# Analyses d'algorithmes pour l'estimation et l'optimisation stochastiques

Aurélien Garivier

CNRS & Telecom ParisTech

28 novembre 2011

# Deux principes d'ingénieur

## Minimum Description Length

Choisis le modèle qui permet la plus courte description des données.

## Paradigme optimiste

Parmi tous les environnements qui rendent les observations suffisamment vraisemblables, fais comme si tu étais dans celui qui t'est le plus favorable.

# Plan de l'exposé


1 Arbres de contextes et mémoire variable
- Noyaux, processus et simulation exacte
- Estimation

2 Quelques inégalités auto-normalisées

3 Apprentissage par renforcements
- Problèmes de bandits classique
- Extensions du modèle

# Mémoire adaptative

- Compression de données:

  **t r y i n g _ v a n i l l a _ q u i e t**

- Linguistique:

  **L o n g t e m p s , j e   m e   s u i s   c o u c h é   d e   b o n n e   h e u r e .   P a r f o i s , . . .**

- Processus de renouvellement:

  **1 0 0 1 0 1 0 0 0 0 1 1 0 0 1 . . .**

- Musique, biologie, optimisation, . . .

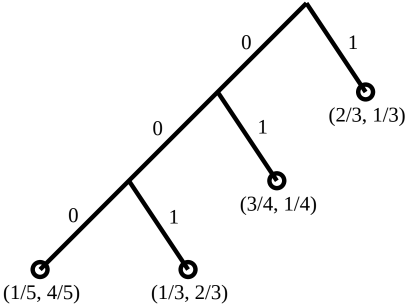
# Arbres de contextes probabilisés

Un arbre de contexte probabilisé (CTS) ou Chaîne de Markov d'ordre variable (VLMC) est une chaine de Markov dont l'ordre est autorisé à dépende des valeurs prises dans le passé.

$$T = \{1, 10, 100, 000\}$$

$$\mathbb{P}(X_1^4 = 00110 | X_{-1}^0 = 10)$$

$$\begin{aligned} = \quad & \mathbb{P}(X_1 = 0 | X_{-1}^0 = 10) \\ \times \quad & \mathbb{P}(X_2 = 0 | X_{-1}^1 = 100) \\ \times \quad & \mathbb{P}(X_3 = 1 | X_{-1}^2 = 1000) \\ \times \quad & \mathbb{P}(X_4 = 1 | X_{-1}^3 = 10001) \\ \times \quad & \mathbb{P}(X_5 = 0 | X_{-1}^4 = 100011) \end{aligned}$$



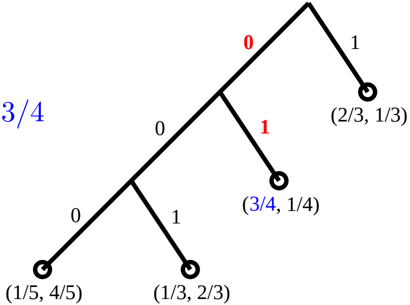
# Arbres de contextes probabilisés

Un arbre de contexte probabilisé (CTS) ou Chaîne de Markov d'ordre variable (VLMC) est une chaine de Markov dont l'ordre est autorisé à dépende des valeurs prises dans le passé.

$$T = \{1, 10, 100, 000\}$$

$$\mathbb{P}(X_1^4 = 00110 | X_{-1}^0 = 10)$$

$$\begin{aligned}
= \quad & \mathbb{P}(X_1 = 0 | X_{-1}^0 = 10) \qquad 3/4 \\
\times \quad & \mathbb{P}(X_2 = 0 | X_{-1}^1 = 100) \\
\times \quad & \mathbb{P}(X_3 = 1 | X_{-1}^2 = 1000) \\
\times \quad & \mathbb{P}(X_4 = 1 | X_{-1}^3 = 10001) \\
\times \quad & \mathbb{P}(X_5 = 0 | X_{-1}^4 = 100011)
\end{aligned}$$



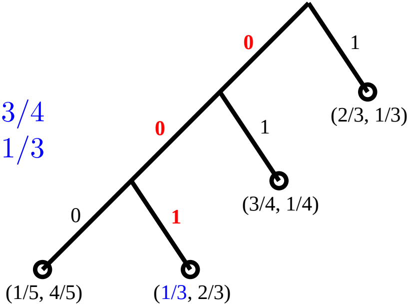
# Arbres de contextes probabilisés

Un arbre de contexte probabilisé (CTS) ou Chaîne de Markov d'ordre variable (VLMC) est une chaine de Markov dont l'ordre est autorisé à dépende des valeurs prises dans le passé.

$$T = \{1, 10, 100, 000\}$$

$$\mathbb{P}(X_1^4 = 00110 | X_{-1}^0 = 10)$$

$$\begin{aligned}
= \quad & \mathbb{P}(X_1 = 0 | X_{-1}^0 = 10) & \quad 3/4 \\
\times \quad & \mathbb{P}(X_2 = 0 | X_{-1}^1 = 100) & \quad 1/3 \\
\times \quad & \mathbb{P}(X_3 = 1 | X_{-1}^2 = 1000) & \\
\times \quad & \mathbb{P}(X_4 = 1 | X_{-1}^3 = 10001) & \\
\times \quad & \mathbb{P}(X_5 = 0 | X_{-1}^4 = 100011) &
\end{aligned}$$

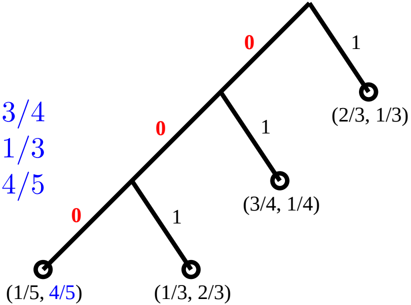


# Arbres de contextes probabilisés

Un arbre de contexte probabilisé (CTS) ou Chaîne de Markov d'ordre variable (VLMC) est une chaine de Markov dont l'ordre est autorisé à dépende des valeurs prises dans le passé.

$$T = \{1, 10, 100, 000\}$$

$$\begin{array}{lll} & \mathbb{P}(X_1^4 = 00110 | X_{-1}^0 = 10) & \\ = & \mathbb{P}(X_1 = 0 | X_{-1}^0 = 10) & 3/4 \\ \times & \mathbb{P}(X_2 = 0 | X_{-1}^1 = 100) & 1/3 \\ \times & \mathbb{P}(X_3 = 1 | X_{-1}^2 = 1000) & 4/5 \\ \times & \mathbb{P}(X_4 = 1 | X_{-1}^3 = 10001) & \\ \times & \mathbb{P}(X_5 = 0 | X_{-1}^4 = 100011) & \end{array}$$



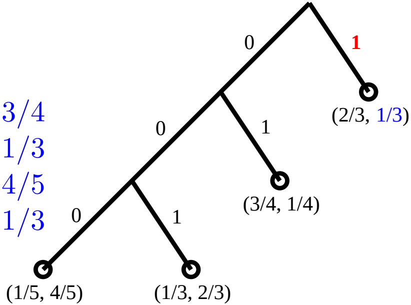
# Arbres de contextes probabilisés

Un arbre de contexte probabilisé (CTS) ou Chaîne de Markov d'ordre variable (VLMC) est une chaine de Markov dont l'ordre est autorisé à dépende des valeurs prises dans le passé.

$$T = \{1, 10, 100, 000\}$$

$$\mathbb{P}(X_1^4 = 00110 | X_{-1}^0 = 10)$$

$$\begin{array}{cll} = & \mathbb{P}(X_1 = 0 | X_{-1}^0 = 10) & 3/4 \\ \times & \mathbb{P}(X_2 = 0 | X_{-1}^1 = 100) & 1/3 \\ \times & \mathbb{P}(X_3 = 1 | X_{-1}^2 = 1000) & 4/5 \\ \times & \mathbb{P}(X_4 = 1 | X_{-1}^3 = 10001) & 1/3 \\ \times & \mathbb{P}(X_5 = 0 | X_{-1}^4 = 100011) & \end{array}$$



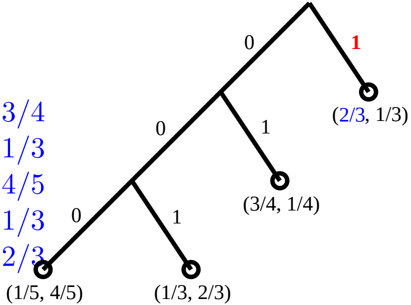
# Arbres de contextes probabilisés

Un arbre de contexte probabilisé (CTS) ou Chaîne de Markov d'ordre variable (VLMC) est une chaine de Markov dont l'ordre est autorisé à dépende des valeurs prises dans le passé.

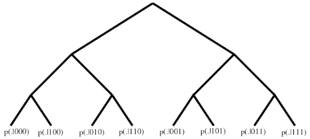
$$T = \{1, 10, 100, 000\}$$

$$\mathbb{P}(X_1^4 = 00110 | X_{-1}^0 = 10)$$

$$\begin{array}{lll} = & \mathbb{P}(X_1 = 0 | X_{-1}^0 = 10) & 3/4 \\ \times & \mathbb{P}(X_2 = 0 | X_{-1}^1 = 100) & 1/3 \\ \times & \mathbb{P}(X_3 = 1 | X_{-1}^2 = 1000) & 4/5 \\ \times & \mathbb{P}(X_4 = 1 | X_{-1}^3 = 10001) & 1/3 \\ \times & \mathbb{P}(X_5 = 0 | X_{-1}^4 = 100011) & 2/3 \end{array}$$



# Noyaux




**Alphabet** fini $\mathcal{A}$, $\mathcal{A}^* = \cup_{n\geq 0}\mathcal{A}^k$.

**Passé** $\underline{w} = w_{-\infty:-1} \in \mathcal{A}^{-\mathbb{N}}$

**Distance ultra-métrique** $\delta(\underline{w}, \underline{z}) = 2^{\sup\{k<0: w_k \neq z_k\}}$

**Boules** $B \subset \mathcal{A}^{-\mathbb{N}}$ est une boule (ouverte et fermée) si

$$B = \left\{\underline{z}s : \underline{z} \in \mathcal{A}^{-\mathbb{N}}\right\} \text{ pour un certain } s \in \mathcal{A}^*$$

**Arbres, racines** $B = \mathcal{T}(s)$, $s = \mathcal{R}(B)$

**Noyau** $P : \mathcal{A}^{-\mathbb{N}} \to \mathfrak{M}_1(\mathcal{A})$, on note $P(a|\underline{w}) = P(\underline{w})(a)$.

**Continu** comme application de $(\mathcal{A}^{-\mathbb{N}}, \delta)$ dans $(\mathfrak{M}_1(A), d_{TV})$.

**Oscillation** de $P$ sur la boule $\mathcal{T}(s)$

$$\eta(s) = \sup\left\{\left|P(\cdot|\underline{w}) - P(\cdot|\underline{z})\right|_{TV} : \underline{w}, \underline{z} \in \mathcal{T}(s)\right\}.$$

# Processus

Processus stationnaire $(X_t)_{t\in\mathbb{Z}}$ de loi $\mathbb{P}$ sur $\mathcal{A}^{\mathbb{Z}}$ est dit *compatible* avec le noyau $P$ si ce dernier est une version régulière de ses lois conditionnelles au passé :

$$\mathbb{P}\left(X_i = a | X_{i+j} = w_j, j \in -\mathbb{N}^*\right) = P(a|\underline{w})$$

pour tout $i \in \mathbb{Z}, a \in \mathcal{A}$ et pour $\nu$-presque tout $\underline{w}$.

Problème 1 : étant donné un noyau $P$, existe-t-il un processus compatible avec lui ? Est-il unique ?

Problème 2 : puis-je alors simuler des trajectoires $(X_t)_{1\leq t\leq n}$ ?

$\Longrightarrow$ [Comets, Fernandez, Ferrari '02] oui aux deux questions ensemble si le module de continuité de $P$

$$\sup\{\eta(s) : s \in \mathcal{A}^{-k}\}$$

décroît assez vite vers $0$ quand $k$ tend vers l'infini et sous des hypothèses de régénération assez fortes - en particulier $\eta(\epsilon) > 0$.

# Chaînes de Markov d'ordre variable

Arbre complet de suffixes (CSD) = ensemble $T \subset A^*$ qui définit une partition de $\mathcal{A}$:

$$\mathcal{A}^{-\mathbb{N}} = \bigsqcup_{s \in T} \mathcal{T}(s)$$

Source à arbre de context = processus compatible avec un noyau $P_T$ constant sur chaque boule associée aux feuilles d'un CSD $T$

$T$ fini = chaîne de Markov d'ordre $\mathrm{prof}(T)$

Vraisemblance comme pour les chaînes de Markov:

$$P_T(x_1^n | x_{-\infty}^0) = \prod_{i=1}^{n} P_T(x_i | x_{i-L_i}^{i-1}) = \prod_{s \in T} \prod_{i \in I_s} P_T(x_i | s),$$

où $I_s = \{i \in \{1, \ldots, n\} : x_{i-|s|}^{i-1} = s\}$.

# L'algorithme de Propp-Wilson pour les chaînes de Markov

Objectif : simulation d'une variable $X_0$ suivant la loi stationnaire $\pi$ d'une chaîne de Markov de noyau $Q : \mathcal{A} \to \mathfrak{M}_1(\mathcal{A})$.

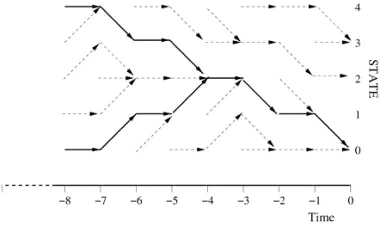
Chaîne auxiliaire : $(F_t)_{t\geq 0}$ à valeur dans $\mathcal{A}^{\mathcal{A}}$ donné par $F_0 = \mathrm{id}$ et

$$F_{t+1} = F_t \circ f_t$$

où les $f_t$ sont i.i.d. de loi $P(f_t(a) = b) = Q(a; b)$.

Temps d'attente : $\tau = \inf\{t \geq 1 : F_t \text{ constante }\}$

Sortie : si $\mathrm{Im}(F_\tau) = \{X_0\}$, alors $X_0 \sim \pi$




$\mathbb{E}[\tau] \approx$ temps de mélange de la chaîne

# Extension au cas des chaînes à mémoire variable [G.]

Objectif : simulation d'un morceau de trajectoire $X_{1:n}$ sous la loi stationnaire $\nu$ d'un processus compatible avec le noyau $P : \mathcal{A}^{-\mathbb{N}} \to \mathfrak{M}_1(\mathcal{A})$.

Chaîne auxiliaire $(F_t)_{t\geq 0}$ à valeur dans $\mathcal{A}^{\mathcal{A}^{-\mathbb{N}}}$ donné par $F_0 = \mathrm{id}$ et

$$F_{t+1} = F_t \circ f_t$$

où les $f_t$ sont i.i.d. de loi $P(f_t(\underline{w}) = b) = P(\underline{w}; b)$.

Temps d'attente : $\tau = \inf \{t \geq 1 : \Pi^n \circ F_t \text{ constante } \}$ où $\Pi^n(\underline{w}) = w_{-n:-1}$.

Idée : par *couplage*, on se ramène à un système dynamique sur l'ensemble des arbres étiquetés

Intérêt : couplage plus rapide que CFF, efficace même pour les noyaux de Markov d'ordre $k$

# Plan de l'exposé

1 Arbres de contextes et mémoire variable
- Noyaux, processus et simulation exacte
- Estimation

2 Quelques inégalités auto-normalisées

3 Apprentissage par renforcements
- Problèmes de bandits classique
- Extensions du modèle

# Algorithme Context

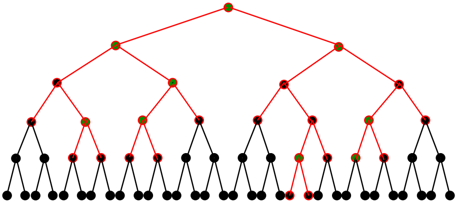
- Introduit par Rissanen en 1981, même principe que CART.
- Pour tout $s \in A^*$, on calcule la mesure de distortion

$$\delta(s) = \max_{a \in A} \left\| \hat{P}(\cdot|s) - \hat{P}(\cdot|as) \right\| .$$

- On garde tous les $s \in A^*$ tels que

$$\exists u \in A^* : \delta(us) \geq \epsilon(n)$$

comme noeuds internes de $\hat{T}_C$.



# Maximum de vraisemblance pénalisée

- On choisit

$$\hat{T}_{pml} = \arg\max_{T} \log \hat{P}_T(x_1^n | x_{-\infty}^0) + \mathrm{pen}(n, T),$$

  où $\mathrm{pen}(n, T)$ = fonction de pénalité croissante en $n$ et $|T|$.

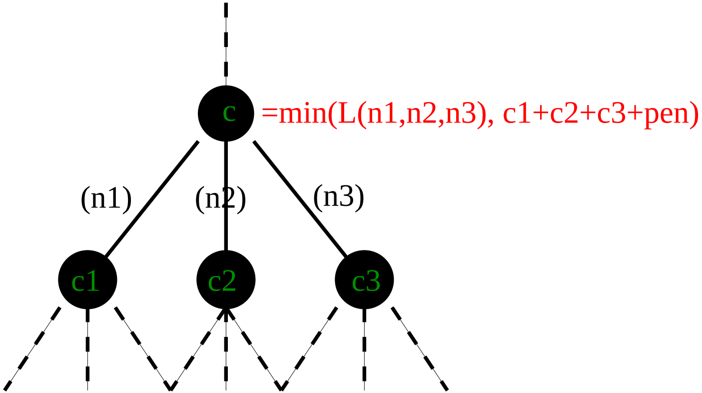
- Pénalité BIC

$$pen(n, T) = \frac{|T|(|A| - 1)}{2} \log n.$$

- Interprétation Minimum Description Length: choisit le modèle qui donne la plus courte description des données (= qui permet de mieux les compresser)

# Calcul de $\hat{T}_{pml}$

Procédure récursive "Context Tree Maximization" : un noeud $s$ est dit actif si $x_{I(s)}$ se code mieux avec de la mémoire.




En partant du sommet, on ne garde que les noeuds actifs.

# Comparaison des deux estimateurs

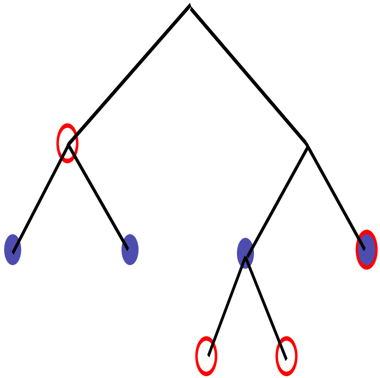
[G. & Leonardi '11]

- - Pour l'algorithme Context, l'activité d'un noeud se mesure uniquement dans ce noeud.
  - Pour PML, l'activité d'un noeud prend en compte tout ce qui est sous ce noeud.
- - L'algorithme Context garde une branche dès que son noeud le plus profond est actif.
  - PLM ne garde que des noeuds actifs.
- $\Longrightarrow$ pour des choix de paramètres comparables, on montre que l'agorithme Context sélectionne systématiquement des arbres plus grands que PLM.

# Sous-estimation et sur-estimation

Deux erreurs d'estimation sont possibles:

1. sous-estimation: $\exists s \in T_0 : s \notin \hat{T}$ $\implies$ "facilement" évitée (régime de grandes déviations) à vitesse exponentielle
2. sur-estimation: $\exists s \in \hat{T} : s \notin T_0$ $\implies$ plus délicat, pas de taux exponentiels [Finesso '92]



# Résultats asymptotiques

- [Rissanen '81, Bühlmann&Wyner '99...] pour un arbre fini $T_0$, si $\epsilon(n) = C\log(n)/n$, alors quand $n \to \infty$:

$$\mathbb{P}(\hat{T}_C \neq T_0) \to 0.$$

- [Csiszár & Talata '06, G. '06]: Si $K \in \mathbb{N}^*$ et si $\hat{T}_{pml}$ maximise la vraisemblance pénalisée parmi les arbres de hauteur $D(n) = o(\log n)$, alors

$$\hat{T}_{pml}^{|K} = T_0^{|K}$$

presque sûrement pour $n$ assez grand. Pour un arbre fini $T_0$, pas besoin de restreindre la maximisation.

- Quelques résultats non asymptotiques : [Galves,Maume-Deschamps&Schmitt '05, Leonardi '08...] avec hypothèses plus ou moins plaisantes

# Estimation conjointe de deux sources partiellement partagées

- $X = (X_n)_{n \in \mathbb{Z}}$ et $Y = (Y_n)_{n \in \mathbb{Z}}$ sont des CTS indépendantes de noyaux respectifs $P_X$ et $P_Y$
- $P_X$ et $P_Y$ partagent certains contextes et certaines lois conditionnelles, mais pas toutes
- Motivation : Linguistique (portugais européen vs brésilien)
- On note $\tau_X = \tau_0 \cup \tau_1$ le CSD de $P_X$, et $\tau_Y = \tau_0 \cup \tau_2$ le CSD de $P_Y$
- On veut estimer $\tau_X$ et $\tau_Y$: est-ce possible de faire mieux qu'en traitant les deux problèmes séparément ?

# Estimation jointe par maximum de vraisemblance pénalisé

- La vraisemblance jointe s'écrit:

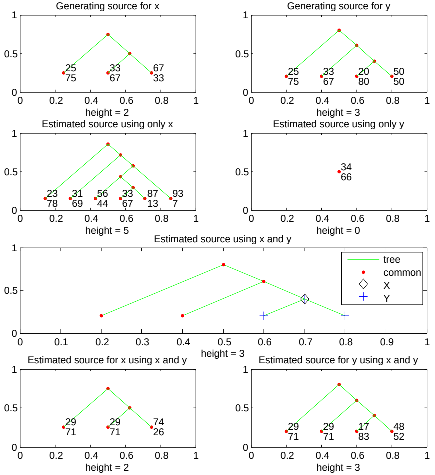
$$\sum_{s\in\tau_1}\sum_{a\in A} N_{n,X}(s,a)\log\left(\frac{N_{n,X}(s,a)}{N_{n,X}(s)}\right)$$
$$+\sum_{s\in\tau_2}\sum_{a\in A} N_{m,Y}(s,a)\log\left(\frac{N_{m,Y}(s,a)}{N_{m,Y}(s)}\right)$$
$$+\sum_{s\in\tau_0}\sum_{a\in A}\left[N_{n,X}(s,a)+N_{m,Y}(s,a)\right]\log\left(\frac{N_{n,X}(s,a)+N_{m,Y}(s,a)}{N_{n,X}(s)+N_{m,Y}(s)}\right)$$

- [Galves, G. & Gassiat]
  - estimation jointe par maximum de vraisemblance pénalisée
  - procédure récursive "à la Context Tree Maximization"
  - consistance pour une pénalité de type BIC

# Exemple de résultats

100 répétitions, échantillons de taille $n_x = n_y = 400$

- $\tau_X$ est bien estimé avec $X$ seul 96 fois, et avec notre algo 80 fois : la performance est donc dégradée.
- mais $\tau_Y$ est bien estimé avec $Y$ seul 47 fois, et avec notre algo 85 fois
- surtout, $\tau_X$ et $\tau_Y$ sont séparément bien estimés 45 fois, et conjointement 67 fois.



# Etude non asymptotique [G. & Leonardi '11]

- Pour l'algorithme Context on doit contrôler

$$\|\hat{P}_t(\cdot|s) - P(\cdot|s)\|$$

- Pour le maximum de vraisemblance pénalisée, il faut majorer

$$KL\left(\hat{P}_t(\cdot|s), P(\cdot|s)\right).$$

- Dans les deux cas, on se ramène à l'étude des maxima d'une martingale "moyenne normalisée" du type:

$$Z_t = \frac{1}{\sqrt{N_t(s)}} \sum_{u=1}^{t} (\mathbb{1}_{\{X_u=a} - P(a|s)\})\mathbb{1}_{\{X_{u-|s|}^{u-1}=s\}}$$

- Des inégalités de déviations pour

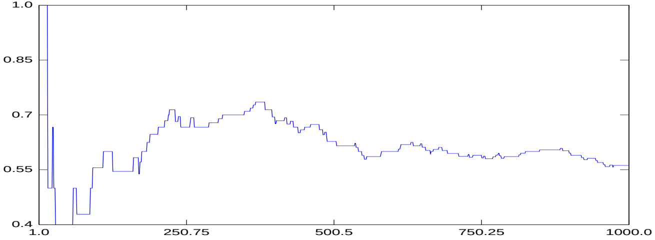
$$W_t = N_t(s) KL\left(\hat{P}_t(\cdot|s); P(\cdot|s)\right)$$

permettent de se passer d'hypothèses "intrinsèquement inutiles" du genre $\forall a \in \mathcal{A}, P(a|s) = 0$ ou $P(s;a) > \epsilon$.

# Quelle est l'activité normale d'un noeud ?

Pour tout contexte possible $s$, l'estimateur du maximum de vraisemblance de la loi conditionnelle est :

$$\forall a \in A, \hat{P}(a|s) = \frac{1}{n} \sum_{k=1}^{n} \mathbb{1}_{\{X_k = a\}} \mathbb{1}_{\{X_{k-|s|}^{k-1} = s\}}$$



# Plan de l'exposé


1 Arbres de contextes et mémoire variable
- Noyaux, processus et simulation exacte
- Estimation

2 Quelques inégalités auto-normalisées

3 Apprentissage par renforcements
- Problèmes de bandits classique
- Extensions du modèle

# Hypothèses

**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 $\phi : ]\lambda_1, \lambda_2[ \to \mathbb{R}$ telle que pour tout $\lambda \in ]\lambda_1, \lambda_2[$ et pour $t \geq 1$,

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

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

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

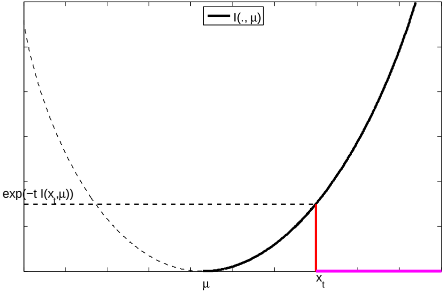
**Cas sous-gaussien** $\phi(\lambda) = \sigma^2\lambda^2/2$, cf. aussi [De La Peña & al. '04]

# Déviations et borne de Chernoff

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

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 :

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

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

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

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

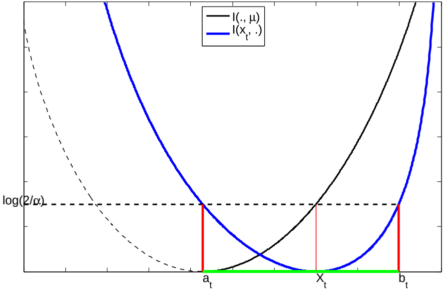
$$[a_t, b_t] = \left\{\mu : tI(\bar{X}_t;\mu) \leq \log\frac{2}{\alpha}\right\}$$

# Déviations et borne de Chernoff

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

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 :

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

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

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

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

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

# Théorèmes [G. & Leonardi '11]

## Borne générale

Pour tout $\delta > 0$,

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

## Cas log-concave

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

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

## Schéma de preuve

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

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

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

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

  où $W_t^k = \exp\left(\lambda_k S_t - t\phi(\lambda_k)\right)$ 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 $\eta = 1/(\delta - 1)$

# Version auto-normalisée pour observation optionnelle

Pour tout $t \in \{1, \dots, n\}$, supposons que $\varepsilon_t \in \{0, 1\}$ soit $\mathcal{F}_{t-1}$-mesurables et telle que l'estimée courante au temps $n$ de la moyenne soit

$$\bar{X}(n) = \frac{S(n)}{N(n)} \,, \quad \text{où } S(n) = \sum_{t=1}^{n} \varepsilon_t X_t \quad \text{et } N(n) = \sum_{t=1}^{n} \varepsilon_t$$

## Borne générale : version auto-normalisée

Pour tout $\delta > 0$,

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

# Observations non stationnaires [G. & Moulines '11]

- $(X_t)_t$ indépendantes et bornées par $B$, d'espérances $\mu_t$ 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 $M_\gamma(n) = \sum_{t=1}^n \gamma^{n-t}\varepsilon_t \mu_t$,

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

- 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 $P_\theta$ admet la densité

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

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

**Divergence**

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

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

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

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

**Intervalles de confiance**

$$\left\{\theta \in \Theta : N(n)\, \mathrm{KL}\left(P_{\mu^{-1}(\bar{X}(n))}; P_\theta\right) \leq \delta\right\}$$

$$= \left\{\theta \in \Theta : I\left(\bar{X}(n); \mu(\theta)\right) \leq \frac{\delta}{N(n)}\right\} .$$

# Lois multinomiales [G. & Leonardi '11]

## Lemme: réduction aux lois de Bernoulli

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

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

## Corollaire: Voisinages KL pour multinomiales

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

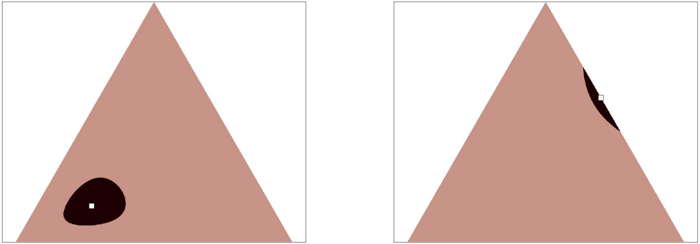
$$P\left(\exists t \in \{1, \ldots, n\} : \quad \mathrm{KL}\left(\hat{P}_t; P_0\right) \geq \frac{\delta}{t}\right) \leq 2e\left(\delta \log(n) + |\mathcal{A}|\right) \exp\left(-\frac{\delta}{|\mathcal{A}|}\right)$$

# 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 :

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

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



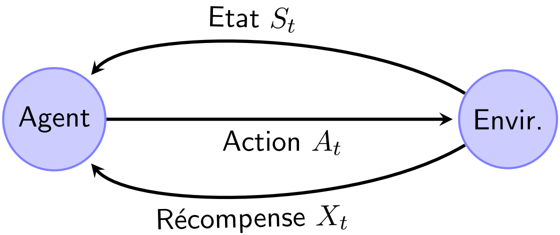
# Plan de l'exposé

1 Arbres de contextes et mémoire variable
- Noyaux, processus et simulation exacte
- Estimation

2 Quelques inégalités auto-normalisées

3 Apprentissage par renforcements
- Problèmes de bandits classique
- Extensions du modèle

# Apprentissage par renforcement




dilemme
exploration
|
exploitation

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

# 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)

# Le "problème de bandits multibras"

Environnement : 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 : $\pi = (\pi_1, \pi_2, \dots)$ telle que

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

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

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

# Performance, regret

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

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

  où $\mu_a = E[P_a]$
- Objectif équivalent : minimiser le *regret*

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

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

# Principe d'optimisme

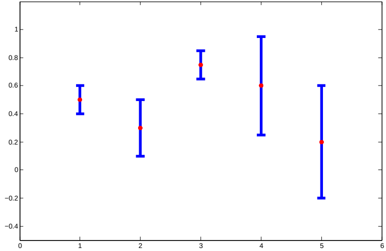
Algorithmes **optimistes** : [Lai&Robins '85; Agrawal '95]

*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 '07]

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



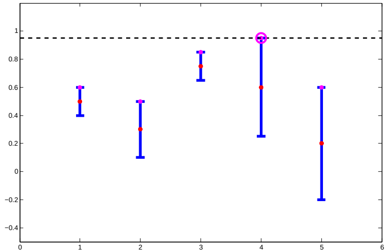

- Avantage : comportement facilement interprétable et "acceptable"

$$\Rightarrow \text{le regret grandit comme } C\log(n)$$

  où $C$ dépend des $P_a, a \in \mathcal{A}$
- *Politique d'indice* : on calcule un indice par bras et on choisit celui qui est le plus élevé, cf. [Gittins '79]

# Stratégies "Upper Confidence Bound" [Auer&al '02; Audibert&al '07]

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

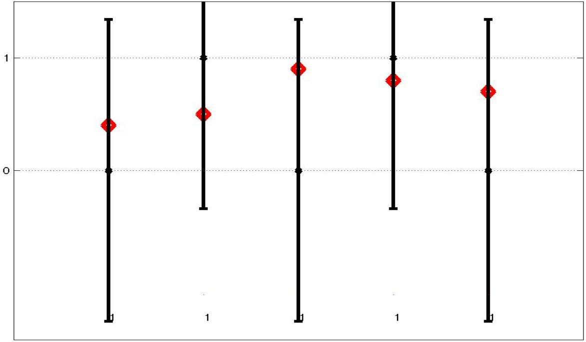



- Avantage : comportement facilement interprétable et "acceptable"

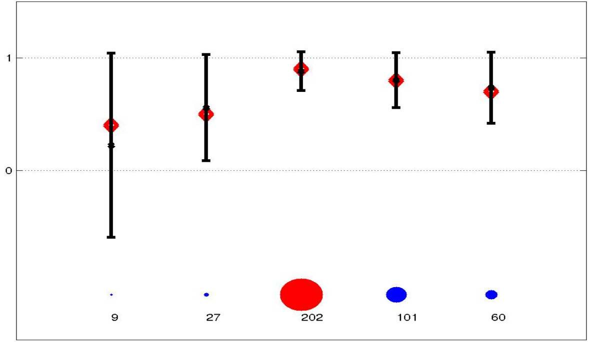
$$\Rightarrow \text{le regret grandit comme } C\log(n)$$

  où $C$ dépend des $P_a, a \in \mathcal{A}$
- *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



# 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

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

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 :

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

Premier terme : contrôlé par les inégalités auto-normalisées car

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

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 $P_\theta$ 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**

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

**Alternative** bayésienne [Kaufmann, Cappé & Garivier] : quantiles des lois a posteriori

# Application : récompenses bornées [G. Cappé '11]

## Borne de regret

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

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

- 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 $\mathcal{F}$

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

# Autre choix pour les récompenses bornées

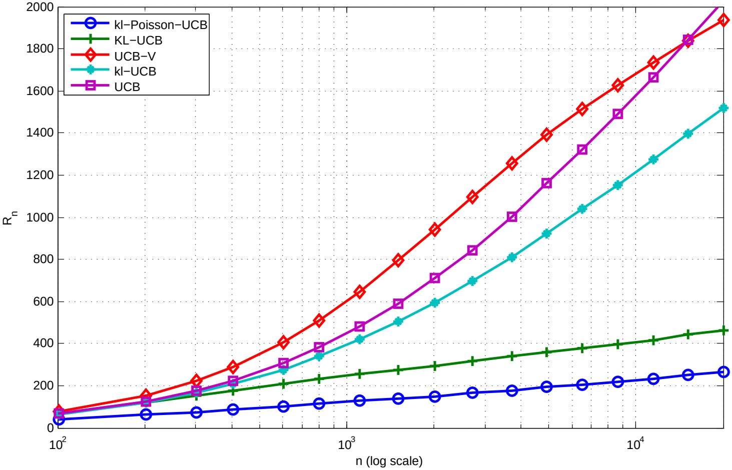
Quand $\mathrm{Var}[P_a] \ll \mu_a(1-\mu_a)$, la borne de confiance utilisée est pessimiste.
Estimation totalement non paramétrique : $\mathcal{F} = \mathfrak{M}_1([0,1])$ et $\Pi_{\mathcal{F}} = id$.

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

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

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

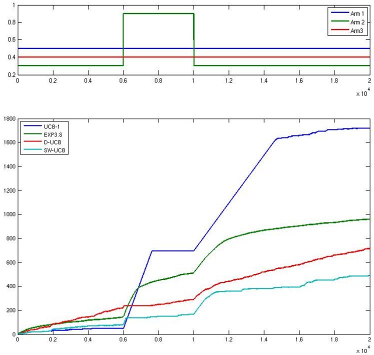
# Comparatif sur un exemple




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

# 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

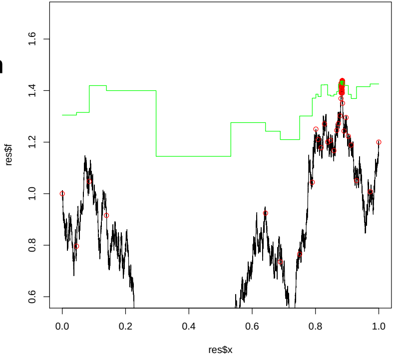
- Exemple : pour des récompenses binaires

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

- 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 :

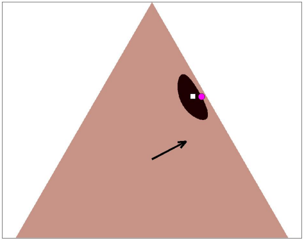
$$S_{t+1} \sim P(\cdot; S_t, A_t) \text{ et } R_t = r(S_t, A_t) + \epsilon_t$$

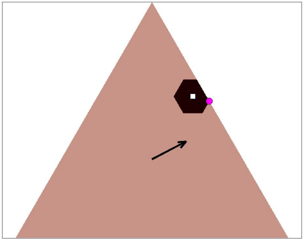
Meilleur modèle pour les communications numériques, mais aussi pour :

- 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 :

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

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

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

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

# 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 :

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

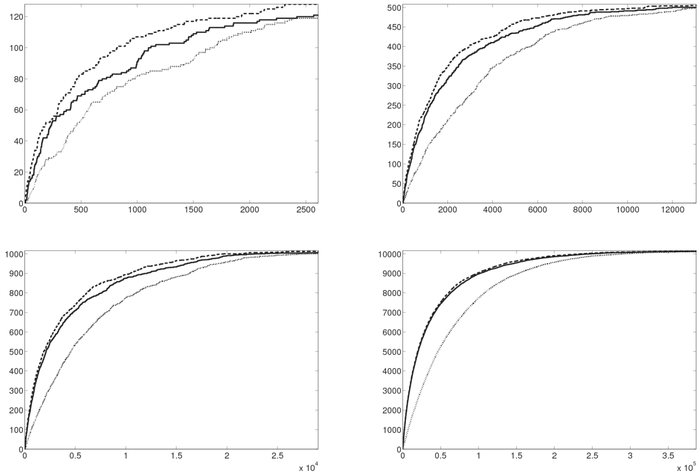
Limite macroscopique :

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

## 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.