

Apprentissage par renforcement et déviations auto-normalisées

Aurélien Garivier

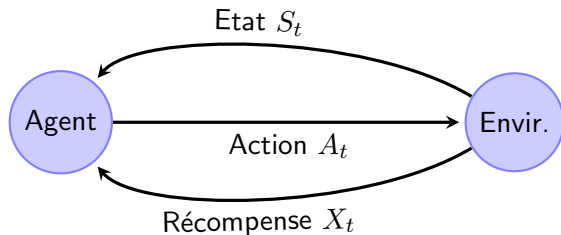
CNRS & Telecom ParisTech

24 janvier 2012

Plan de l'exposé

- 1 Apprentissage par renforcements
- 2 Inégalités auto-normalisées
- 3 Application en apprentissage par renforcement
 - Bandits classiques
 - Extensions du modèle

Apprentissage par renforcement



dilemme
exploration
|
exploitation

RL \neq apprentissage classique (notion de récompense)

RL \neq théorie des jeux (environnement indifférent)

Exemples: essais cliniques, marketing en ligne, recherche d'information, proposition de contenu, yield management, contrôle stochastique. . .

Essais cliniques séquentiels

On considère le cas de figure suivant :

- des patients atteints d'une certaine maladie 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)

Le "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 $(X_{a,t})_{a \in \mathcal{A}, t \geq 1}$ est indépendante

Règle d'allocation dynamique : $\pi = (\pi_1, \pi_2, \dots)$ telle que

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

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 : $S_n = X_1 + \dots + X_n$, $n \geq 1$
- Objectif: choisir π de manière à maximiser

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

où $\mu_a = E[P_a]$

- Objectif équivalent : minimiser le *regret*

$$R_n = n\mu^* - E[S_n] = \sum_{a: \mu_a < \mu^*} (\mu^* - \mu_a) \mathbb{E}[N_a(n)]$$

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

Principe d'optimisme

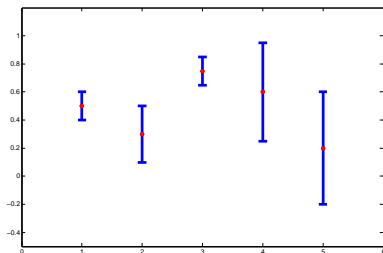
Algorithmes **optimistes** : [Lai&Robins '85; Agrawal '95]

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]

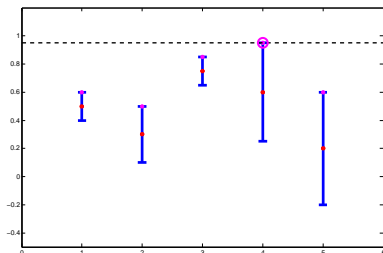
UCB (Upper Confidence Bound)
 = établir une borne supérieure de
 l'intérêt de chaque action, et choisir
 celle qui est la plus prometteuse



- **Avantage** : comportement facilement interprétable et "acceptable"
- *Politique d'indice* : on calcule un indice par bras et on choisit celui qui est le plus élevé, comme la politique optimale dans la modélisation bayésienne de [Gittins '79]

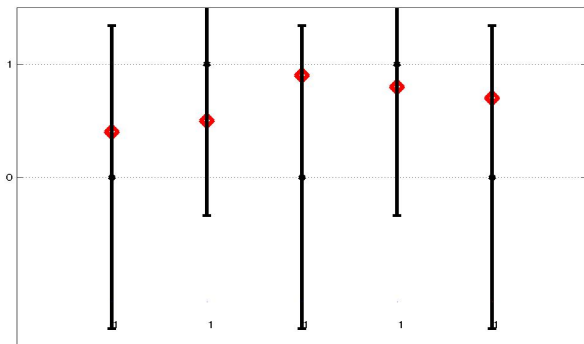
Stratégies "Upper Confidence Bound" [Auer&al '02; Audibert&al '09]

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

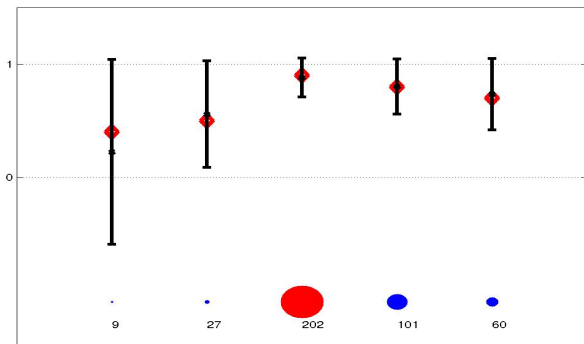


- **Avantage** : comportement facilement interprétable et "acceptable"
- *Politique d'indice* : on calcule un indice par bras et on choisit celui qui est le plus élevé, comme la politique optimale dans la modélisation bayésienne de [Gittins '79]

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 :

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

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

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

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

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

Plan de l'exposé

- 1 Apprentissage par renforcements
- 2 Inégalités auto-normalisées
- 3 Application en apprentissage par renforcement
 - Bandits classiques
 - Extensions du modèle

Cadre

L'étude du regret des algorithmes de type UCB conduit sur chaque bras au problème suivant :

Incréments $(X_t)_{t \geq 0}$ i.i.d., $\mathcal{F}_t = \sigma(X_1, \dots, X_t)$

Processus $S(n) = \sum_{t=1}^n \varepsilon_t X_t$

Option prévisible ε_t est borné, \mathcal{F}_{t-1} -mesurable

Borne $\mathbb{E}[\exp(\lambda X_t) | \mathcal{F}_{t-1}] \leq \exp(\phi(\lambda))$

Cas sous-gaussien centré : $\phi(\lambda) = \sigma^2 \lambda^2 / 2$

Crochet $V(n) = \sigma^2 \sum_{t=1}^n \varepsilon_t^2$.

Objectif obtenir des bornes de déviation pour $S(n)$
correctement renormalisé

Heuristique : le cas sous-gaussien déterministe

Dans le cas sous-gaussien centré,

$$\forall \lambda, W^\lambda(n) \doteq \exp\left(\lambda S(n) - \frac{\lambda^2}{2} V(n)\right)$$

est une sur-martingale positive.

Si $V(n)$ déterministe, on prend $\lambda = \sqrt{2x/V(n)}$ et on obtient :

$$P\left(\frac{S(n)}{\sqrt{V(n)}} > \sqrt{2x}\right) \leq \exp(-x)$$

On veut étendre ce résultat sans trop perdre sur la concentration.

Mélange : borne de [De La Peña & al. '04]

Idée : en "mélangeant" toutes les valeurs de λ

$$W(n) = \int W^\lambda(n) d\mathcal{N}(\lambda)$$

on a toujours $\mathbb{E}[W(n)] \leq 1$, et $W(n)$ se calcule explicitement.
Cela donne par exemple pour une gaussienne centrée réduite :

$$P \left(\frac{S(n)}{\sqrt{(V(n) + \sigma^2) \left(1 + \frac{1}{2} \log(V(n) + \sigma^2)\right)}} > \sqrt{2x} \right) \leq \exp(-x)$$

[De La Peña & al. '07] : le mélange est pris "aussi uniforme que possible" pour couvrir toutes les plages possibles de λ .

Autre approche : peeling [G.&Leonardi '11, G.&Moulines '11,...]

Au lieu d'intégrer en λ , on choisit *quelques* valeurs de λ , et on regarde ce que l'on perd *autour* de ces valeurs.

Le choix des valeurs retenues est fait pour "uniformiser les pertes" : pour la plage $(1 + \eta)^{k-1} < V(n) \leq (1 + \eta)^k$ on prend


$$\lambda_k = \sqrt{\frac{2x}{(1 + \eta)^{k-1/2}}}$$

On trouve pour tout $\delta \geq 1$, si $V(n) \leq n$:

$$P\left(\frac{S(n)}{\sqrt{V(n)}} \geq \sqrt{2x}\right) \leq \lceil \sqrt{ex} \log(n) \rceil \exp(-x)$$

ou bien :

$$P\left(\frac{S(n)}{\sqrt{V(n)}} \geq \sqrt{2x + 2c \log \log(V(n)) + \log(x) + \square}\right) \leq \zeta(c) \exp(-x)$$

Revient à un mélange avec une loi discrète (non plate) 

Comparatif

- Avantage du mélange : extension facile aux dimensions supérieures : si $S(n) = \sum_{t=1}^n \varepsilon_t X_t$ avec ε_t \mathcal{F}_{t-1} -mesurable à valeur dans \mathbb{R}^d et $\mathbb{E}[\exp(\lambda X_t) | \mathcal{F}_{t-1}] \leq \exp(\sigma^2 \lambda^2 / 2)$, alors en notant $V(n) = \sigma^2 \sum_{t=1}^n \varepsilon_t \varepsilon_t^T$ on a

$$\forall \Lambda \in \mathbb{R}^d, \exp\left(\langle \Lambda, S(n) \rangle - \frac{1}{2} \|\Lambda\|_{V(n)}^2\right) \text{ sur-martingale,}$$

et par exemple [Abbasi-Yadkori&al '11] :

$$P\left(\|S(n)\|_{V(n)^{-1}} \geq \sqrt{2 \log\left(\frac{\sqrt{\det(V(n) + v) / \det(v)}}{\delta}\right)}\right) \leq \delta$$

- Avantage du peeling : pas spécifique au cas sous-gaussien.

Hypothèses

Processus $S(n) = \sum_{t=1}^n \varepsilon_t X_t$, \mathcal{F}_t -mesurable

Option prévisible ε_t à valeur dans $\{0, 1\}$, \mathcal{F}_{t-1} -mesurable

Incréments il existe une fonction $\phi :]\lambda_1, \lambda_2[\rightarrow \mathbb{R}$ telle que pour tout $\lambda \in]\lambda_1, \lambda_2[$ et pour $t \geq 1$,

$$\mathbb{E} [\exp(\lambda X_t) | \mathcal{F}_{t-1}] \leq \exp(\phi(\lambda))$$

où ϕ est convexe $\mathcal{C}^\infty(] \lambda_1, \lambda_2[)$, $\phi'(\mu) = 0$

Transformée de Fenchel-Legendre $I(\cdot; \mu)$ définie par

$$I(x; \mu) = \sup_{\lambda \in \mathbb{R}} \{ \lambda x - \phi(\lambda) \} ;$$

convexe, \mathcal{C}^∞ sur \mathcal{D}_I contenant 0, tq $I(\mu; \mu) = 0$, et $\forall x, I(x; \mu) < \infty \implies \exists \lambda(x) \in]\lambda_1, \lambda_2[$ tq

$$\phi'(\lambda(x)) = x \quad \text{et} \quad I(x; \mu) = \lambda(x)x - \phi(\lambda(x))$$

Cas sous-gaussien centré $\phi(\lambda) = \sigma^2 \lambda^2 / 2$

Formulation alternative

Processus $(S_t)_{t \geq 0}$ réel temps discret tq $S_0 = 0$ adapté à $(\mathcal{F}_t)_{t \geq 0}$

Incréments $X_t = S_t - S_{t-1}$ dominés : il existe une fonction $\phi :]\lambda_1, \lambda_2[\rightarrow \mathbb{R}$ telle que pour tout $\lambda \in]\lambda_1, \lambda_2[$ et pour $t \geq 1$,

$$\mathbb{E}[\exp(\lambda X_t) | \mathcal{F}_{t-1}] \leq \exp(\phi(\lambda))$$

où ϕ est convexe $C^\infty(] \lambda_1, \lambda_2[)$, $\phi'(\mu) = 0$

Transformée de Fenchel-Legendre $I(\cdot; \mu)$ définie par

$$I(x; \mu) = \sup_{\lambda \in \mathbb{R}} \{ \lambda x - \phi(\lambda) \} ;$$

convexe, C^∞ sur \mathcal{D}_I contenant 0, tq $I(\mu; \mu) = 0$, et $\forall x, I(x; \mu) < \infty \implies \exists \lambda(x) \in]\lambda_1, \lambda_2[$ tq

$$\phi'(\lambda(x)) = x \quad \text{et} \quad I(x; \mu) = \lambda(x)x - \phi(\lambda(x))$$

Cas sous-gaussien centré $\phi(\lambda) = \sigma^2 \lambda^2 / 2$

Objectif

Soit μ la moyenne commune des X_t

Formulation 1 On veut obtenir une borne pour les déviations normalisées de $\bar{X}(n) = \frac{S(n)}{N(n)}$ autour de μ où $N(n) = \sum_{t=1}^n \varepsilon_t$:

$$\frac{|\bar{X}(n) - \mu|}{\sqrt{N(n)}}$$

Formulation 2 On veut obtenir une borne **le maximum** des déviations normalisées de S_t autour de μ :

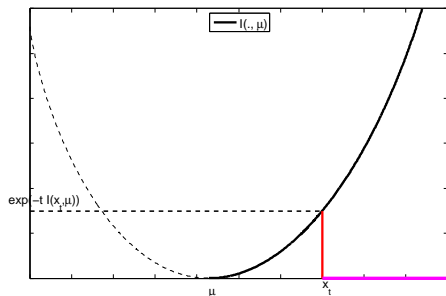
$$\max_{t \leq n} \frac{|\bar{X}_t - \mu|}{\sqrt{t}}$$

Déviations et borne de Chernoff

$$\mathbb{E} [\exp (\lambda S_t - t\phi(\lambda))] \leq 1$$

si $\bar{X}_t = S_t/t$, et $x_t \geq \mu$, donne
pour $\lambda = \lambda(x_t)$:

$$P(\bar{X}_t \geq x_t) \leq \exp(-tI(x_t; \mu))$$



Autre formulation :

$$P(I(\bar{X}_t; \mu) \geq I(x_t; \mu), \bar{X}_t \geq \mu) \leq \exp(-tI(x_t; \mu))$$

soit, en posant $\delta = tI(x_t; \mu)$,

$$P(tI(\bar{X}_t; \mu) \geq \delta, \bar{X}_t \geq \mu) \leq \exp(-\delta)$$

Intervalles de confiance de risque α : I -voisinage de \bar{X}_t

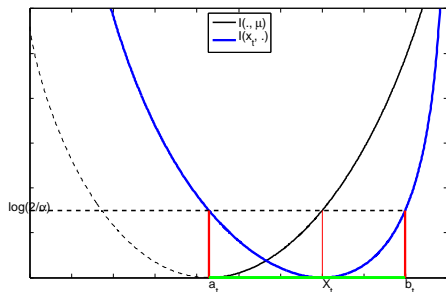
$$[a_t, b_t] = \left\{ \mu : tI(\bar{X}_t; \mu) \leq \log \frac{2}{\alpha} \right\}$$

Déviations et borne de Chernoff

$$\mathbb{E} [\exp (\lambda S_t - t\phi(\lambda))] \leq 1$$

si $\bar{X}_t = S_t/t$, et $x_t \geq \mu$, donne
pour $\lambda = \lambda(x_t)$:

$$P(\bar{X}_t \geq x_t) \leq \exp(-tI(x_t; \mu))$$



Autre formulation :

$$P(I(\bar{X}_t; \mu) \geq I(x_t; \mu), \bar{X}_t \geq \mu) \leq \exp(-tI(x_t; \mu))$$

soit, en posant $\delta = tI(x_t; \mu)$,

$$P(tI(\bar{X}_t; \mu) \geq \delta, \bar{X}_t \geq \mu) \leq \exp(-\delta)$$

Intervalles de confiance de risque α : I -voisinage de \bar{X}_t

$$[a_t, b_t] = \left\{ \mu : tI(\bar{X}_t; \mu) \leq \log \frac{2}{\alpha} \right\}$$

Résultats

Borne générale : version auto-normalisée

Pour tout $\delta > 0$,

$$P \left(I(\bar{X}(n); \mu) \geq \frac{\delta}{N(n)} \right) \leq 2e^{\lceil \delta \log(n) \rceil} \exp(-\delta)$$

Cas log-concave

Si $I(\cdot; \mu)$ est log-concave,

$$P \left(I(\bar{X}(n); \mu) \geq \frac{\delta}{N(n)} \right) \leq 2\sqrt{e} \left\lceil \frac{\sqrt{\delta}}{2} \log(n) \right\rceil \exp(-\delta)$$

Résultats : formulation 2

Borne générale

Pour tout $\delta > 0$,

$$P(\exists t \in \{1, \dots, n\} : tI(\bar{X}_t; \mu) \geq \delta) \leq 2e^{\lceil \delta \log(n) \rceil} \exp(-\delta)$$

Cas log-concave

Si $I(\cdot; \mu)$ est log-concave,

$$P(\exists t \in \{1, \dots, n\} : tI(\bar{X}_t; \mu) \geq \delta) \leq 2\sqrt{e} \left\lceil \frac{\sqrt{\delta}}{2} \log(n) \right\rceil \exp(-\delta)$$

Schéma de preuve

- Si $t_k = \lfloor (1 + \eta)^k \rfloor$ et $D = \lceil \log(n) / \log(1 + \eta) \rceil$,

$$P \left(\bigcup_{t=1}^n \{tI(\bar{X}_t; \mu) \geq \delta\} \right) \leq \sum_{k=1}^D P \left(\bigcup_{t=t_{k-1}+1}^{t_k} \{tI(\bar{X}_t; \mu) \geq \delta\} \right)$$

- Si λ_k tq $I(x_{t_k}; \mu) = \lambda_k x_{t_k} - \phi(\lambda_k)$, pour $t_{k-1} < t \leq t_k$:

$$tI(\bar{X}_t; \mu) \geq \delta \text{ et } \bar{X}_t \geq \mu \implies W_t^k \geq \exp\left(\frac{\delta}{1 + \eta}\right)$$

où $W_t^k = \exp(\lambda_k S_t - t\phi(\lambda_k))$ est une sur-martingale

- Or par l'inégalité maximale

$$P \left(\bigcup_{t=t_{k-1}+1}^{t_k} \left\{ W_t^k \geq \exp\left(\frac{\delta}{1 + \eta}\right) \right\} \right) \leq \exp\left(-\frac{\delta}{1 + \eta}\right)$$

- On conclut en choisissant $\eta = 1/(\delta - 1)$

Observations non stationnaires [G. & Moulines '11]

- $(X_t)_t$ indépendantes et bornées par B , d'espérances μ_t ne variant pas trop vite (ou pas trop souvent).
- Estimateur escompté : pour $\gamma \in]0, 1[$,

$$\bar{X}_\gamma(n) = S_\gamma(n)/N_\gamma(n)$$

où $S_\gamma(n) = \sum_{t=1}^n \gamma^{n-t} \varepsilon_t X_t$ et $N_\gamma(n) = \sum_{t=1}^n \gamma^{n-t} \varepsilon_t$

- Décomposition biais-variance : si $M_\gamma(n) = \sum_{t=1}^n \gamma^{n-t} \varepsilon_t \mu_t$,

$$\bar{X}_\gamma(n) - \mu_n = \underbrace{\bar{X}_\gamma(n) - \frac{M_\gamma(n)}{N_\gamma(n)}}_{\text{}} + \frac{M_\gamma(n)}{N_\gamma(n)} - \mu_n$$

- Contrôle du terme de variance : pour tout $\eta > 0$,

$$P \left(\frac{S_\gamma(n) - M_\gamma(n)}{\sqrt{N_{\gamma^2}(n)}} \geq \delta \right) \leq \left\lceil \frac{\log \nu_\gamma(n)}{\log(1 + \eta)} \right\rceil \exp \left(-\frac{2\delta^2}{B^2} \left(1 - \frac{\eta^2}{16} \right) \right)$$

où $\nu_\gamma(n) = \sum_{t=1}^n \gamma^{n-t} < \min\{(1 - \gamma)^{-1}, n\}$.

Modèle exponentiel canonique [G. & Cappé '11]

Modèle $P_{\theta_0} \in \{P_{\theta} : \theta \in \Theta\}$, où P_{θ} admet la densité

$$p_{\theta}(x) = \exp(x\theta - b(\theta) + c(x)) .$$

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

Divergence

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

Exemple 1 loi de Poisson : $I(x, y) = y - x + x \log \frac{x}{y}$

Exemple 2 loi bornée $X_t \in [0, 1]$:

$$I(x, y) = \text{kl}(x, y) = x \log \frac{x}{y} + (1-x) \log \frac{1-x}{1-y} \geq 2(x-y)^2$$

Intervalle de confiance

$$\begin{aligned} & \left\{ \theta \in \Theta : N(n) \text{KL} \left(P_{\mu^{-1}(\bar{X}(n))}; P_{\theta} \right) \leq \delta \right\} \\ & = \left\{ \theta \in \Theta : I(\bar{X}(n); \mu(\theta)) \leq \frac{\delta}{N(n)} \right\} . \end{aligned}$$

Lois multinomiales [G. & Leonardi '11]

Lemme: réduction aux lois de Bernoulli

Si $P, Q \in \mathfrak{M}_1(\mathcal{A})$,

$$\text{KL}(P; Q) \leq \sum_{x \in \mathcal{A}} \text{kl}(P(x); Q(x))$$

Corollaire: Voisines KL pour multinomiales

Si $X_1, \dots, X_n \sim P_0 \in \mathfrak{M}_1(\mathcal{A})$ iid, et $\hat{P}_t(k) = \sum_{s=1}^t \mathbb{1}\{X_s = k\}/t$

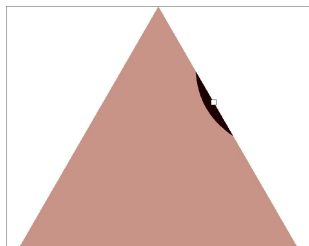
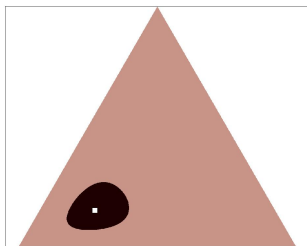
$$P \left(\exists t \in \{1, \dots, n\} : \text{KL}(\hat{P}_t; P_0) \geq \frac{\delta}{t} \right) \\ \leq 2e (\delta \log(n) + |\mathcal{A}|) \exp \left(-\frac{\delta}{|\mathcal{A}|} \right)$$

KL-balls [Filippi, G. & Cappé '10]

Suite $(R_t)_{t \leq n}$ de régions de confiance “de type Sanov” pour P_0 simultanément valides avec probabilité $1 - \alpha$ en choisissant des voisinages de Kullback-Leibler du maximum de vraisemblance :

$$R_t = \left\{ Q \in \mathfrak{M}_1(\mathcal{A}) : \text{KL}(\hat{P}_t; Q) \leq \frac{\delta}{t} \right\},$$

avec δ tel que $2e(\delta \log(n) + |\mathcal{A}|) \exp(-\delta/|\mathcal{A}|) = \alpha$.



Plan de l'exposé

- 1 Apprentissage par renforcements
- 2 Inégalités auto-normalisées
- 3 Application en apprentissage par renforcement
 - Bandits classiques
 - Extensions du modèle

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 $\Pi : \mathfrak{M}_1(\mathbb{R}) \rightarrow \mathcal{F}$. L'algorithme KL-UCB sur consiste à choisir

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

avec

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

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 :

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

Premier terme : contrôlé par les inégalités auto-normalisées car

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

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

$$p_{\theta}(x) = \exp(x\theta - b(\theta) + c(x)) .$$

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

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

Divergence

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

Indice

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

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,

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

- 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}

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

Autre choix pour les récompenses bornées

Quand $\text{Var}[P_a] \ll \mu_a(1 - \mu_a)$, la borne de confiance utilisée est pessimiste.

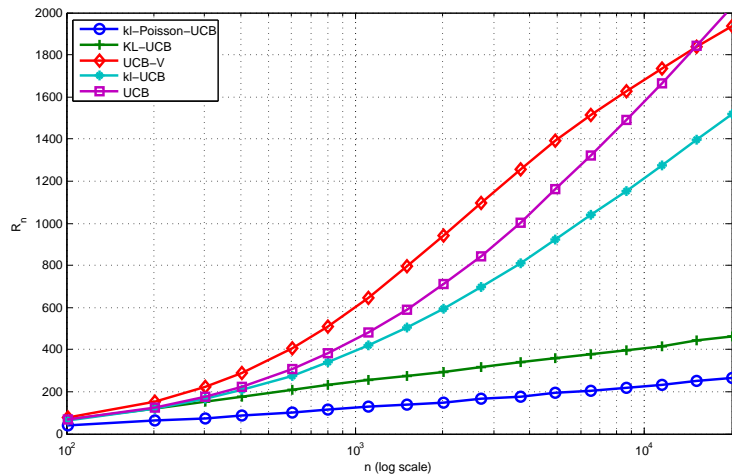
Estimation totalement non paramétrique : $\mathcal{F} = \mathfrak{M}_1([0, 1])$ et $\Pi_{\mathcal{F}} = id$.

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

- problème d'optimisation numériquement simple
- l'idée "intermédiaire" d'estimer la variance n'est pas facile à mettre en oeuvre efficacement, cf Bernstein et UCB-V :

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

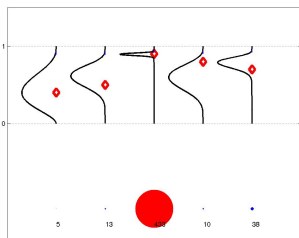
Comparatif sur un exemple



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

Approche bayésienne [Kaufmann, Cappé & G.]

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

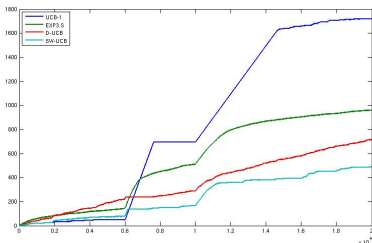
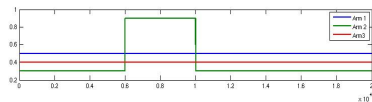


Avantage : peut être mis en oeuvre dans beaucoup de contextes par simulation (ex: bandits linéaires sparses).

Approche de Thompson : on prend simplement pour indice une valeur tirée sous l'a posteriori !

Bandits non stationnaires [G. Moulines '11]

- **Changepoint** : les distributions des récompenses *variant 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 $\mu : \mathbb{R} \rightarrow \mathbb{R}$ est la fonction de lien dans un modèle linéaire généralisé

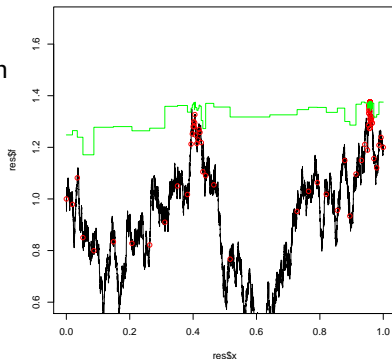
- Exemple : pour des récompenses binaires

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

- 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 $f : C \subset \mathbb{R}^d \rightarrow \mathbb{R}$ 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 :

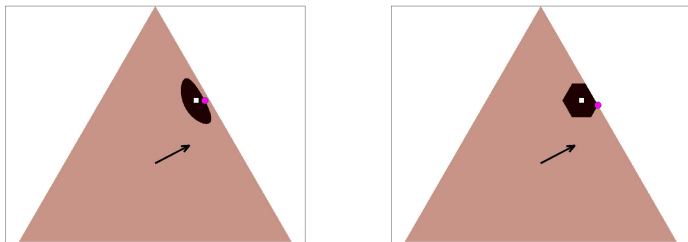
$$S_{t+1} \sim P(\cdot; S_t, A_t) \text{ et } R_t = r(S_t, A_t) + \varepsilon_t$$

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

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 $a \in \mathcal{A}$

Requêtes : 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 requêtes :

$$F_n = \text{Card} (B \cap \{X_1, \dots, X_n\})$$

≠ bandit : trouver deux fois le même élément ne sert à rien !

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

$$A_{t+1}^* = \arg \max_{a \in \mathcal{A}} P_a (B \setminus \{X_1, \dots, X_t\})$$

Estimation de la masse manquante

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

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 :

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

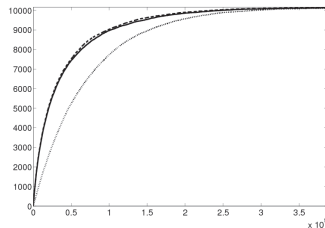
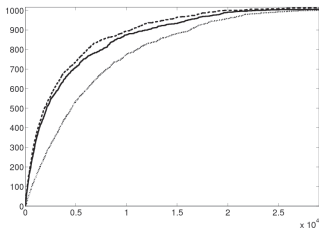
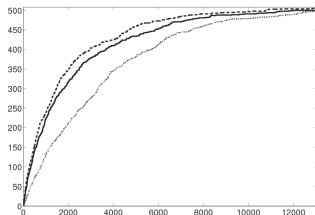
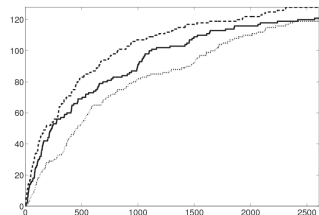
Limite macroscopique :

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

Optimalité macroscopique

Quand N tend vers l'infini, la performance de Good-UCB au cours du processus de découverte $t \mapsto F([Nt])$ 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

- [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)
- [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.
- [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
- [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.
- [Bubeck, Ernst&G. '11]** Sébastien Bubeck, Damien Ernst, and Aurélien Garivier. Good-UCB : an optimistic algorithm for discovering unseen data, 2011.
- [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.
- [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.
- [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.
- [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.
- [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.
- [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.
- [Lai&Robins '85]** T.L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1) :4-22, 1985.