Les bandits stochastiques à récompenses d'espérance non-définie

Adam Cohen, Maxime Genest, Vincent Masse

24 novembre 2020

Rappel sur les bandits stochastiques classiques

- Ensemble de K actions (bras, machines).
- Chaque action k est associée à un paramètre inconnu μ_k tel que $X_{k_t} \sim \nu(\mu_k)$ où $\nu(\mu_k)$ est une distribution d'espérance μ_k.

Rappel sur les bandits stochastiques classiques

- Ensemble de K actions (bras, machines).
- Chaque action k est associée à un paramètre inconnu μ_k tel que $X_{k_t} \sim \nu(\mu_k)$ où $\nu(\mu_k)$ est une distribution d'espérance μ_k.

Dans le jeu des bandits stochastiques, à chaque pas de temps $t = 1, 2, \ldots, T$, l'agent :

- Sélectionne une action $k_t \in \{1, 2, \ldots K\}$
- On observe une récompense (reward) $r_t \sim \nu(\mu_{k_t})$.

But : Déterminer une politique d'action qui maximisera $\mathbb{E}\left[\sum_{t=1}^{T} r_t\right]$

Mesure de performance empirique pour les bandits stochastiques

Dans cette situation, à chaque pas de temps $t = 1, 2, \ldots, T$, l'agent cumule un regret :

$$\Delta_{k_t} = \mu^\star - \mu_{k_t}$$

À la fin de l'épisode, on peut calculer le regret cumulatif empirique :

$$R(T) = \sum_{t=1}^{T} \Delta_{k_t}$$

Cela nous permet de comparer empiriquement la performance de plusieurs politiques, en simulant plusieurs épisodes et en comparant le regret cumulatif moyen sur ces épisodes.

L'hypothèse d'existence de l'espérance

Le jeu des bandits stochastiques ainsi présenté sous-entend que la distribution des rewards associés aux bras du bandit est d'espérance qui existe. Or, plusieurs distributions de probabilité ont une distribution d'espérance non-définie.

L'hypothèse d'existence de l'espérance

Le jeu des bandits stochastiques ainsi présenté sous-entend que la distribution des rewards associés aux bras du bandit est d'espérance qui existe. Or, plusieurs distributions de probabilité ont une distribution d'espérance non-définie.

Par exemple, La loi de Cauchy ou certaines configuration de la loi de Pareto.

L'hypothèse d'existence de l'espérance

Le jeu des bandits stochastiques ainsi présenté sous-entend que la distribution des rewards associés aux bras du bandit est d'espérance qui existe. Or, plusieurs distributions de probabilité ont une distribution d'espérance non-définie.

Par exemple, La loi de Cauchy ou certaines configuration de la loi de Pareto.

Mettre des graphiques pour jaser un peu des queues des distribution

La loi de Cauchy

La loi de Cauchy est une loi continue de fonction de densité

$$f(x; L; a) = \frac{1}{\pi\, a \left[1 + \left(\frac{x-L}{a}\right)^2\right]} = \frac{1}{\pi}\left[\frac{a}{(x-L)^2 + a^2}\right]$$

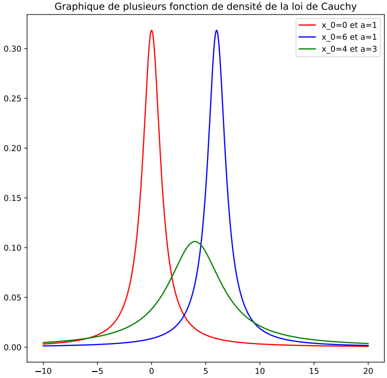
où $l \in \mathbb{R}$ est un paramètre de localisation et $a > 0$ est un paramètre d'échelle.

La loi de Cauchy

La loi de Cauchy est une loi continue de fonction de densité

$$f(x; L; a) = \frac{1}{\pi\, a \left[1 + \left(\frac{x-L}{a}\right)^2\right]} = \frac{1}{\pi}\left[\frac{a}{(x-L)^2 + a^2}\right]$$

où $l \in \mathbb{R}$ est un paramètre de localisation et $a > 0$ est un paramètre d'échelle.



Les bandits de Cauchy

À chaque pas de temps $t = 1, 2, \ldots, T$, l'agent :

- Sélectionne une action $k_t \in \{1, 2, \ldots, K\}$
- Observe une reward $r_t \sim \mathrm{Cauchy}(L_{k_t}, a)$

Les bandits de Cauchy

À chaque pas de temps $t = 1, 2, \dots, T$, l'agent :

- Sélectionne une action $k_t \in \{1, 2, \dots, K\}$
- Observe une reward $r_t \sim \mathrm{Cauchy}(L_{k_t}, a)$

L'action optimale et la localisation optimale sont définis à partir de la localisation des différents bras :

Les bandits de Cauchy

À chaque pas de temps $t = 1, 2, \ldots, T$, l'agent :

- Sélectionne une action $k_t \in \{1, 2, \ldots, K\}$
- Observe une reward $r_t \sim \mathrm{Cauchy}(L_{k_t}, a)$

L'action optimale et la localisation optimale sont définis à partir de la localisation des différents bras :

$$L^\star := \max_k L_k \qquad \text{et} \qquad k^\star := \underset{k}{\mathrm{argmax}} L_k$$

Les bandits de Cauchy

À chaque pas de temps $t = 1, 2, \dots, T$, l'agent :

- Sélectionne une action $k_t \in \{1, 2, \dots, K\}$
- Observe une reward $r_t \sim \mathrm{Cauchy}(L_{k_t}, a)$

L'action optimale et la localisation optimale sont définis à partir de la localisation des différents bras :

$$L^\star := \max_k L_k \qquad \text{et} \qquad k^\star := \underset{k}{\mathrm{argmax}} L_k$$

le gap (regret) associé à l'action k devient $\Delta_k = L^\star - L_k$

Les bandits de Cauchy

À chaque pas de temps $t = 1, 2, \ldots, T$, l'agent :

- Sélectionne une action $k_t \in \{1, 2, \ldots, K\}$
- Observe une reward $r_t \sim \mathrm{Cauchy}(L_{k_t}, a)$

L'action optimale et la localisation optimale sont définis à partir de la localisation des différents bras :

$$L^\star := \max_k L_k \qquad \text{et} \qquad k^\star := \underset{k}{\mathrm{argmax}} L_k$$

le gap (regret) associé à l'action k devient $\Delta_k = L^\star - L_k$

Mesure de performance empirique d'un agent : $R(T) = \sum_{t=1}^{T} \Delta_{k_t}$

Les algorithmes classiques

À faire : Montrer le graphique du regret d'une expérience basée sur un exemple d'algorithme classique basé sur la moyenne empirique

Les algorithmes classiques

À faire : Montrer le graphique du regret d'une expérience basée sur un exemple d'algorithme classique basé sur la moyenne empirique

Cause de la mauvaise performance : la moyenne empirique $\hat{\mu}_k(t)$ n'est pas un bon estimateur de la localisation L_k de la loi de Cauchy du bras no k.

Les algorithmes classiques

À faire : Montrer le graphique du regret d'une expérience basée sur un exemple d'algorithme classique basé sur la moyenne empirique

Cause de la mauvaise performance : la moyenne empirique $\hat{\mu}_k(t)$ n'est pas un bon estimateur de la localisation L_k de la loi de Cauchy du bras no k.

À faire, montrer une expérience montrant le comportement chaotique de l'estimateur $\hat{\mu}_k(t)$.

Estimateurs de la localisation L d'une loi de Cauchy

Présentation des différents estimateurs de localisation pour le paramètre L

Adaptation des algorithmes

Présentation des différentes adaptations des algorithmes (etc, epsilon_greedy,...) pour les bandits Cauchy.

La loi de Pareto

La loi de Pareto est une loi continue dont la fonction de densité est donnée par

$$f(x; x_m, k) = \begin{cases} \dfrac{k x_m^k}{x^{k+1}} & \text{si } x \geq x_m \\ 0 & \text{sinon} \end{cases}$$

Dans le cas particulier où $k = 1$, on obtient que

$$f(x; x_m) = \begin{cases} \dfrac{x_m}{x^2} & \text{si } x \geq x_m \\ 0 & \text{sinon} \end{cases}$$

Dans ce cas particulier, la loi de Pareto possède une espérance non-définie (infinie).