Incentivizing exploration

T agents come sequentially

Choose from a set of actions and get stochastic reward

Principal gets their feedback on rewards and decide what information to send to future agents

Goal: maximize the total rewards of all agents

Challenge: benefits of exploration only accrued for future agents, resulting in misaligned incentives

Prior work: use information asymmetry

No constraints on the format of information sent from principal to agents (it suffices to only use direct recommendation)

Our model and results

Selective data disclosure:

- Subhistory: actions and rewards from a subset of previous rounds
- Unbiased: subsets chosen beforehand, transitive

Flexible frequentist agents:

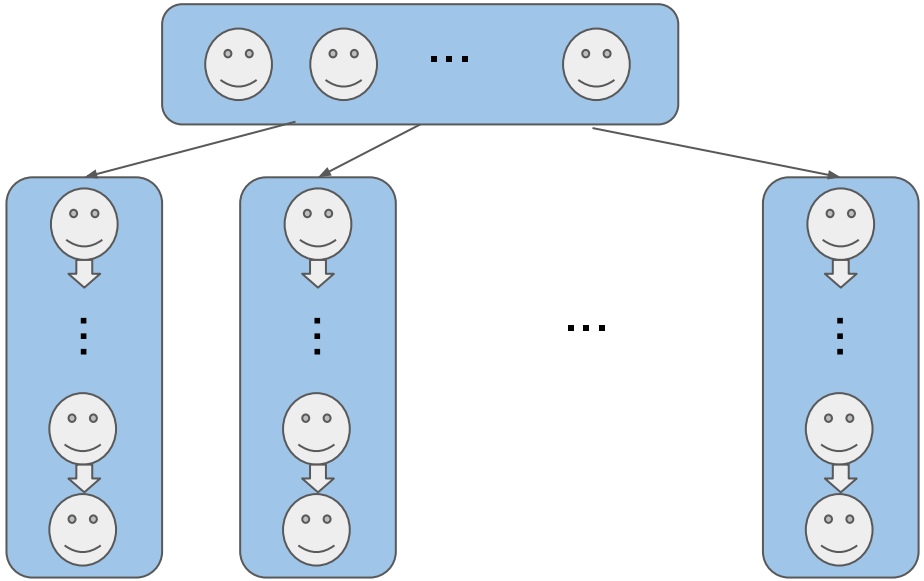
- Empirical mean reward of each action in the subhistory
- Take the action with (approximately) highest empirical mean

Results: our policy obtains regret $\tilde{O}(T^{1/2})$ for 2 actions.

Warm-up: 2-level policy

Level 2

Level 1

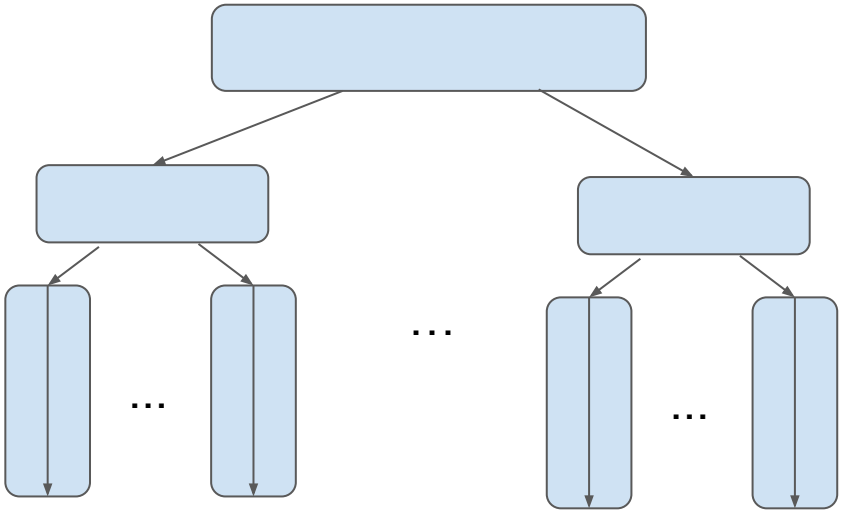


T- $cT^{2/3}\log(T)$ agents

$T^{2/3}\log(T)$ groups of size constant c

Theorem: Regret $\tilde{O}(T^{2/3})$

3-level policy



Level 3: exploitation
$\Theta(T)$ agents

Level 2: explore more if the gap is not too big
$\Theta(T^{6/7})$ agents

Level 1: exploration
$\Theta(T^{4/7})$ agents

Theorem: Regret $\tilde{O}(T^{4/7})$

More levels: L-level policy gets regret $O\left(T^{2^{L-1}/(2^L-1)}\log^2(T)\right)$