Incentivized Learning in Principal - Agent Bandit Games

A. Scheid , D. Tiapkin , E. Boursier , A. Capitaine , E. Moulines , M. Jordan , E. El Mhamdi , A. Durmus   Ecole Polytechnique , INRIA , U.C. Berkeley , ENS

Take home message

- Principal's strategy: decouple the problem and learn first the optimal incentives with a binary search and then run any bandit subroutine.
- The cost to learn the optimal incentives is very small as compared to any subroutine regret.
- Extension to the linear contextual bandit setting.

Setting and objectives

- **Setting:** Two players: the principal and the agent with a bandit instance $(\nu_a)_{a\in\mathcal{A}}$ for the principal and known rewards $\mathrm{s} = (\mathrm{s}_1, \dots, \mathrm{s}_K) \in \mathbb{R}_+^K$ for the agent. Set of actions $\mathcal{A} = [K]$.
- **Game:** over $T \in \mathbb{N}^\star$ rounds. At any step $t \in [T]$, the principal proposes a transfer $\pi(t)$ to the agent associated with an action $a_t \in \mathcal{A}$. During the round, agent picks action $A_t \in \mathcal{A}$ and their utilities are

$$X_{A_t}(t) - \mathbb{1}a_t(A_t)\pi(t) \text{ for the principal, where } X_{A_t}(t) \sim \nu_{A_t} ,$$
$$\mathrm{s}_{A_t} + \mathbb{1}a_t(A_t)\pi(t) \text{ for the agent .}$$

- **Agent's behaviour:** We assume that the agent is myopic and always maximises his instantaneous utility, hence the choice of A_t

$$A_t \in \mathrm{argmax}_{a\in\mathcal{A}}\{\mathrm{s}_a + \mathbb{1}a_t(a)\pi(t)\} .$$

- **Principal's objective:** Maximize his utility and solve

$$\begin{aligned} &\text{maximize } \int x\nu_a(dx) - \pi \text{ over } \pi \in \mathbb{R}_+, a \in [K] \\ &\text{such that } a \in \mathrm{argmax}_{a'\in[K]} \{\mathrm{s}_{a'} + \mathbb{1}_a(a')\pi\} , \end{aligned} \tag{1}$$

which is equivalent to minimizing her regret (where $\mu^\star$ solution of (1))

$$\mathfrak{R}(T) = T\,\mu^\star - \sum_{t=1}^{T} \mathbb{E}[X_{A_t}(t) - \mathbb{1}_{a_t}(A_t)\pi(t)] .$$

Questions:

- From the principal's side, how can we define the optimal incentives to guide the agent's behaviour?
- How can we learn the optimal incentives as well as playing on the bandit instance?

If minimal incentives $\pi_a^\star = \max_{a'\in[K]} \mathrm{s}_{a'} - \mathrm{s}_a$ to enforce $A_t = a$ are known, the problem is reduced to a shifted bandit instance, hence the idea of `IPA`:

- First learn the optimal incentives with a precision $1/T$ through a binary search like procedure: $O(K\log_2(T))$ rounds.
- Then run any bandit subroutine on the shifted bandit instance.

→ Separate the learning of the optimal incentives from the bandit game to get the optimal regret bound.

Principal's strategy

Decouple the problem between a first binary search phase to learn the optimal incentives and the run of a bandit subroutine.

- **Binary search steps:** We show that a binary search procedure can be run in $K\lceil\log_2 T\rceil$ rounds such that we obtain $\hat{\pi}_a$ at the end and

$$|\hat{\pi}_a - \pi_a^\star| \le 2/T \text{ for any } a \in \mathcal{A} .$$

as well $\hat{\pi}_a > \pi_a^\star$, and therefore, for any step $t \ge K\lceil\log_2 T\rceil$

$$\text{if } (a_t, \pi(t)) = (a, \hat{\pi}_a), \text{ then } A_t = a .$$

→ after the binary search, the principal can guide the agent's action with an extra cost of at most $2/T$.

- **Bandit subroutine:** Then, `ALG` (which can be `UCB` or `ETC` for instance) is run the bandit instance with rewards $(X_a(1) - \hat{\pi}_a)_{a\in\mathcal{A}} \sim \rho$.

Regret bound for the principal-agent game

Theorem. `IPA` run over T rounds has an overall regret $\mathfrak{R}(T)$ such that

$$\mathfrak{R}(T) \le \mathcal{O}(\sqrt{KT\log(T)}) ,$$

with `Alg` = `UCB` as the principal's subroutine on the shifted multi-armed bandit after the binary search.

Extension to the contextual case

- Set of possible actions $\mathcal{A}_t \subseteq \mathrm{B}(0,1)$, where $\mathrm{B}(0,1)$ stands for the unit closed ball in $\mathbb{R}^d$, family of zero-mean distributions $(\tilde{\nu}_a)_{a\in\mathrm{B}(0,1)}$ such that

$$\text{for any } a \in \mathrm{B}(0,1), t \in [T], \eta_a(t) \sim \tilde{\nu}_a .$$

- Principal's reward: family $\{(X_a(t))_{a\in\mathrm{B}(0,1)} : t \in [T]\}$ of independent random variables such that for any $t \in [T], a \in \mathrm{B}(0,1)$,

$$X_a(t) := \langle\theta^\star, a\rangle + \eta_a(t) ,$$

and agent's reward: $(\langle\mathrm{s}^\star, a\rangle)_{a\in\mathrm{B}(0,1)}$. At each step t, the principal offers a transfer $\kappa(t,\cdot)$ and aims to design $\kappa(t,\cdot)$ to find

$$\begin{aligned} &\text{maximize } \langle\theta^\star, a\rangle - \kappa(t,a) \text{ over } \kappa(t,\cdot) : \mathcal{A}_t \to \mathbb{R}_+, \\ &\text{such that } a \in \mathrm{argmax}_{a'\in\mathcal{A}_t}\{\langle\mathrm{s}^\star, a'\rangle + \kappa(t,a')\} . \end{aligned}$$

- **Differences with the multi-armed case:** Although the problem seems very similar, the binary search cannot be run as before due to the set $\mathcal{A}_t$ which changes over the steps, hence our definition of the event

$$\mathcal{E}_t := \left\{ \max_{a_t^1 \ne a_t^2 \in \mathcal{A}_t} \text{diam}\left(\mathcal{S}_t, \frac{a_t^1 - a_t^2}{\|a_t^1 - a_t^2\|}\right) < \frac{1}{T} \right\} ,$$

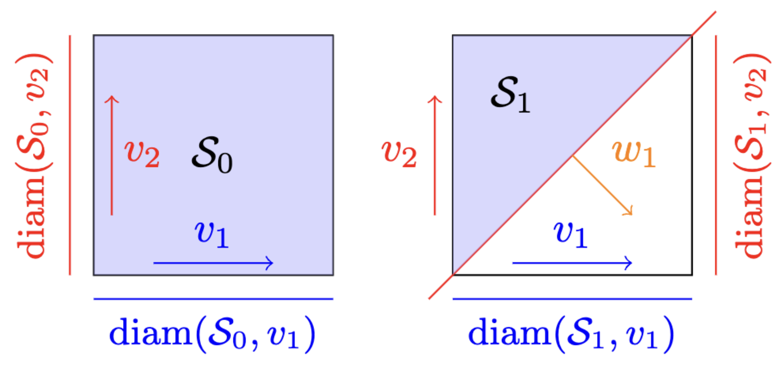
to decide wheter `Contextual IPA` runs a multidimensional binary search or a contextual bandit subroutine.

Recovering a regret bound

Theorem. If `Contextual IPA` is run with the corruption robust subroutine `CW-OFUL` (He et al., 2022), the regret of `Contextual IPA` is bounded as

$$\mathfrak{R}(T) \le \mathcal{O}(d\log(dT) + d\sqrt{T}\log(T)) .$$

Main technical difficulty: the multidimensional binary search that we achieve with the bound from Lobel et al., 2018. Without any extra cost nor assumption on $\mathcal{A}_t$, we converge towards the optimum!



Volume of $\mathcal{S}_0$ cut along a direction w_1 while the diameter is not reduced along v_1 nor v_2..

Experiments

Cumulative regret of `IPA` on a 5 arms, 1-subgaussian rewards bandit.