

# Social Learning in Multi Agent Multi Armed Bandits

Abishek Sankararaman

May 2020

## Joint Work with

- Sanjay Shakkottai, UT Austin
- Ayalvadi Ganesh, University of Bristol

# Multi Armed Bandit Problem

---



1



2



3



4

Drugs with a-priori unknown cure rates

Explore/Exploit Tradeoff for each new patient [Thompson' 33]

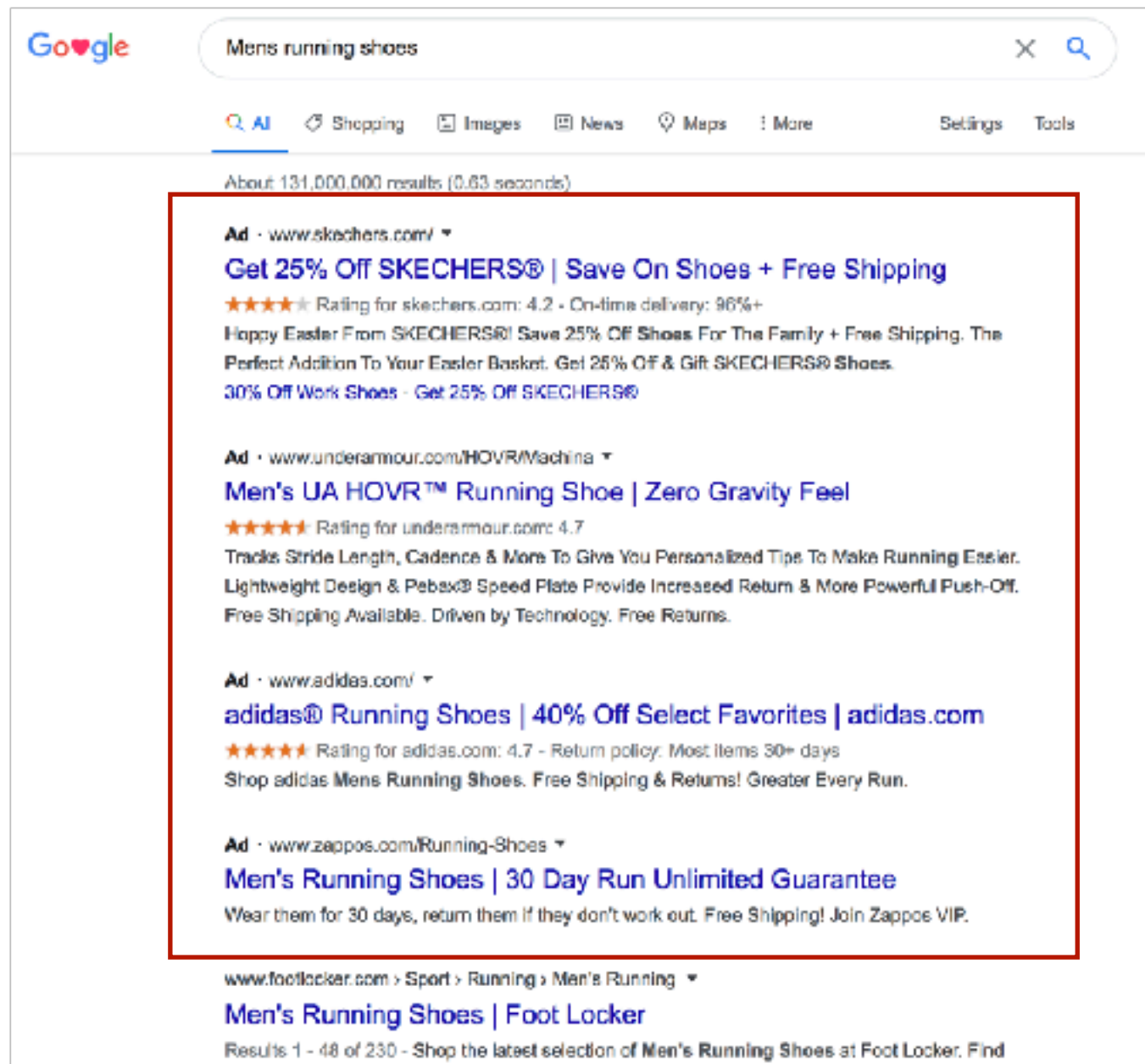
*Exploit* Prescribe a drug that has shown the best promise so far

*Explore* Try a new drug to discover more promising alternatives

Run a risk of not curing these patients

# Explore/Exploit Phenomena

Online Advertising - Which of the large collection of ads to display ?



The screenshot shows a Google search for "Mens running shoes". The search bar at the top contains the text "Mens running shoes" and a magnifying glass icon. Below the search bar, there are tabs for "All", "Shopping", "Images", "News", "Maps", and "More". The "All" tab is selected. Below the tabs, it says "About 131,000,000 results (0.63 seconds)". A red box highlights four advertisements:

- Ad - www.skechers.com/**  
**Get 25% Off SKECHERS® | Save On Shoes + Free Shipping**  
★★★★★ Rating for skechers.com: 4.2 - On-time delivery: 96%+  
Happy Easter From SKECHERS®! Save 25% Off Shoes For The Family + Free Shipping. The Perfect Addition To Your Easter Basket. Get 25% Off & Gift SKECHERS® Shoes.  
30% Off Work Shoes - Get 25% Off SKECHERS®
- Ad - www.underarmour.com/HOVR/Machina**  
**Men's UA HOVR™ Running Shoe | Zero Gravity Feel**  
★★★★★ Rating for underarmour.com: 4.7  
Tracks Stride Length, Cadence & More To Give You Personalized Tips To Make Running Easier. Lightweight Design & Pebax® Speed Plate Provide Increased Return & More Powerful Push-Off. Free Shipping Available. Driven by Technology. Free Returns.
- Ad - www.adidas.com/**  
**adidas® Running Shoes | 40% Off Select Favorites | adidas.com**  
★★★★★ Rating for adidas.com: 4.7 - Return policy: Most items 30+ days  
Shop adidas Mens Running Shoes. Free Shipping & Returns! Greater Every Run.
- Ad - www.zappos.com/Running-Shoes**  
**Men's Running Shoes | 30 Day Run Unlimited Guarantee**  
Wear them for 30 days, return them if they don't work out. Free Shipping! Join Zappos VIP.

Below the red box, there is a breadcrumb trail: "www.footlocker.com > Sport > Running > Men's Running". Below that is another advertisement: "Men's Running Shoes | Foot Locker". At the bottom, it says "Results 1 - 48 of 230 - Shop the latest selection of Men's Running Shoes at Foot Locker. Find".

Exploit - What has worked in the past ?

Explore - Discover a more relevant ad

**Internet Advertising accounts for 1 Trillion in revenue  
6% of US GDP ! [HBS Report 2019]**

# Outline

---

1. Single Agent MAB
2. The Multi-Agent Setup
3. Social Learning Algorithm
4. Insights



# Multi Armed Bandit Problem

---

At each time,  $t \in \{1, \dots, T\}$  an agent

- chooses an arm  $I_t \in \{1, \dots, K\}$
- receives a stochastic reward  $X_t \in \{0, 1\}$

$\mathbb{P}[X_t = 1 | I_t] = \mu_{I_t}$  independent of everything else

Each arm corresponds to a drug in the previous example

Goal - Maximize total reward  $\mathbb{E}[\sum_{t=1}^T X_t]$

# Multi Armed Bandit Problem

---

At each time,  $t \in \{1, \dots, T\}$  an agent

- chooses an arm  $I_t \in \{1, \dots, K\}$
- receives a stochastic reward  $X_t \in \{0, 1\}$

$\mathbb{P}[X_t = 1 | I_t] = \mu_{I_t}$  independent of everything else

Each arm corresponds to a drug in the previous example

Goal - Maximize total reward  $\mathbb{E}[\sum_{t=1}^T X_t]$

Challenge Arm-means  $(\mu_i)_{i=1}^K$  initially unknown

# Multi Armed Bandit Problem

---

As we play arms, can learn  $(\mu_i)_{i=1}^K$

# Multi Armed Bandit Problem

---

As we play arms, can learn  $(\mu_i)_{i=1}^K$

## Explore-Exploit Tradeoff

**Exploit**      Play the arm that has been best so far

**Explore**      Play an arm played few times so as to see if it is good

# Multi Armed Bandit Problem

---

As we play arms, can learn  $(\mu_i)_{i=1}^K$

## Explore-Exploit Tradeoff

**Exploit**      Play the arm that has been best so far

**Explore**      Play an arm played few times so as to see if it is good

Performance Metric - Regret

$$R_T = \mu^* T - \mathbb{E}\left[\sum_{t=1}^T X_t\right]$$
$$\mu^* = \max\{\mu_1, \dots, \mu_K\}$$

How much loss due to lack of knowledge ?

# Upper Confidence Bound (UCB) Algorithm

---

UCB Algorithm [Auer et.al. '02]

At time  $t$ , choose arm  $I_t \in \arg \max_k \left( \hat{\mu}_k(t-1) + \sqrt{\frac{4\alpha \log(t)}{N_k(t-1)}} \right)$

$\hat{\mu}_k(t-1)$  Empirical Mean of arm  $k$  at time  $t-1$

$N_k(t-1)$  Number of times arm  $k$  has been played

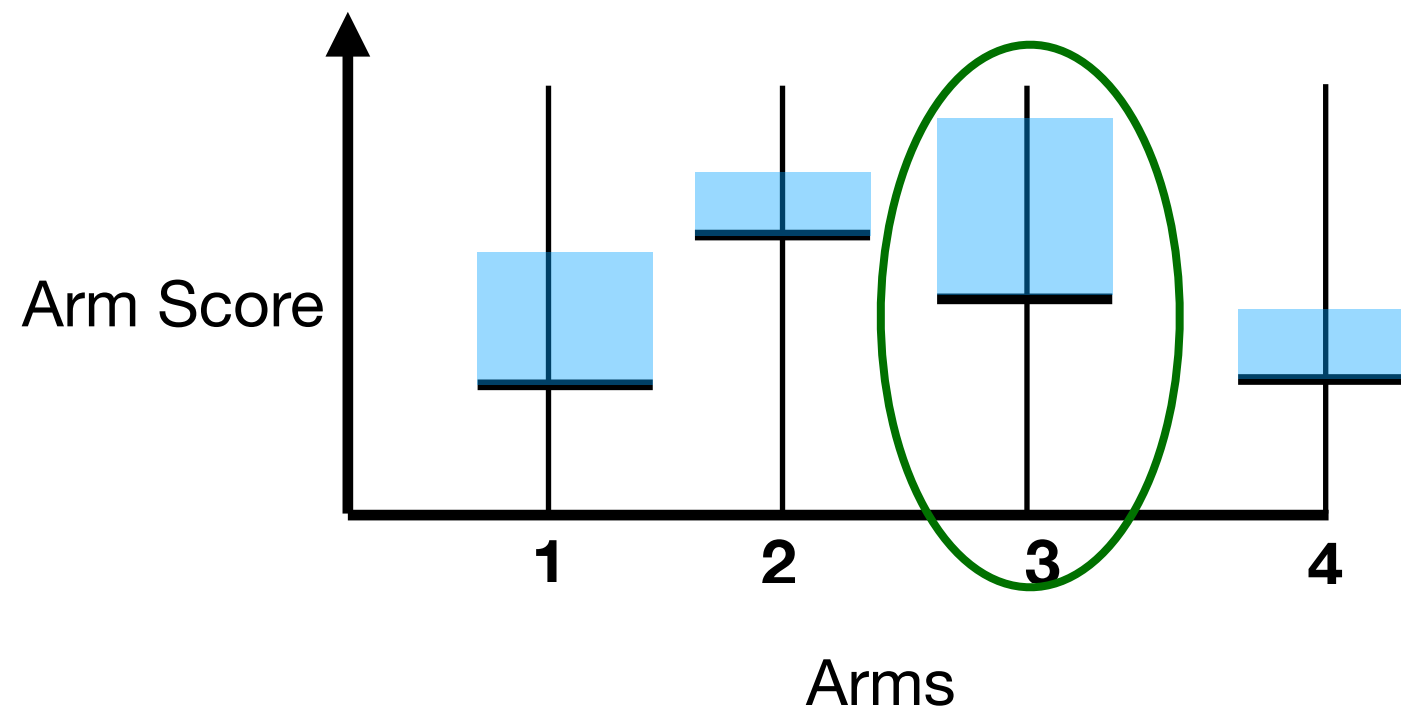
# Upper Confidence Bound (UCB) Algorithm

UCB Algorithm [Auer et.al. '02]

At time  $t$ , choose arm  $I_t \in \arg \max_k \left( \hat{\mu}_k(t-1) + \sqrt{\frac{4\alpha \log(t)}{N_k(t-1)}} \right)$

$\hat{\mu}_k(t-1)$  Empirical Mean of arm  $k$  at time  $t-1$

$N_k(t-1)$  Number of times arm  $k$  has been played



*Optimism in the Face of Uncertainty*

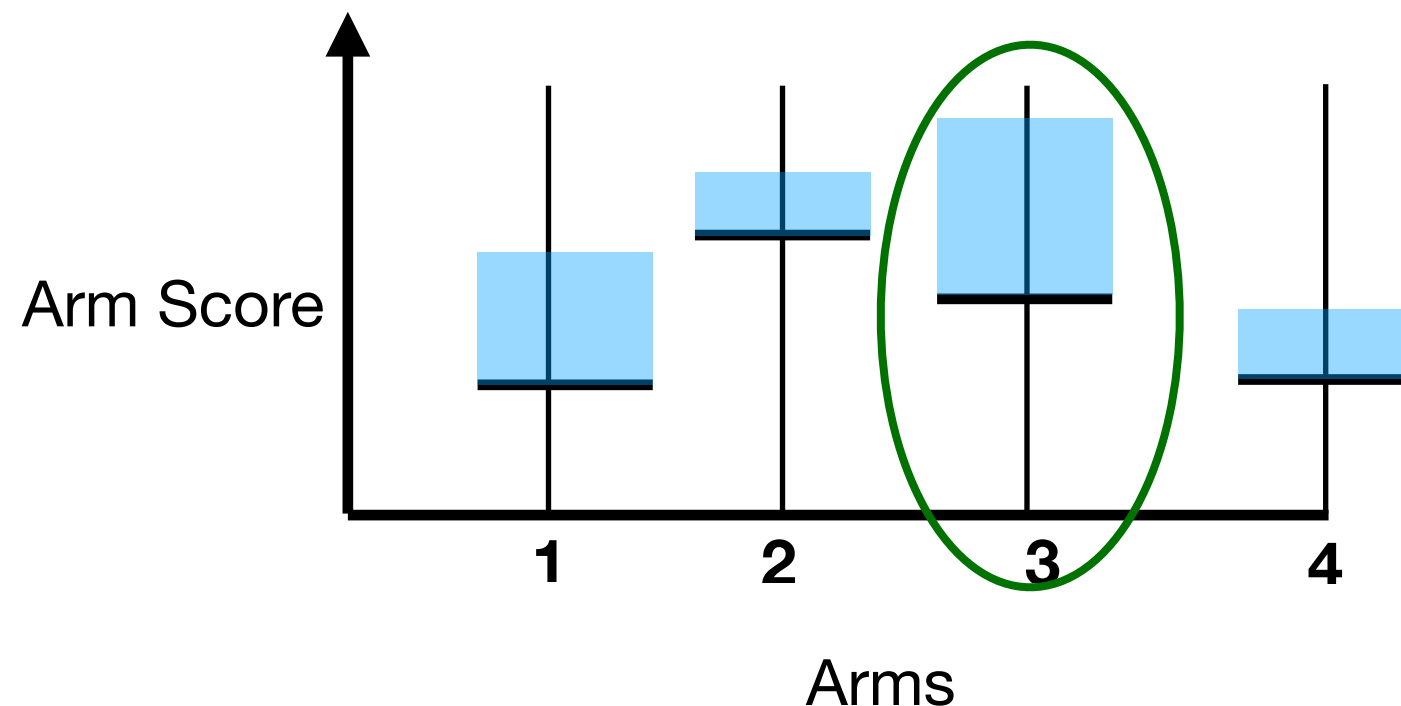
# Upper Confidence Bound (UCB) Algorithm

UCB Algorithm [Auer et.al. '02]

At time  $t$ , choose arm  $I_t \in \arg \max_k \left( \hat{\mu}_k(t-1) + \sqrt{\frac{4\alpha \log(t)}{N_k(t-1)}} \right)$

$\hat{\mu}_k(t-1)$  Empirical Mean of arm  $k$  at time  $t-1$

$N_k(t-1)$  Number of times arm  $k$  has been played



*Optimism in the Face of Uncertainty*

Theorem  $R_T \leq O\left(\frac{K}{\Delta} \log(T)\right)$

$\Delta$  Difference in arm mean between best and second best arm



# Multi Agent Setup

---



[shorturl.at/huO57](https://shorturl.at/huO57)

What if multiple agents play the same MAB instance ?

Can they collaborate and jointly reduce their individual regret ?

# Multi Agent Setup - Motivation

---

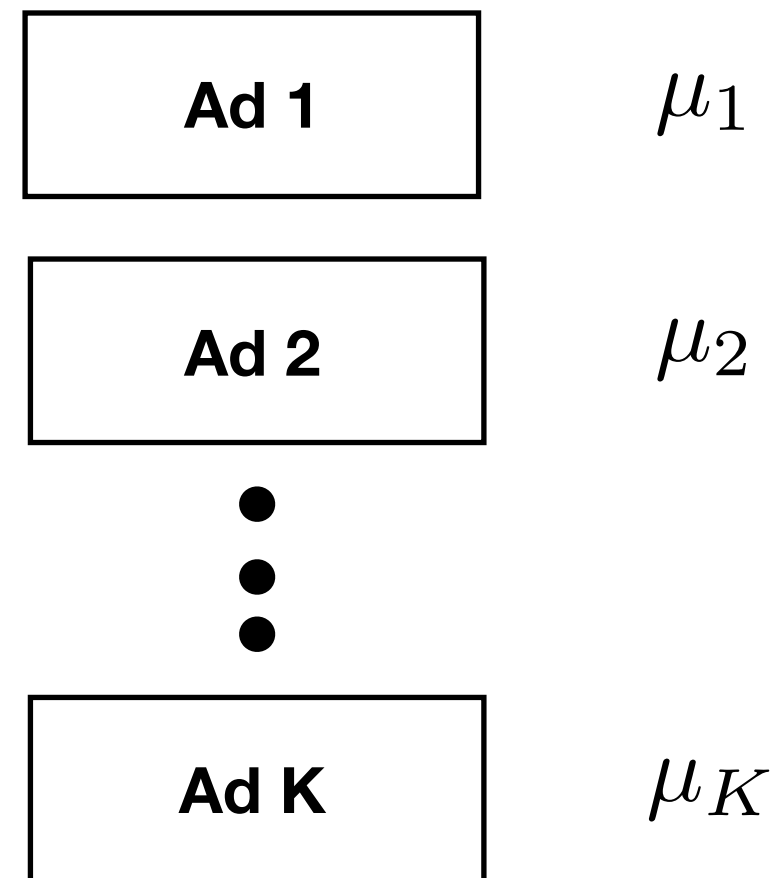
One server is serving ads for a fixed keyword

At each search request, server can choose to display one ad

*Choice of an arm to pull*



[shorturl.at/foRV5](https://shorturl.at/foRV5)



At the end, receives a stochastic reward

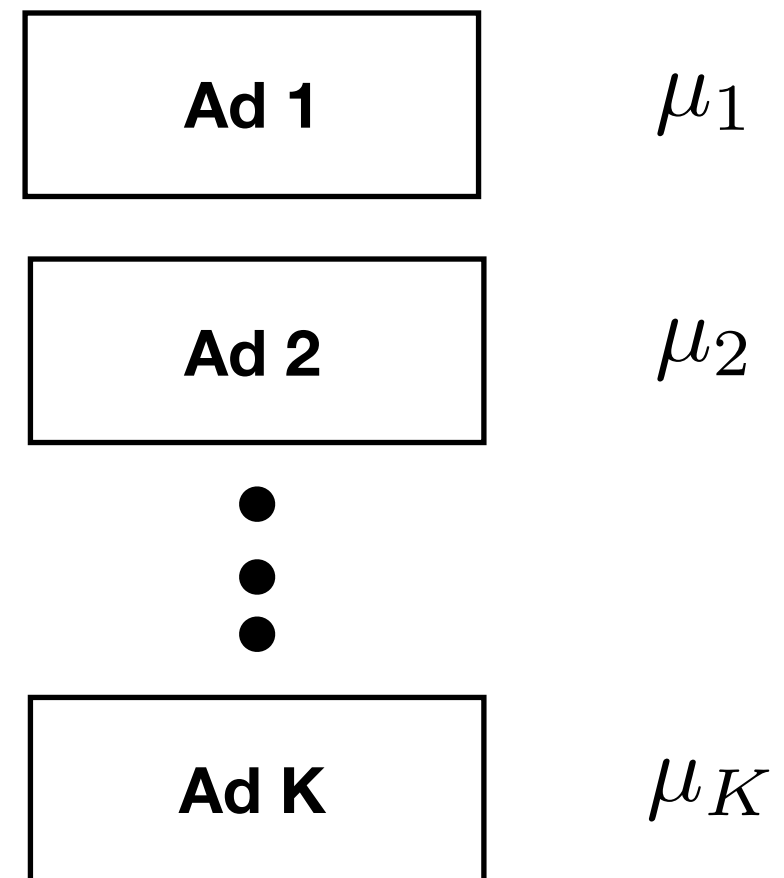
Goal is to maximize revenue (minimize regret)

# Multi Agent Setup - Motivation

Multiple servers serving ads for a fixed keyword

*Each search request, routed to a server*

Each server chooses to display one ad when routed to it.

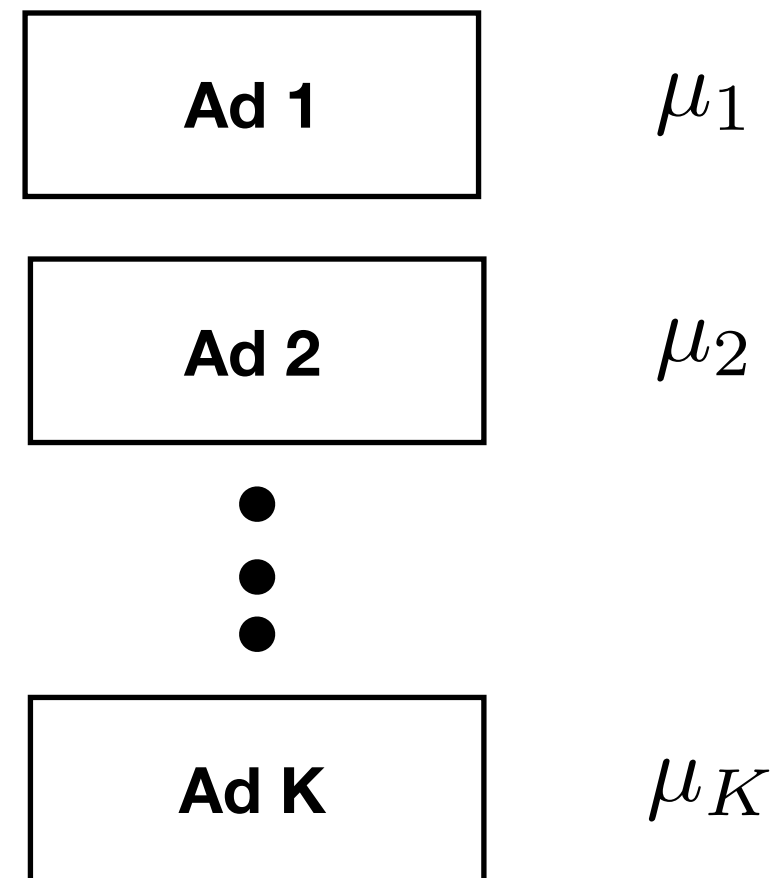
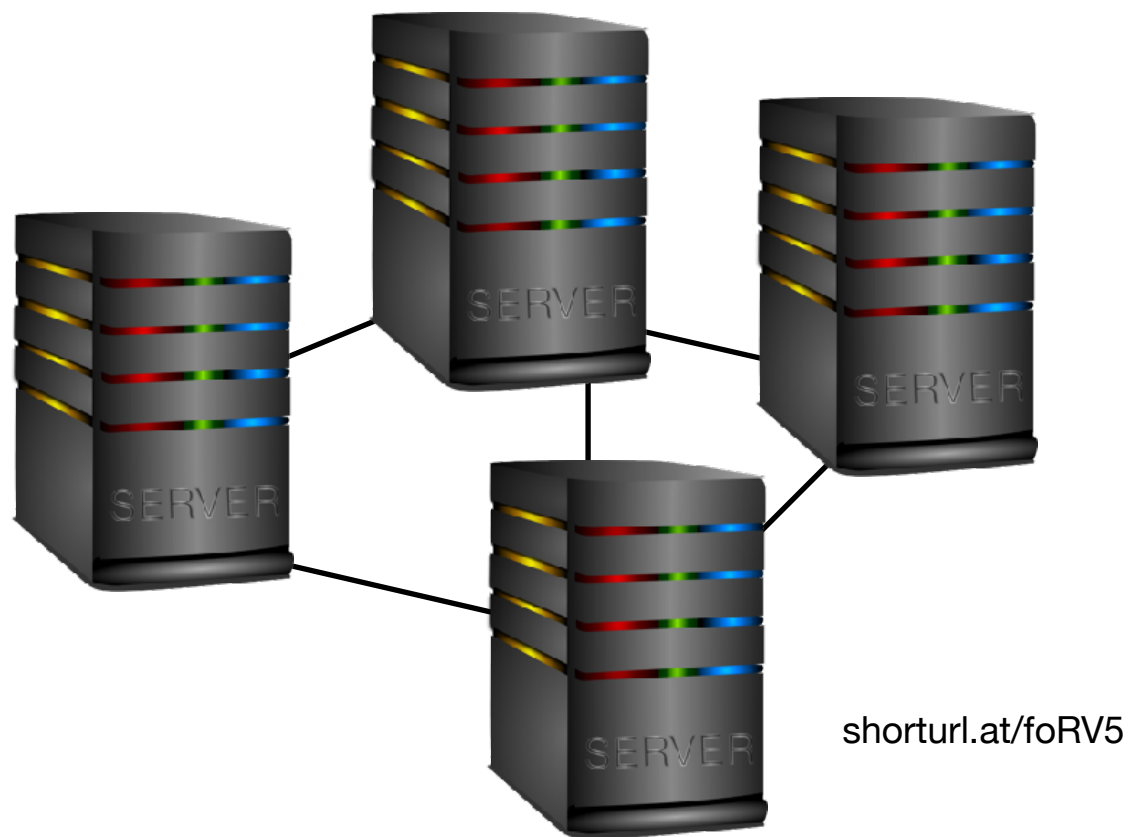


# Multi Agent Setup - Motivation

Multiple servers serving ads for a fixed keyword

*Each search request, routed to a server*

Each server chooses to display one ad when routed to it.



Servers can potentially collaborate and learn from each other's experience.

*Managed by the same company*

# Multi Agent Setup - Motivation

---

At each time, every server makes a decision from  $K$  alternatives

*Large volume of search queries*



# Multi Agent Setup - Motivation

---

At each time, every server makes a decision from  $K$  alternatives

*Large volume of search queries*

## 1. No Communication -

Individual Server Regret -  $O\left(\frac{K}{\Delta} \log(T)\right)$

Communication Resources - **0**



# Multi Agent Setup - Motivation

---

At each time, every server makes a decision from  $K$  alternatives

*Large volume of search queries*



## 1. No Communication -

Individual Server Regret -  $O\left(\frac{K}{\Delta} \log(T)\right)$

Communication Resources - **0**

## 2. Full Interaction -

Individual Server Regret -  $O\left(\frac{1}{N} \cdot \frac{K}{\Delta} \log(T)\right)$

*Overall system can be abstracted as a single agent*

Communication Resources - **T broadcasts per agent !**

# Multi Agent Setup - Motivation

At each time, every server makes a decision from  $K$  alternatives

*Large volume of search queries*



## 1. No Communication -

Individual Server Regret -  $O\left(\frac{K}{\Delta} \log(T)\right)$

Communication Resources - **0**

## 2. Full Interaction -

Individual Server Regret -  $O\left(\frac{1}{N} \cdot \frac{K}{\Delta} \log(T)\right)$

*Overall system can be abstracted as a single agent*

Communication Resources - **T broadcasts per agent !**

*Best of both situations ??*



# The Multi Agent Problem

$K$  arms,  $N$  agents,

Asynchronous System - Each agent  $j$  has an i.i.d. Poisson clock  $\mathcal{C}_j(\cdot)$

When clock  $\mathcal{C}_j(\cdot)$  rings for the  $t^{\text{th}}$  time, agent  $j$

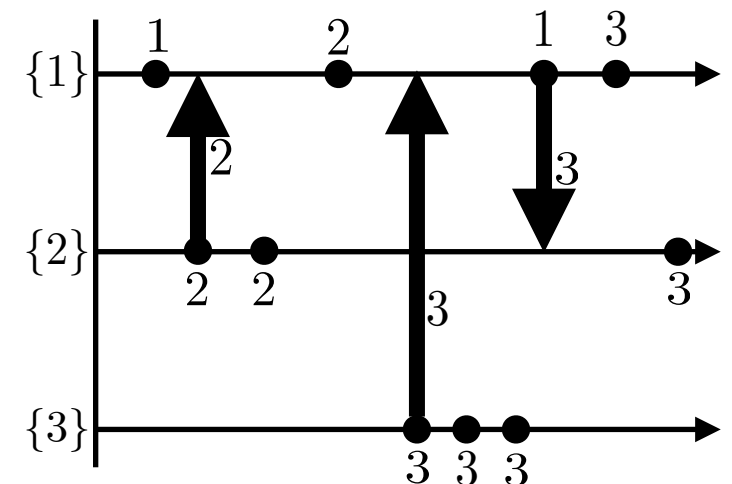
1. Plays an arm  $I_j(t) \in \{1, \dots, K\}$  and receives reward  $X_j(t) \in \{0, 1\}$
2. Can **choose to send information** to any other agent of choice

$\mathbb{P}[X_j(t) = 1 | I_j(t)] = \mu_{I_j(t)}$  *Independent rewards across agents*

Decentralized Algorithms -

Choice of arm of an agent only a function of its observed history.

Minimize Individual Regret -  $R_T^{(j)} := \mu^* T - \mathbb{E} \left[ \sum_{t=1}^T \mu_{I_j(t)} \right]$

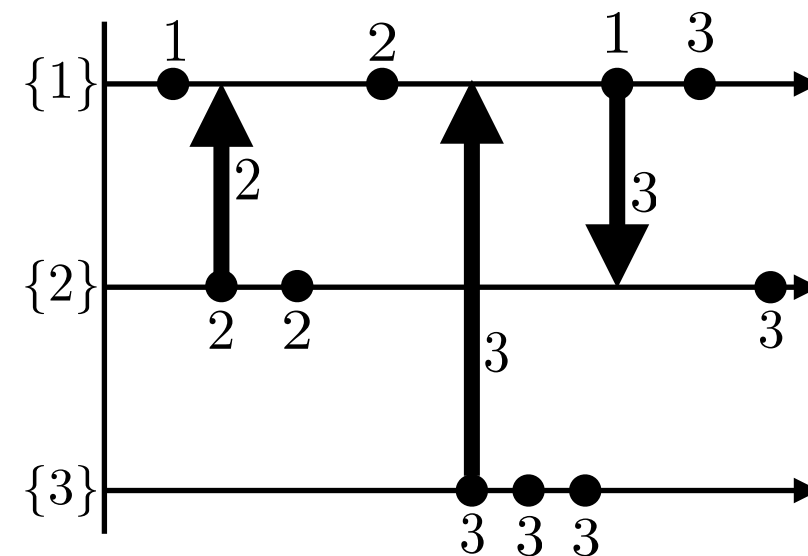


# The Multi Agent Problem

## Communication Constraints

### 1. Pairwise communication

Desirable for practical implementations



### 2. Fixed number of bits per message

Can't communicate arm-means to arbitrary precision.

Number of bits must not depend on arm-gap or other parameters of the instance.



### 3. Each agent communicates only $o(T)$ times

In the search engine example, each server gets a huge volume of search queries.

If servers communicate as frequently as they make decisions, it will congest the server system

# Related Work - Multi Agent Bandits

---

Multi-Agent Bandits are receiving wide attention -

## 1. Competitive Agents -

*Multiple agents pull same arm, no-one (or only a subset) get rewards*

[Anandkumar et.al '11],[Kalathil et.al. '14],[Rosenski et.al. '16],[Bistritz et.al. '18]

## 2. Collaborative Agents - Neighbors can observe all samples

[Buccapatnam et.al. '15][Landgren. et.al '16][Kolla et.al. '18] [Martinez-Rubio et.al '18]

# Main Result

---

Theorem (Informal) - We give a social algorithm, where regret of any agent is

$$O\left(\left(\frac{K}{N} + \log(N)\right) \frac{\log(T)}{\Delta}\right) + \underbrace{O\left(\frac{\log^3(N)}{\Delta^2} \log \log(N)\right)}_{\text{Constant Independent of time}}$$

1. Each agent communicates only  $O(\log(T))$  times
2. Each communication exchanges  $O(\log(K))$  bits (just arm-ids)
3. Each agent communicates with an agent chosen at random (Gossip)

# Main Result

---

In order to derive insight, suppose  $K=N$

	<u>No-Interaction</u>	<u>Social Learning</u>	<u>Full Interaction</u>
<u>Regret</u>	$O\left(N\frac{\log(T)}{\Delta}\right)$	$O\left(\log(N)\frac{\log(T)}{\Delta}\right)$	$O\left(\frac{\log(T)}{\Delta}\right)$
<u>Communication</u>	0	$O(\log(T))$	$T$

**Even a minimal collaboration helps in reducing regret**

# Key Idea - Gossip the Best Arm

---

1. Agents use communication to recommend arms

*Each message is  $\log(K)$  bits*

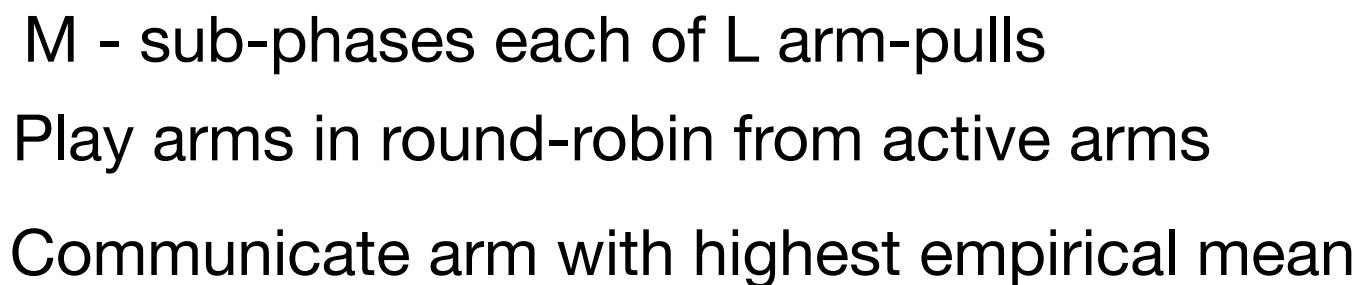
2. Agents only play from an active set of arms

*Initially every agent has a small set of active arms that forms a partition of all arms*

$$(K/N)+1$$

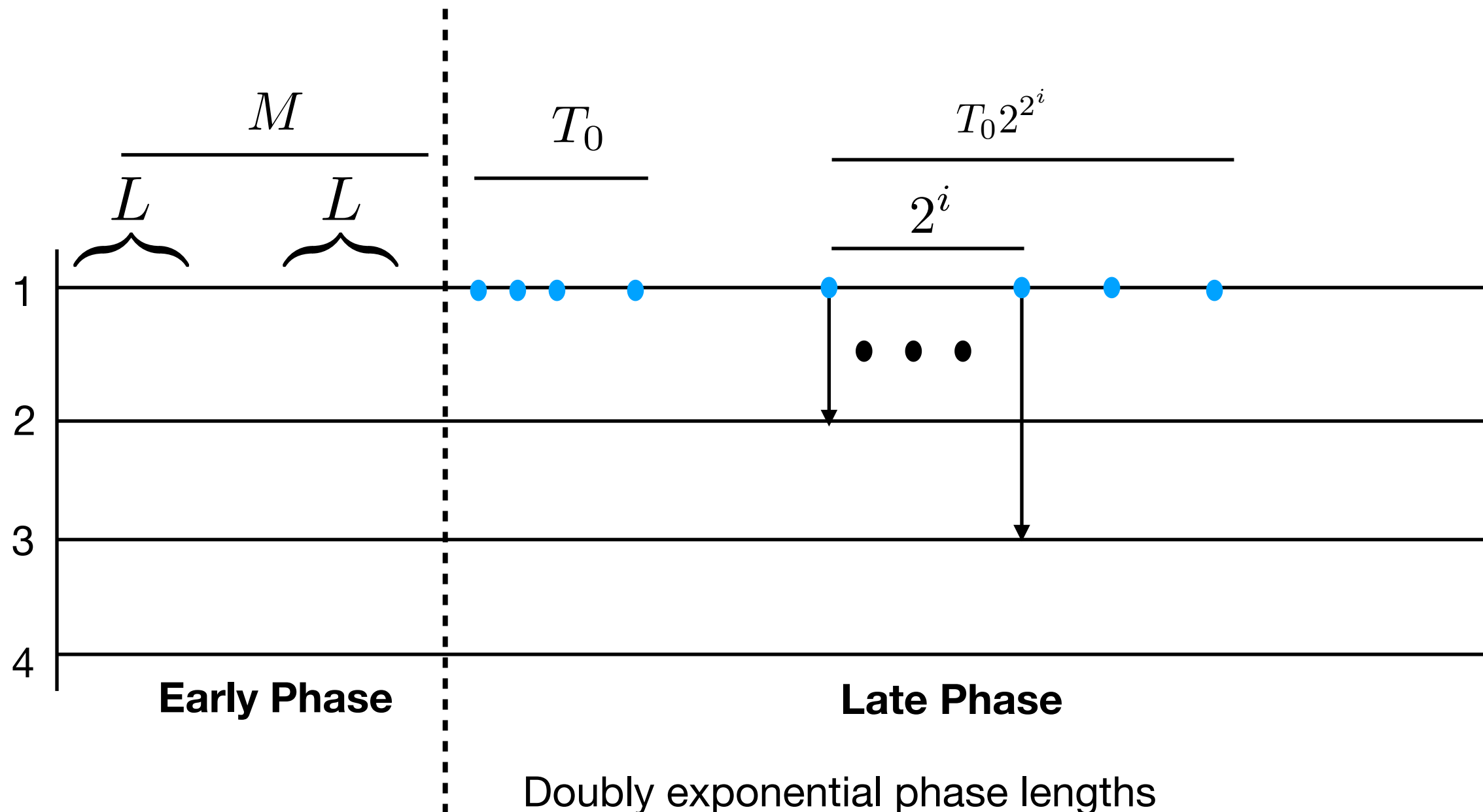
*The set of active arms grow with increasing recommendations (monotone)*

## Two Phases to algorithm evolution



# Algorithm Details

Two Phases to algorithm evolution



Doubly exponential phase lengths

Play arms according to UCB from active arms

Communicate arm the most played arm in the previous phase



# Performance Guarantees

---

## Theorem

Suppose all agents execute the algorithm with parameters

$$M = O(\log(N)), \quad 0 < \varepsilon \leq \Delta, \quad L = O\left(\frac{\log^2(N)}{\varepsilon^2} \log \log(N)\right) \quad \text{and}$$

$$T_0 = O\left(\frac{\max(K^2, N)}{\varepsilon^2} \log\left(\frac{1}{\varepsilon}\right)\right), \text{ then the regret of any agent } i \text{ is}$$

$$O\left(\left(\frac{K}{N} + \log(N)\right) \frac{\log(T)}{\Delta}\right) + \underbrace{O\left(\frac{\log^3(N)}{\varepsilon^2} \log \log(N)\right)}_{\text{Constant Independent of time}}$$

The total number of communications by any agent is  $M + O(N \log(T))$

# Performance Guarantees

---

## Proof Ingredients

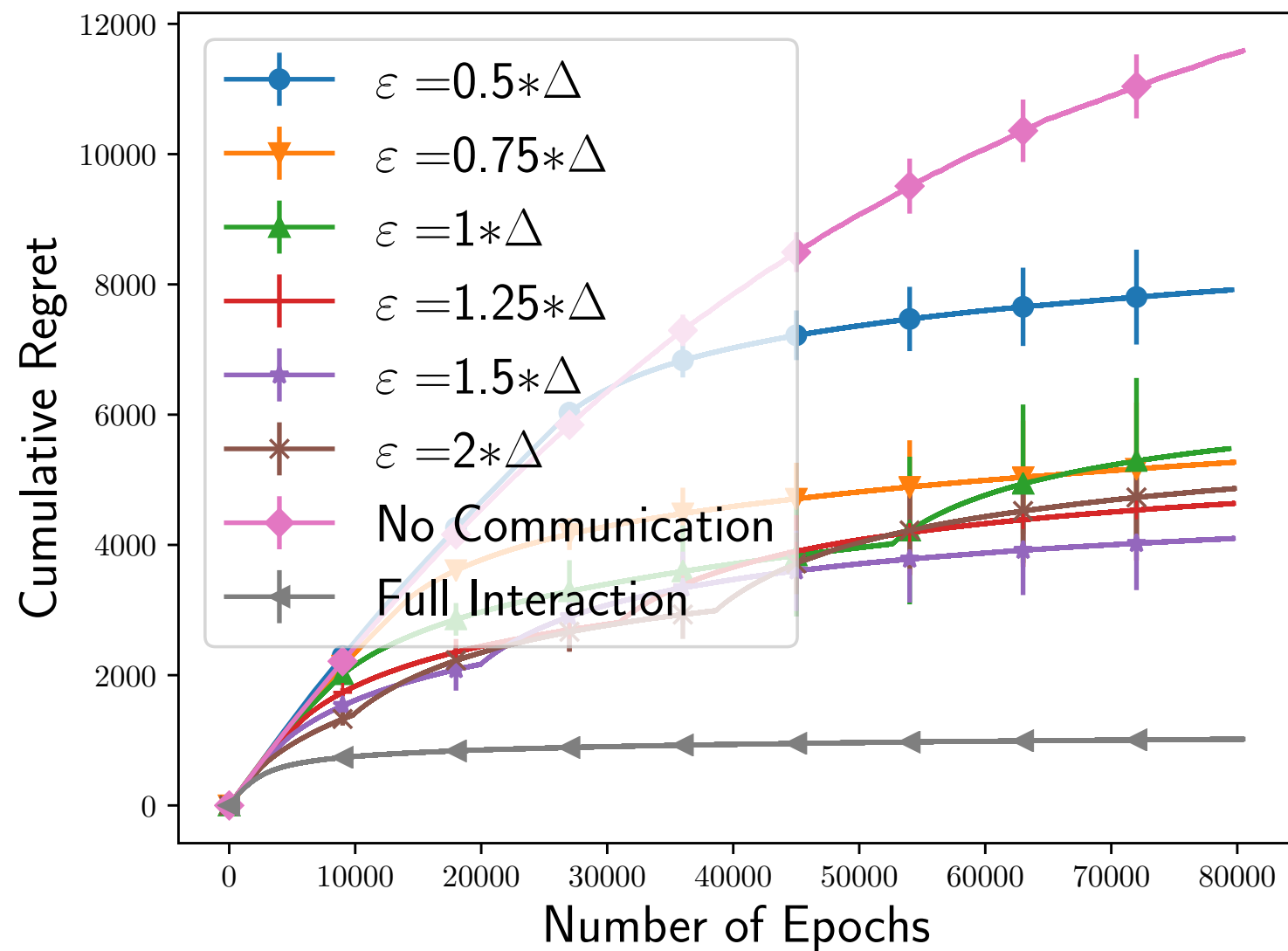
1. With high probability, at the end of the early phase, every agent's active arms
  - contains the best arm
  - total number of active arms is  $O(\log(N) + K/N)$
2. With high probability, at every late phase communication of every agent, the best arm is communicated.

## Algorithm Intuition

Not every agent needs to play and figure out a bad arm

Amortize exploration cost of bad arms across the network

# Empirical Performance



A representative plot with 20 agents and 50 Arms

# Conclusions

---

A new collaborative multi-agent MAB setting

Social Learning based algorithm

Each agent plays from a reduced set of arms

Cost of exploration of a bad arm is amortized over the network