

# Incentivized Exploration

IJCAI 2021 tutorial

Alex Slivkins (Microsoft Research NYC)

<https://www.microsoft.com/en-us/research/people/slivkins/>

Based on my survey “Exploration & Persuasion” (2021)

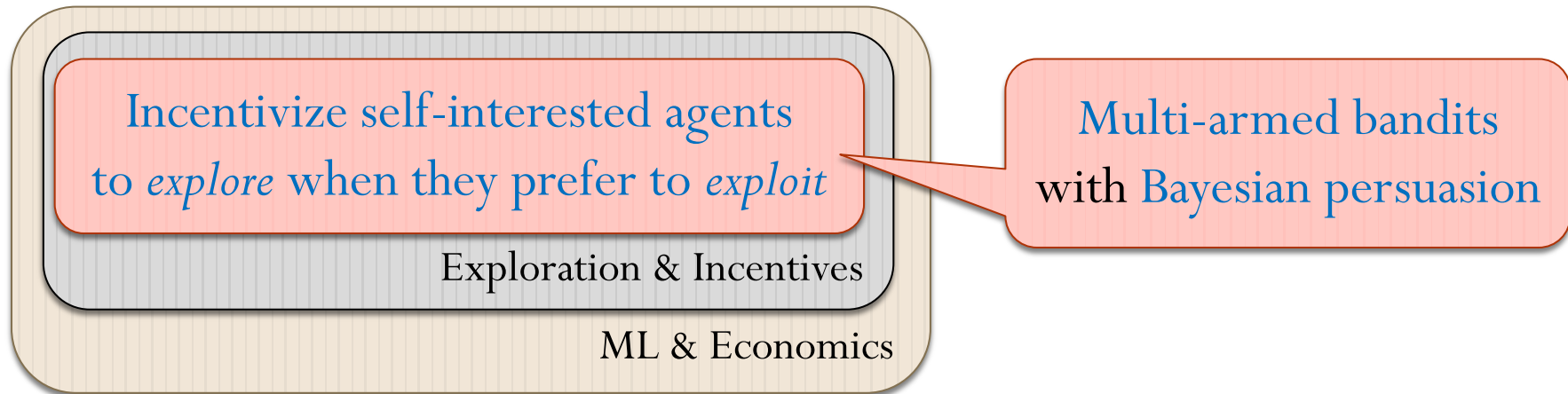
<http://slivkins.com/work/ExplPers.pdf>

See also Chapter 11 of my bandits book (<https://arxiv.org/abs/1904.07272>)

# Abstract (from the program)

How do you incentivize self-interested agents to explore when they prefer to exploit? In contrast with traditional formulations of exploration-exploitation tradeoff, agents control the choice of actions, whereas an algorithm can only issue recommendations. This problem space combines (algorithmic) exploration and (strategic) communication. The tutorial will be self-contained, providing sufficient background on both.

# Our scope: incentivized exploration



## Outline:

- ❑ (brief) background on multi-armed bandits
- ❑ deep-dive into “incentivized exploration”
  - ❑ (brief) background on Bayesian persuasion

# Bandits: examples

- **Dynamic pricing.**

You release a song which customers can download for a price. What price will maximize profit?

- Customers arrive one by one, you can update the price

- **Web advertisement.**

Every time someone visits your site, you display an ad.

Many ads to choose from. Which one maximizes #clicks?

- you can update your selection based on the clicks received

# Basic model

	<i>arms</i>	<i>rewards</i>
<i>pricing</i>	prices	payments
<i>web ads</i>	ads	clicks

- $K$  actions (“arms”),  $T$  rounds
- In each round  $t = 1 \dots T$  algorithm chooses an arm  $a_t$ , and observes the reward  $r_t \in [0,1]$  for the chosen arm
- “Bandit feedback”: no other rewards are observed!
- IID rewards: reward for each arm is drawn independently from a fixed distribution specific to this arm



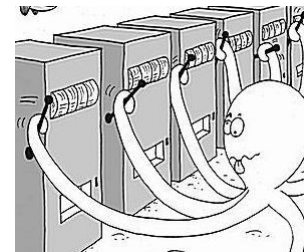
$\mu = .6$



$\mu = .2$



$\mu = .4$



# Basic model

	<i>arms</i>	<i>rewards</i>
<i>pricing</i>	prices	payments
<i>web ads</i>	ads	clicks

- $K$  actions (“arms”),  $T$  rounds
- In each round  $t = 1 \dots T$  algorithm chooses an arm  $a_t$ , and observes the reward  $r_t \in [0,1]$  for the chosen arm
- “Bandit feedback”: no other rewards are observed!
- IID rewards: reward for each arm is drawn independently from a fixed distribution specific to this arm

- **Regret**  $R(T) = T\mu^* - \sum_{t \in [T]} r_t$   
 $\mu_a \in [0,1]$ : mean reward of arm  $a$  (fixed over time)  
best arm benchmark:  $\mu^* = \max_a \mu_a$
- *Bayesian bandits*: ( $\mu_a$ : arms  $a$ ) drawn from known prior  
**Bayesian regret**:  $E_{\text{prior}}[R(T)]$

# Exploration vs Exploitation

- Explore: try out new arms to get more info  
... perhaps playing low-paying arms
- Exploit: play arms that seem best based on current info  
... but maybe there is a better arm we don't know about
- Bandits: fundamental model for explore-exploit tradeoff
- Studied since 1933 in OR, Econ, CS, Stats,  
various versions and extensions



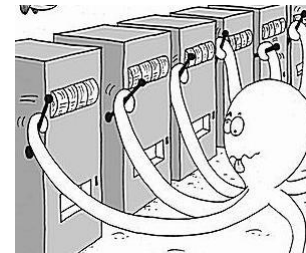
$\mu = .6$



$\mu = .2$



$\mu = .4$



# More examples

Example	Action	Rewards / costs
medical trials	drug to give	health outcomes
internet ads	which ad to display	bid value if clicked, 0 othw
content optimization	e.g.: font color or page layout	#clicks
sales optimization	which products & prices to offer	\$\$\$
recommender systems	suggest a movie, restaurants, etc.	user satisfaction
computer systems	which server(s) to route the job to	job completion time
crowdsourcing systems	assign tasks to workers	quality of completed work
	which price to offer?	#completed tasks
wireless networking	which frequency to use?	#successful transmissions
robot control	a “strategy” for a given state & task	#tasks successfully completed
game playing	an action for a given game state	#games won

# Many “problem dimensions”

**Non-IID rewards:** e.g., chosen by an adversary  
(constrained adversary: rewards cannot change too much or too often)

**Context** observed before each round (e.g.: user profile/features)

**Known structure:** e.g.: arms are points in  $[0,1]^d$ ,  
rewards are linear/concave/Lipschitz function of the arm

**Bayesian prior** (problem instance comes from known distribution)

**Global constraints:** e.g.: limited #items to sell

**Complex decisions:** a *slate* of articles, prices for *several* products

**Books** on bandits: Gittins et al. (2011), Bubeck & Cesa-Bianchi (2012),  
[more current] **Slivkins** (2019-2021), Lattimore & Szepesvari (2020)

# Example: Two-armed bandits

**Non-adaptive exploration** (does not adapt to observations)

- try each arm  $N$  times (*explore*), choose the best one & *exploit*
- concentration  $\Rightarrow |\mu_a - \hat{\mu}_a| < \frac{\log 1/\delta}{\sqrt{N}}$  w/prob  $1 - \delta$
- lose  $\sim \sqrt{N}$  per round in exploit,  $\sim 1$  /round in explore,  
optimize  $N \Rightarrow$  regret  $R(T) = \tilde{O}(T^{2/3})$

optimal for non-adaptive exploration

**Adaptive exploration**

- alternate arms until one of them is better w.h.p., then *exploit*
- concentration:  $G := |\mu_1 - \mu_2| < \tilde{O}\left(\frac{1}{\sqrt{t}}\right) \quad \forall \text{round } t \text{ in exploration}$
- regret  $R(T) = \tilde{O}(\min(\sqrt{T}, 1/G))$

optimal

# Outline

✓ (brief) background on bandits

Deep-dive into “incentivized exploration”

*How to incentivize self-interested agents to explore if they prefer to exploit?*

# Motivation: recommender systems

- Watch this movie
- Dine in this restaurant
- Vacation in this resort
- Buy this product
- Drive this route
- See this doctor



# Info flow in recommender system

- user arrives, needs to choose a product
- receives recommendation (& extra info)
- chooses a product, leaves feedback

consumes info  
from prior users

produces info  
for future users

For common good, user population should balance

- **exploration**: *trying out various alternatives to gather info*
- **exploitation**: *making best choices given current info*

Example: coordinate via system's recommendations.

# Misaligned incentives

**Problem:** self-interested users (*agents*) favor exploitation

- **Under-exploration:** some actions explored at sub-optimal rate

Ex: best action remains unexplored if it seems worse initially

- **Selection bias:** chosen action & outcome depend on agents' type

Ex: you may only see people who are likely to like this movie

- rarely see some sub-population  $\Rightarrow$  learn slowly, at best
- data is unreliable at face value

# Model: incentivized exploration

- T rounds, K actions (“arms”). In each round  $t$  : “GREEDY algorithm”  
new agent arrives, observes *something* ( $\text{msg}_t$ ), chooses an arm, and reports her reward  $\in [0,1]$
- IID rewards: reward of arm  $a$  drawn from distribution  $D_a$   
Distributions fixed but unknown; common Bayesian prior  
Objective: social welfare (= cumulative reward)

default: full history

*Rational choice*:  $\operatorname{argmax}_{\text{arms } a} E[\mu_a | \text{msg}_t]$

deterministic rewards

# What goes wrong with GREEDY

$a_t \in \operatorname{argmax}_a E[\mu_a | H_t]$ ,  $H_t$  is history @ round  $t$  (*exploitation-only*)

- 2 arms,  $G := E[\mu_1 - \mu_2] > 0$  expectation over the prior
- Round 1: arm 1 is chosen,  $\mu_1$  is observed
- If  $\mu_1 > E[\mu_2]$  then arm 2 is never chosen exploration fails”

# What goes wrong with GREEDY

$a_t \in \operatorname{argmax}_a E[\mu_a | H_t]$ ,  $H_t$  is history @ round  $t$  (*exploitation-only*)

- 2 arms,  $G := E[\mu_1 - \mu_2] > 0$

expectation over the prior

- Thm:  $\Pr[\text{arm 2 is never chosen}] \geq G$

exploration fails”

- Proof: Let  $\tau$  first time arm 2 is chosen,  $T+1$  othw

- $Z_t = E[\mu_1 - \mu_2 | H_t]$

$Z_t > 0 \Rightarrow \text{arm 1}$

- $E[Z_t | H_{t-1}] = Z_{t-1}$

(Doob) martingale

- $\tau$  is a “stopping time” determined by  $H_t$

Optional Stopping Theorem

- $G = E[Z_1] = E[Z_\tau]$

$$= \Pr[\tau \leq T] E[Z_\tau | \tau \leq T] + \Pr[\tau > T] E[Z_\tau | \tau > T]$$

$a_\tau = 2$ , so  $Z_\tau \leq 0$

“arm 2 never chosen

$\leq 1$

# Incentivize exploration without payments

How to incentivize agents to try seemingly sub-optimal actions?

based on agents' biases and/or system's current info)

“External” incentives:

- monetary payments / discounts
- promise of a higher social status
- people's desire to experiment

prone to selection bias;  
not always feasible

## Recommendation systems

Watch this movie

NETFLIX

Dine in this restaurant

yelp

Vacation in this resort

tripadvisor

Buy this product

amazon.com

Drive this route

waze

See this doctor

suggest  
doctor

# Incentivize exploration without payments

How to incentivize agents to try seemingly sub-optimal actions?

based on agents' biases and/or system's current info)

“External” incentives:





- monetary payments / discounts
- promise of a higher social status
- people's desire to experiment

prone to selection bias;  
not always feasible

**Recommendation systems**

**Our approach:** *create info asymmetry by not revealing full history*

One in this restaurant  
Vacation in this resort  
Buy this product  
Drive this route  
See this doctor

# Incentivized Exploration

- T rounds, K actions (“arms”). In each round  $t$  :  
new agent arrives, observes *something* ( $\text{msg}_t$ ),  
chooses an arm, and reports her reward  $\in [0,1]$
- IID rewards: reward of arm  $a$  drawn from distribution  $D_a$   
Distributions fixed but unknown; common Bayesian prior  
Objective: social welfare (= cumulative reward)

chosen by algorithm

*Rational choice*:  $\operatorname{argmax}_{\text{arms } a} \mathbb{E}[\mu_a | \text{msg}_t]$

w.l.o.g.  $\text{msg}_t$  is a suggested arm, &  
algorithm is *Bayesian Incentive-Compatible* (BIC):  
 $\mathbb{E}[\mu_a - \mu_b | \text{msg}_t = a] \geq 0 \quad \forall t, \text{arms } a, b$

bandit algorithm  
with BIC constraint

compare BIC algs  
vs. optimal algs

# Paper trail (by first pub)

Kremer, Mansour, Perry (2013)

Che & Horner (w.p. 2013)

Mansour, Syrgkanis, **Slivkins** (2015)

Papanastasiou, Bimpikis, Savva (w.p. 2015)

Mansour, Syrgkanis, **Slivkins**, Wu (2016)

Bahar, Smorodinsky, Tennenholtz (2016)

Schmit & Riquelme (2018)

Immorlica, Mao, **Slivkins**, Wu (2019)

Immorlica, Mao, **Slivkins**, Wu (2020)

Bahar, Smorodinsky, Tennenholtz (2019)

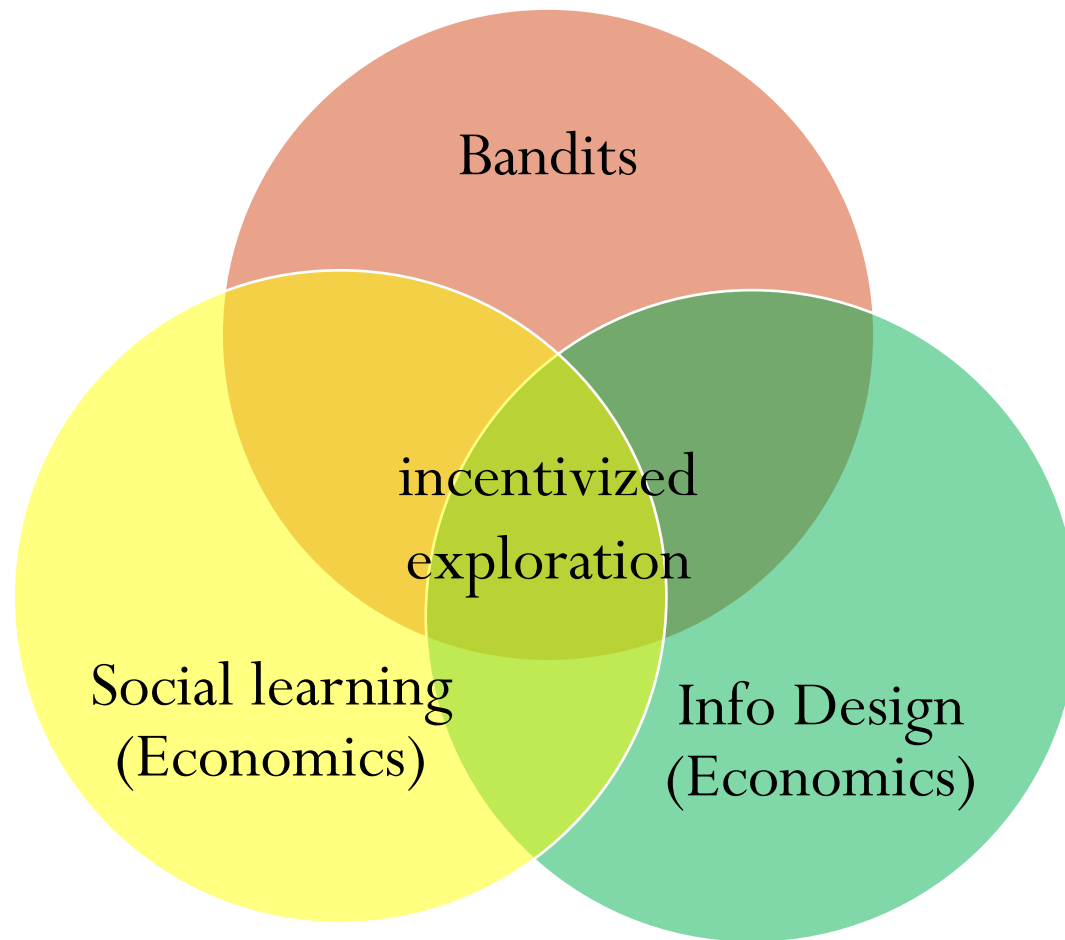
Cohen & Mansour (2019)

Sellke & **Slivkins** (2021)

**Slivkins** & Simchowitiz (2021)

Home community:  
economics & computation  
(ACM EC)

# “Zoom out”



# Outline

- ✓ (brief) background on bandits

Deep-dive into “incentivized exploration”

- ✓ Motivation & model

- Focus on a single round: Bayesian Persuasion

# One round: Bayesian Persuasion

## Game protocol

- principal receives a **signal**, recommends an arm **rec**
- agent observes **rec** and chooses an arm  $a_*$
- rewards:  $\mu_a$  for the agent,  $u_a$  for the principal

## What's known

- Bayesian prior on (reward vectors  $\mu, u$ ; signal)
- principal's policy: signal  $\rightarrow$  recommendation

## Rational agent

$$a_* = \operatorname{argmax}_{\text{arms } a} E[\mu_a | \text{rec}]$$

## Principal's goal

Choose policy to maximize  $E[u_{a_*}]$

wlog  $E[\mu_1] > E[\mu_2]$

## In "incentivized exploration"

- Signal: algorithm's history;  
e.g.,  $u = (0, 1)$  (*principal's goal: explore arm 2*)
- Example:  $T = 2$  & deterministic rewards  
In round 2: **Bayesian Persuasion with signal  $\mu_1$**

# Ex: Bayesian Persuasion

2 arms,  $E[\mu_1] > E[\mu_2]$   
Signal  $\mu_1 \in \{L, H\}$   
Principal's reward  $u = (0,1)$

Exact solution

$$E[u_{a_*}] = \frac{H - E[\mu_1]}{H - E[\mu_2]}$$

Under “full revelation”:  $E[\mu_2] \rightarrow L$

Technique

foundational in BP

1. Belief  $B_\pi = \Pr(\mu_1 = H \mid \text{rec})$  given policy  $\pi$   
RV on  $[0, 1]$  with expectation  $\Pr[\mu_1 = H]$   
realization determines the agent's choice
2. Any *consistent* RV is realized as  $B_\pi$  for some  $\pi$
3. Maximize directly over all *consistent beliefs*

“Consistent  
belief”

Recap: completely solved (a special case of)

Incentivized Exploration with  $T = 2$  & deterministic rewards

# Outline

✓ (brief) background on bandits

Deep-dive into “incentivized exploration”

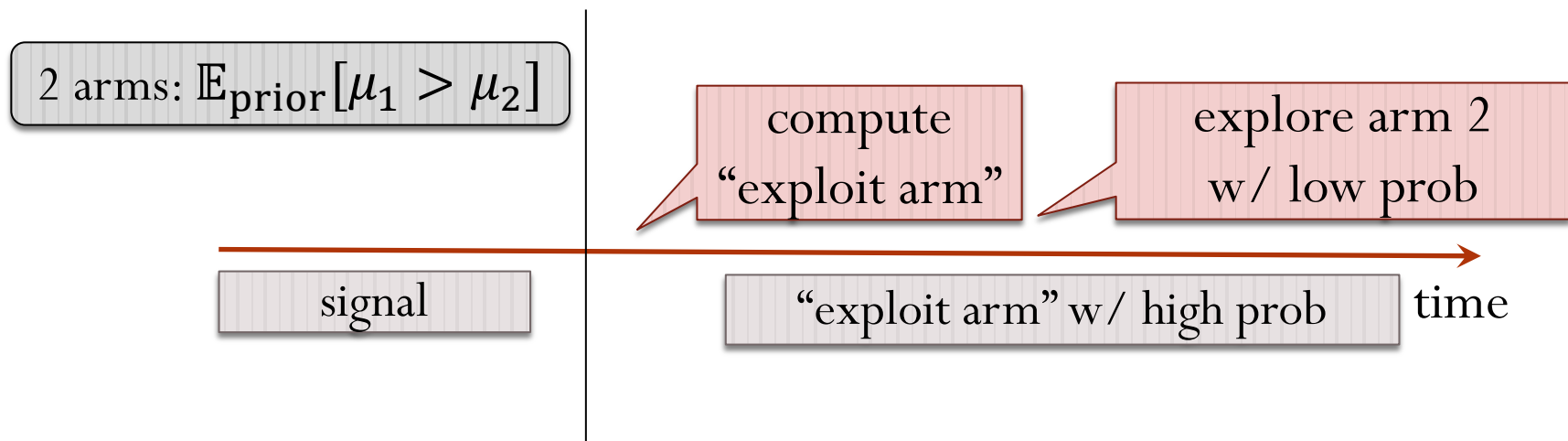
✓ Motivation & model

✓ One round: Bayesian Persuasion

□ A general solution

# Hidden exploration

**Key idea:** Hide exploration in a large pool of exploitation



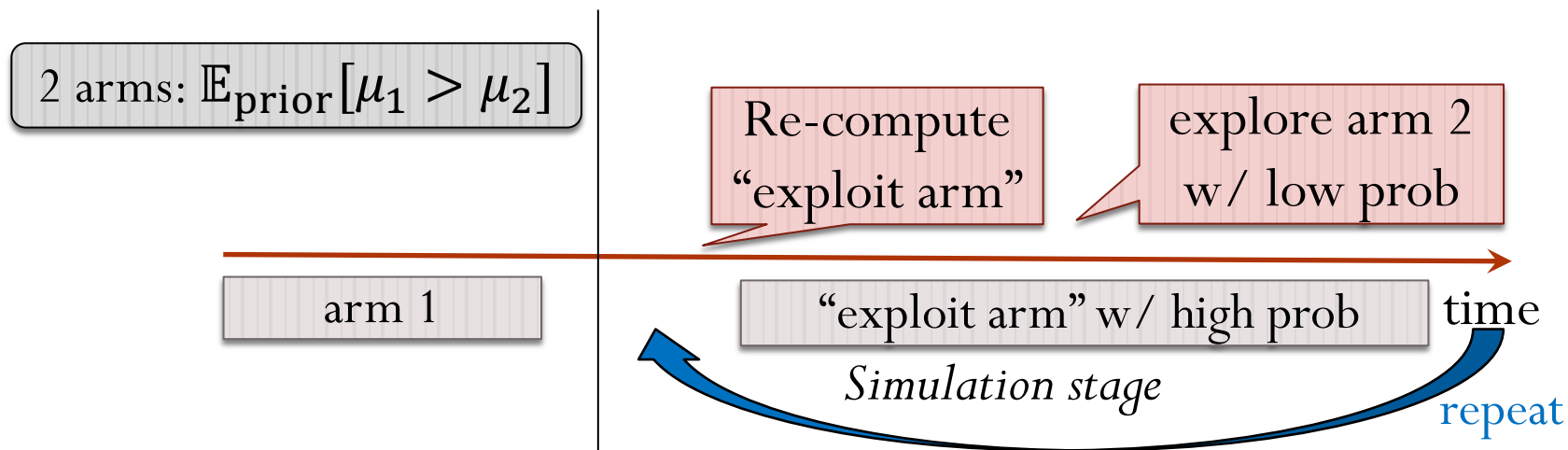
Enough initial samples  $\Rightarrow$  any arm could be the exploit arm!

Agent does not know if it is exploitation or exploration

Explore prob. low enough  $\Rightarrow$  follow recommendation.

# Repeated Hidden exploration

**Key idea:** Hide exploration in a large pool of exploitation



"Explore" prob. low enough  $\Rightarrow$  follow recommendation.

*Performance:* pick arm 2 with (small) const prob in each round

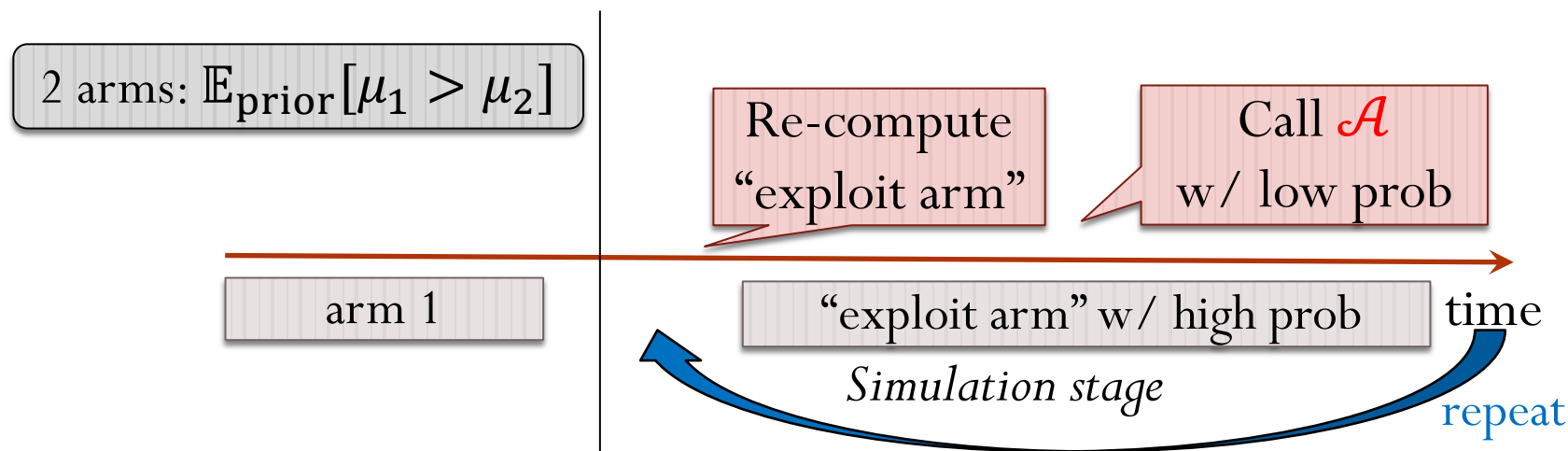
*Non-adaptive exploration* (can exploit after fixed #rounds)

Repeated

Simulate bandit algorithm  $\mathcal{A}$

# Hidden exploration

**Key idea:** Hide exploration in a large pool of exploitation



"Explore" prob. low enough  $\Rightarrow$  follow recommendation.

Performance:  $\mathbb{E}_{\text{prior}}[\text{reward}]$  of exploit arm  $\geq$  that of  $\mathcal{A}$

Bayesian regret: *match  $\mathcal{A}$  up to a prior-dependent factor*

# Assumptions on the prior

- Hopeless in general: e.g., if  $\mu_1$  and  $\mu_1 - \mu_2$  are independent

- **Sufficient condition:**

*Arm 2 can become “exploit arm” after enough samples of arm 1.*

- $G_n := \mathbb{E}[\mu_2 - \mu_1 | n \text{ samples of arm 1}]$  (“posterior gap”)

$$\exists n: \mathbb{P}(G_n > 0) > 0$$

- This condition is **necessary** to sample arm 2 in any round  $t$

- Proof:  $E[\mu_2 - \mu_1 | \text{rec}_t = 2] = E[G_t | \text{rec}_t = 2] \leq 0$

Law of iterated expectation & induction on  $t$

if the condition is false

- Similar condition suffices for  $> 2$  arms

Includes: *independent priors, bounded rewards, full support on  $[L, H]$*

# Outline

✓ (brief) background on bandits

Deep-dive into “incentivized exploration”

✓ Motivation & model

✓ One round: Bayesian Persuasion

✓ A general solution: Hidden Exploration

□ Extensions in the basic model

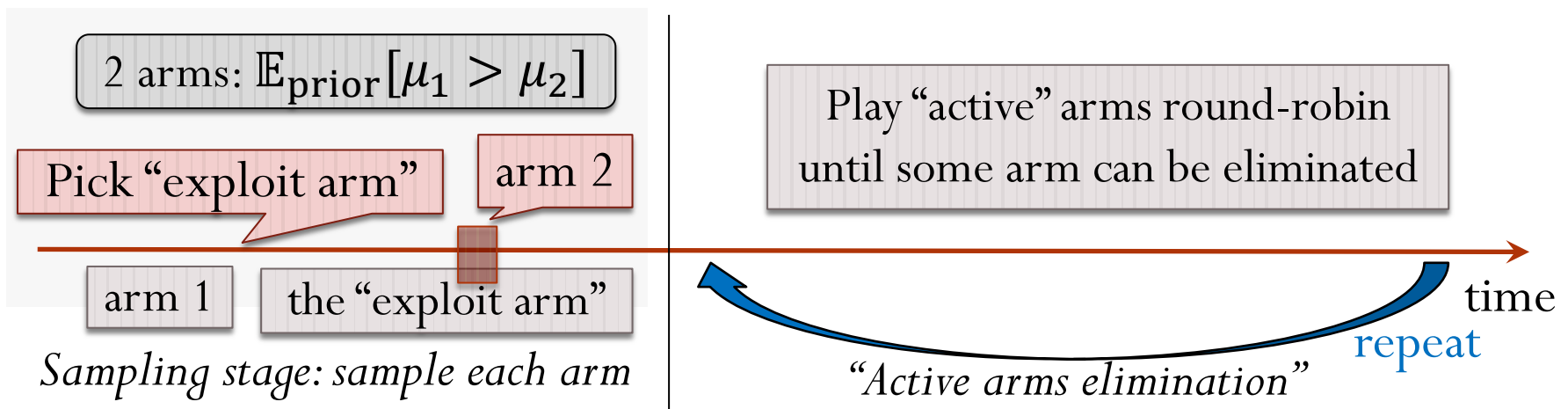
□ Beyond the basic model

□ Concluding remarks

# Beyond Bayesian regret

- “Exploit arm” computed via Bayesian update  
only good in *expectation over the prior*  $\Rightarrow$  only *Bayesian regret*
- Regret bounds *for each realization of the prior*?  
Different algorithm, (only) uses sample average rewards
- This algorithm is “detail-free”
  - instead of the full prior, inputs (only)  
two numerical parameters, and only approximately
  - agents can have different beliefs, “consistent” with the inputs

# The detail-free algorithm



Define "exploit arm" & "elimination condition" via sample averages.  
For BIC, connect sample averages to Bayesian posteriors (tricky!).  
Enough initial samples  $\Rightarrow$  "Active arms elimination" is BIC

$$R(T) = T\mu^* - \sum_{t \in [T]} r_t$$

# Optimal regret bounds

For each realization of the prior  $\mathcal{P}$ :

Constant # arms

$$\text{Regret}(T) = O\left(\underset{\text{BIC}}{c_{\mathcal{P}}} \min\left(\frac{\log T}{\text{Gap}}, \sqrt{T \log T}\right)\right)$$

Depends on  $\mathcal{P}$ .  
“Price” for **BIC**.

**Gap** between best & 2nd-best arm  
Optimal regret for given **Gap**.

optimal regret  
in the worst case

# Price of Incentives

## Problem

**Sample complexity:** #rounds to explore each arm once  
Independent priors:  $K$  arms, all arms' priors from family  $\mathcal{F}$

## Results

#rounds is **linear** or **exponential** in  $K$ , depending on  $\mathcal{F}$

For Beta priors and truncated Gaussian priors,

- #rounds is **linear** in  $K$
- exponential in “strength of beliefs”:  $1/\min_{\mathcal{P} \in \mathcal{F}} \text{Var}(\mathcal{P})$

## Algorithm

Probabilistically chooses between three branches:  
exploration, exploitation & “secret sauce” combining both;  
Exploration prob increases exponentially over time

# “Natural” BIC algorithms

- Thompson Sampling: standard, optimal bandit algorithm
- Thompson sampling is BIC given a “warm-start”:  
 $N$  samples from each arm, where  $N$  determined by the prior
  - assume independent priors
  - $N$  is linear in  $K = \text{\#arms}$ , and  $O(\log K)$  for Beta priors
  - alt: collect the  $N$  samples exogenously (e.g., pay)
- “Price of Incentives”: performance loss due to the warm-start
  - Bayesian regret  $\leq \text{\#rounds}$ ,
  - use “sample complexity” results to bound  $\text{\#roundsc}$
- Similar results for other “natural” bandit algorithms ???

# Optimal BIC algorithms

$$E[\mu_1] > E[\mu_2]$$

## Result

Optimal BIC algorithm for 2 arms & deterministic rewards  
(first result on incentivized exploration: Kremer, Mansour, Perry '13)

## algorithm

- in round 1, sample arm 1, observe  $\mu_1$
- place  $\mu_1$  among thresholds  $0 = \theta_1 < \theta_2 < \theta_3 < \dots$   
let  $n$  be such that  $\theta_n \leq \mu_1 < \theta_{n+1}$
- first time choose arm 2 in round  $n$ ,  
choose the better arm ever after

## Analysis outline

1. There is an optimal BIC algorithm which is “threshold-based”
2. Optimize among “threshold-based” algorithms

# Outline

✓ (brief) background on bandits

Deep-dive into “incentivized exploration”

✓ Motivation & model

✓ One round: Bayesian Persuasion

✓ A general solution: Hidden Exploration

✓ Extensions in the basic model

❑ Beyond the basic model

❑ Concluding remarks

# Beyond the basic model

## Extend the ML model

- auxiliary feedback before and/or after each round
- large, structured problems, *e.g.*, incentivized RL

## Extend the Econ model

- heterogenous agents (public or private types)
- multiple agents playing a game
- inevitable revelation: some history observed no matter what
- a common theme: explore all “explorable” arms (some arms aren’t)
- relax rationality assumptions

# [Relaxing] rationality assumptions

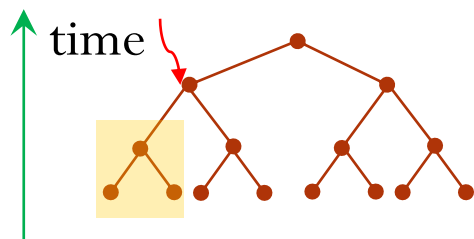
- “Power to commit” to the algorithm: do I know the algorithm?  
Do I trust the platform to implement it?
- Cognitive limitations: e.g., can/would I do a Bayesian update?
- Rational choice: would I just optimize expected utility?
  - Risk aversion, SoftMax vs HardMax
  - “experimentation aversion”

## How to ensure predictable user behavior?

Immorlica, Mao, Slivkins, Wu (2020)

# Unbiased histories

- Users want full history; let's give them the next best thing
- Principal only chooses partial order (DAG) on rounds



of the relevant sub-algorithm

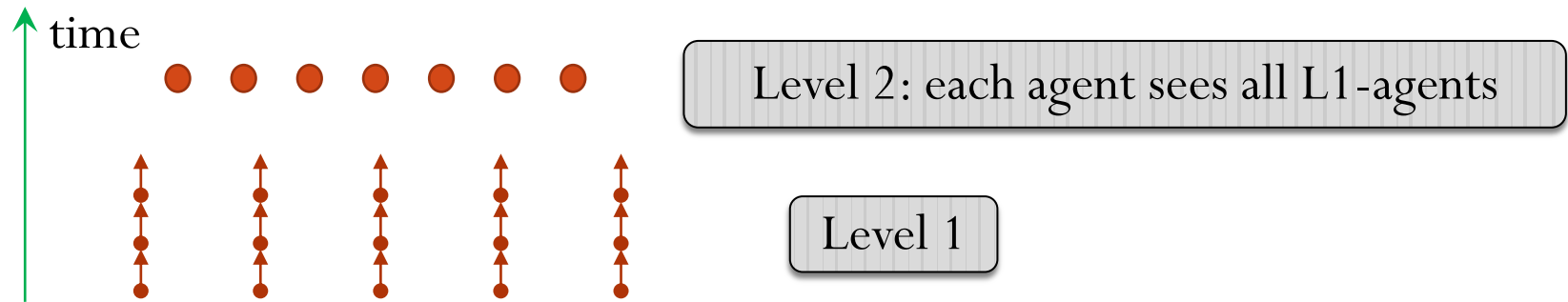
- Each user sees full history *of her branch*

“Unbiased history”: data-independent, e.g., no sub-sampling

- Economics foundation: assumptions only on users that see full history
  - HardMax or SoftMax? anything consistent with confidence intervals

# Design the partial order

Each agent is “locally greedy”, and yet it works!



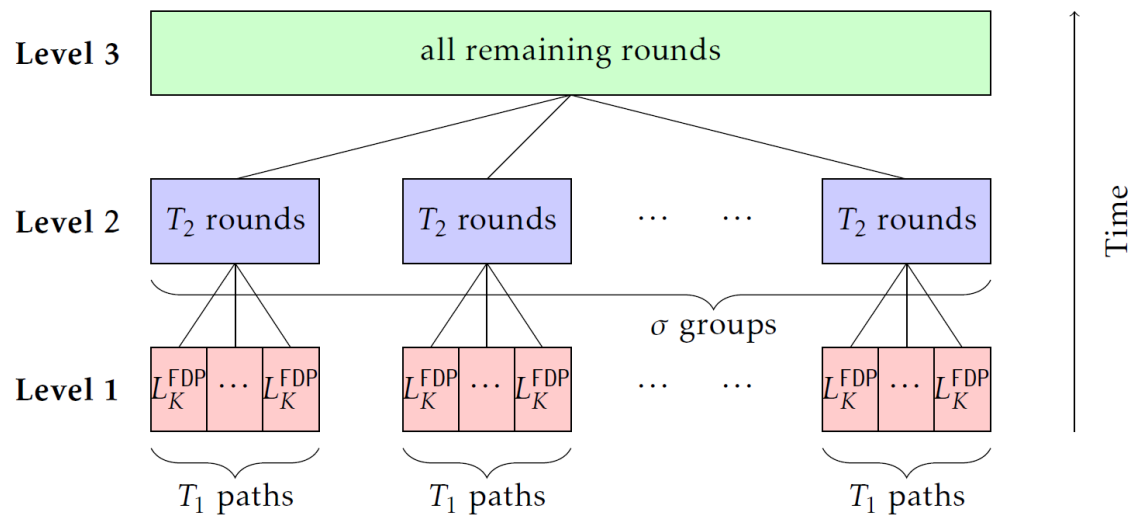
Simple construction (2 arms): regret  $T^{2/3}$

Two “levels”: implements non-adaptive exploration

Can we get  $\sqrt{T}$  regret?

# Adaptive exploration

Beat the  $T^{2/3}$  barrier:  $T^{4/7}$  regret with 3 levels



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

# Adaptive exploration

$\sqrt{T}$  regret with  $\log T$  levels (for constant #arms)

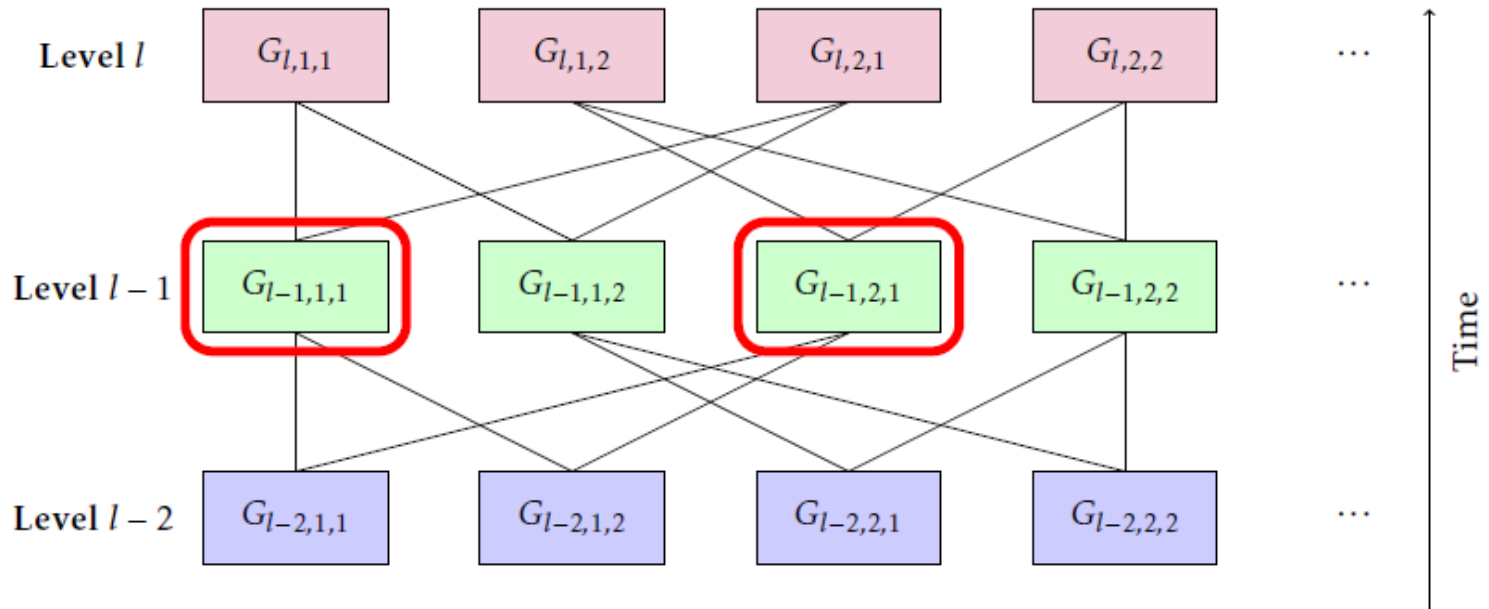


Figure 3: Interlacing connections between levels for the  $L$ -level policy.

# Outline

- ✓ (brief) background on bandits

Deep-dive into “incentivized exploration”

- ✓ Motivation & model
- ✓ One round: Bayesian Persuasion
- ✓ A general solution: Hidden Exploration
- ✓ Extensions in the basic model
- ✓ Beyond the basic model
- Concluding remarks

# Perhaps “full revelation” suffices?

- Does greedy algorithm work?

Yes, for linear bandits with [smoothed/diverse contexts](#)

Bastani, Bayati, Khosravi '18

$\sqrt{T}$  regret: (Kannan, Morgenstern, Roth, Waggoner, Wu '18)

$T^{1/3}$  Bayesian regret: (Raghavan, Slivkins, Vaughan, Wu; '18)

- Maybe different people just try out different things?

*Probably not enough*: want best action for each type

(and exploring all what's explorable was very tricky!)

*Yes, under strong assumptions*

Schmit & Riquelme, '18; Acemoglu, Makhdoumi, Malekian, Ozdaglar, '17

All directions very open, despite  
substantial prior work on some

# Open questions

- Re relaxed economic assumptions:  
Do we have the “right” ones? (and what does “right” mean?)  
Make the constructions simpler/ more general / more robust
- [Adapting to] partially known priors
- Long-lived agents
- Inevitable observations:  
some aspects of the history are always observed
- Heterogenous agents: regret bounds?  
Can we use diversity to help BIC exploration?

# Connection to medical trials

Medical trial as a bandit algorithm: for each patient, choose a drug

- one of original motivations for bandits
- basic design: new drug vs. placebo (blind, randomized)  
“advanced” designs studied & used (adaptive,  $>2$  arms, contexts)

- Participation incentives: why take less known drug?  
Major obstacle, esp. for wide-spread diseases & cheap drugs.
- Medical trial as a BIC recommendation algorithm!
  - minimal info disclosure is OK for medical trials

# Bandits & agents

- agents choose **actions**  $\Rightarrow$  incentivized exploration  
via info asymmetry (our scope) and/or with money
- agents choose **bids**  $\Rightarrow$  repeated auctions  
dynamic auctions (ex: Athey & Segal '13, Bergemann & Valimaki '10)  
ad auctions with unknown CTRs (ex: Babaioff, Kleinberg, Slivkins '10)
- agents only affect **rewards**  
dynamic {pricing, assortment, contract design}
- agents (users) choose between **bandit algorithms**  
Bandit algorithms compete for users (e.g., Google vs Bing)  
(ex: Aridor, Mansour, Slivkins, Wu '20)