Gradient Bandits

On note $H_t : A \longmapsto \mathbb{R}$ la fonction de préférences sur les bandits à l'instant $t \in \mathbb{N}$.
On suppose que l'agent choisit un bandit selon sa fonction de préférence en utilisant la loi de probabilité suivante (soft max / Boltzmann / Gibbs).

$$\mathbb{P}(A_t = a) = \frac{e^{H_t(b)}}{\sum_{b=1}^{m} e^{H_t(b)}} = \pi_t(a)$$

Initialement, on a $H_0(a) = 0$. On utilise une montée de gradient stochastique pour mettre à jour H_t (i.e calculer H_{t+1} en fonction de H_t et du résultat du choix à t).

$$H_{t+1}(a) = H_t(a) + \alpha \frac{\partial \mathbb{E}[R_t]}{\partial H_t(a)}$$ où la fonction objectif est la récompense moyenne

$$\mathbb{E}[R_t] = \sum_{b=1}^{m} \pi_t(b)\, q_*(b).$$

$$\frac{\partial \mathbb{E}[R_t]}{\partial H_t(a)} = \frac{\partial}{\partial H_t(a)} \left[\sum_{b=1}^{m} \pi_t(b) q_*(b) \right] = \sum_{b=1}^{m} q_*(b) \frac{\partial \pi_t(b)}{\partial H_t(a)}$$

Ici, la somme des dérivées partielles sur les bandits est nulle : $\sum_{b=1}^{m} \frac{\partial \pi_t(b)}{\partial H_t(a)} = 0$
(on veut rester dans le simplex, donc les changements doivent s'équilibrer)

On peut donc ajouter sans changer l'égalité un terme de référence B_t indépendant de H_t :

$$\frac{\partial \mathbb{E}[R_t]}{\partial H_t(a)} = \sum_{b=1}^{m} (q_*(b) - B_t) \frac{\partial H_t(b)}{\partial H_t(a)}$$

$$= \sum_{b=1}^{m} \pi_t(b) (q_*(b) - B_t) \frac{\partial H_t(b)}{\partial H_t(a)} * \frac{1}{\pi_t(b)} \quad \text{en } * \frac{\pi_t(b)}{\pi_t(b)}$$

On peut maintenant voir cette expression comme une espérance

$$\frac{\partial \mathbb{E}[R_t]}{\partial H_t(a)} = \mathbb{E}\left[\frac{q_*(A_t) - B_t}{\pi_t(A_t)} \frac{\partial H_t(A_t)}{\partial H_t(a)} \right]$$

On a $\mathbb{E}[R_t \mid A_t] = q_*(A_t)$ donc on peut remplacer $q_*(A_t)$ par R_t

On peut choisir $B_t = \bar{R}_t$ (la moyenne des récompenses obtenues jusqu'ici)

$$\frac{\partial \mathbb{E}[R_t]}{\partial H_t(a)} = \mathbb{E}\left[\frac{R_t - \bar{R}_t}{\pi_t(A_t)} \cdot \frac{\partial H_t(A_t)}{\partial H_t(a)} \right] \quad (*)$$

$$\frac{\partial H_t(b)}{\partial H_t(a)} = \frac{\partial}{\partial H_t(a)} \left[\frac{e^{H_t(b)}}{\sum_{c=1}^{m} e^{H_t(c)}} \right] = \frac{\frac{\partial H_t(b)}{\partial H_t(a)} e^{H_t(b)} \sum_{c=1}^{m} e^{H_t(c)} - e^{H_t(b)} e^{H_t(a)}}{\left(\sum_{c=1}^{m} e^{H_t(c)} \right)^2}$$

$$= \frac{\mathbb{1}_{a=b} e^{H_t(b)} \sum_{c=1}^{m} e^{H_t(c)} - e^{H_t(b)} e^{H_t(a)}}{\left(\sum_{c=1}^{m} e^{H_t(c)} \right)^2}$$

$$= \mathbb{1}_{a=b} \frac{e^{H_t(b)}}{\sum_{c=1}^{m} e^{H_t(c)}} - \pi_t(b)\pi_t(a)$$

$$= \mathbb{1}_{a=b}\, \pi_t(b) - \pi_t(b)\pi_t(a)$$

donc $\frac{\partial H_t(b)}{\partial H_t(a)} = \pi_t(b) \left[\mathbb{1}_{a=b} - \pi_t(a) \right]$

On utilise cette expression dans $(*)$

$$\frac{\partial \mathbb{E}[R_t]}{\partial H_t(a)} = \mathbb{E}\left[\frac{R_t - \bar{R}_t}{\pi_t(A_t)} \cdot \frac{\partial H_t(A_t)}{\partial H_t(a)} \right]$$

$$= \mathbb{E}\left[\frac{R_t - \bar{R}_t}{\pi_t(A_t)} \pi_t(A_t) \left[\mathbb{1}_{a=A_t} - \pi_t(a) \right] \right]$$

$$\frac{\partial \mathbb{E}[R_t]}{\partial H_t(a)} = \mathbb{E}\left[(R_t - \bar{R}_t) \left(\mathbb{1}_{A_t=a} - \pi_t(a) \right) \right]$$

En espérance, on pourra utiliser un échantillon de $(R_t - \bar{R}_t)\left(\mathbb{1}_{A_t=a} - \pi_t(a) \right)$ pour approximer le gradient

On pourra donc mettre à jour $H_t(a)$ par :

$$H_{t+1}(a) = \begin{cases} H_t(a) - \alpha (R_t - \bar{R}_t) \pi_t(a) & \text{Si } A_t \neq a \\ H_t(a) + \alpha (R_t - \bar{R}_t)(1 - \pi_t(a)) & \text{Si } A_t = a \end{cases}$$