

The Price of Incentivizing Exploration: A Characterization via Thompson Sampling and Sample Complexity

Mark Sellke (Stanford University) & Aleksandrs Slivkins (Microsoft Research NYC): EC 2021, <https://arxiv.org/abs/2002.00558>

1 Incentivized Exploration: multi-armed bandits with incentives

- K arms a_1, \dots, a_K . IID rewards, independent prior \mathcal{P}_i over each mean reward μ_i .
Each round t : new agent arrives, ALG recommends arm A_t ,
agent chooses an arm which maximizes $\mathbb{E}_{\mathcal{P}}[\mu_a | A_t]$
- Info flow: ALG observes chosen arms & rewards, each agent only observes A_t
(if ALG can reveal an arbitrary message, WLOG this message is A_t)
- Agent obeys ALG if **Bayesian Incentive-Compatible (BIC)**:
 $\mathbb{E}[\mu_a - \mu_b | A_t = a] \geq 0 \quad \forall t, \text{ arms } a, b$.
- Long line of work starting from [Kremer, Mansour, Perry: EC'13, JPE'14]
- Our angle: what is the *Price of Incentives (Pol)* vs. ordinary bandits?
 - Bayesian regret: $\mathbb{E}_{\mathcal{P}}[T \cdot \mu_{a^*} - \sum_t \mu_{A_t}]$
 - Sample complexity: time T_{SC} to explore every arm.*
- [Mansour-Slivkins-Syrgkanis EC'15, OpRe'20]: reduction from any bandit algorithm
 - Bayesian regret $BReg(T) \leq c_{\mathcal{P}} \sqrt{T}$.
 - $c_{\mathcal{P}}$ and T_{SC} can be exponential in K and $1/\sigma_{\min}$, where $\sigma_{\min}^2 = \min_a \text{Var}(\mathcal{P}_a)$

*Some arms may be unexplorable by BIC algorithm. Restrict to the explorable arms.

2 Thompson Sampling (TS) is BIC After Warm Start

TS is a standard bandit algorithm: $\mathbb{P}[A_t = a | \text{history}] = \mathbb{P}[a \text{ is best arm} | \text{history}]$.

Theorem: if $N_{\mathcal{P}}$ samples of each arm collected by time t , then TS is BIC after time t .

Corollary: Bayesian regret $\leq N_{\mathcal{P}} T_{SC} + O(\sqrt{T}) \Rightarrow$ additive Pol $N_{\mathcal{P}} T_{SC}$

Moreover: if $\mathbb{E}[\mu_1] = \dots = \mathbb{E}[\mu_K]$ then $N_{\mathcal{P}} = 0$.

Arms from finite set \mathcal{C} of types: $N_{\mathcal{P}} \leq O_{\mathcal{C}}(K)$ always.

Upper bound on $T_{SC} \rightarrow$ upper bound on Bayesian regret.

Neat proof that TS is BIC when $\mathbb{E}[\mu_1] = \dots = \mathbb{E}[\mu_K]$:

- Want to show $\mathbb{E}[\mu_i - \mu_j | A_t = a_i] \geq 0$. Bayes and defn of TS imply:

$$\mathbb{E}[\mu_i - \mu_j | A_t = a_i] = \frac{\mathbb{E}[\mathbb{E}^t[\mu_i - \mu_j] \cdot \mathbb{P}^t[A^* = a_i]]}{\mathbb{P}[A^* = a_i]}.$$

- **Red/blue** terms are **martingales that always move the same direction**.

\Rightarrow **Product** is submartingale \Rightarrow numerator increases from 0 $\Rightarrow \mathbb{E}[\mu_i - \mu_j | A_t = a_i] \geq 0$. ■

3 BIC Algorithm for Initial Sampling: An Over-Simplified Outline

Goal: obtain 1 sample of each arm with a BIC algorithm.

To explore each arm $j = 1 \dots K$ (in decreasing order of $\mathbb{E}[\mu_1] \geq \dots \geq \mathbb{E}[\mu_K]$)

Getting started: if all $j - 1$ previous arms are terrible, sample arm j

Loop: in k -th iteration, $p_k := \mathbb{P}[\text{have explored arm } j]$,

$\mathbb{P}[\text{explore arm } j \text{ in this iteration}] = \lambda \cdot p_k \Rightarrow p_{k+1} \leftarrow (1 + \lambda)p_k$.

Choose between 3 branches: explore arm j , exploit, and **j -exploit** (*the secret sauce*)

j -exploit: if arm j already explored, carefully choose when to recommend it, **maximizing**

$$\lambda := \min_{i < j} \mathbb{E}[\mu_j - \mu_i | A_t = j \text{ in } j\text{-exploitation}].$$

counterbalances $\lambda \cdot p_k$ amount of fresh exploration.

Exponential growth: $p_k \sim p_0 e^{\lambda k} \Rightarrow T_{SC} \approx \lambda^{-1}$.

Optimal λ via minimax theorem: $\lambda = \min_{j \in [K], q \in \Delta_{j-1}} \mathbb{E}[(\mu_j - \mu_q)_+]$, $\mu_q := \sum_i q_i \mu_i$

4 Consequences for Sample Complexity

Lower bound: $T_{SC} \geq L := \max_{j \in [K], q \in \Delta_K} \mathbb{E}[\mu_q - \mu_j] / \mathbb{E}[(\mu_j - \mu_q)_+]$.

- Proof idea: need to play arm j at least once, beat μ_q on average.

Theorem: $T_{SC}(ALG) \leq \text{poly}(L, \sigma_{\min}^{-1}, K)$. Resolves poly vs. exp dependence on K and σ_{\min}

Dependence on the prior for Beta priors:

Optimal $T_{SC} \sim \sigma_{\min}^{-2} m^{O(Km)}$, where $m = 1/[\text{second smallest } \text{Var}(\mathcal{P}_a)]$

Easy for *one* well-known arm (important special case!), difficult for ≥ 2

Dependence on K : Linear vs Exp Dichotomy if all priors \mathcal{P}_i lie in finite set \mathcal{C}

(a) If $\mathbb{P}[\mu_j > \mathbb{E}[\mu_i] + \delta] > \delta$ for all $\mu_i, \mu_j \in \mathcal{C}$: $T_{SC} = O_{\delta}(K)$.

Typical case

(b) If $\mathbb{P}[\mu_j > \mathbb{E}[\mu_i] - \delta] = 0$ for some $\mu_i, \mu_j \in \mathcal{C}$: $T_{SC} \geq \exp_{\delta}(K)$.