The KL-UCB Algorithm for Bounded Stochastic Bandits and Beyond

Online learning and aggregation

Thomas Levy, Zakarya Ali

École nationale
de la statistique
et de l'administration
économique

3 Avril 2018

Plan

Bandits stochastiques

Dans le problème des bandits, on a K actions possibles (parmi K bras).
A chaque date $t = 1, 2, ..., T$:

- On choisit une action $I_t \in \{1, .., K\}$
- Chaque action I_t conduit a un gain X_{I_t}

Bandits stochastiques (bornés):

- $X_{i,t}$ sont tirés (de manière i.i.d) suivant une loi ν_i bornée (par ex: sur [0,1])
- E[$X_{i,t}$]=μ_i
- Avec ν_i et μ_i inconnu.

Le but est de maximiser le gain (ou de minimiser le regret)!

UCB

Notations :

- $S_i(t)$: gain total du bras i à la date t
- $T_i(t)$: nombre de fois où le bras i a été choisi à la date t
- $\widehat{\mu}_{i,T_i(t-1)} = \frac{S_i(t-1)}{T_i(t-1)}$: moyenne empirique des gains pour le bras i à $t-1$

Choix de l'action à la date t:

- $I_t = \underset{i=1,..,k}{argmax} \, [\widehat{\mu}_{i,T_i(t-1)} + \sqrt{\frac{\alpha.log(t)}{2T_i(t-1)}}]$

KL-UCB

Notation :

- $K(p,q)$: divergence de Kullback-Leibler
  Forme générale :

$$K(p,q) = \int_{-\infty}^{+\infty} p(X) log(\frac{p(X)}{q(X)}) dX$$

Choix de l'action à la date t:

$$\underset{i=1,..,k}{argmax}[\max_{q\in[0,1]}\{T_i(t-1).K(\widehat{\mu}_{i,T_i(t-1)},q) \leq log(t) + c.log(log(t))\}]$$

- Pour une convergence optimale, on choisit c=0

Etude du Pseudo-Regret

Forme générale:

- $\bar{R}_T = \sum_{i=1}^{k} \Delta_i x E[T_i(t)]$
- Avec: $\Delta_i = \mu^* - \mu_i$

Borne pour UCB:

- $\bar{R}_T \leq \sum_{i:\Delta_i \neq 0} [\frac{2\alpha \log T}{\Delta_i} + \frac{\alpha}{\alpha - 2}]$
- Avec: $\alpha > 2$

Pour KL-UCB:

- $\bar{R}_T \leq \sum_{i:\Delta_i \neq 0} [\frac{\Delta_i \alpha \log T}{K(\mu_i, \mu^*)} + C]$
- Optimal pour gain suivant Bernoulli

Scénario 1

2 bras avec gains suivant Bernoulli ($\mu_1 = 0.8,\ \mu_2 = 0.9$)

- Cas simple
- Simulation (10 000 fois) des algorithmes jusqu'à $t = 5000$
  - Pour UCB : on fait varier le paramètre α
  - Pour KL-UCB, on utilise :
    - différentes méthodes de maximisation de q
    - $K(p, q) = p.log(\frac{p}{q}) + (1 - p).log(\frac{1-p}{1-q})$, forme de la divergence pour deux distributions de Bernoulli
- On observe :
  - les gains obtenus
  - l'évolution moyenne du Pseudo-Regret en fonction du temps
  - la distribution du Pseudo-Regret à $t = 5000$

Notebook Scénario 1

Scénario 2

10 bras avec gains faibles (Bernoulli)

- Simulation (500 fois) des algorithmes jusqu'à $t = 50000$
- Représentatif de cas rencontrés dans l'Internet advertising...
- Pour le bras optimal: $\mu_9 = 0.1$
- Les 9 autres bras, $\mu = 0.05$, 0.02 ou 0.01.
- On observe :
  - les gains obtenus
  - l'évolution moyenne du Pseudo-Regret en fonction du temps
  - la distribution du Pseudo-Regret à $t = 5000$

Notebook Scénario 2

Scénario 3

5 bras avec gains exponentiels tronqués $(\frac{1}{5}, \frac{1}{4}, \frac{1}{3}, \frac{1}{2}, 1)$

- Les gains ne sont plus bornés sur [0,1] mais [0,10]
- Simulation (1 000 fois) des algorithmes jusqu'à $t = 5000$
- Mêmes UCB et KL-UCB qu'aux scénarios précédents
- On introduit KL-UCB Exponentiel avec

$$K(p, q) = \frac{p}{q} - 1 - log(\frac{p}{q})$$

Notebook Scénario 3

Bibliographie

Sébastien Bubeck and Nicolo Cesa-Bianchi. *Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems*. 2012.

Olivier Cappé and Aurélien Garivier. L'algorithme kl-ucb pour les bandits bornés, et au delà, 2011.

Olivier Cappé and Aurélien Garivier. The kl-ucb algorithm for bounded stochastic bandits and beyond. 2013.

Sarah Filippi, Olivier Cappé, and Aurélien Garivier. Optimism in reinforcement learning and kullback-leibler divergence. September 2011.

KL-UCB : Upper Confidence Bound par dichotomie

On cherche q tel que $K(\widehat{\mu}_{i,T_i(t-1)}, q) - \frac{log(t)}{T_i(t-1)} \simeq 0$

Pour n=0, on pose : $upperbound = u = \frac{log(t)}{T_i(t-1)}$ et $l = \widehat{\mu}_{i,T_i(t-1)}$

Algorithm 1 Upper Confidence Bound par dichotomie

```
while n < maxiterations AND u + l > epsilon do
    q = (l+u)/2
    if K(μ̂_{i,T_i(t−1)}, q) > upperbound then
        u = q
    else
        l = q
    end if
    n = +n1
end while
q = (l+u)/2
```

KL-UCB - Trouver l'Upper Confidence Bound avec la méthode de Newton

Pour chaque bras i, l'upper-confidence bound est :

$$\max_{q \in [0,1]} \{T_i(t-1).K(\widehat{\mu}_{i,T_i(t-1)}, q) \leq log(t) + c.log(log(t))\}, \text{ avec } c = 0$$

$q \mapsto K(x, q)$ est convexe, donc $q \mapsto K(\widehat{\mu}_{i,T_i(t-1)}, q) - \frac{log(t)}{T_i(t-1)}$ aussi

On cherche q tel que $f : q \mapsto K(\widehat{\mu}_{i,T_i(t-1)}, q) - \frac{log(t)}{T_i(t-1)} \simeq 0$

- On pose $q = p + h$, $h > 0$
- Par développement limité, on cherche $f(q) = f(p+h) = f(p) + f'(p)(q-p) + \frac{f''(p)}{2}(q-p)^2 = 0$

KL-UCB - Trouver l'Upper Confidence Bound avec la méthode de Newton

- or, $f''(p) = \frac{1}{p(1-p)}$, $f'(p) = 0$ et $f(p) = -\frac{log(t)}{T_i(t-1)}$
- donc, $q = p + \sqrt[2]{\frac{2 * log(t)}{f''(p) * T_i(t-1)}}$

On cherche par récurrence, $q_n = q_{n-1} + \sqrt[2]{\frac{2 * p(1-p) * log(t)}{T_i(t-1)}}$ tel que $f(q_n) \simeq 0$