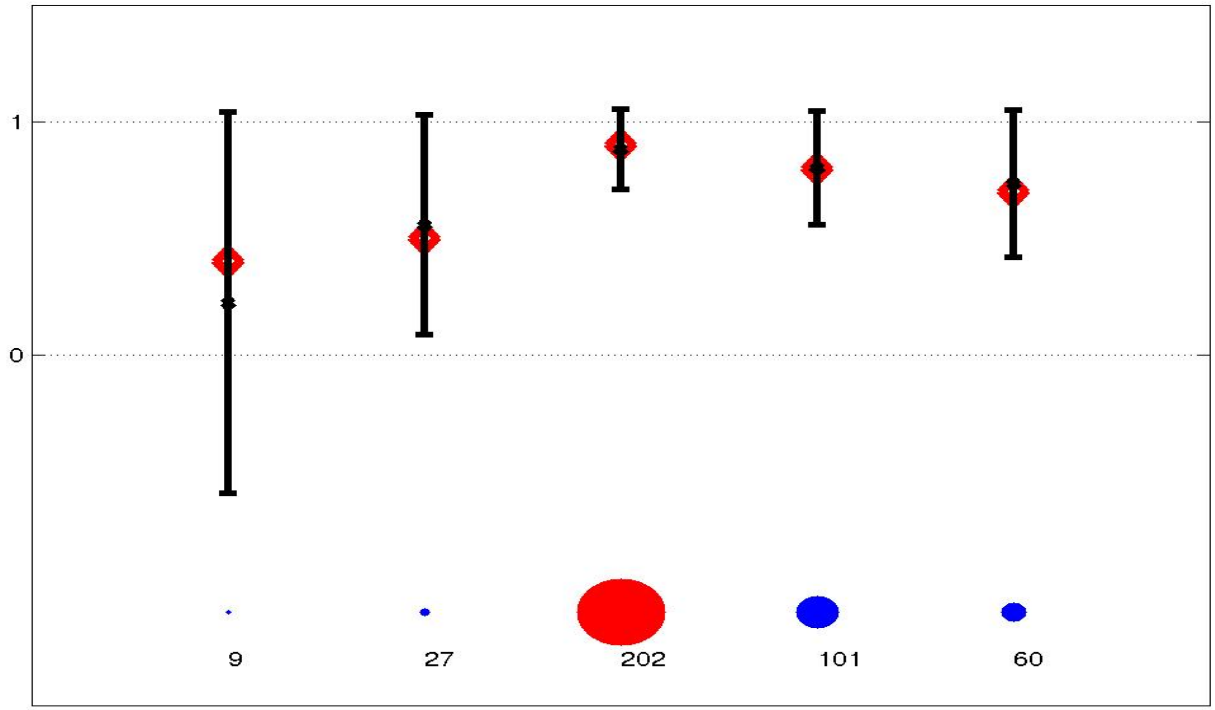
$$\mathbb{E}[R_n] \lesssim \sum_{a:\theta_a<\theta^*} \frac{1}{2(\theta^* - \theta_a)} \log(n)$$

Politique d'indice : on calcule un indice par bras et on choisit celui qui est le plus élevé, cf. [Gittins '79]

UCB en action



UCB en action



Début

KL-UCB

Require: n (horizon), K (nb de bras), REWARD (récompenses)

for $t = 1$ **to** K **do**
  $N[t] \leftarrow 1$
  $S[t] \leftarrow \text{REWARD}(\text{arm} = t)$
end for
for $t = K + 1$ **to** n **do**

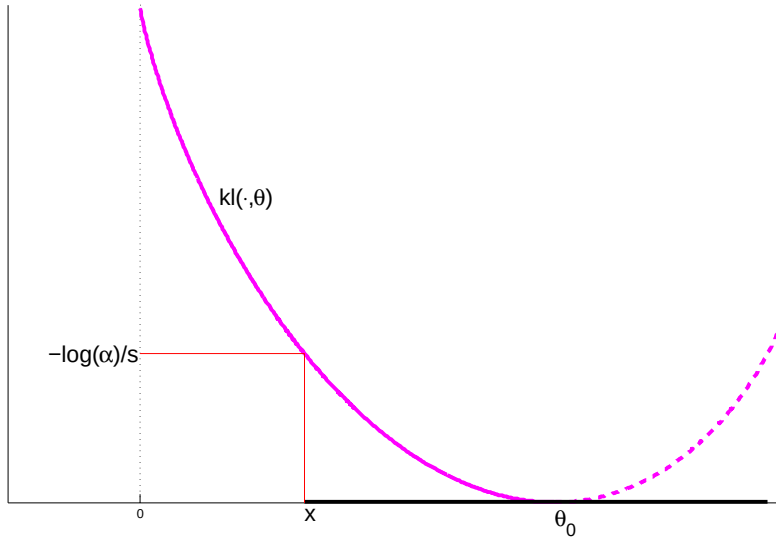
$$a \leftarrow \arg\max_{1 \leq a \leq K} \max \left\{ q \in \Theta : N[a] \ \text{kl}\left(\frac{S[a]}{N[a]}, q\right) \leq \log(t) \right\}$$

  $r \leftarrow \text{REWARD}(\text{arm} = a)$
  $N[a] \leftarrow N[a] + 1$
  $S[a] \leftarrow S[a] + r$
end for

Région de confiance KL

Si $Z_1, \ldots, Z_s \overset{iid}{\sim} \mathcal{B}(\theta_0)$, et si $\hat{p}_s = (Z_1 + \cdots + Z_s)/s$, alors

$$\mathbb{P}\left(\hat{p}_s < x\right) \leq \exp\left(-s\,\mathrm{kl}(x, \theta_0)\right)$$



Autrement dit, si $\alpha = \exp\left(-s\,\mathrm{kl}(x, \theta_0)\right)$:

$$\mathbb{P}\left(\hat{p}_s < x\right) = \mathbb{P}\left(\mathrm{kl}(\hat{p}_s, \theta_0) > -\frac{\log(\alpha)}{s},\ \hat{p}_s < \theta\right) \leq \alpha$$

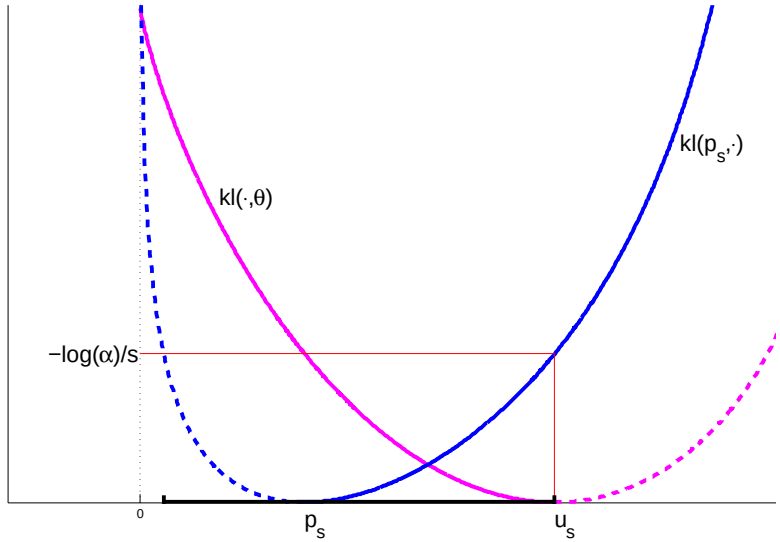
D'où une borne de confiance supérieure pour p au risque α :

$$u_s = \sup\left\{\theta > \hat{p}_s : \mathrm{kl}(\hat{p}_s, \theta) \leq -\frac{\log(\alpha)}{s}\right\}$$

Région de confiance KL

Si $Z_1, \ldots, Z_s \overset{iid}{\sim} \mathcal{B}(\theta_0)$, et si $\hat{p}_s = (Z_1 + \cdots + Z_s)/s$, alors

$$\mathbb{P}\left(\hat{p}_s < x\right) \leq \exp\left(-s\,\mathrm{kl}(x, \theta_0)\right)$$



Autrement dit, si $\alpha = \exp\left(-s\,\mathrm{kl}(x, \theta_0)\right)$:

$$\mathbb{P}\left(\hat{p}_s < x\right) = \mathbb{P}\left(\mathrm{kl}(\hat{p}_s, \theta_0) > -\frac{\log(\alpha)}{s},\ \hat{p}_s < \theta\right) \leq \alpha$$

D'où une borne de confiance supérieure pour p au risque α :

$$u_s = \sup\left\{\theta > \hat{p}_s : \mathrm{kl}(\hat{p}_s, \theta) \leq -\frac{\log(\alpha)}{s}\right\}$$

Borne de regret

Théorème[G., Cappé '11] :

Soit $\theta \in [0,1]^K$, et soit $b \in \{1, \dots, K\}$ tel que $\theta_b < \theta^*$. Pour tout $\epsilon > 0$ il existe $C_1, C_2(\epsilon)$ et $\beta(\epsilon)$ tels que

$$\mathbb{E}[N_n^{KL-UCB}(b)] \leq \frac{\log(n)}{\mathrm{kl}(\theta_b, \theta^*)}(1+\epsilon) + C_1 \log(\log(n)) + \frac{C_2(\epsilon)}{n^{\beta(\epsilon)}} \ .$$

Corollaire :

$$R_n(\theta) \lesssim \sum_{b:\theta_b<\theta^*} \frac{\theta^* - \theta_b}{\mathrm{kl}(\theta_b, \theta^*)} \log(n)$$

$\Longrightarrow$ KL-UCB est asymptotiquement optimal, et on dispose d'une borne pour son regret en temps fini.

Ingrédient essentiel : déviations auto-normalisées

Pour l'analyse, il faut contrôler les *déviations auto-normalisées, mesurées dans la bonne métrique, de la moyenne empirique :*

Théorème

Soit $(X_t)_t \geq 1$ une suite de v.a. indépendantes de loi $\mathcal{B}(\theta)$ sur $(\Omega, \mathcal{F}, \mathbb{P})$. Soit $\mathcal{F}_t$ be une suite croissante de tribus de $\mathcal{F}$ tq $\forall t, \sigma(X_1 \ldots, X_t) \subset \mathcal{F}_t$ et pour $s > t$, X_s est indépendante de $\mathcal{F}_t$. Soit$(\epsilon_t)_{t\geq 1}$ une suite prévisible de variables de Bernoulli. On définit, pour tout $\delta > 0$:

$$S(n) = \sum_{s=1}^{n} \epsilon_s X_s \ , \qquad N(n) = \sum_{s=1}^{n} \epsilon_s \ , \qquad \hat{\theta}(n) = \frac{S(n)}{N(n)} \ ,$$

$$u(n) = \max \left\{ q > \hat{\theta}_n : N(n) d\left(\hat{\theta}(n), q\right) \leq \delta \right\} \ .$$

Alors

$$\mathbb{P}\left(u(n) < \theta\right) \leq e \left\lceil \delta \log(n) \right\rceil \exp(-\delta)$$

$$\mathbb{P}\Big(N(n) d(\hat{\mu}(n), \theta) > \delta\Big) \leq 2e \left\lceil \delta \log(n) \right\rceil \exp(-\delta)$$

Simulations : scénario à deux bras

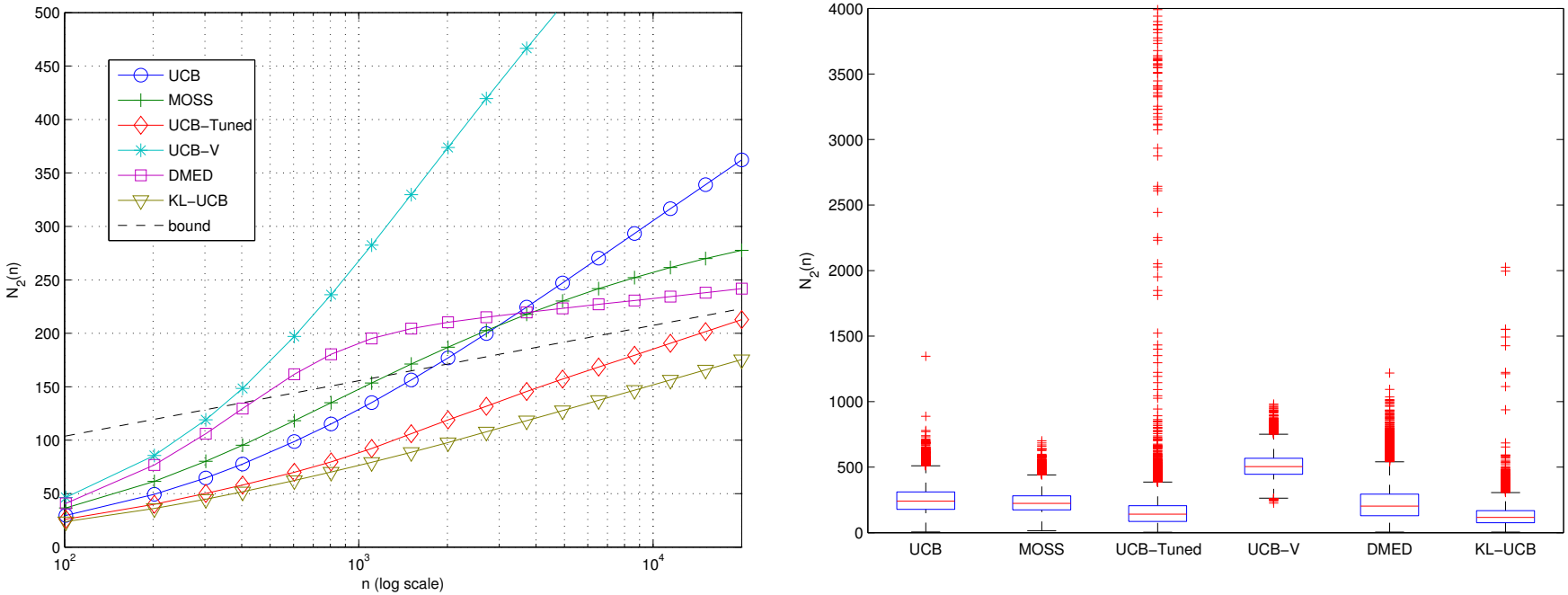


FIG.: Performance de différents algorithmes dans le scénario à deux bras où $\theta = (0.9, 0.8)$. A gauche : nombre moyen de tirages du bras sous-optimal en fonction du temps. A droite : distribution du nombre de tirages du bras 2 au temps $n = 5000$. Résultats basés sur 50000 expériences indépendantes.

Simulations : scénario à récompenses faibles

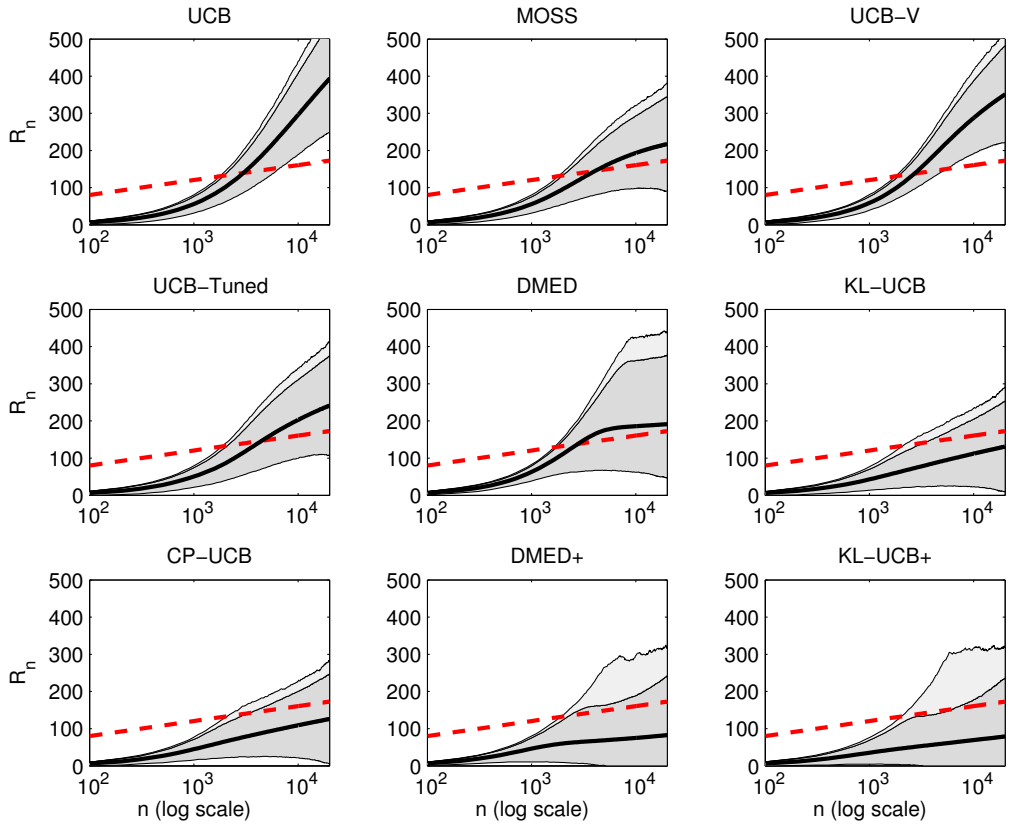


FIG.: Regrets de différents algorithmes en fonction du temps pour un scénario à dix bras où $\theta = (0.1, 0.05, 0.05, 0.05, 0.02, 0.02, 0.02, 0.01, 0.01, 0.01)$. Ligne rouge pointillée : borne inférieure de Lai&Robbins. Ligne épaisse : regret moyen. Régions grisées : région centrale à 99% et le quantile à $99,95\%$.

Plan de l'exposé

Récompenses bornées

Il suffit de ramener les récompenses dans $[0, 1]$, et on peut utiliser le *même algorithme* KL-UCB et obtenir les *mêmes* bornes de regret grâce au

Lemme :

soit X une variable aléatoire à valeur dans $[0, 1]$, et soit $\mu = \mathbb{E}[X]$. Alors, pour tout $\lambda \in \mathbb{R}$,

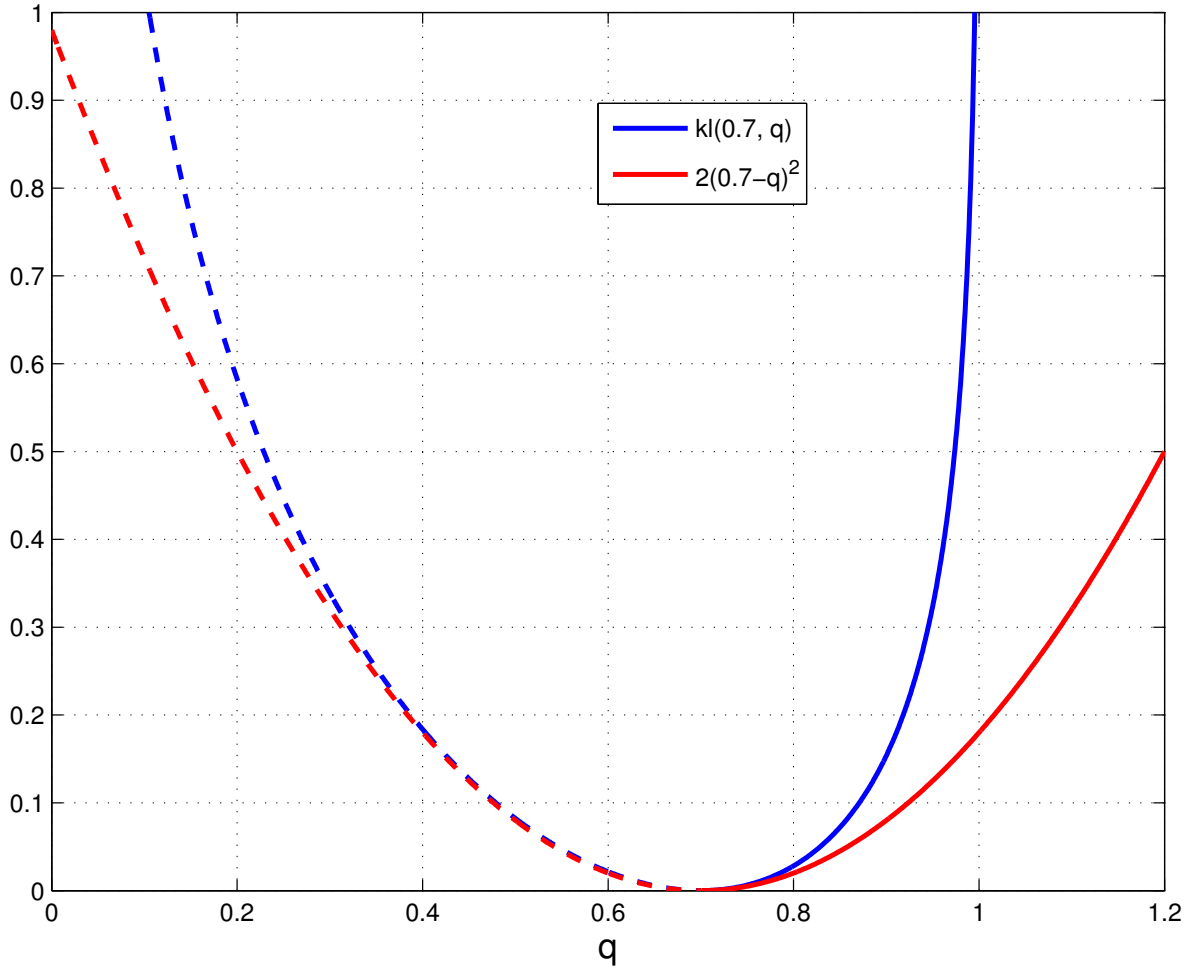
$$E\left[\exp(\lambda X)\right] \leq 1 - \mu + \mu \exp(\lambda) \ .$$

KL-UCB fait toujours mieux que UCB : Inégalité de Pinsker

$$\mathrm{kl}(\mu_1, \mu_2) \geq 2(\mu_1 - \mu_2)^2$$

Toutefois, il peut y avoir mieux à faire si les distributions des récompenses ont une faible variance par rapport à la loi de Bernoulli correspondante.

Comparaison UCB vs KL-UCB



Simulations : exponentielles bornées

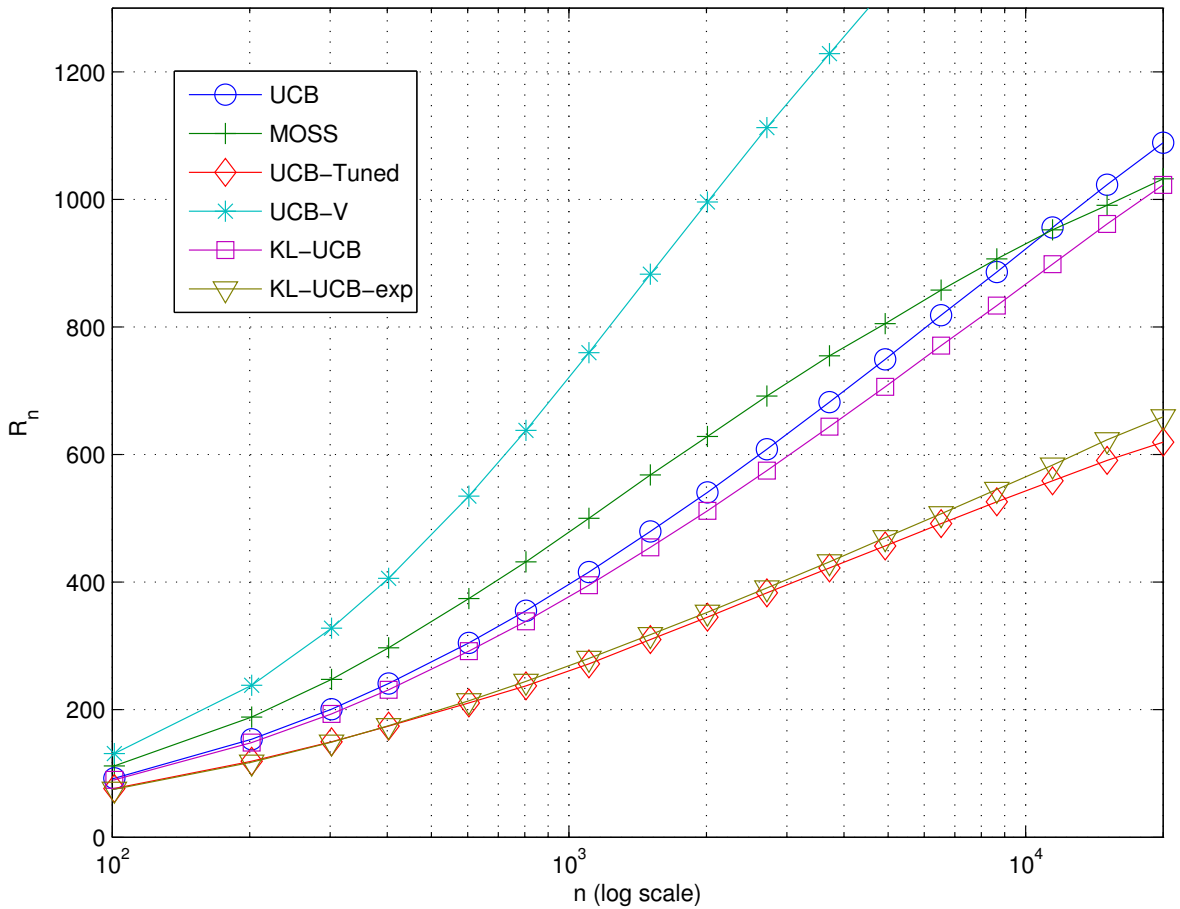


FIG.: Regret de différentes politiques en fonction du temps, sur le scénario des exponentielles bornées.

Récompenses dans la famille exponentielle

- La même preuve se généralise directement à des récompenses dont les distributions admettent par rapport à une mesure dominante une densité pouvant s'écrire

$$p_{\theta_a}(x) = \exp\big(x\theta_i - b(\theta_a) + c(x)\big), \quad 1 \leq a \leq K$$

  pour un certain paramètre $\theta \in \mathbb{R}^K$
- L'algorithme reste le même, seule la définition de la fonction kl est modifiée - par exemple, pour des récompenses de loi exponentielle :

$$\mathrm{kl}(x, y) = y - x + x \log \frac{x}{y}$$

- Une inégalité de déviation analogue se prouve alors de la même façon, et conduit aux même bornes de regret

Bandits non stationnaires

- On autorise les distributions des récompenses à *varier brutalement* au cours du temps
- L'objectif est alors de faire *poursuivre le meilleur bras*
- Application : dans un scanner à effet tunel, la qualité de l'image dépend d'un réglage mais les distributions peuvent brutalement changer en cas de déplacement inopiné de la pointe
- On étudie alors D-UCB et SW-UCB [G. Moulines '08], variantes qui incluent un *oubli* (progressif) du passé
- On montre des bornes de regret en $O(\sqrt{n \log n})$, qui sont (presque) optimales

Bandits linéaires / linéaires généralisés

- Modèle de bandit avec information contextuelle :

$$\mathbb{E}[X_t|A_t] = \mu(m'_{A_t}\theta_*)$$

où $\theta_* \in \mathbb{R}^d$ désigne un paramètre inconnu et où $\mu : \mathbb{R} \to \mathbb{R}$ est la fonction de lien dans un modèle linéaire généralisé

- Exemple : pour des récompenses binaires

$$\mu(x) = \frac{\exp(x)}{1 + \exp(x)}$$

- Application : publicité ciblée sur internet
- GLM-UCB [Filippi, Cappé, G. '10], borne de regret dépendant de d et pas du nombre d'actions possibles

Optimisation stochastique

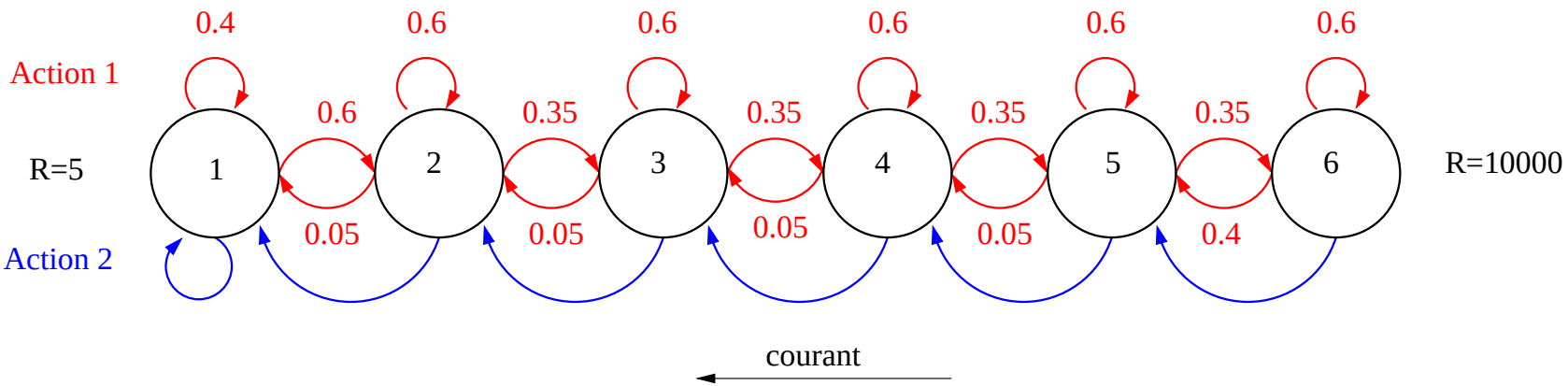
- Objectif : trouver le maximum (ou les quantiles) d'une fonction $f : C \subset \mathbb{R}^d \to \mathbb{R}$ observée dans du bruit (ou pas)
- Application : exposition aux ondes électro-magnétiques (indice DAS $=$ SAR)
- Modélisation : f est la réalisation d'un processus Gaussien, ou alors fonction de faible norme dans le RKHS associé au noyau de ce processus
- GP-UCB : jouer le point $x \in C$ pour lequel l'intervalle de confiance est le plus haut

Processus de Décision Markoviens

Le système est dans un état S_t qui évolue de façon markovienne :

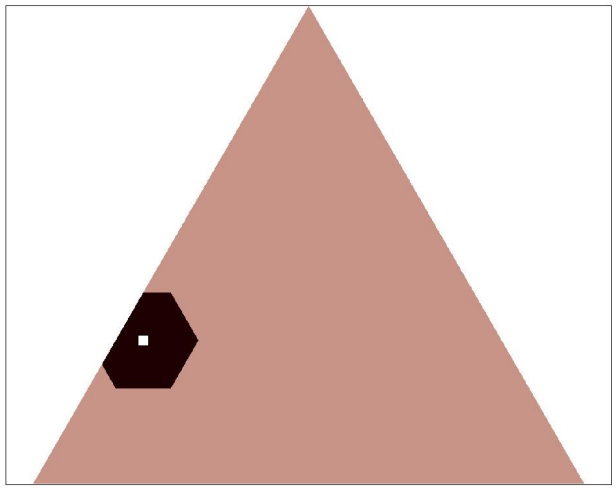
$$S_{t+1} \sim P(\cdot; S_t, A_t) \text{ et } R_t = r(S_t, A_t) + \epsilon_t$$

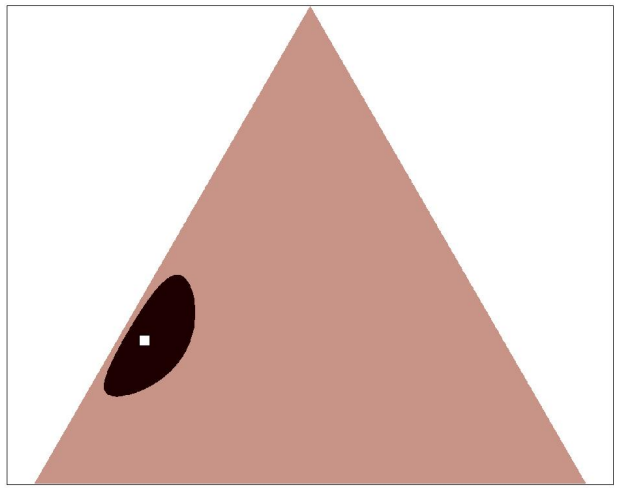
Exemple / Benchmark : RiverSwim [Strehl&Littman'08]



Optimisme pour les MDP

Le paradigme optimiste conduit à la recherche d'une matrice de transition "la plus avantageuse" dans un voisinage de son estimateur de maximum de vraisemblance.





L'utilisation de voisinages de Kullback-Leibler, autorisée par des inégalités de déviations semblables à celles montrées plus haut, conduisent à des algorithmes plus efficaces ayant de meilleures propriétés