University of Colorado Boulder

Inria

Conference on Decision and Control
CDC2025
Rio de Janeiro, Brazil. Dec. 09 - 12

Incentivized Lipschitz Bandits

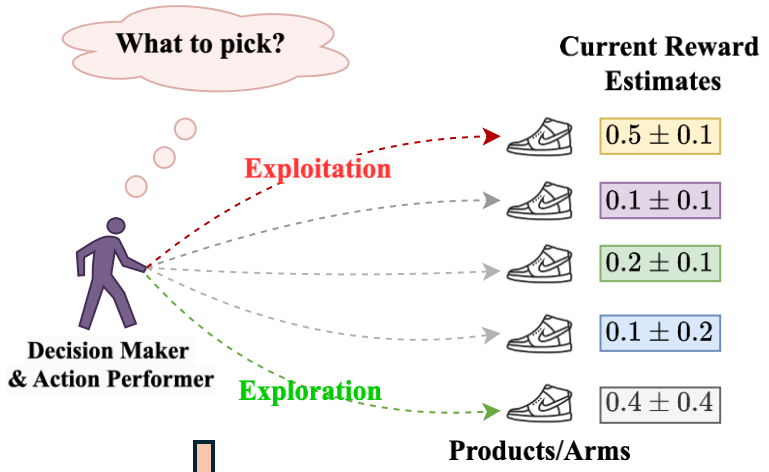
Sourav Chakraborty[†*], Amit Kiran Rege[†*], Claire Monteleoni[‡*] and Lijun Chen[*],

[*]Dept. Of Computer Science, University of Colorado Boulder.

[‡]INRIA, Paris.

What is this about?

Stochastic Multi-armed Bandit (Finite Arms)



limitation?

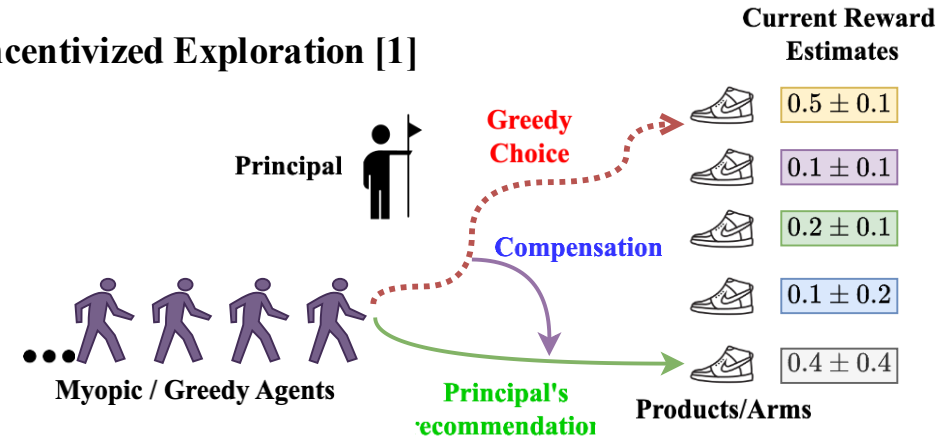
Decision Maker and Action Performer may have misaligned interests

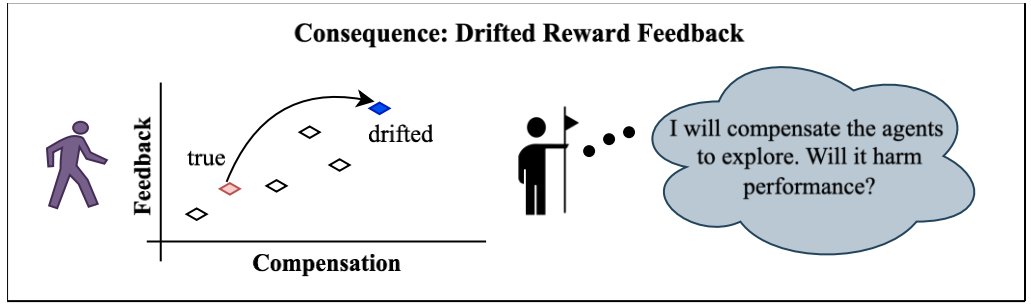
Decision Maker / Principal: Let the agent converge to the overall best arm.

Agent / Action Performer: Let's just be greedy

Solution Attempt

Incentivized Exploration [1]





Central Question: Can we achieve effective Incentivized Exploration with a **massive** (possibly infinite) number of arms and drifted reward feedback?

[1] Liu et. al. - Incentivized Exploration for Multi-Armed Bandits under Reward Drift. arXiv:1911.05142

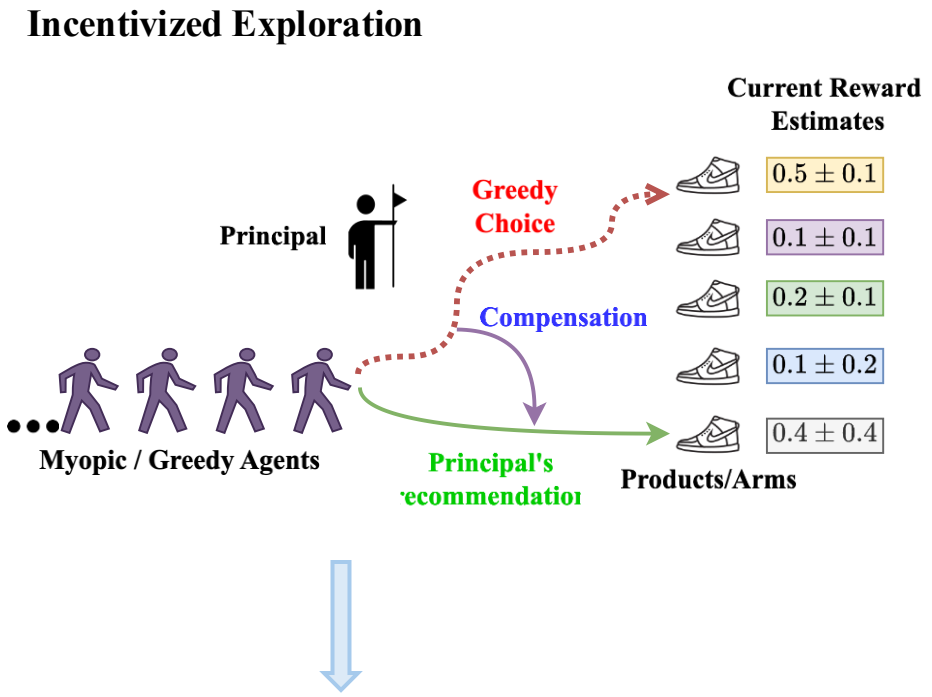
Is it useful?

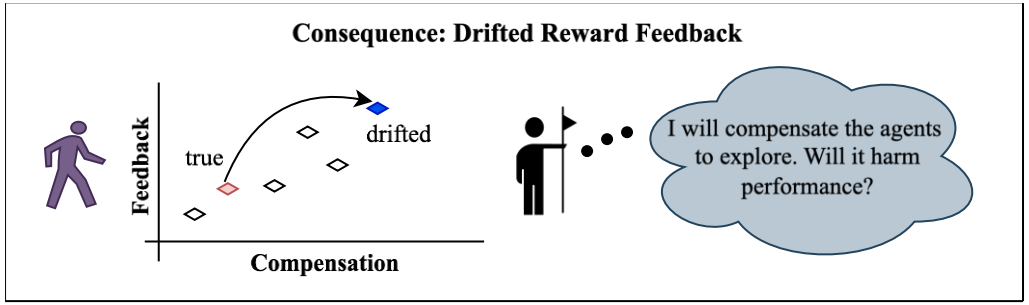
Scenario:

- **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*).
- The inventory can have a massive number of product listings (*arms*), say **100M.**

Bonus:

- What if the rewards are affected by some **contextual** information, like user preference?





Model: Problem Formulation

Problem Instance

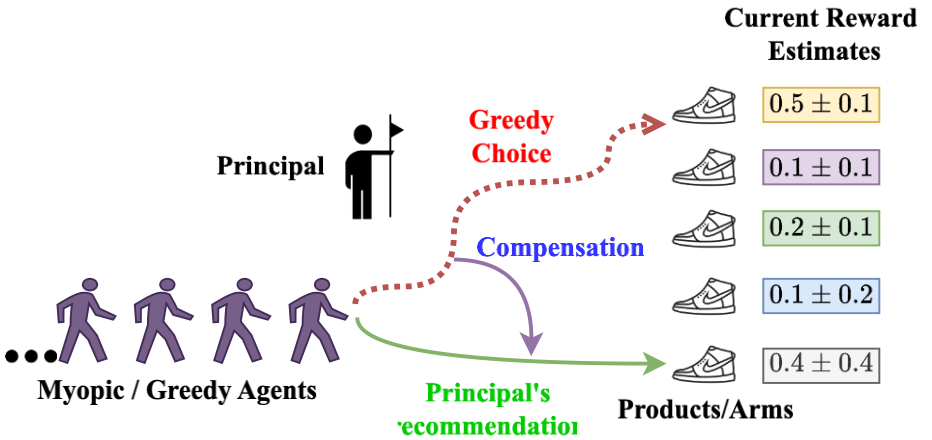
$$(A, \Phi, \mu)$$

Set of arms.
For example;
$a \in A \subseteq [0,1]^m$
$m \in \mathbb{N}$

Any metric space;
for example ℓ_1, ℓ_2

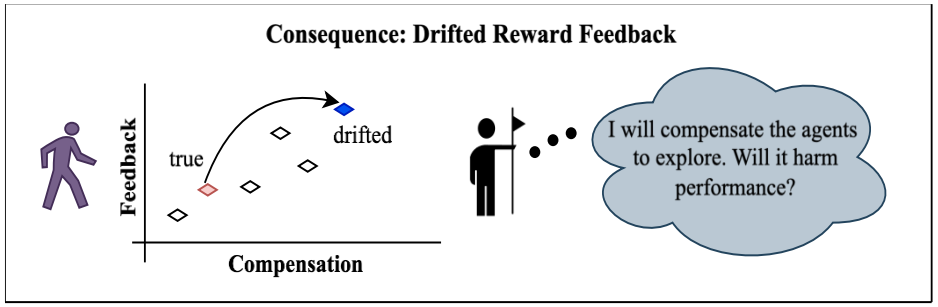
Expected reward vector for each arm, supported on $[0,1]$

Incentivized Exploration



Assumption

$$|\mu(a) - \mu(a')| \leq L_s \cdot \Phi(a, a'), \quad \forall a, a' \in A,$$



How do we measure performance (Stochastic)?

Regret. Difference between the optimal (oracle) performance and the algorithm's performance (*i.e., accumulated rewards*) after some known time horizon *T*.

$$\mathbb{E}[\mathcal{R}_T(A)] = \mathbb{E}\left[\sum_{t=0}^{T}(\mu^* - \rho_t(a_t))\right]$$

Pull arms till horizon *T*

Expectation of the cumulative **regret** for algorithm *A*, after *T* time steps.

Expected reward for the **best** arm.

Realized reward for the **chosen arm at time *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*.

$$\mathbb{E}\left[\mathcal{C}_T(A)\right] = \mathbb{E}\left[\sum_{t=1}^{T}\left(\hat{\mu}_t(g_t) - \hat{\mu}_t(a_t)\right)\right]$$

Expectation of the cumulative **compensation** for algorithm *A*, after *T* time steps.

Gap between the **estimated** means of the **greedy**(agent's choice) and the **principal's choice.**

Refined Question:

Can we design algorithm strategies for Incentivized Exploration with Infinitely-many arms and drifted reward feedback, where
the achieved ***expected*** regret and compensation is **sublinear** in
the time horizon *T*?

What's the solution?

Discretize the continuum with a finite set A_0 such that,

$$\mathrm{diam}(S_i) \le \psi, \quad A \subseteq \bigcup_i S_i$$

Algorithm 1 Incentivized Exploration in Infinite-Armed Stochastic Bandits

Require: Metric space for arms (A, Φ), time horizon T, Lipschitz constant L

Set: $\psi \leftarrow O\left(T^{-1/(d+2)} L_s^{-2/(d+2)} (\log T)^{1/(d+2)}\right)$

Construct a ψ-covering $A_0 \subseteq A$

for $t = 1, 2, \ldots, T$ **do**

  Select arm: $a_t \leftarrow \arg\max_{a \in A_0} \hat{\mu}_t(a) + \sqrt{\frac{2 \log t}{N_t(a)}}$

  Agent picks: $g_t \leftarrow \arg\max_{a \in A_0} \hat{\mu}_t(a)$

  Compensation: $\kappa_t \leftarrow \hat{\mu}_t(g_t) - \hat{\mu}_t(a_t)$

  Pull a_t, observe reward: $r_t = \rho_t(a_t) + \gamma_t(\kappa_t)$

  Update: $\hat{\mu}_{t+1}(a_t) \leftarrow$ Update based on r_t

end for

Received reward is ***biased*** or ***drifted***.

$\rho_t(a_t)$ is the pure reward and $\gamma_t(a_t)$ is the drift.

How did we do?

Theorem 1. *Consider Algorithm 1 run on a metric space (A, Φ), where arms are discretized using a ψ-covering $A_0 \subseteq A$ satisfying $|A_0| \leq O(1/\psi^d)$, with d denoting the covering dimension of (A, Φ). If $\psi = O\left(T^{-1/(d+2)} L_s^{-2/(d+2)} (\log T)^{1/(d+2)}\right)$, then the expected regret and compensation satisfy:*

$$\mathbb{E}[\mathcal{R}_T] = \tilde{O}\left(L_s^{\frac{d}{d+2}} T^{\frac{d+1}{d+2}}\right), \quad \mathbb{E}[\mathcal{C}_T] = \tilde{O}\left(L_s^{\frac{d}{d+2}} T^{\frac{d+1}{d+2}}\right),$$

where L_s is the Lipschitz constant of the reward function $\mu(a)$, and the hidden constants in the $O(\cdot)$ and $\tilde{O}(\cdot)$ notations depend on the drift parameter ℓ.

Assumption 1. *The reward drift function $\gamma_t(a)$ is nondecreasing, satisfies $\gamma_t(0) = 0$, and is Lipschitz continuous, i.e., there exists a constant ℓ_t such that*

$$|\gamma_t(a) - \gamma_t(a')| \leq \ell_t |a - a'|, \quad \forall a, a'.$$

Case with Contextual Information?

Problem Instance

$$(A, X, \Phi_X, \Phi_A, \nu)$$

Set of contexts, for example;
$x \in X \subseteq [0,1]^k$
$k \in \mathbb{N}$

Metric spaces for contexts and arms respectively

Reward based arm and context $\nu(a, x)$

Assumption

$$\left|\nu(a,x) - \nu(a',x')\right| \leq L_c \cdot \left[\Phi_X(x,x') + \Phi_A(a,a')\right].$$

Algorithm 2 Incentivized Exploration in Infinite-Armed Contextual Bandits

Require: Metric spaces $(A, \Phi_A), (X, \Phi_X)$, snapping function $\xi : X \to X_0$, horizon T

Set: $\psi \leftarrow O\left(T^{-1/(d+2)} L_c^{-2/(d+2)} (\log T)^{1/(d+2)}\right)$

Construct ψ-covering $A_0 \times X_0 \subset A \times X$

for $t = 1, 2, \ldots, T$ **do**

  Observe context $x_t \in X$, snap: $x_0 \leftarrow \xi(x_t)$

  Principal picks arm:

$$a_t \leftarrow \arg\max_{a_0 \in A_0} \hat{\nu}_t(a_0, x_0) + \sqrt{\frac{2 \log t}{N_t(a_0, x_0)}}$$

  Agent picks: $g_t \leftarrow \arg\max_{a_0 \in A_0} \hat{\nu}_t(a_0, x_0)$

  Compensation: $\kappa_t \leftarrow \hat{\nu}_t(g_t, x_0) - \hat{\nu}_t(a_t, x_0)$

  Reward: $r_t = \rho_t(a_t, x_t) + \gamma_t(\kappa_t)$

  Update: $\hat{\nu}_{t+1}(a_t, x_0) \leftarrow$ Update based on r_t

end for

How did we do in the contextual case?

Theorem 2. *Consider Algorithm 2 executed on a metric space $(A \times X, \Phi)$, where a ψ-covering discretization $A_0 \times X_0 \subseteq A \times X$ is used with $|A_0 \times X_0| \leq O(1/\psi^d)$, and $d = d_a + d_x$ denotes the covering dimension of $(A \times X, \Phi)$. If $\psi = O\left(T^{-1/(d+2)} L_c^{-2/(d+2)} (\log T)^{1/(d+2)}\right)$, then the expected regret and compensation satisfy:*

$$\mathbb{E}[\mathcal{R}_T] = \tilde{O}\left(L_c^{\frac{d}{d+2}} T^{\frac{d+1}{d+2}}\right), \quad \mathbb{E}[\mathcal{C}_T] = \tilde{O}\left(L_c^{\frac{d}{d+2}} T^{\frac{d+1}{d+2}}\right),$$

where L_c is the Lipschitz constant of the reward function $\nu(a, x)$, and the hidden constants in the $O(\cdot)$ and $\tilde{O}(\cdot)$ notations depend on the drift parameter ℓ.

Assumption 1. *The reward drift function $\gamma_t(a)$ is nondecreasing, satisfies $\gamma_t(0) = 0$, and is Lipschitz continuous, i.e., there exists a constant ℓ_t such that*

$$|\gamma_t(a) - \gamma_t(a')| \leq \ell_t |a - a'|, \quad \forall a, a'.$$

Thank you!

For questions:
Sourav Chakraborty - sourav.chakraborty@colorado.edu
Amit Kiran Rege – amit.rege@colorado.edu