Incentivizing Exploration with Unbiased Histories

Nicole Immorlica[1], Jieming Mao[2], Aleksandrs Slivkins[1], Zhiwei Steven Wu[3]

Microsoft Research[1], University of Pennsylvania[2], University of Minnesota[3]

Problem

- Incentivizing Exploration:
  - Agents arrive sequentially in a social learning setting
  - The principal balances exploration and exploitation by controlling the information each agent receives
- Prior work achieves much progress but relies heavily on trust and rationality assumptions (e.g. direct recommendation of an action)
- We would like to retain the trustworthiness of revealing the full history but it is impossible without using information asymmetry
- Achieve low-regret with messages called *unbiased subhistories*
  - Actions and rewards from a subsequence of past agents
  - The subsequence is chosen ahead of time

Model

A game consists of T rounds between a principal and T agents. Each round $t \in [T]$ proceeds as follows:

- Agent t receives message m_t from the principal
- Agent t chooses action $a_t \in \mathcal{A}$, $|\mathcal{A}| =$ constant K
- Agent t obtains reward $r_t \in \{0, 1\}$ sampled from Bernoulli $\mu_{a_t} \in [1/3, 2/3]$
- Agent t reports r_t back to the principal

Unbiased subhistories: the subhistory for a subset of rounds $S \subset [T]$

$$\mathcal{H}_S = \{(s, a_s, r_s) : s \in S\}$$

with two properties:

- Unbiased: S_t is chosen before round 1
- Transitive: $t \in S_{t'} \Rightarrow S_t \subset S_{t'}$

Agents' behavior: Agent t forms an estimate $\hat{\mu}_a^t$ for each arm a and chooses the one with the highest estimate. Estimates are close to empirical averages in the subhistories:

$$|\hat{\mu}_a^t - \bar{\mu}_a^t| < \frac{C}{\sqrt{N_{t,a}}}.$$

Regret:

$$Reg(T) = T \cdot \max_{a \in \mathcal{A}} \mu_a - \sum_{t \in [T]} \mathbb{E}[\mu_{a_t}].$$

Warm-up: Two-level Policy

Full-disclosure Path: reveal the full history in each round
Lemma: a full-disclosure path of some constant length samples each arm at least once with constant probability

Two-level Policy:

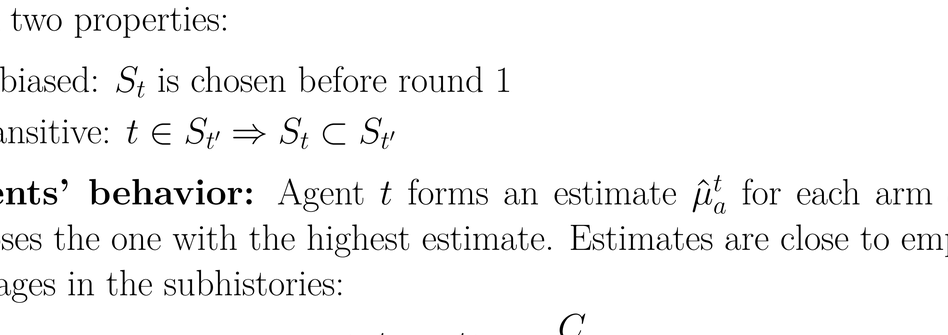

Figure 1: Info-graph for the 2-level policy.

Theorem:

$$Reg(T) \leq O_K(T^{2/3}(\log(T))^{1/3})$$

Three-level Policy

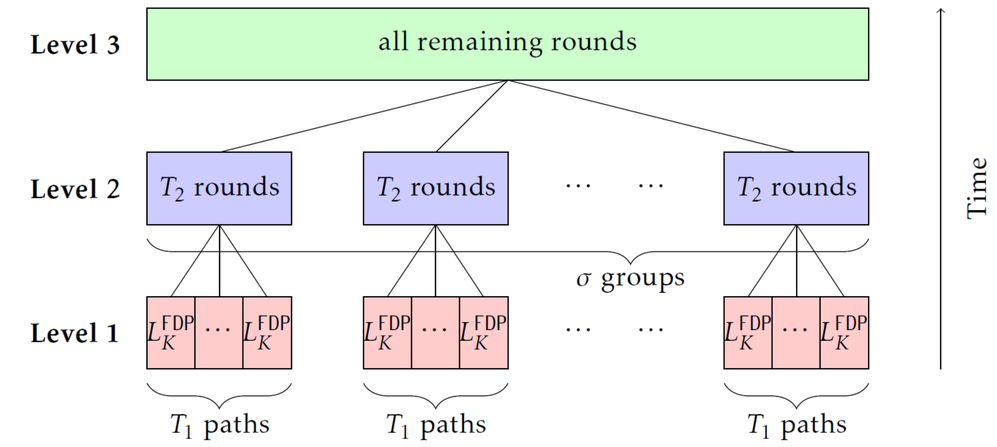

Figure 2: Info-graph for the three-level policy. Each red box in level 1 corresponds to T_1 full-disclosure paths of length L_K^{FDP} each.

Theorem:

$$Reg(T) \leq O_K(T^{4/7} \log(T))$$

Proof Sketch (for 2 arms): Wlog $\mu_1 \geq \mu_2$, $\Delta = \mu_1 - \mu_2$
$T_1 = T^{4/7} \log^{-1/7}(T)$, $T_2 = T^{6/7} \log^{-5/7}(T)$, $\sigma = \log(T)$
Case analysis based on Δ:

- (Negligible gap) $\Delta < T^{-3/7} \log^{6/7}(T)$:
  picking any arm gives small regret

- (Large gap) $\Delta > \sqrt{\log(T)/T_1} = T^{-2/7} \log^{4/7}(T)$:
  concentration in the first level
  $\Rightarrow$ all agents in the second and the third levels pull arm 1

- (Small gap) $\Delta \in (T^{-3/7} \log^{6/7}(T), \sqrt{1/T_1})$:
  anti-concentration in the first level
  $\Rightarrow$ both arms are pulled $\geq T_2$ times, concentration in the second level
  $\Rightarrow$ all agents in the third level pull arm 1

- (Medium gap) $\Delta \in (\sqrt{1/T_1}, \sqrt{\log(T)/T_1})$:
  concentration in the first level
  $\Rightarrow$ all agents in the third level pull arm 1

L-level Policy

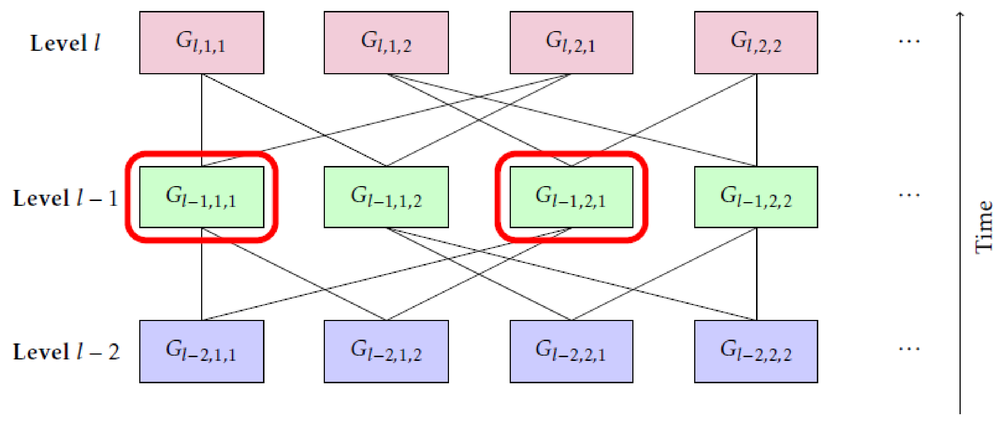

Figure 3: Interlacing connections between levels for the L-level policy.

Interlacing connection: Agents in Group $G_{l,u,v}$ sees history of agents in group $G_{l-1,v,w}$ for all $u, v, w \in [\log(T)]$

Theorem: L-level policy:

$$Reg(T) \leq O_K(T^{2^{L-1}/(2^L-1)} \text{poly}\log(T))$$

$O(\log(T)/\log\log(T))$-level policy:

$$Reg(T) \leq O_K(T^{1/2} \text{poly}\log(T))$$