Le bandit stochastique à K bras

Considérons K bras (actions, choix) définis par des distributions $(\nu_k)_{1\leq k\leq K}$ à valeurs dans $[0,1]$, de loi inconnues. A chaque instant, l'agent choisit un bras $I_t \in \{1, \ldots, K\}$ et observe une récompense conditionnellement indépendante des récompenses passées générée selon la loi du bras I_t. Son objectif est de maximiser l'espérance de somme des récompenses reçues. Nous notons $\{X_{k,i}\}_{i=1}^{\infty}$ la suite des récompenses (inconnues) associées à chacun des bras (ce sont des variables aléatoires indépendantes et identiquement distribuées).

Notons $\{\mu_k\}_{k=1}^{N}$ l'espérance de récompense de chaque bras, et $\mu^* = \max_k \mu_k$ l'espérance du meilleur bras. Si l'agent connaissait les lois, il choisirait alors le meilleur bras à chaque instant et obtiendrait une récompense moyenne μ^* Comme il ne connait pas l'espérance des différents bras, il doit explorer les différents bras pour acquérir de l'information (exploration) ; cette connaissance lui servira ensuite pour agir optimalement (exploitation). Cette stratégie illustre le compromis *exploration-exploitation*.

Pour évaluer la performance d'une stratégie donnée, on va définir à quelle vitesse cette stratégie permet d'atteindre un taux de récompense moyen optimal. Pour cela on définit le *regret cumulé* à instant n :

$$R_n = n\mu^* - \sum_{t=1}^{n} X_{I_t,t},$$

qui représente la différence en récompenses cumulées entre ce qu'il a obtenu et ce qu'il aurait pu obtenir en moyenne s'il avait joué le bras optimal à chaque itération du jeu. On va étudier une stratégie en cherchant à caluler son regret cumulé moyen : $\mathbb{E}[R_n]$.

1. – Calculer l'espérance du regret en fonction de $\Delta_k = \mu^* - \mu_k$ (la différence entre la performance moyenne du k-ième bras et du bras optimal) et $T_k(n) = \sum_{t=1}^{n} \mathbb{1}\{I_t = k\}$ le nombre de fois où le k-ième bras est choisi.

Un bon algorithme de bandit devra tirer peu souvent les bras sous-optimaux. Pour analyser les algorithmes de bandits, il est nécessaire de disposer de bornes précises sur les fluctuations des sommes.

Soit Z une variable aléatoire réelle. Pour $\lambda \geq 0$, l'inégalité de Markov implique que

$$\mathbb{P}\{Z \geq t\} \leq \mathrm{e}^{-\lambda t}\mathbb{E}\left[\mathrm{e}^{\lambda Z}\right]$$

Comme cette inégalité est satisfaite pour tout $\lambda \geq 0$, on peut choisir la valeur de λ qui minimise cette borne supérieure. Considérons le logarithme de la fonction génératrice des moments

$$\psi_Z(\lambda) = \ln \mathbb{E}\left[e^{\lambda Z}\right] \text{ pour tout } \lambda \geq 0,$$

et en appelant

$$\psi_Z^*(t) = \sup_{\lambda \geq 0}(\lambda t \; - \psi_Z(\lambda)) \; ,$$

nous obtenons l'inégalité de Chernoff

$$P\{Z \geq t\} \leq \exp(-\psi_Z^*(t))$$

La fonction ψ_Z^* est appelée la *transformée de Cramer* de Z. Comme $\psi_Z(0) = 0$, ψ_Z^* est une fonction positive. Si $\mathbb{E}[Z]$ existe, la convexité de la fonction exponentielle et l'inégalité de Jensen impliquent que $\psi_Z(\lambda) \geq \lambda \mathbb{E}[Z]$ et donc pour toutes les valeurs négatives de λ, $\lambda t - \psi_Z(\lambda) \leq 0$ whenever $t \geq EZ$. Cela signifie que l'on peut prendre le suprêmum sur $\lambda \in \mathbb{R}$ dans la définition de la transformée de Cramer :

$$\psi_Z^*(t) = \sup_{\lambda \in \mathbb{R}}(\lambda t \; - \psi_Z(\lambda))$$

L'expression apparaissant dans le terme de droite de l'identité précédent est appelée la *fonction duale de Fenchel-Legendre* de ψ_Z. Pour tout $t \geq \mathbb{E}[Z]$, la transformée de Cramér $\psi_Z^*(t)$ coincide avec la fonction duale de Fenchel-Legendre.

Bien entendu, l'inégalité de Chernoff est triviale si $\psi_Z^*(t) = 0$. C'est le cas lorsque $\psi_Z(\lambda) = \infty$ pour tout λ ou si $t \leq \mathbb{E}[Z]$ (en utilisant encore l'inégalité $\psi_Z(\lambda) \geq \lambda \mathbb{E}[Z]$). Pour obtenir des bornes non triviales, nous supposons dans la suite qu'il existe $\lambda > 0$ tel que $\mathbb{E}\left[e^{\lambda Z}\right] < \infty$.

2. 1. Montrer que l'ensemble des valeurs de $\lambda \in \mathbb{R}^+$ telles que $\mathbb{E}\left[e^{\lambda Z}\right] < \infty$ est un intervalle de la forme $[0, b)$ où $0 < b \leq \infty$

2. Montrer que ψ_Z est convexe (et même strictement convexe si la variable Z n'est pas constante presque-sûrement) et est infiniment différentiable sur $I = (0, \; b)$.
3. On suppose que $\mathbb{E}[Z] = 0$. Montrer que ψ_Z est continûment différentiable sur $[0, \; b)$ et que $\psi_Z'(0) = \psi_Z(0) = 0$
4. Montrer que l'on peut alors écrire la transformée de Cramér $\psi_Z^*(t) = \sup_{\lambda \in I}(\lambda t - \psi_Z(\lambda))$.
5. Montrer que
$$\psi_Z^*(t) = \lambda_t t - \psi_Z(\lambda_t)$$
où λ_t est tel que $\psi_Z'(\lambda_t) = t$.

6. Montrer que la fonction ψ'_Z admet une fonction inverse croissante $(\psi'_Z)^{-1}$ sur l'intervalle $\psi'_Z(I) := (0,\ B)$ et donc que, pour tout $t \in (0,\ B)$,
$$\lambda_t = (\psi'_Z)^{-1}(t)$$
7. Soit Z une variable gaussienne de moyenne nulle et de variance σ^2. Montrer que
$$P\{Z \geq t\} \leq \mathrm{e}^{-t^2/(2\sigma^2)} .$$
8. Reprendre la question précédente avec une variable de Poisson de paramètre ν.
9. Reprendre la question précédente avec une variable de Bernoulli de paramètre p.

Nous supposons dans la suite de l'énoncé que la distribution des récompenses vérifie
$$\ln \mathbb{E}\left[\mathrm{e}^{\lambda(X-\mathbb{E}[X])}\right] \leq \psi(\lambda) , \tag{1}$$
$$\ln \mathbb{E}\left[\mathrm{e}^{\lambda(\mathbb{E}[X]-X)}\right] \leq \psi(\lambda) . \tag{2}$$

3. Soient $\hat{\mu}_{i,s}$ les moyennes empiriques des récompenses associées au bras i lorsque le bras est tiré s fois.

- Montrer que
$$\mathbb{P}(\mu_i - \hat{\mu}_{i,s} > \varepsilon) \leq \mathrm{e}^{-s\psi^*(\varepsilon)}$$
- Montrer que pour tout $\delta \in (0,1)$, avec une probabilité $1-\delta$, nous avons
$$\hat{\mu}_{i,s} + (\psi^*)^{-1}(\frac{1}{s}\ln\frac{1}{\delta}) > \mu_i.$$

La stratégie $(\alpha,\ \psi)$-UCB, où $\alpha > 0$ est un paramètre à ajuster, consiste à choisir, lors de l'épisode t,
$$I_t \in \mathrm{argmax}_{i=1,\ldots,K}[\hat{\mu}_{i,T_i(t-1)} + (\psi^*)^{-1}(\frac{\alpha \ln t}{T_i(t-1)})]$$

4. 1. Montrer que l'évènement $\{I_t = i\} = A_{i,t} \cup B_{i,t} \cup C_{i,t}$, où
$$A_{i,t} := \left\{\hat{\mu}_{i^*,T_{i^*}(t-1)} + (\psi^*)^{-1}(\frac{\alpha \ln t}{T_{i^*}(t-1)}) \leq \mu^*\right\}$$
$$B_{i,t} := \left\{\hat{\mu}_{i,T_i(t-1)} > \mu_i + (\psi^*)^{-1}(\frac{\alpha \ln t}{T_i(t-1)})\right\}$$
$$C_{i,t} := \left\{T_i(t-1) < \frac{\alpha \ln n}{\psi^*(\triangle_i/2)}\right\}$$

2. On pose
$$u = \lceil \frac{\alpha \ln n}{\psi^*(\triangle_i/2)} \rceil$$
Montrer que
$$\mathbb{E}\left[T_i(n)\right] \leq u + \sum_{t=u+1}^{n} \{\mathbb{P}(A_{i,t}) + \mathbb{P}(B_{i,t})\} \ .$$
3. Montrer que
$$\mathbb{P}(A_{i,t}) \leq \sum_{s=1}^{t} \frac{1}{t^\alpha} \sim \frac{1}{t^{\alpha-1}}.$$
4. Etablir une borne similaire pour $\mathbb{P}(B_{i,t})$.
5. Montrer que
$$\overline{R}_n \leq \sum_{i:\Delta_i>0} (\frac{\alpha \Delta_i}{\psi^*(\Delta_i/2)} \ln n + \frac{\alpha}{\alpha-2}) \ .$$