Finite Continuum-Armed Bandits

Solenne Gaucher

MATHEMATIQUES ORSAY
université PARIS-SACLAY
FACULTÉ DES SCIENCES D'ORSAY
NEURAL INFORMATION PROCESSING SYSTEMS

Problem: Sequential allocation of T resources between N actions described by covariates a_i. Each action can only be completed once, and results in a stochastic reward with mean $m(a_i)$. The aim is to maximize the cumulative reward.

Results: When $T \propto N$, the regret is $O\big(T^{1/3}\log(T)^{4/3}\big)$. When T/N decreases, the regret increases. When $T/N \sim N^{-1/3}\log(N)^{2/3}$, the problem becomes similar to a Continuum-Armed Bandit and the regret increases up to $O\big(T^{1/2}\log(T)\big)$.

A budget allocation problem

In the Finite Continuum-Armed Bandits (FCAB) problem, the agent has a **budget T** to spend on **N actions** described by covariates $\{a_1, a_2, \dots, a_N\} \in [0,1]^N$, where $T < N$. Each action i can only be **completed once**. At each time $t \leq T$, the agent

- selects an action with covariates $a_{\varphi(t)}$
- receives a corresponding stochastic reward in $[0,1]$ with mean $m(a_{\varphi(t)})$.

She aims at minimizing the regret

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

where φ^* is the oracle strategy : $m(a_{\varphi^*(1)}) \geq m(a_{\varphi^*(2)}) \geq \dots \geq m(a_{\varphi^*(N)})$.

Motivations

- Pair matching problem : one aims at discovering edges in graphs by conducting sequential tests (protein-protein interaction networks, social online networks). Side information on pairs of nodes is available.
- Allocation of scarce resources between competing candidates described by covariates (scholarship for students, medical supply for patients, financial help for households).
- Advertisement with pay-per-impression constraints and limited budget, when side information on potential customers is available.

Main challenges

The FCAB is more constrained than the classical Continuum-Armed Bandits (CAB) : it leads to lower cumulative rewards... but also **lower regrets**.

- Restricted choice of actions.
- The agent cannot select good actions indefinitely (but neither can the oracle strategy !)
- The agent must identify and select many good actions.

The ratio $p = T/N$ governs the difficulty of the problem:

- When $p \to 1$ the problem becomes trivial (any strategy must select all actions).
- When $p \to 0$, the problem becomes similar to a classical Continuum-Armed Bandit (good actions are always available).

Assumptions

Distribution of the covariates:

(A1) *For* $i = 1, \dots, N$, $a_i \sim \mathcal{U}([0,1])$ *i.i.d.*

$\boldsymbol{M = min\{A : \lambda(\{x : m(x) \geq A\}) < p\}}$ is the expected reward for selecting $\boldsymbol{\varphi^*(T)}$.

Weak Lipschitz assumption:

(A2) : There exists $L > 0$ such that for all $(x, y) \in [0,1]^2$,

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

Margin assumption:

(A3) : There exists $Q > 0$ such that for all $\varepsilon \in (0,1)$,

$$\lambda(\{x : |M - m(x)| \leq \varepsilon\}) \leq Q\varepsilon$$

Strategy

Discretize the problem as a Finite Multi-Armed Bandit (FMAB) problem

- Divide $[0,1]$ into K same size intervals.
- Each interval I_k can be selected at most N_k times, where N_k is the number of actions in I_k.
- For $k \in \{1, \dots, K\}$, $m_k = K \int_{I_k} m(a)da$ is the expected reward for selecting an action in I_k.

Apply the UCB algorithm on the FMAB problem

Discard an interval once all its actions have been selected.

Upper Confidence Bound for FCAB (UCBF) Algorithm

Parameters : K, δ

Initialization :

- Divide $[0,1]$ into K same size intervals I_k
- Discard intervals with no actions
- Select one action uniformly at random in each interval
- Discard those actions

For $t = K+1, \dots, T$:

- Discard intervals with no actions
- Select $k \in argmax\ \widehat{m}_k\big(n_k(t-1)\big) + \sqrt{\frac{\log(T/\delta)}{2n_k(t-1)}}$
- Select an action uniformly at random among the actions in I_k
- Discard this action

Upper bounds on the regret

under (A1), (A2) (A3)

Regime : $\boldsymbol{T = pN}$ for $p \in (0,1)$

Assume that $(p^{-1} \vee (1-p)^{-1}) < \lfloor N^{1/3}\log(N)^{-2/3} \rfloor$. There exists $C_{L,Q}$ depending only on L and Q such that for the choice $K = \lfloor N^{1/3}\log(N)^{-2/3} \rfloor$, and $\delta = N^{-4/3}$, with probability $O(N^{-1})$,

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

Regime : $\boldsymbol{T = 0.5N^{\alpha}}$ for $\alpha \in (2/3 + \epsilon_N, 1]$ $\qquad \epsilon_N = \left(\frac{2}{3}\log\log(N) + \log(2)\right)/\log(N)$

There exists $C_{L,Q}$ depending only on L and Q such that for the choice $K = \lfloor \alpha^{2/3}\,(2T)^{1/(3\alpha)}\log(2T)^{-2/3} \rfloor$ and $\delta = N^{-4/3}$, with probability $O(N^{-1})$,

$$\boldsymbol{R_T \leq C_{L,Q}T^{1/(3\alpha)}\log(T)^{4/3}}.$$

Lower bounds on the regret

(A4) : $a_i = i/N$ for $i = 1, \dots, N$ (equally spaced actions)
(A5) : reward for selecting a_i is $Bernoulli(m(a_i))$ (special case of FCAB)
$\mathfrak{F}_{p,Q,L}$: functions verifying assumptions (A2) and (A3)

Regime : $\boldsymbol{T = pN}$ for $p \in (0,1)$

For all $p \in (0,1)$, all $L > 0$, all $Q > (6/L \vee 12)$, there exists a constant C_L depending on L such that under (A4) and (A5), for all $N \geq C_L(p^{-3} \vee (1-p)^{-3})$,

$$\boldsymbol{inf_{\varphi}sup_{m \in \mathfrak{F}_{p,Q,L}}\mathbb{P}\big(R_T^{\varphi}(m) \geq 0.01T^{1/3}p^{-1/3}\big) \geq 0.1}.$$

Regime : $\boldsymbol{T = 0.5N^{\alpha}}$ for $\alpha \in (2/3 + C_L/\log(N), 1]$

For all $L > 0$, all $Q > (6/L \vee 12)$, there exists a constant C_L depending on L such that under (A4) and (A5), for all $N \geq \exp(3C_L)$ and all $T = 0.5N^{\alpha}$ for some $\alpha \in \left(\frac{2}{3} + \frac{C_L}{\log(N)}, 1\right]$,

$$\boldsymbol{inf_{\varphi}sup_{m \in \mathfrak{F}_{0.5N^{\alpha},Q,L}}\mathbb{P}\big(R_T^{\varphi}(m) \geq 0.01T^{1/(3\alpha)}\big) \geq 0.1}.$$

References

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

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.