Algorithmes du bandit

Damien Lamberton
Université Paris-Est Marne-la-Vallée

Journée de formation, 23 janvier 2018


ONE ARMED
BANDIT
GO FOR THE JACKPOT PAYOUT
JACKPOT
BAR
JACKPOT ON ALL BAR AND
7
COMBOS
ONE ARMED
BANDIT


MULTI-ARMED
BANDIT

Plan

Plan

Plan

L'algorithme suivant a été introduit indépendamment par Norman (1968) et Shapiro et Narendra (1969).

L'algorithme suivant a été introduit indépendamment par Norman (1968) et Shapiro et Narendra (1969).
On considère un bandit à deux bras, notés A et B. L'action du bras A (resp. B) donne lieu à un gain aléatoire suivant une loi de Bernoulli de paramètre p_A (resp. p_B). Les paramètres p_A et p_B sont inconnus et le joueur cherche à actionner le bras le plus favorable. La stratégie proposée consiste à choisir le bras au hasard, mais en modifiant au cours du temps la loi de probabilité du choix du bras de la façon suivante.
Si à l'étape n, le bras actionné avait une probabilité x d'être choisi et s'il produit un gain, on lui affecte à l'étape suivante la probabilité

$$x + \gamma(1 - x),$$

où $\gamma \in]0, 1[$.

Pour formaliser de manière plus précise, notons X_n la probabilité de choisir le bras A à l'étape n. La suite $(X_n)_{n\in\mathbb{N}}$ est une suite de variables aléatoires à valeurs dans $[0,1]$, avec $X_0 = x_0 \in [0,1]$ (par exemple $X_0 = 1/2$) et, pour $n \geq 0$,

$$X_{n+1} = X_n + \begin{cases} \gamma_{n+1}(1 - X_n) \text{ si } A \text{ est actionné et gagnant} \\ \\ \end{cases}$$

Pour formaliser de manière plus précise, notons X_n la probabilité de choisir le bras A à l'étape n. La suite $(X_n)_{n\in\mathbb{N}}$ est une suite de variables aléatoires à valeurs dans $[0,1]$, avec $X_0 = x_0 \in [0,1]$ (par exemple $X_0 = 1/2$) et, pour $n \geq 0$,

$$X_{n+1} = X_n + \begin{cases} \gamma_{n+1}(1 - X_n) \text{ si } A \text{ est actionné et gagnant} \\ \\ -\gamma_{n+1} X_n \text{ si } B \text{ est actionné et gagnant} \end{cases}$$

Pour formaliser de manière plus précise, notons X_n la probabilité de choisir le bras A à l'étape n. La suite $(X_n)_{n\in\mathbb{N}}$ est une suite de variables aléatoires à valeurs dans $[0,1]$, avec $X_0 = x_0 \in [0,1]$ (par exemple $X_0 = 1/2$) et, pour $n \geq 0$,

$$X_{n+1} = X_n + \begin{cases} \gamma_{n+1}(1 - X_n) \text{ si } A \text{ est actionné et gagnant} \\ \\ -\gamma_{n+1} X_n \text{ si } B \text{ est actionné et gagnant} \end{cases}$$

Cela s'écrit aussi, en introduisant une suite de variables aléatoires $(U_n)_{n\in\mathbb{N}}$ indépendantes et uniformément distribuées sur $[0,1]$ et des événements $(A_n, B_n)_{n\in\mathbb{N}}$ de probabilités respectives p_A et p_B, indépendants, et indépendants de la suite (U_n),

$$X_{n+1} = X_n + \gamma_{n+1}(1-X_n)\mathbf{1}_{A_{n+1}}\mathbf{1}_{\{U_{n+1}\leq X_n\}} - \gamma_{n+1}X_n\mathbf{1}_{B_{n+1}}\mathbf{1}_{\{U_{n+1}>X_n\}}$$

On a, en notant $\mathcal{F}_n$ la tribu engendrée par les variables aléatoires $X_0, X_1, \ldots, X_n$,

$$\mathbb{E}\left(X_{n+1} \mid \mathcal{F}_n\right) = X_n +$$

On a, en notant $\mathcal{F}_n$ la tribu engendrée par les variables aléatoires $X_0, X_1, \ldots, X_n$,

$$\mathbb{E}\left(X_{n+1} \mid \mathcal{F}_n\right) = X_n + \gamma_{n+1}(1 - X_n)X_n\left(p_A - p_B\right).$$

On a, en notant $\mathcal{F}_n$ la tribu engendrée par les variables aléatoires $X_0, X_1, \ldots, X_n$,

$$\mathbb{E}\left(X_{n+1} \mid \mathcal{F}_n\right) = X_n + \gamma_{n+1}(1 - X_n)X_n\left(p_A - p_B\right).$$

Si $p_A > p_B$, la suite $(X_n)_{n\in\mathbb{N}}$ est une sous-martingale,
Si $p_A < p_B$, la suite $(X_n)_{n\in\mathbb{N}}$ est une sur-martingale,
Si $p_A = p_B$, la suite $(X_n)_{n\in\mathbb{N}}$ est une martingale.
Les théorèmes de convergence des sous-martingales permettent d'affirmer que la suite $(X_n)_{n\in\mathbb{N}}$ converge presque sûrement vers une limite X_∞ vérifiant $0 \leq X_\infty \leq 1$ et on a, par convergence dominée,

$$\lim_{n\to\infty} \mathbb{E}(X_n) = \mathbb{E}(X_\infty).$$

On a, en notant $\mathcal{F}_n$ la tribu engendrée par les variables aléatoires $X_0, X_1, \ldots, X_n$,

$$\mathbb{E}\left(X_{n+1} \mid \mathcal{F}_n\right) = X_n + \gamma_{n+1}(1 - X_n)X_n\left(p_A - p_B\right).$$

Si $p_A > p_B$, la suite $(X_n)_{n\in\mathbb{N}}$ est une sous-martingale,
Si $p_A < p_B$, la suite $(X_n)_{n\in\mathbb{N}}$ est une sur-martingale,
Si $p_A = p_B$, la suite $(X_n)_{n\in\mathbb{N}}$ est une martingale.
Les théorèmes de convergence des sous-martingales permettent d'affirmer que la suite $(X_n)_{n\in\mathbb{N}}$ converge presque sûrement vers une limite X_∞ vérifiant $0 \leq X_\infty \leq 1$ et on a, par convergence dominée,

$$\lim_{n\to\infty} \mathbb{E}(X_n) = \mathbb{E}(X_\infty).$$

Par ailleurs

$$\mathbb{E}\left(X_{n+1}\right) = \mathbb{E}(X_n) + \gamma_{n+1}\mathbb{E}\left[(1 - X_n)X_n\right]\left(p_A - p_B\right)$$

Si $p_A \neq p_B$ on peut supposer (par symétrie) $p_A > p_B$, et on espère $\mathbb{P}(X_\infty = 1) = 1$. Il est clair que si la série $\sum \gamma_n$ diverge, on a $\mathbb{E}\left[(1 - X_\infty)X_\infty\right] = 0$ et par conséquent $\mathbb{P}\left(X_\infty \in \{0, 1\}\right) = 1$.

Si $p_A \neq p_B$ on peut supposer (par symétrie) $p_A > p_B$, et on espère $\mathbb{P}(X_\infty = 1) = 1$. Il est clair que si la série $\sum \gamma_n$ diverge, on a $\mathbb{E}\left[(1 - X_\infty)X_\infty\right] = 0$ et par conséquent $\mathbb{P}\left(X_\infty \in \{0, 1\}\right) = 1$.

Théorème

On suppose $\gamma_n = \frac{C}{n+C}$, *avec* $C > 0$, *et* $x_0 \in]0, 1[$. *Alors, on a* $\mathbb{P}(X_\infty = 1) = 1$ *si et seulement si* $C \leq 1/p_B$.

Référence : DL, G. Pagès, P. Tarrès (2004). Pour le cas $C = 1$, cf P. Tarrès (2000).

Bandits à K bras

Le bandit à K bras peut être modélisé de la façon suivante. A chacun des K bras, correspond une suite de variables aléatoires indépendantes et équidistribuées (à valeurs dans $[0,1]$ pour simplifier) : $(X_n^{(k)})_{n\in\mathbb{N}}$, $k = 1, \ldots, K$.
On note

$$\mu_k = \mathbb{E}\left(X_n^{(k)}\right) \quad \text{et} \quad \mu^* = \max_{1\leq k\leq K} \mu_k.$$

Bandits à K bras

Le bandit à K bras peut être modélisé de la façon suivante. A chacun des K bras, correspond une suite de variables aléatoires indépendantes et équidistribuées (à valeurs dans $[0,1]$ pour simplifier) : $(X_n^{(k)})_{n\in\mathbb{N}}$, $k = 1, \ldots, K$.
On note

$$\mu_k = \mathbb{E}\left(X_n^{(k)}\right) \quad \text{et} \quad \mu^* = \max_{1\leq k\leq K} \mu_k.$$

Pour chaque instant $t \in \mathbb{N}$, on note I_t le numéro du bras actionné. Chaque variable aléatoire I_t est donc à valeurs dans $\{1, 2, \ldots, K\}$ et si on note X_t le gain à l'instant t, la loi conditionnelle de X_t sachant $\{I_t = k\}$ est la loi de $X_1^{(k)}$. Le gain cumulé à l'instant T s'écrit

$$S_T = \sum_{t=1}^{T} X_t$$

et on cherche à maximiser $\mathbb{E}(S_T)$, ou, de façon équivalente, à minimiser le regret, défini par $R_T = \mu^* T - \mathbb{E}(S_T)$.

Notons que

$$\begin{aligned}\mathbb{E}(S_T) = \sum_{t=1}^{T} \mathbb{E}(X_t) &= \sum_{t=1}^{T} \sum_{k=1}^{K} \mathbb{E}\left(X_t \mid I_t = k\right) \mathbb{P}(I_t = k) \\ &= \sum_{k=1}^{K} \mu_k \sum_{t=1}^{T} \mathbb{P}(I_t = k) \\ &= \sum_{k=1}^{K} \mu_k \mathbb{E}\left(N_k(T)\right),\end{aligned}$$

où $N_k(T) = \sum_{t=1}^{T} \mathbf{1}_{\{I_t = k\}}$, nombre d'interventions du bras k entre les instants 1 et T.

Notons que

$$\begin{aligned}\mathbb{E}(S_T) = \sum_{t=1}^{T} \mathbb{E}(X_t) &= \sum_{t=1}^{T} \sum_{k=1}^{K} \mathbb{E}\left(X_t \mid I_t = k\right) \mathbb{P}(I_t = k) \\ &= \sum_{k=1}^{K} \mu_k \sum_{t=1}^{T} \mathbb{P}(I_t = k) \\ &= \sum_{k=1}^{K} \mu_k \mathbb{E}\left(N_k(T)\right),\end{aligned}$$

où $N_k(T) = \sum_{t=1}^{T} \mathbf{1}_{\{I_t = k\}}$, nombre d'interventions du bras k entre les instants 1 et T. On en déduit, en écrivant $T = \sum_{k=1}^{K} \mathbb{E}\left(N_k(T)\right)$,

$$\begin{aligned}R_T &= \mu^* T - \mathbb{E}(S_T) \\ &= \sum_{k=1}^{K} \left(\mu^* - \mu_k\right) \mathbb{E}\left(N_k(T)\right)\end{aligned}$$

Sous des hypothèses techniques, Lai et Robbins (1985) démontrent que la croissance du regret est au moins logarithmique et établissent une borne inférieure explicite pour $\liminf_{T\to\infty} R_T/\log T$. Cela conduit à la notion de stratégie asymptotiquement optimale, et à la recherche de stratégies asymptotiquement optimales explicites.

Sous des hypothèses techniques, Lai et Robbins (1985) démontrent que la croissance du regret est au moins logarithmique et établissent une borne inférieure explicite pour $\liminf_{T\to\infty} R_T/\log T$. Cela conduit à la notion de stratégie asymptotiquement optimale, et à la recherche de stratégies asymptotiquement optimales explicites.
La stratégie suivante, proposée par Auer, Cesa-Bianchi et Fischer (2002) est une stratégie dite *optimiste* basée sur les bornes supérieures des intervalles de confiance pour l'estimation de la moyenne. Pour chaque $k = 1, \ldots, K$, on pose

$$B_k(t) = \frac{S_k(t)}{N_k(t)} + \sqrt{\frac{2\log t}{N_k(t)}},$$

où $S_k(t) = \sum_{s=1}^{t} X_s \mathbf{1}_{\{I_s=k\}}$ est la somme des gains dus à l'action du bras k entre 1 et t.
La stratégie consiste à poser $I_{t+1} = \text{argmax}_{k=1,\ldots K} B_k(t)$

Intervalle de confiance basé sur l'inégalité de Hoeffding

Theorem

Soit $(Y_n)_{n\geq 1}$ une suite de variables aléatoires indépendantes équidistribuées à valeurs ans $[0,1]$ et soit $\mu = \mathbb{E}(Y_1)$. On a, pour tout entier strictement positif N,

$$\mathbb{P}\left(\left|\frac{1}{N}\sum_{i=1}^{N} Y_i - \mu\right| > \varepsilon\right) \leq 2e^{-2N\varepsilon^2}$$

En prenant $\varepsilon = \sqrt{\frac{2\log t}{N}}$, on voit que, si on pose $\hat{\mu}_N = \sum_{i=1}^{N} Y_i/N$, l'intervalle $\left[\hat{\mu}_N - \sqrt{\frac{2\log t}{N}}, \hat{\mu}_N + \sqrt{\frac{2\log t}{N}}\right]$ est un intervalle de confiance de niveau $1 - 2t^{-4}$ pour μ.

Références

- E. Kaufmann (2016), *Matapli* **109**, 51-64.
- R. Munos, *Apprentissage par renforcement*, cours du master MVA, ENS de Cachan, http://researchers.lille.inria.fr/ munos/master-mva/
- T. L. Lai, H. Robbins (1985), Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, **6**, 4-22.
- P. Auer, N. Cesa-Bianchi, and P. Fischer (2002), Finite-time analysis of the multi-armed bandit problem. *Machine Learning*, **47**(2/3), 235-256.
- M.F. Norman (1968), On linear Models with Two Absorbing Barriers, *J. of Mathematical Psychlogy*, **5**, pp.225-241.
- I.J. Shapiro, K.S. Narendra, Use of Stochastic Automata for Parameter Self-Optimization with Multi-Modal Perfomance Criteria, *IEEE Trans. Syst. Sci. and Cybern.*, SSC-5, 352-360.

- D. Lamberton, G. Pagès, P. Tarrès (2004) When can the two-armed bandit algorithm be trusted? *Annals of Applied Probability* **14**, 1424-1454.
- P. Tarrès (2000), Pièges répulsifs, *C.R.A.S. Acad. Sc. de Paris*, Série I, **330**, pp.125-130.