Bandits finis sur un espace continu

S. Gaucher

Laboratoire de Mathématiques d'Orsay

3 Juin 2019

Problème de bandits finis sur un espace continu

Allocation de ressources:

On alloue des ressources limitées entre différentes actions :

- on choisit séquentiellement T actions **sans remise**;
- entre N actions;
- quand on complète une action, on reçoit un paiement aléatoire;
- notre but est de maximiser la somme des paiements reçus;
- chaque action est décrite par une covariable a_i, la moyenne du paiement est fonction de la covariable décrivant l'action;
- compromis entre exploration et exploitation.

Motivations (1)

Problème d'appariement

Le but est de trouver des "bonnes paires" :

- entre des joueurs pour un site de jeu en ligne;
- entre des utilisateurs pour un site de rencontre;
- entre des protéines, dans le cas de réseaux d'interaction proteine-proteine.

Chaque action (paire de joueurs, d'individus, de protéines) est décrite par des caractéristiques.

Motivations (2)

Allocation de ressources rares

T resources à allouer séquentiellement entre N candidats :

- resources médicales pour des patients
- places pour des étudiants
- aides contre la pauvreté.

Formulation du problème

Bandit fini sur un espace continu (FCAB)

- N actions, chaque action i est décrite par une covariable $a_i \in \mathcal{X}$
- On choisit séquentiellement T actions **sans remise**
- Si on choisit l'action (de covariable) a_i, on reçoit la récompense $y_i \in [0, 1]$ telle que $\mathbb{E}[y_i | a_i] = m(a_i)$ où m est la fonction de payment.

Objectif

On veut maximiser le gain total de notre stratégie ϕ. De manière équivalente, on cherche à minimiser le **regret**

$$R_T(\phi) = \sum_{1 \leq t \leq T} m(a_{\phi^*(t)}) - \sum_{1 \leq t \leq T} m(a_{\phi(t)})$$

où ϕ^* est la meilleure stratégie si on connaît m.

Remarques préliminaires

Ce problème ressemble à un problème de bandit sur un espace continu [Kleinberg, 2004, Auer et al., 2007].

Bandit sur un espace continu (CAB):

- à chaque tour t, on choisit une action $a_t \in \mathcal{X}$
- on reçoit un payment $y_t \in [0, 1]$ tel que $\mathbb{E}[y_t|a_t] = m(a_t)$

Différences entre (FCAB) et (CAB)

- même structure de dépendance entre les paiements
- le choix des actions pour est plus restreint dans (FCAB);
- dans (FCAB), on ne peut choisir chaque action qu'une fois;
- compromis exploration-exploitation différent;
- problème plus difficile?
- on peut obtenir des regrets plus faibles dans le cadre (FCAB)!

Hypothèses (1)

Budget

Le budget est une fraction $p \in (0,1)$ du nombre d'actions :

$$T = pN.$$

Hypothèse sur les actions

H1 : $\forall i \in \{1, ..., N\}$, $a_I \overset{i.i.d.}{\sim} \mathcal{U}([0,1])$.

Hypothèses (2)

Remarque

La stratégie oracle ϕ^* est telle que

$$m(a_{\phi^*(1)}) \geq m(a_{\phi^*(2)}) \geq ... \geq m(a_{\phi^*(N)}).$$

Elle choisit toutes les actions a_i tels que $m(a_i) \geq m(a_{\phi^*(T)})$. Sous **H1**, $\mathbb{E}\left[m(a_{\phi^*(T)})\right] = M$, où

$$M = \min\left\{A : \lambda\left(\{x : m(x) \geq A\}\right) < p\right\}$$

Condition lipschitzienne locale

H2 : $|m(x) - m(y)| \leq \max\{|M - m(x)|, L\,|x - y|\}$.

Condition de marge

H3 :$\lambda\left(\{x : |M - m(x)| \leq \epsilon\}\right) \leq Q\epsilon$.

Algorithme Upper Confidence Bound pour des bandits Finis (UCBF)

Paramètres : K, δ

- Diviser $[0, 1]$ en K intervalles I_k de même taille. N_k désigne le nombre d'actions dans I_k.
- Choisir une action dans chaque intervalle I_k tel que $N_k \geq 1$.
- Pour $t = K + 1, ..., T$:
  - $n_k(t-1)$ désigne le nombre d'actions choisies dans I_k avant t, et $\widehat{m}_k(n_k(t-1))$ le paiement moyen de ces actions;
  - Choisir un interval k_t maximisant

$$\widehat{m}_k(n_k(t-1)) + \sqrt{\frac{\log(T/\delta)}{2n_k(t-1)}}$$

    parmi les intervals tels que $N_k > n_k(t-1)$;
  - choisir une action uniformément dans I_{k_t}.

Contrôle du regret de UCBF (1)

En divisant l'espace $[0, 1]$ en K intervals, on se ramène au problème suivant :

Problème de bandit à K bras fini (FMAB)

- K actions
- on peut choisir la k-ième action N_k fois
- on note $m_k = \int_{a \in I_k} m(a)da$ le paiment moyen reçu si on choisit l'action k.

Stratégie oracle pour le problème discétisé $\phi^{(d)}$

- On suppose que $m_1 \geq m_2 \geq ... \geq m_K$
- et que $T = N_1 + N_2 + ... + N_f$
- alors $\phi^{(d)}$ choisit les actions de I_1, puis celles de I_2, ..., puis celles de I_f et

$$\sum_{1 \leq t \leq T} m(a_{\phi^{(d)}(t)}) \approx \sum_{1 \leq k \leq f} N_k m_k.$$

Contrôle du regret de UCBF (2)

Décomposition du regret

$$R_T(\phi) = \underbrace{\sum_{1\leq t\leq T} m(a_{\phi^*(t)}) - \sum_{1\leq t\leq T} m(a_{\phi^d(t)})}_{= R_T^{(d)}} + \underbrace{\sum_{1\leq t\leq T} m(a_{\phi^d(t)}) - \sum_{1\leq t\leq T} m(a_{\phi(t)})}_{= R_T^{(FMAB)}}$$

.

- $R_T^{(d)}$ correspond à l'erreur de discrétisation;
- $R_T^{(FMAB)}$ au regret du problème de bandits finis à K bras;
- il faut choisir K pour équilibrer ces deux termes.

Contrôle de l'erreur de discrétisation $R_T^{(d)}$

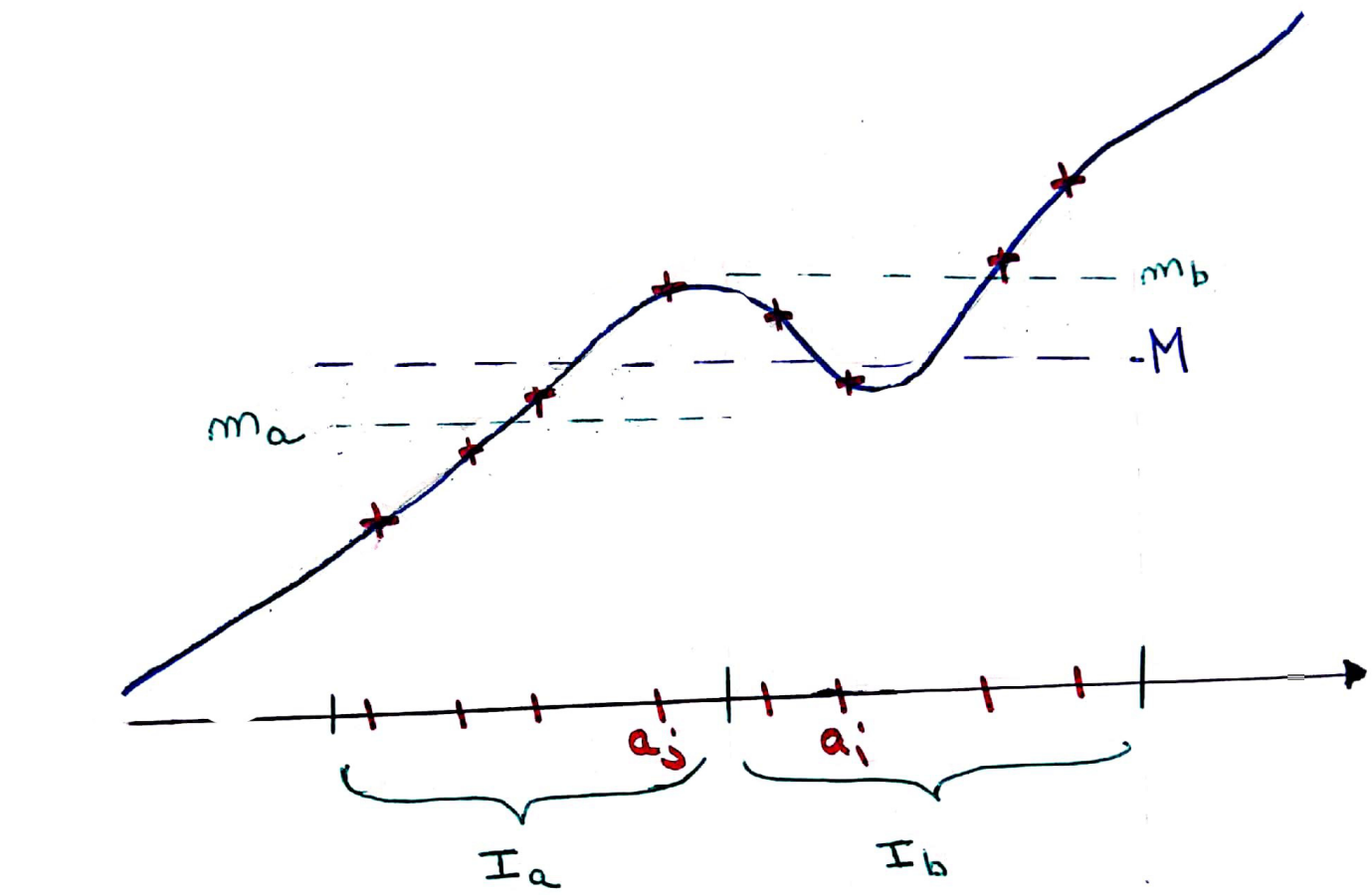
$$R_T^{(d)} = \sum_{1 \leq t \leq T} m(a_{\phi^*(t)}) - \sum_{1 \leq t \leq T} m(a_{\phi^d(t)}).$$

Il y a beaucoup de termes nuls!

Idée pour contrôler $R_T^{(d)}$ (1)

- ϕ^* choisit les actions a_i tels que $m(a_i) \geq m(a_{\phi^*(T)}) \simeq M$;
- $\phi^{(d)}$ choisit les actions a_i tels que $a_i \in I_k$ et $m_k \geq m_f \simeq M$.

Contrôle de l'erreur de discrétisation $R_T^{(d)}$



ϕ^d choisit

ϕ^* choisit

Contrôle de l'erreur de discrétisation $R_T^{(d)}$

$$R_T^{(d)} = \sum_{1 \le t \le T} m(a_{\phi^*(t)}) - \sum_{1 \le t \le T} m(a_{\phi^d(t)}).$$

Les actions a_i et a_j, choisies réciproquement par ϕ^d mais pas par ϕ^*, et inversement, ont des paiements dans $[M - L/K, M + L/K]$. Par **H3**, il y a $O(N/K)$ paires d'actions de ce type, donc

$$R_T(\phi) = O(N/K^2) = O(T/K^2).$$

Contrôle de l'erreur $R_T^{(FMAB)}$

$$R_T^{(FMAB)} \approx \sum_{1\leq k\leq f} N_k m_k - \sum_{1\leq k\leq K} n_k(T) m_k.$$

On montre que

$$\left.\begin{aligned} R_T^{(FMAB)} \approx & \sum_{1\leq k\leq f} (N_k - n_k(T))(m_k - M) \\ & + \sum_{k>f} n_k(T)(M - m_k) \end{aligned}\right\} = R_{opt} + R_{subopt}$$

$$= O(T/K^2 + K\log(T)\log(K)).$$

Contrôle du regret de UCBF (3)

On a donc

$$R_T(\phi) = O(T/K^2 + K \log(T) \log(K))$$

On choisit

$$K = N^{1/3} / \log(N)^{2/3}$$

Contrôle du regret de UCBF (4)

On suppose que $N^{1/3}/\log(N)^{2/3}\rfloor > p^{-1} \vee (1-p)^{-1}$. On choisit $K = \lfloor N^{1/3}/\log(N)^{2/3}\rfloor$ et $\delta = N^{-4/3}$.

Théorème

*Sous les hypothèses **H1**, **H2** et **H3**, il existe une constant $C_{L,Q,p}$ dépendant de L, Q et p telle que*

$$R_T(\phi) \leq C_{L,Q,p}\, T^{1/3} \log(T)^{4/3}$$

avec probabilité au moins $1 - 12(N^{-1} \vee e^{-N^{-1/3}/3})$.

Comparaison avec (CAB)

- sous les hypothèses **H1**, **H2** et **H3**, dans le cadre (FCAB) le regret R_T est de l'ordre de $T^{1/3}$
- Sous des hypothèses similaires, dans le cadre (CAB) le regret R_T est de l'ordre de $T^{1/2}$
- Dans (FCAB), le choix optimal est $K \approx T^{1/3}$
- Dans (CAB), le choix optimal est $K \approx T^{1/2}$
- moins d'intervalles pour une exploitation plus longue

Borne inférieure sur le regret

Hypothèse déterministe sur les actions

H4 $a_i = i/N$.

Paiement Bernoulli

H5 $y_i|a_i \sim$Bernoulli$(m(a_i))$.

Théorème

Pour tout $p \in (0,1)$, $L > 0$, $Q > (6/L \vee 12)$, *il existe* $N_{L,p}$ *dépendant de* L *et* p *tel que for all* $N \geq N_{L,p}$, *sous* ***H4*** *et* ***H5***

$$\inf_{\phi} \sup_{m \in \mathcal{F}_{Q,L}} \mathbb{P}\left(R_T^{\phi}(m) \geq 0.01\, T^{1/3} p^{-1/3}\right) \geq 0.1.$$

où $\mathcal{F}_{p,L,Q}$ est l'ensemble des fonctions satisfaisant **H2** et **H3**.

Références

Auer, P., Ortner, R., and Szepesvári, C. (2007).
Improved rates for the stochastic continuum-armed bandit problem.
In Bshouty, N. H. and Gentile, C., editors, *Learning Theory*, pages 454–468, Berlin, Heidelberg. Springer Berlin Heidelberg.

Kleinberg, R. (2004).
Nearly tight bounds for the continuum-armed bandit problem.
In *Proceedings of the 17th International Conference on Neural Information Processing Systems*, NIPS'04, page 697–704, Cambridge, MA, USA. MIT Press.