L'algorithme UCB, et comment aller plus loin

I. Rappel : Borne de Chernoff-Hoeffding pour les variables bornées

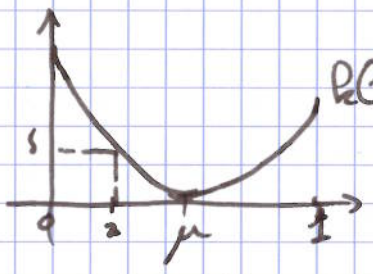
Soit $X_1, \dots, X_n$ iid $\sim \nu \in \mathcal{M}_1([0,1])$ — on note $\mu = \mathbb{E}_\nu[X_1] = E(\nu)$

Log-Laplace : $\phi_\nu(\lambda) = \ln \mathbb{E}_\nu[e^{\lambda X_1}] \leq \ln(1 - \mu + \mu e^\lambda)$

→ si $x < \mu$, on utilise Markov + optimise en λ et on obtient

$$\mathbb{P}_\nu(\bar{X}_n < x) \leq \exp(-n\, kl(x;\mu)) \quad (1)$$

où $kl : (p,q) \to p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q}$ = divergence sur $[0,1]$



← point de vue "probas" : $\bar{X}_n$ est probable et donc voisine de μ pr $kl(\cdot,\mu)$

posons $\delta = kl(x;\mu)$.

(1) peut se ré-écrire $\mathbb{P}_\nu(\bar{X}_n < \mu,\ kl(\bar{X}_n,\mu) > \delta) \leq \exp(-n\delta)$

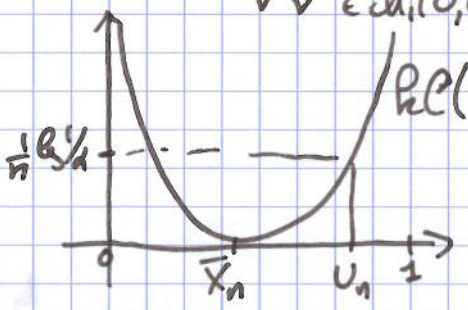
et donc $\forall \varepsilon > 0,\ \mathbb{P}_\nu(\bar{X}_n < \mu,\ n\,kl(\bar{X}_n,\mu) > \varepsilon) \leq \exp(-\varepsilon)$

D'où la borne supérieure de confiance

$$U_n = \max\{\mu' \in [0,1] : n\,kl(\bar{X}_n,\mu') \leq \log 1/\alpha\}$$

Elle vérifie en effet

$$\forall \nu \in \mathcal{M}_1([0,1]),\ \mathbb{P}_\nu(U_n \geqslant E(\nu)) \geqslant 1-\alpha$$



← point de vue "stats" : la moyenne est dans un voisinage de $\bar{X}_n$ pr la 2e marginale de kl

Relaxation de Hoeffding :

si inégalité de Pinsker $kl(p,q) \geqslant 2(p-q)^2$ (bon si $p \geq q \geq 1/2$)

$$\to \forall x < \mu,\ \mathbb{P}_\nu(\bar{X}_n < x) \leq \exp(-2n(x-\mu)^2)$$

$$\mathbb{P}_\nu(\bar{X}_n < \mu,\ 2n(\bar{X}_n - \mu)^2 > \varepsilon) \leq \exp(-\varepsilon)$$

$$\tilde{U}_n = \max\{\mu' \in [0,1] : 2n(\bar{X}_n - \mu) \leq \log 1/\alpha\} = \bar{X}_n + \sqrt{\frac{\log 1/\alpha}{2n}}$$

Bonne approximation si $\mu \approx 1/2$ et n grand

II Application à l'allocation dynamique de ressources : algorithme UCB

On présente dans cette partie les résultats principaux de

[1] Auer, Cesa-Bianchi, Fischer (2002)
Finite Time Analysis of the Multiarmed Bandit Problem

1. Rappel du cadre :

K entier positif. $(X_{a,n})_{\substack{1\le a\le K\\ n\ge 1}}$ famille de v.a. indépendantes, $X_{a,n} \sim \nu_a$ sur $[0,1]$

On note $\mu_a = E(\nu_a) = \mathbb{E}[X_{a,1}]$, $\overline{X}_{a,n} = \frac{1}{n}\sum_{k=1}^{n} X_{a,k}$

On définit une règle d'allocation dynamique π avec

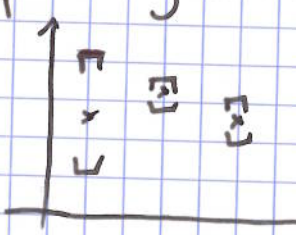
$$\pi = (\pi_t)_{t\ge 1} \qquad \pi_t : \left(\{1,\dots K\}\times[0,1]\right)^{t-1} \to \{1,\dots K\}$$

Par récurrence, cela définit les suites de variables aléatoires

$$A_t = \pi_t(A_1, Y_1, \dots, A_{t-1}, Y_{t-1})$$
$$Y_t = X_{t,N_a(t)}$$
$$N_a(t) = \sum_{s\le t} \mathbb{1}_{\{A_s = a\}} = N_a(t-1) + \mathbb{1}_{\{A_t=a\}}$$
$$S_a(t) = \sum_{s\le t} \mathbb{1}_{\{A_s=a\}} Y_s = \sum_{n=1}^{N_a(t)} X_{a,n}$$
$$\overline{X}_a(t) = \frac{S_a(t)}{N_a(t)} = \overline{X}_{a,N_a(t)}$$

2. Paradigme optimiste et algo UCB

"Parmi tous les environnements suffisamment vraisemblables, fais comme si tu te trouvais dans celui qui t'est le plus favorable".



$\longrightarrow$ Politique UCB

$$\to A_t = \underset{1\le a\le K}{\arg\max}\ \overline{X}_a(t) + \sqrt{\frac{4\log(t)}{2N_a(t)}}$$

(le facteur 4 est nécessaire dans la preuve de [1]), interprétation d'$\frac{1}{t}$

Théorème ([1]) : sous la politique UCB, pour tout bras a tq $\mu_a < \mu^* = \max_{1\le b\le K}\mu_b$ pour tout $T \ge 1$

$$\mathbb{E}[N_a(T)] \le \frac{8\log(T)}{(\mu^*-\mu_a)^2} + \frac{\pi^2}{3} + 1$$

3- Preuve du théorème

On pose $n_T = \left\lceil \frac{8 \ln(T)}{(\mu^* - \mu_a)^2} \right\rceil$ et a^* est un bras optimal $1 \le a \le K$ μ_a

$$\mathbb{E}[N_a(T)] \le n_T + \sum_{t=K}^{T-1} \mathbb{P}(A_{t+1}=a, N_a(t) \ge n_T)$$

$$\le n_T + \sum_{t=K}^{T-1} \mathbb{P}(L_a(t) > \mu_a) \bullet$$

$$+ \sum_{t=K}^{T-1} \mathbb{P}(U_{a^*}(t) < \mu_{a^*})$$

$$+ \sum_{t=K}^{T-1} \mathbb{P}(A_{t+1}=a, N_a(t) \ge n_T, L_a(t) \le \mu_a, U_a(t) \ge \mu^*)$$

où $[L_a(t), U_a(t)] = \overline{X}_a(t) \pm \sqrt{\frac{4 \ln t}{2 N_a(t)}}$

- $\sum_{t=1}^{T} \mathbb{P}(L_a(t) > \mu_a) = \sum_{t=1}^{T} \sum_{n=1}^{t} \mathbb{P}(L_a(t) > \mu_a, N_a(t) = n)$

$$= \sum_{t=1}^{T} \sum_{n=1}^{t} \mathbb{P}\left(\overline{X}_n > \mu + \sqrt{\frac{4 \ln t}{2n}}, N_a(t) = n\right)$$

$$\le \sum_{t=1}^{T} \sum_{n=1}^{t} 1/t^4 \le \sum_{t=1}^{T} t \times 1/t^4 < \sum_{t=1}^{\infty} 1/t^2 \le \pi^2/6$$

et, de même, $\sum_{t=1}^{T} \mathbb{P}(U_{a^*}(t) < \mu_{a^*}) \le \pi^2/6$

- en outre, si $L_a(t) < \mu_a$, $U_{a^*}(t) \ge \mu^*$ et $A_{t+1} = a$ alors

$$\mu^* < U_{a^*}(t) \le U_a(t) = L_a(t) + 2\sqrt{\frac{4 \ln t}{2 N_a(t)}} < \mu_a + \sqrt{\frac{8 \ln t}{N_a(t)}}$$

et donc $N_a(t) < \frac{8 \ln t}{(\mu^* - \mu_a)^2} \le n_T$

et donc $\forall t$, $\mathbb{P}(A_{t+1}=a, N_a(t) \ge n_T, L_a(t) \le \mu_a, U_a(t) \ge \mu^*) = 0$

d'où le résultat.

4- Commentaires :

$$\mathbb{E}[N_a(T)] \le \frac{16 \ln T}{2(\mu^* - \mu_a)^2} (+c)$$

← 4 pour la borne de l'union
← 4 parce qu'on a attendu trop longtemps
← Pinsker : $kl(\mu, \mu^*) \ge 2(\mu^* - \mu_a)^2$

Version asymptotique : $\mathbb{E}[N_a(T)] \ge \frac{\ln T}{kl(\mu_a, \mu^*)} (1 - o(1))$

5_ Améliorations

1. Utiliser U_n et pas $\tilde{U}_n$:

$$A_{t+1} = \underset{1 \le a \le K}{\arg\max} \left\{ \mu' \in [0,1] : N_a(t)\, kl(\hat{X}_a(t), \mu') \le f(t) \right\}$$

avec $f(t) = \log t \left(+ 3 \log \log t \right)$

2. Éviter la borne de l'union : inégalité auto-normalisée

$$\mathbb{P}\left(N_a(t)\, kl(\hat{X}_a(t), \mu_a) > f(t) \right) \le e \lceil f(t) \log t \rceil e^{-f(t)}$$

obtenu en prouvant que

$$\forall \varepsilon > 0, \quad \mathbb{P}\left(\bigcup_{n \le t} \left\{ n\, kl(\hat{X}_{a,n}, \mu_a) > \varepsilon \right\} \right) \le e \lceil \varepsilon \log t \rceil e^{-\varepsilon}$$

Preuve : reprendre celle de LLI par martingales

3. Décomposition plus fine :

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

↳ on obtient $\mathbb{E}[N_a(T)] \le \dfrac{\log T}{kl(\mu_a, \mu^*)} + o(\log T)$

↑ complètement explicite

cf Cappé, Garivier, Maillard, Munos, Stoltz
Kullback-Leibler Upper Confidence Bounds
for Optimal Sequential Allocation
Annals of Statistics, 2013

Idée : d'où vient $\dfrac{\log T}{kl(\mu_a, \mu^*)}$? → c'est le premier n pour lequel $U_a(t) > \mu^*$ correspond à une déviation à droite par $\hat{X}_{a,n}$