Problèmes de bandits et applications

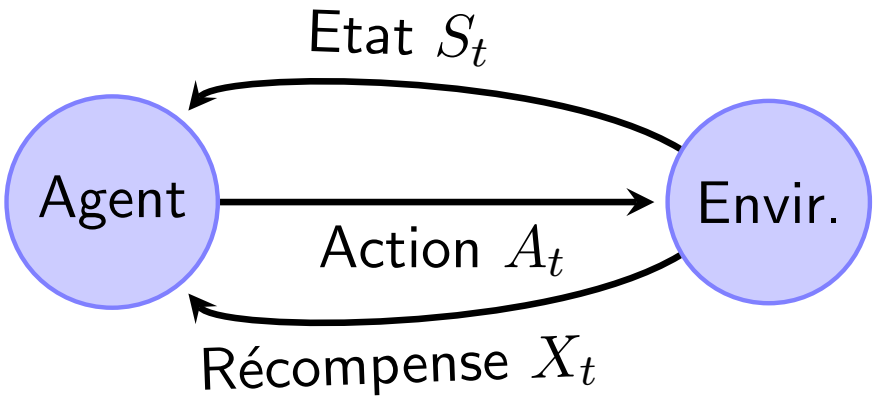
Aurélien Garivier, avec Sarah, Eric, Olivier, Emilie...

CNRS & Telecom ParisTech

Mardi 3 juillet 2012

Plan de l'exposé

Apprentissage par renforcement



RL $\neq$ apprentissage classique

RL $\neq$ théorie des jeux

- Allocation séquentielle de ressources
- Contrôle optimal à temps discret pour les processus de décision markovients (MDP)
- Application à l'optimisation de fonctions bruitées coûteuses à évaluer

Champs d'application classiques: essais cliniques, marketing en ligne, recherche d'information, proposition de contenu, yield management, robotique...

Paradigme: Essais cliniques séquentiels

On considère le cas de figure suivant :

- des patients atteints d'une certaine maladies sont diagnostiqués au fil du temps
- on dispose de plusieurs traitements mal dont l'efficacité est a priori inconnue
- 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)

Formalisation: "problème de bandits multibras"

Environment : ensemble de bras $\mathcal{A}$; le choix du bras $a \in \mathcal{A}$ à l'instant t donne la récompense

$$X_t = X_{a,t} \sim P_a \in \mathfrak{M}_1(\mathbb{R})$$

et la famille $\left(X_{a,t}\right)_{a\in\mathcal{A},t\geq 1}$ est indépendante

Règle d'allocation dynamique : $\phi = (\phi_1, \phi_2, \dots)$ telle que

$$A_t = \phi_t(X_1, \dots, X_{t-1})$$

Nombre de tirages du bras $a \in \mathcal{A}$ à l'instant $t \in \mathbb{N}$:

$$N_a(t) = \sum_{s\leq t} \mathbb{1}\{A_s = a\}$$

Performance, regret

- Récompense cumulée : $S_n = X_1 + \cdots + X_n, \quad n \geq 1$
- Objectif: choisir ϕ de manière à maximiser

$$\begin{aligned}\mathbb{E}\left[S_n\right] &= \sum_{t=1}^{n} \sum_{a \in \mathcal{A}} \mathbb{E}\Big[\mathbb{E}\left[X_t \mathbb{1}\{A_t = a\} | X_1, \ldots, X_{t-1}\right]\Big] \\ &= \sum_{a \in \mathcal{A}} \mu_a \mathbb{E}\left[N_a(n)\right]\end{aligned}$$

où $\mu_a = E[P_a]$

- Objectif équivalent : minimiser le *regret*

$$R_n\big((P_a)_{a \in \mathcal{A}}\big) = n\mu^* - E\left[S_n\right] = \sum_{a : \mu_a < \mu^*} (\mu^* - \mu_a)\mathbb{E}\left[N_a(n)\right]$$

où $\mu^* = \max\left\{\mu_a : a \in \mathcal{A}\right\}$

Plan de l'exposé

Le modèle bayésien

Cadre paramétrique: la distribution du bras a est paramétrée par $\theta_a \in \mathbb{R}^d$, typiquement la famille exponentielle canonique

A priori $\pi = \otimes_{a \in \mathcal{A}} \pi_a$ sur les paramètres:

$$\theta_a \overset{iid}{\sim} \pi_a$$

Exemple: $P_a = \mathcal{B}(\theta_a)$, $\pi_a(\theta) = Beta(1,1)$

Regret Bayésien:

$$R_n(\pi) = \int R_n\big((P_a)_{a \in \mathcal{A}}\big) d\pi$$

Objectif: trouver la règle d'allocation dynamique ϕ qui minimise $R_n(\pi)$.

La solution de [Gittins '79]

Idée (Bellman): contrôle optimal, programmation dynamique

Réduction: Gittins se ramène à l'étude d'**un bras aléatoire contre un bras constant**

Arrêt optimal: il montre qu'alors la politique optimale consiste à **jouer le bras aléatoire tant qu'il garde une chance d'être le meilleur**

Indice de Gittins $IG_a(t) =$ valeur du bras constant qui rend *indifférent* le choix entre le bras constant et le bras de loi $\pi_a(\cdot|X_1, \ldots, X_{t-1})$

Politique d'indice: Dans le problème initial, on fait appel au bras ayant l'**indice de priorité** le plus grand :

$$A_t = \phi_t(X_1, \ldots, X_{t-1}) = \arg\max_{a \in \mathcal{A}} IG_a(t)$$

Propriétés de la politique de Gittins

optimale...

- c'est la politique qui **minimise** $R_n(\pi)$
- pour des récompenses dans la famille exponentielle, **algorithme** pour calculer les indices
- **comportement satisfaisant** : explore surtout au début, exploite beaucoup à la fin

...mais :

- le calcul des indices n'est **pas toujours possible**
- quand il est possible, il est **lourd et coûteux** (horizons limités)
- la politique **dépend beaucoup de l'horizon** (pas 'anytime')
- **pas de garantie** théorique sur son regret pour un problème donné

Approximations de l'indice

Références: [Lai '85] dans le cas binaire et [Burnetas et Katehakis '03] dans le cas exponentiel canonique (avec des techniques différentes)

Méthode: relaxations du problème d'arrêt optimal

Asymptotique: le temps courant t est grand mais loin de l'horizon

Approximation: l'indice de Gittins ressemble à une **borne supérieure de confiance** de risque $1/n$:

$$IG(a) \approx \frac{S_a(t)}{N_a(t)} + \sqrt{\frac{\sigma_a^2 \log(t)}{N_a(t)}}$$

Plan de l’exposé

Principe d'optimisme

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

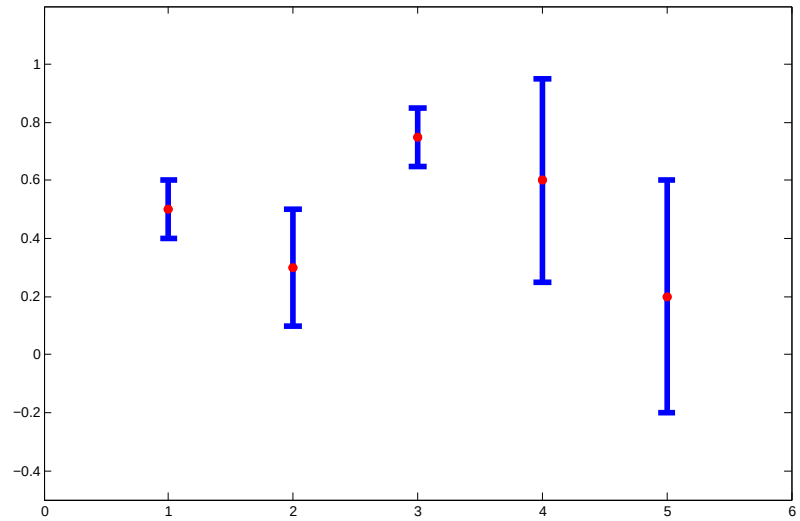
Fais comme si tu te trouvais dans l'environnement qui t'est le plus favorable parmi tous ceux qui rendent les observations suffisamment vraisemblables

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

Stratégies "Upper Confidence Bound" [Auer&al '02; Audibert&al '09]

UCB (Upper Confidence Bound) = établir une borne supérieure de l'intérêt de chaque action, et choisir celle qui est la plus prometteuse

$$U_a(t) = \frac{S_a(t)}{N_a(t)} + \sqrt{\frac{c\log(t)}{2N_a(t)}}$$

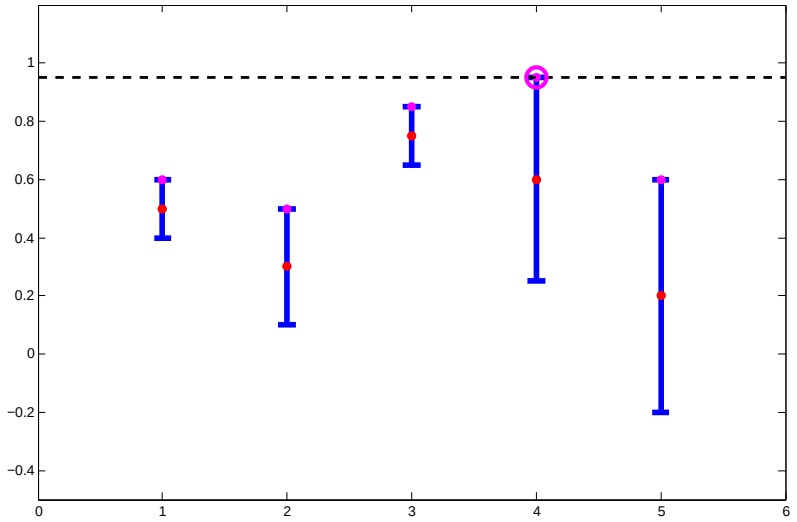


- Politique d'indice basée sur l'inégalité de Hoeffding

Stratégies "Upper Confidence Bound" [Auer&al '02; Audibert&al '09]

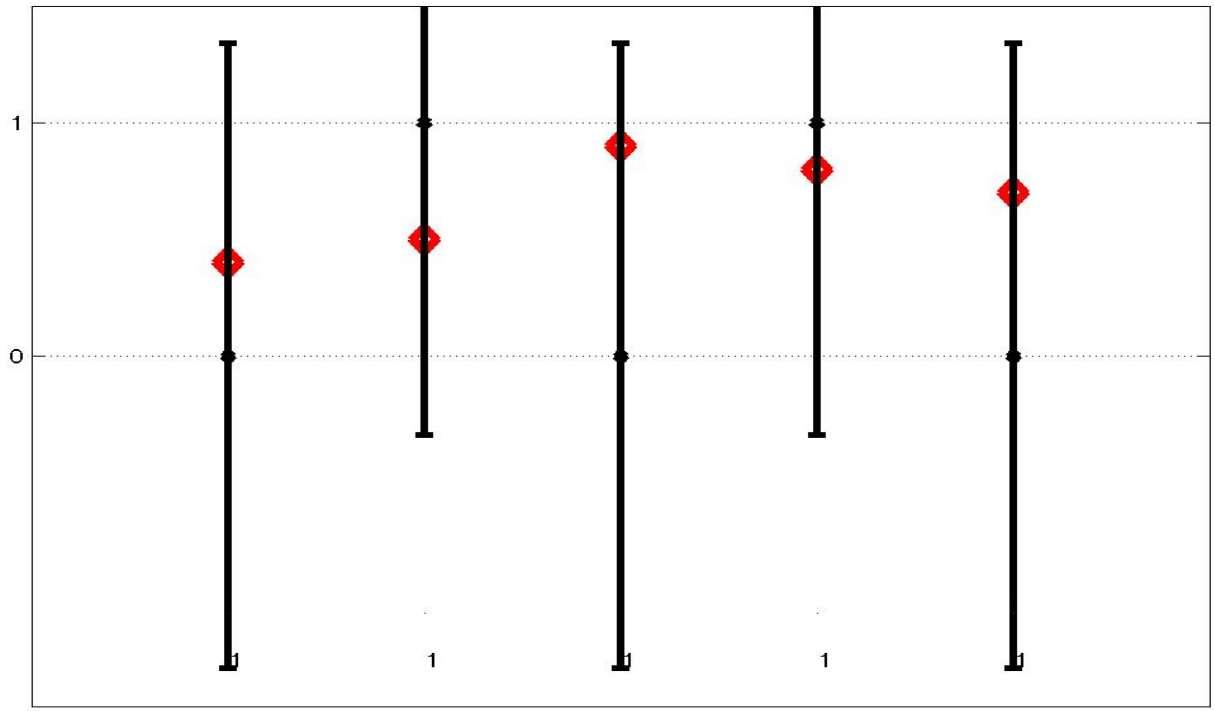
UCB (Upper Confidence Bound) = établir une borne supérieure de l'intérêt de chaque action, et choisir celle qui est la plus prometteuse

$$U_a(t) = \frac{S_a(t)}{N_a(t)} + \sqrt{\frac{c\log(t)}{2N_a(t)}}$$

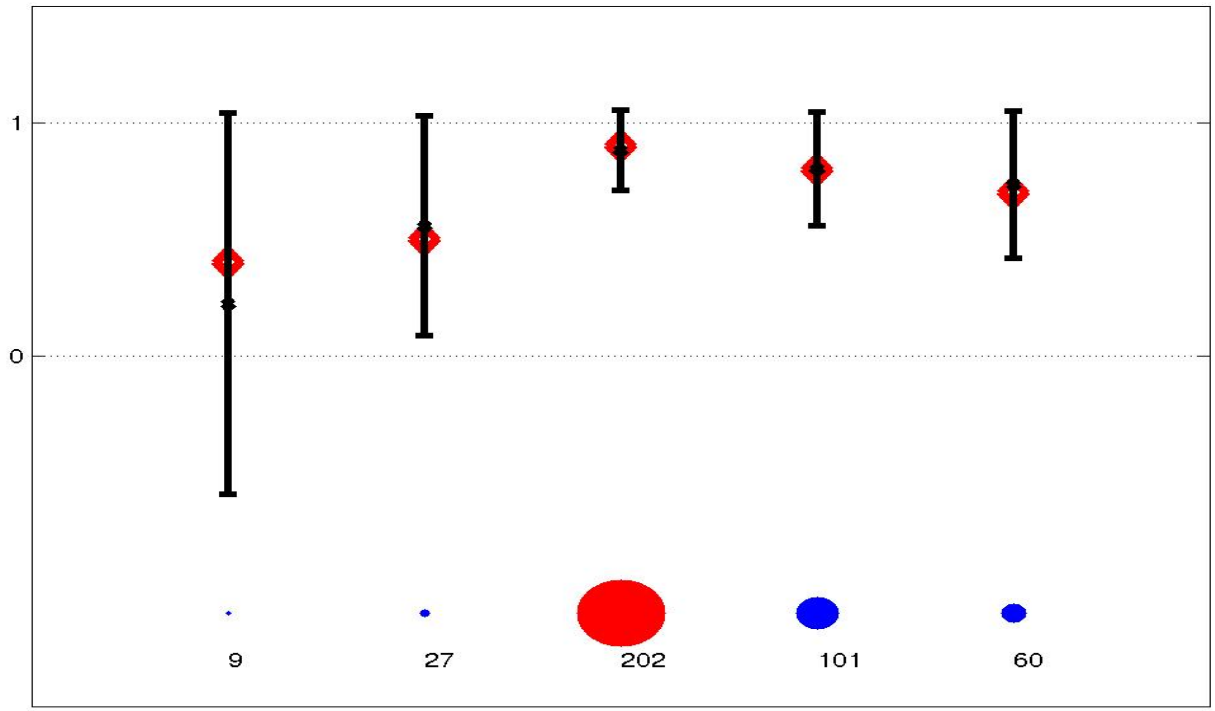


- Politique d'indice basée sur l'inégalité de Hoeffding
- Avantage : comportement facilement interprétable et "acceptable", étude non asymptotique du regret

UCB en action



UCB en action



Performance

Hypothèse : les récompenses X_t sont bornées, disons entre 0 et 1.

- [Auer&al 2002] le nombre de tirage du bras sous-optimal a satisfait :

$$E[N_a(n)] \leq \frac{8}{(\mu^* - \mu_a)^2} \log(n) + C$$

- sous-optimal : [Lai&Robbins '85] pour toute stratégie uniformément bonne,

$$\mathbb{E}[N_a(n)] \geq \left(\frac{1}{\mathrm{kl}\left(\mu_a, \mu^*\right)} + o(1) \right) \log(n),$$

où $\mathrm{kl}(p,q) = p\log(p/q) + (1-p)\log((1-p)/(1-q))$ est la divergence de Kullback-Leibler binaire.

- en particulier, mauvais pour les petites récompenses $\mu^* \ll 1$

Quelles bornes supérieures de confiance ?

Il faut des bornes non asymptotiques, auto-normalisées

Idée 0: Hoeffding = UCB, sous-optimal car variance sur-estimée

Idée 1: Bernstein = [Audibert & al '07] MAIS terme de reste catastrophique

$$U_a(t) = \bar{X}_a(t) + \sqrt{\frac{2\hat{V}_a(t)\log(t)}{N_a(t)}} + \frac{3\log(t)}{N_a(t)}$$

Idée 2: Cramer = possible dans la famille exponentielle canonique, cf [Garivier, Cappé '11]

Idée 3: vraisemblance empirique

Idées bayésiennes: quantiles d'un a posteriori [Kaufmann, Cappé, Garivier '11], ou même simplement tirage sous l'a posteriori [Thompson '33; Kaufmann, Korda, Munos '12] !

KL-UCB [Cappé, G., Maillard, Munos, & Stoltz]

Soit $\hat{P}_a(t) \in \mathfrak{M}_1(\mathbb{R})$ la mesure empirique des observations du bras a à l'instant t :

$$\hat{P}_a(t) = \frac{1}{N_a(t)} \sum_{s \leq t : A_s = a} \delta_{X_{a,t}}$$

Soit $\mathcal{F} \subset \mathfrak{M}_1(\mathbb{R})$ une classe de loi de probabilités, et soit $\Pi : \mathfrak{M}_1(\mathbb{R}) \to \mathcal{F}$. L'algorithme KL-UCB sur consiste à choisir

$$A_{t+1} = \arg\max_{a \in \mathcal{A}} U_a(t)$$

avec

$$U_a(t) = \max \left\{ E[P] : P \in \mathcal{F},\ \mathrm{KL}\left(\Pi_{\mathcal{F}} \left(\hat{P}_a(t) \right), P \right) \leq \frac{f(t)}{N_a(t)} \right\}$$

où, typiquement, $f(t) \approx \log(t)$.

Borne de regret

Pour borner le nombre $N_a(n)$ de tirages du bras sous-optimal $a \in \mathcal{A}$, on écrit pour tout $t \leq n$ où il a été tiré :

Décomposition :

$$\{A_{t+1} = a\} \subset \{U_{a^*}(t) < \mu^*\} \bigcup \{U_a(t) \geq \mu^*\}$$

Premier terme : contrôlé par les inégalités auto-normalisées car

$$U_{a^*}(t) < \mu^* \implies \mathrm{KL}\left(\Pi_{\mathcal{F}}\left(\hat{P}_{a^*}(t)\right), P_{a^*}\right) > \frac{f(t)}{N_{a^*}(t)}$$

Deuxième terme : implique avec grande proba que $N_a(t)$ est petit car $E[P_a] < \mu^*$

Exemple paramétrique : famille exponentielle canonique

Modèle $\mathcal{F} = P_{\theta_0} \in \{P_\theta : \theta \in \Theta\}$, où P_θ admet la densité

$$p_\theta(x) = \exp\left(x\theta - b(\theta) + c(x)\right)$$

et a pour espérance $\mu(\theta) = \dot{b}(\theta)$

Projection $\Pi_{\mathcal{F}}(Q) = P_{\mu^{-1}(E[Q])}$

Divergence

$$\mathrm{KL}(P_\beta; P_\theta) = I(\mu(\beta); \mu(\theta)) = b(\theta) - b(\beta) - \dot{b}(\beta)(\theta - \beta)$$

Indice

$$U_a(t) = \max\left\{\mu : I\left(\bar{X}_a(t); \mu\right) \leq \frac{f(t)}{N_a(t)}\right\}$$

Application : récompenses bornées [G. Cappé '11]

Borne de regret

Pour tout $\varepsilon > 0$, il existe $C_1, C_2(\varepsilon)$ et $\beta(\varepsilon)$ telles que pour n'importe que bras sous-optimal a, sous la politique KL-UCB,

$$\mathbb{E}[N_n(a)] \leq \frac{\log(n)}{\mathrm{kl}(\mu_a, \mu^*)}(1+\varepsilon) + C_1 \log(\log(n)) + \frac{C_2(\varepsilon)}{n^{\beta(\varepsilon)}}$$

- kl-UCB meilleur qu'UCB pour le même cadre d'applications
- *asymptotiquement optimal* pour les variables de Bernoulli : cf borne inférieure de Lai&Robbins, Burnetas&Katehakis : dans le modèle $\mathcal{F}$

$$N_a(n) \geq \left(\frac{1}{\inf_{P \in \mathcal{F} : E[P] > \mu^*} \mathrm{KL}\left(P_a, P\right)} + o(1) \right) \log(n),$$

Méthode de la vraisemblance empirique

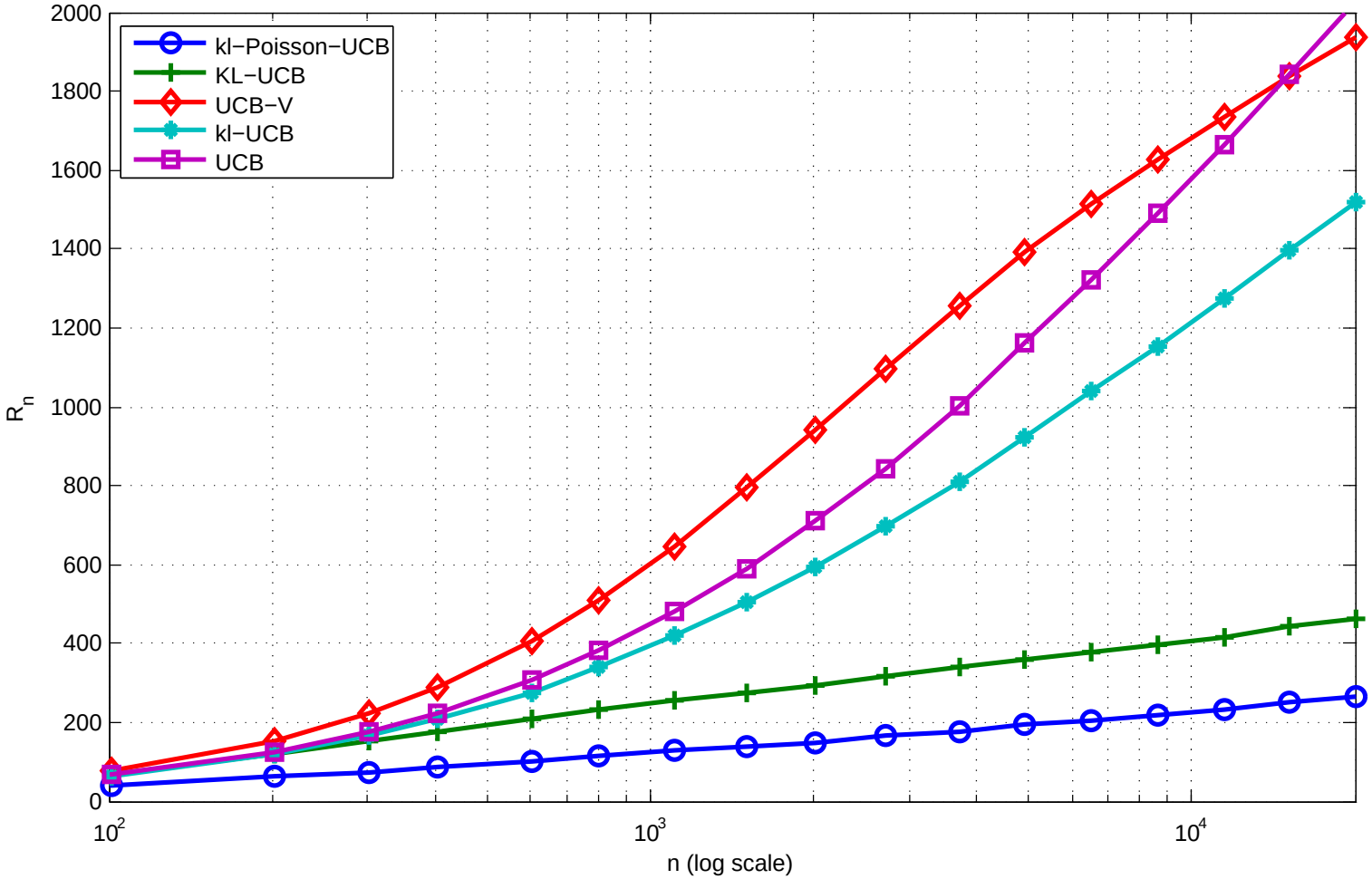
Quand $\mathrm{Var}[P_a] \ll \mu_a(1-\mu_a)$, la borne de confiance utilisée est pessimiste.
Estimation totalement non paramétrique : $\mathcal{F} = \mathfrak{M}_1([0,1])$ et $\Pi_{\mathcal{F}} = id$.

$$U_a(t) = \max\left\{ E[P] : P \in \mathcal{F},\ \mathrm{KL}\Big(\hat{P}_a(t), P\Big) \leq \frac{f(t)}{N_a(t)} \right\}$$

- Problème d'optimisation **numériquement simple**
- Probabilité de **couverture** : pour n'importe quelle loi P à support dans $[0,1]$, après n observations

$$P\big(U_n < E[P]\big) \leq 2e\, n\, e^{-f(n)}$$

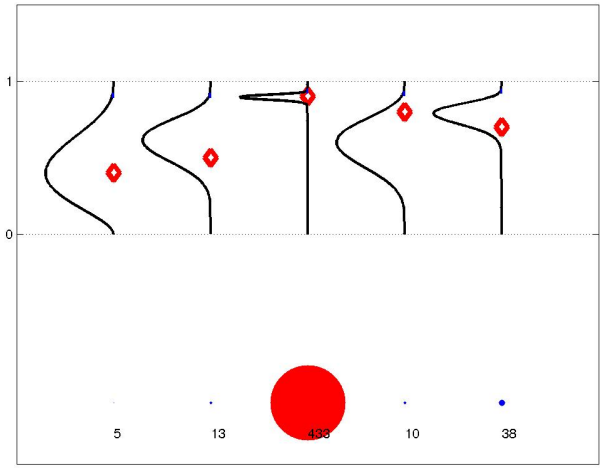
Comparatif sur un exemple



$P_a = \mathcal{P}\left(\frac{1}{2} + \frac{a}{3}\right)$ pour $1 \leq a \leq 6$ bras, tronquée à 10.

Approche bayésienne optimiste [Kaufmann, Cappé & G.]

Optimisme bayésien : on prend pour indice un quantiles des lois a posteriori

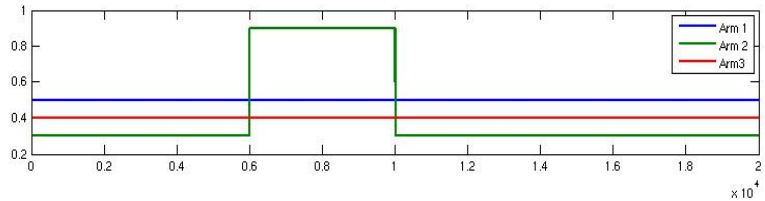


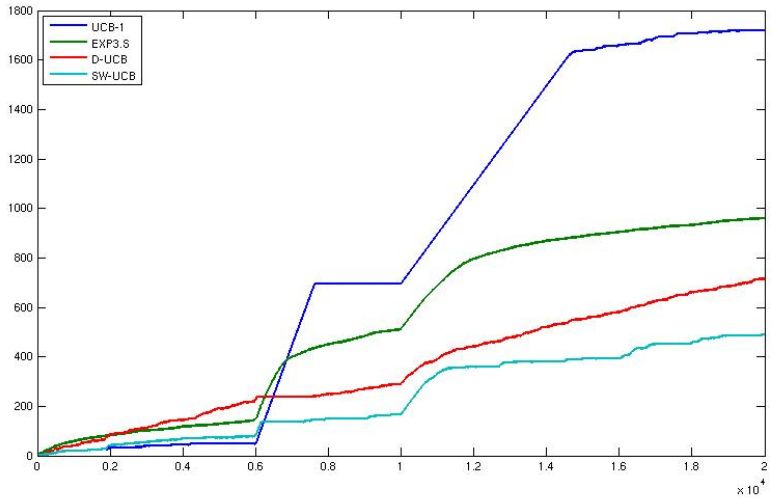
- Dans les contextes connus (récompenses gaussiennes, binaires), on montre les **mêmes garanties** que pour kl-UCB.
- Intérêt : l'a posteriori **s'adapte tout seul** au problème, et la méthode peut être mise en oeuvre dans beaucoup de contextes par simulation (ex: bandits linéaires sparses)
- Approche de Thompson [Thompson '33, Kaufmann, Korda& Munos '12] : on prend simplement pour indice **une valeur tirée sous l'a posteriori !**
- Question: Intérêt (fréquentiste) d'arrêter l'exploration (comme Gittins) quand on approche de l'horizon ?

Plan de l’exposé

Bandits non stationnaires [G. Moulines '11]

- Changepoint : les distributions des récompenses *varient brutalement*
- Objectif : *poursuivre le meilleur bras*
- Application : scanner à effet tunnel





- On étudie alors D-UCB et SW-UCB, 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 [Filippi, Cappé, G. & Szepesvári '10]

- 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

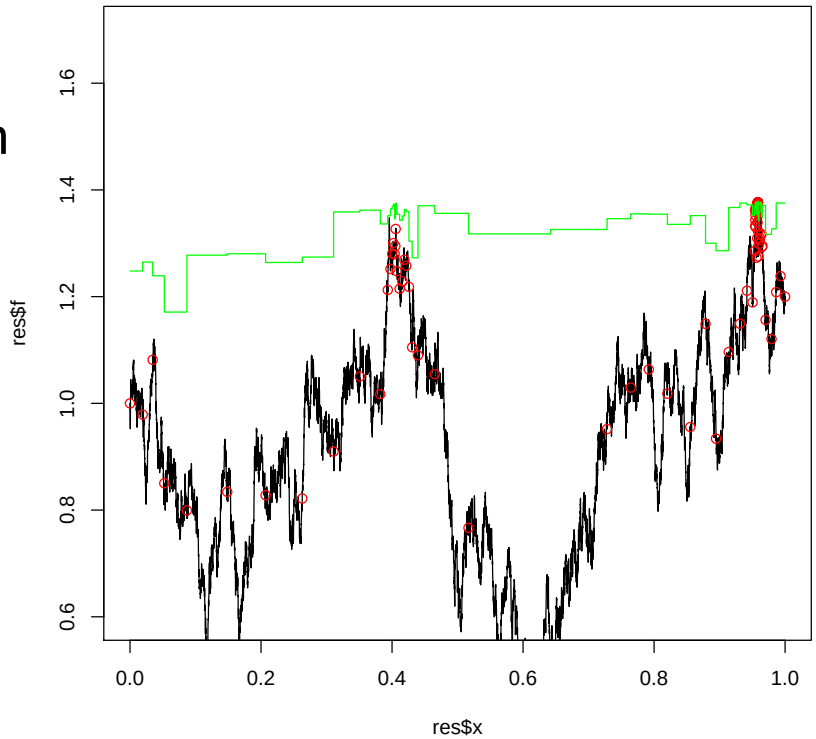
- Exemple : pour des récompenses binaires

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

- Application : publicité ciblée sur internet
- GLM-UCB : borne de regret dépendant de d et pas du nombre d'actions possibles

Optimisation stochastique [G. & Stoltz]

- 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 en cours : thèse de Marjorie Jalla sur l'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 : évaluer f au point $x \in C$ pour lequel l'intervalle de confiance pour $f(x)$ est le plus haut

Processus de Décision Markoviens

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

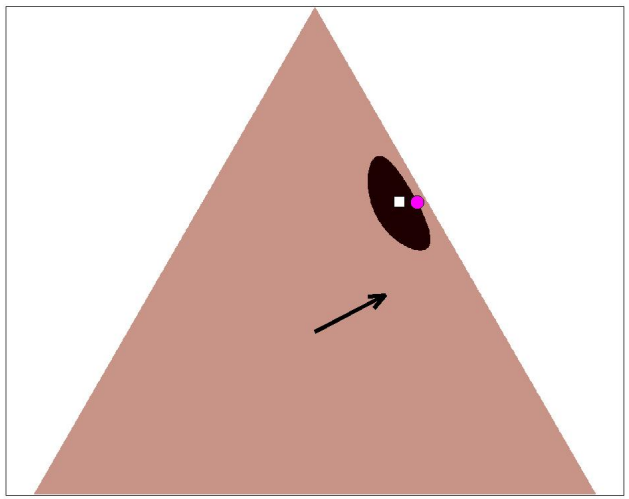
$$S_{t+1} \sim P(\cdot; S_t, A_t) \text{ et } R_t = r(S_t, A_t) + \varepsilon_t$$

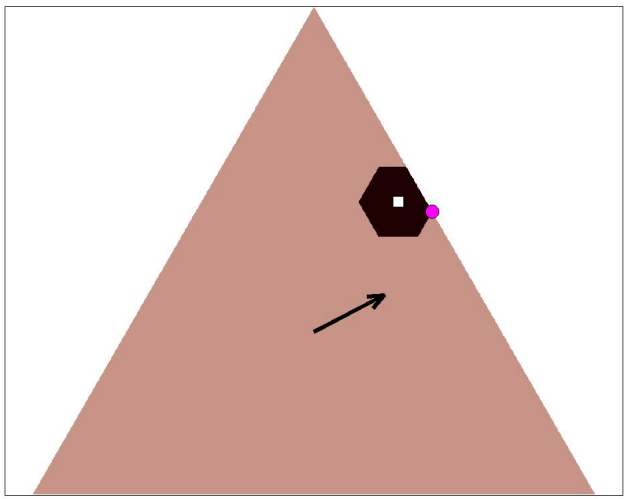
Meilleur modèle pour les communications numériques, mais aussi pour :

- la robotique
- la commande d'une batterie d'ascenseurs
- le routage de paquets sur internet
- l'ordonnancement de tâches
- la maintenance de machines
- les jeux
- le contrôle des réseaux sociaux
- le yield management
- la prévision de charge...

Optimisme pour les MDP [Filippi, Cappé & G. '10]

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

Exploration avec experts probabilistes

Espace de recherche : $B \subset \Omega$ discret

Experts probabilistes : $P_a \in \mathfrak{M}_1(\Omega)$ pour $a \in \mathcal{A}$

Requètes : à l'instant t, l'appel à l'expert A_t donne une réalisation $X_t = X_{A_t,t}$ indépendante de P_a

Objectif : trouver un maximum d'éléments distincts dans B en un minimum de requètes :

$$F_n = \mathrm{Card}\left(B \cap \{X_1, \ldots, X_n\}\right)$$

$\neq$ bandit : trouver deux fois le même élément ne sert à rien !

Oracle : joue l'expert qui a la plus grande "masse manquante"

$$A^*_{t+1} = \arg\max_{a \in \mathcal{A}} P_a\left(B \backslash \{X_1, \ldots, X_t\}\right)$$

Estimation de la masse manquante

Notations :

- $X_t \overset{iid}{\sim} P \in \mathfrak{M}_1(\Omega)$, $O_n(\omega) = \sum_{t=1}^{n} \mathbb{1}\{X_t = \omega\}$
- $Z_n(x) = \mathbb{1}\{O_n(\omega) = 0\}$
- $H_n(\omega) = \mathbb{1}\{O_n(\omega) = 1\}, \quad H_n = \sum_{\omega \in B} H_n(\omega)$

Problème : estimer la masse manquante

$$R_n = \sum_{\omega \in B} P(\omega) Z_n(\omega)$$

Good-Turing : "estimateur" $\hat{R}_n = H_n/n$ tq $\mathbb{E}[\hat{R}_n - R_n] \in [0, 1/n]$.

Concentration : par l'inégalité de McDiarmid, avec proba $1 - \delta$

$$\left| \hat{R}_n - E[\hat{R}_n] \right| \leq \sqrt{\frac{(2/n + p_{\max})^2 \; n \; \log(2/\delta)}{2}}$$

L'algorithme Good-UCB [Bubeck, Ernst & G.]

Algorithme optimiste basé sur l'estimateur de Good-Turing :

$$A_{t+1} = \arg\max_{a \in \mathcal{A}} \left\{ \frac{H_a(t)}{N_a(t)} + c\sqrt{\frac{\log(t)}{N_a(t)}} \right\}$$

- $N_a(t) =$ nombre de tirages de P_a jusqu'à l'instant t
- $H_a(t) =$ nombre d'éléments de B vus une seule fois (en tout) grâce à P_a
- $c =$ constante à régler pour garantir l'estimation simultanée correcte avec grande probabilité

Good-UCB en action

Optimalité macroscopique

Hypothèses :

- $\Omega = \mathcal{A} \times \{1, \dots, N\}$
- $\forall a \in \mathcal{A}, \forall j \in \{1, \dots, N\},\ P_a\left(\{(a, j)\}\right) = 1/N$

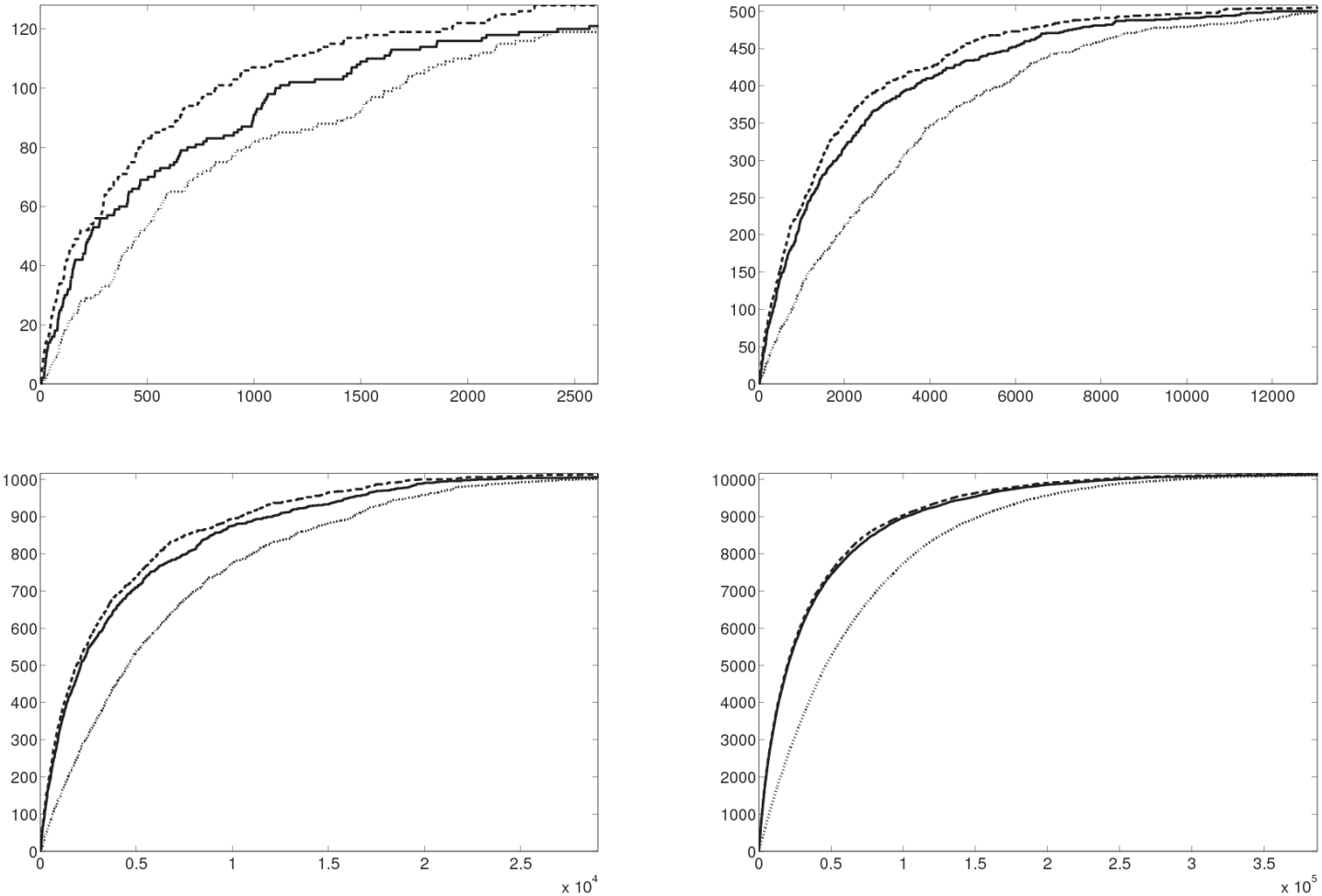
Limite macroscopique :

- $N \to \infty$
- $\forall a \in \mathcal{A},\ \mathrm{Card}\left(B \cap \{a\} \times \{1, \dots, N\}\right)/N \to q_a \in]0, 1[$

Optimalité macroscopique

Quand N tend vers l'infini, la performance de Good-UCB au cours du processus de découvert $t \mapsto F([Nt])$ converge uniformément vers celle de l'oracle $t \mapsto F^*([Nt])$ sur $\mathbb{R}^+$.

Illustration numérique



Nombre d'objets intéressants trouvés par Good-UCB (trait plein), l'oracle (pointillés épais), et par échantillonnage uniforme (pointillé léger) en fonction du temps pour des tailles $N = 128, N = 500$, $N = 1000$ et $N = 10000$, dans un environnement à 7 experts.

Bibliographie

1. **[Abbasi-Yadkori&al '11]** Yasin Abbasi-Yadkori, Dávid Pál, Csaba Szepesvári: Online Least Squares Estimation with Self-Normalized Processes: An Application to Bandit Problems CoRR abs/1102.2670: (2011)
2. **[Agrawal '95]** R. Agrawal. Sample mean based index policies with O(log n) regret for the multi-armed bandit problem. Advances in Applied Probability, 27(4) :1054-1078, 1995.
3. **[Audibert&al '09]** J-Y. Audibert, R. Munos, and Cs. Szepesvári. Exploration-exploitation tradeoff using variance estimates in multi-armed bandits. Theoretical Computer Science, 410(19), 2009
4. **[Auer&al '02]** P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. Machine Learning, 47(2) :235-256, 2002.
5. **[Bubeck, Ernst&G. '11]** Sébastien Bubeck, Damien Ernst, and Aurélien Garivier. Good-UCB : an optimistic algorithm for discovering unseen data, 2011.
6. **[De La Pena&al '04]** V.H. De La Pena, M.J. Klass, and T.L. Lai. Self-normalized processes : exponential inequalities, moment bounds and iterated logarithm laws. Annals of Probability, 32(3) :1902-1933, 2004.
7. **[Filippi, Cappé&Garivier '10]** S. Filippi, O. Cappé, and A. Garivier. Optimism in reinforcement learning and Kullback-Leibler divergence. In Allerton Conf. on Communication, Control, and Computing, Monticello, US, 2010.
8. **[Filippi, Cappé, G.& Szepesvari '10]** S. Filippi, O. Cappé, A. Garivier, and C. Szepesvari. Parametric bandits : The generalized linear case. In Neural Information Processing Systems (NIPS), 2010.
9. **[G.&Cappé '11]** A. Garivier and O. Cappé. The KL-UCB algorithm for bounded stochastic bandits and beyond. In 23rd Conf. Learning Theory (COLT), Budapest, Hungary, 2011.
10. **[G.&Leonardi '11]** A. Garivier and F. Leonardi. Context tree selection : A unifying view. Stochastic Processes and their Applications, 121(11) :2488-2506, Nov. 2011.
11. **[G.&Moulines '11]** A. Garivier and E. Moulines. On upper-confidence bound policies for non-stationary bandit problems. In Algorithmic Learning Theory (ALT), volume 6925 of Lecture Notes in Computer Science, 2011.
12. **[Kaufmann, Cappé & Garivier]**, On Bayesian Upper Confidence Bounds for Bandit Problems, Conference on Artificial Intelligence and Statistics (AISTAT), 2012.
13. **[Lai&Robins '85]** T.L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. Advances in Applied Mathematics, 6(1) :4-22, 1985.