Incentivized Exploration of Non-Stationary Stochastic Bandits

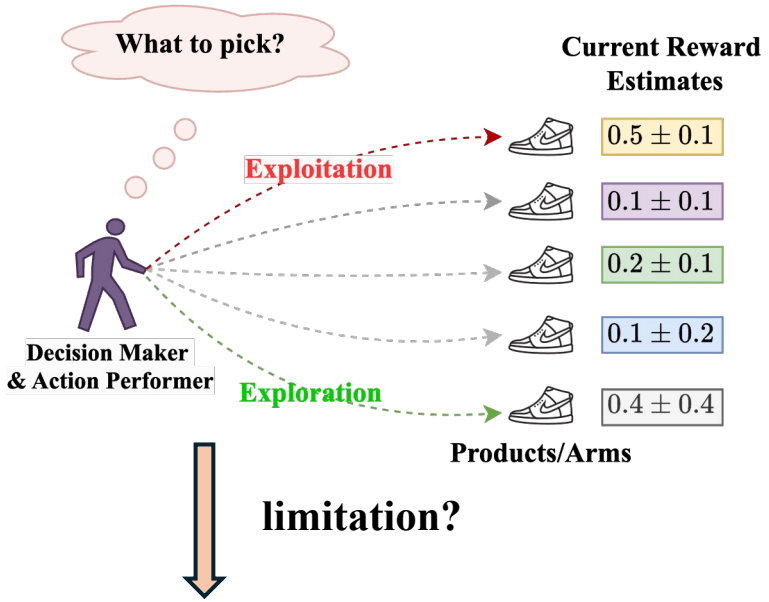
Sourav Chakraborty and Lijun Chen,

Dept. Of Computer Science, University of Colorado Boulder.

IEEE ACC 2024 - Toronto

What are we trying to solve?

Stochastic *(Stationary)* Multi-armed Bandit



limitation?

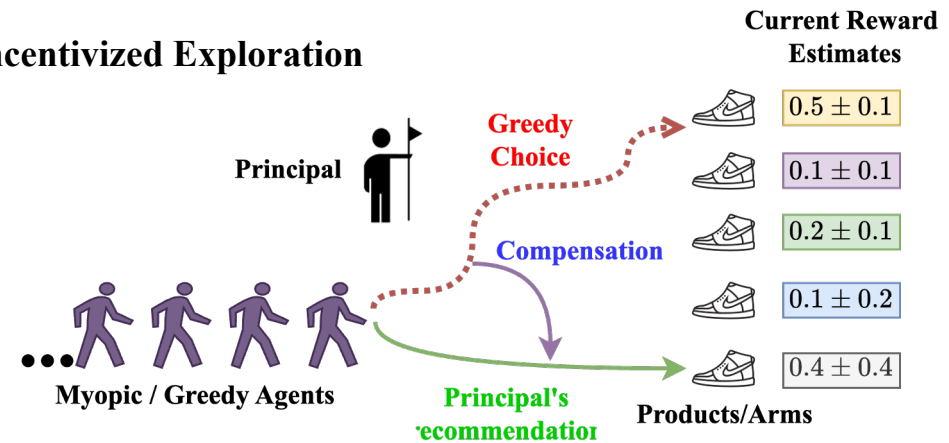
Decision Maker and Action Performer may have misaligned interests

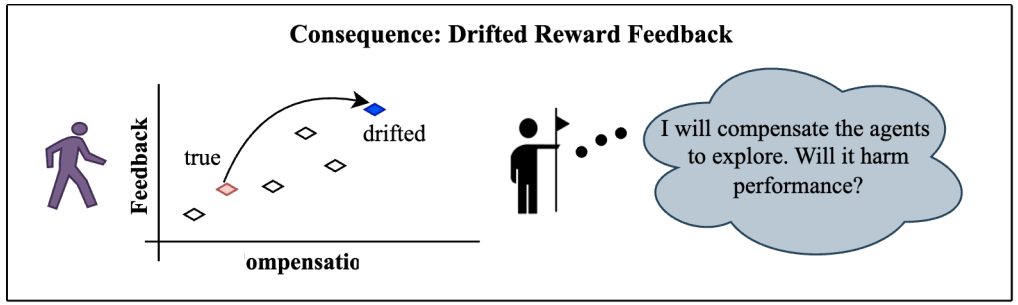
Decision Maker / Principal: Let the agent converge to the overall best arm.

Agent / Action Performer: Let's just be greedy

Solution Attempt

Incentivized Exploration



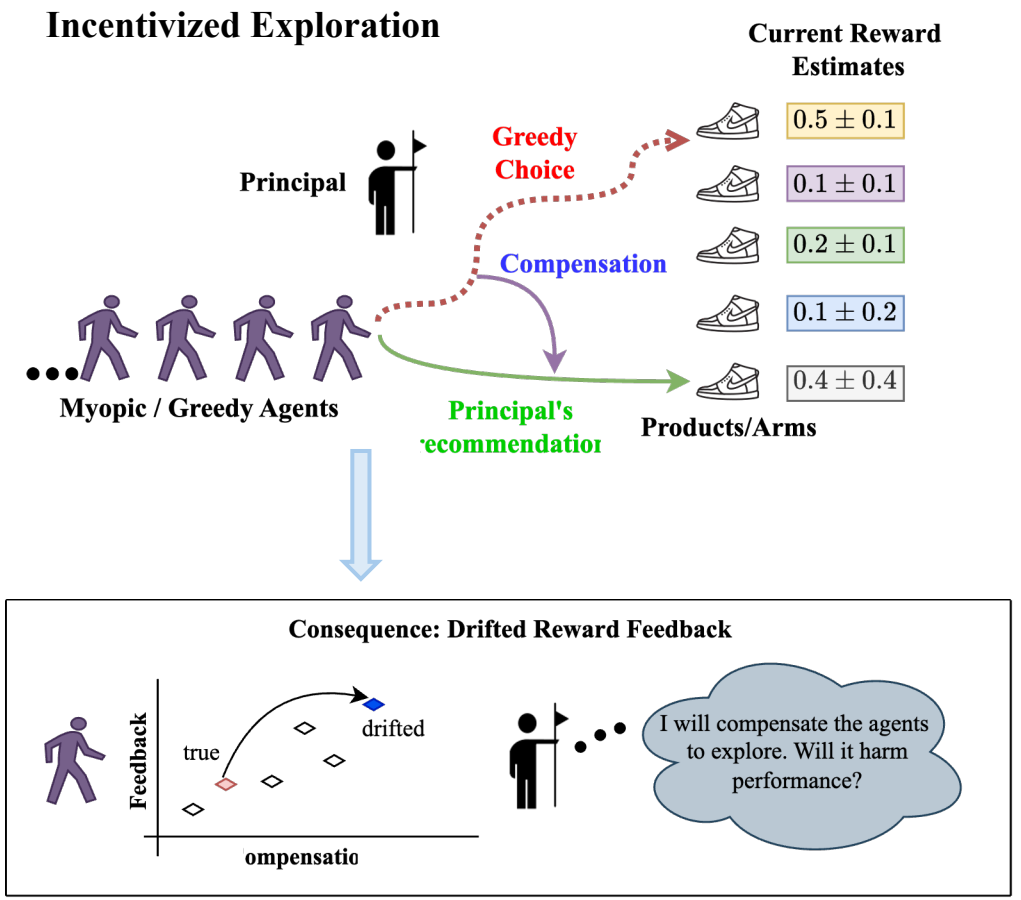


Central Question: Can we achieve effective Incentivized Exploration with Non-stationary rewards and drifted reward feedback?

Is it any useful?

Scenario:

- **Amazon** (*Principal*), **Product** listings (*arms*), and **customers** (*agents*).
- New products added to the inventory need exploration; therefore, compensation (*say discounts or coupons*) is required in cases for customers to buy instead of the highest-rated product (*greedy choice*).
- Products can have different values to users at various times (*non-stationarity in rewards*). For instance, snow boots are more popular in winter, and shorts are more popular in summer.

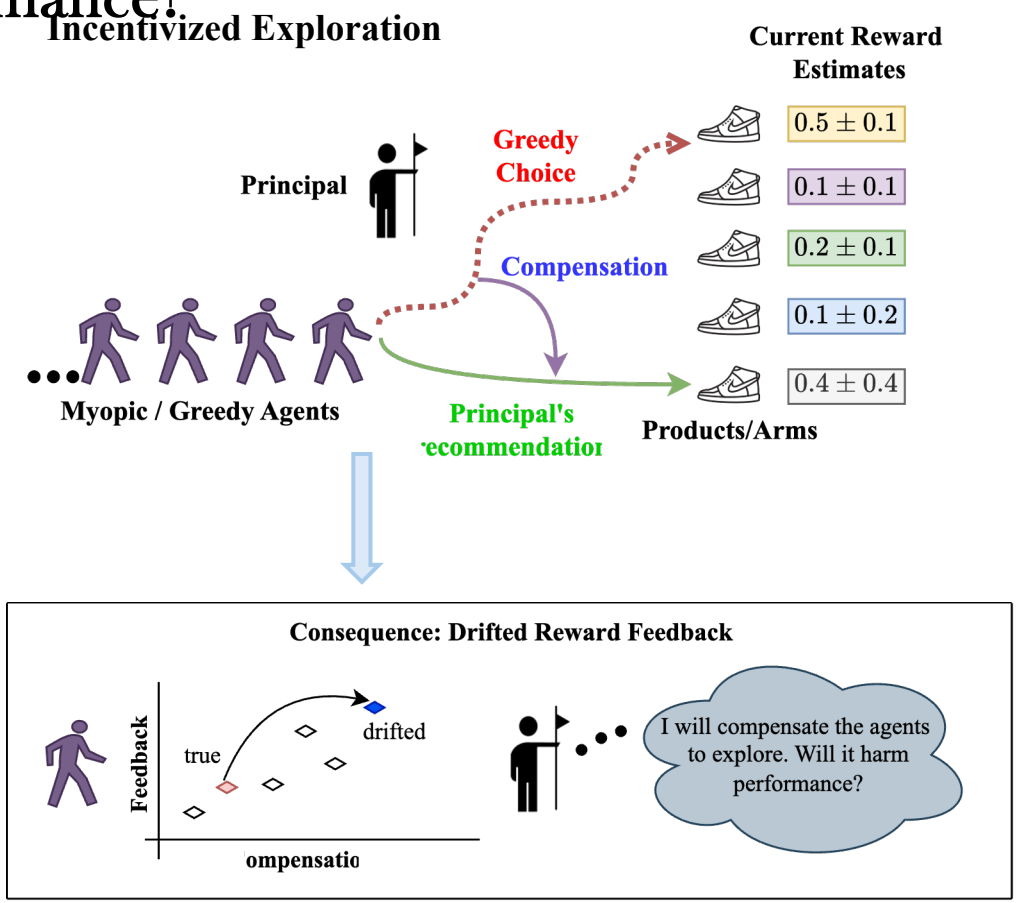


Central Question: Can we achieve effective Incentivized Exploration with Non-stationary rewards and drifted reward feedback?

How do we measure performance?

Measured Metrics

- **Regret.** Difference between the optimal (oracle) performance and the algorithm's performance (*i.e., accumulated rewards*) after some known time horizon T.
- **Compensation**. Total incentives (*i.e., difference between the reward estimates of the greedy arm and the recommended arm*) given by the Principal to the agents after the same time horizon T.

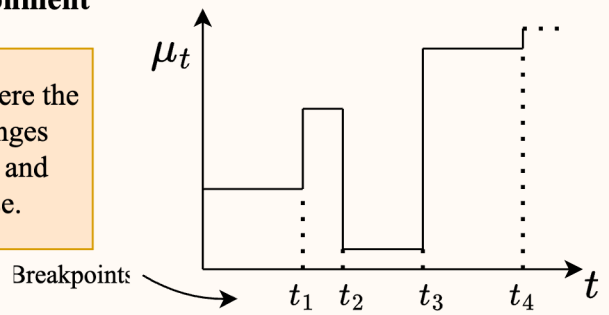


Refined Question: Can we design algorithm strategies for Incentivized Exploration with Non-stationary rewards and drifted reward feedback, where the achieved ***expected*** regret and compensation is **sublinear** in the time horizon T?

How was the performance (Theoretical)?

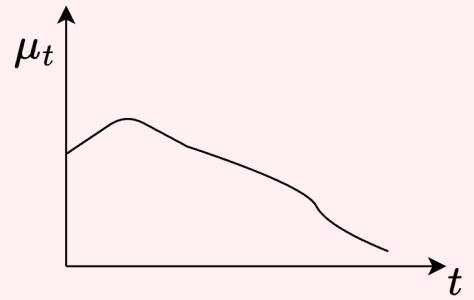
Abruptly Changing Environment

A model of an environment where the mean rewards of an arm changes abruptly during a breakpoint and remain stationary otherwise.



Continuously Changing Environment

A model of an environment where the mean rewards of an arm changes continuously throughout the time horizon within a variation budget.



Algorithm Strategy	Regret	Compensation
Algorithm 3 with discounted UCB	$\tilde{O}(T^{1/2})$	$\tilde{O}(T^{1/2})$
Algorithm 3 with sliding-window UCB	$\tilde{O}(T^{1/2})$	$\tilde{O}(T^{1/4})$

Algorithm Strategy	Regret	Compensation
Algorithm 4 with UCB	$\tilde{O}(T^{2/3})$	$\tilde{O}(T^{2/3})$
Algorithm 4 with eps-Greedy	$\tilde{O}(T^{2/3})$	$\tilde{O}(T^{1/3})$
Algorithm 4 with Thompson Sampling	$\tilde{O}(T^{2/3})$	$\tilde{O}(T^{1/3})$

Takeaway: All the proposed algorithm strategies achieve **sublinear expected regret** and **compensation** for both the non-stationary environments.

Thank you!

For questions:
Sourav Chakraborty - sourav.chakraborty@colorado.edu
Lijun Chen - lijun.chen@colorado.edu