

Bandits and Agents: How to incentivize exploration?

Alex Slivkins (Microsoft Research NYC)

Joint work with

Yishay Mansour (Tel Aviv University)

Vasilis Syrgkanis (MSR-NE)

Steven Wu (Penn)

EC'15, EC'16
working papers
ongoing work

Motivation: recommender systems

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



-
- Take this medicine
 - Use these settings



(medical trials)



(systems)

Exploration

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*

The balance can be coordinated by system's recommendations

Exploration and incentives

Recommender system:

- *agent* 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

Agents make decisions based on available info & initial biases

An alternative that seems worse initially may remain unexplored because agents have no incentives to explore it!

How to incentivize agents to explore?

Exploration and incentives

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

Exploration and incentives

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

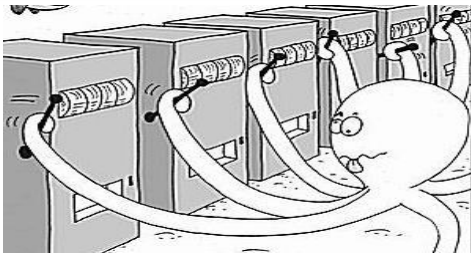
prone to selection bias;
not always feasible

Alternative approach: use *information asymmetry* to create *intrinsic incentives* to follow system's recommendations

Basic model

- K actions; T rounds
- In each round, a new agent arrives: “actions” = “arms”
 - algorithm recommends an action (& extra info)
 - agent chooses an action, reports her reward $\in [0,1]$
- IID rewards: distribution depends only on the chosen action
- Mean rewards are unknown; common Bayesian prior
- Objective: social welfare (= cumulative reward)

If agents follow recommendations \Rightarrow “multi-armed bandits”



classical model in machine learning
for explore-exploit tradeoff

Basic model: BIC bandit exploration

How to account for agents' incentives?

Ensure that following recommendations is in their best interest!

Recommendation algorithm is *Bayesian Incentive-Compatible* (**BIC**) if

$$\mathbb{E}_{\text{prior}}[\text{reward}(a) - \text{reward}(b) \mid \text{rec}_t = a] \geq 0$$

\forall round t , arms a, b

recommendation in round t

Goal: design **BIC** bandit algorithms to maximize performance

Can **BIC** bandit algorithms perform as well as the best bandit algorithms, **BIC or not?**

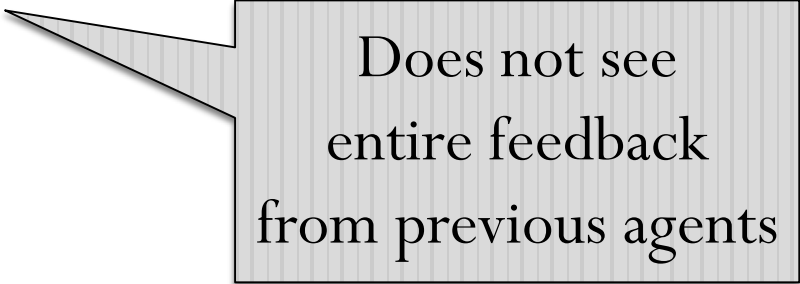
Exploration, exploitation, incentives

Algorithm wants to **balance exploration & exploitation**,
can choose suboptimal arms for the sake of new info

Each agent is myopic: **does not care to explore, only exploits**

... based on what she knows:

- common prior
- recommendation algorithm
- algorithm's recommendation
(& extra info, if any)



Does not see
entire feedback
from previous agents

Information asymmetry

- Revealing all info to all agents does not work

Then algorithm can only exploit \Rightarrow not good.
E.g.: can only pick the “prior best” arm.

- So, algorithm needs to reveal less than it knows.
W.l.o.g., reveal only recommended arm, no extra info

Approach: hide *a little exploration* in *lots of exploitation*.

- Each agent gets “exploitation” with high prob,
“exploration” with low prob, but does not know which

Related work: multi-armed bandits

- Most related: IID rewards, with or without a prior
E.g.: Thompson Sampling, Gittins Index, UCB1 (Auer et al.'02).
- *Best arm prediction*: care about learning rate, not total reward
E.g.: Even-Dar et al.'02, Goel et al.'09, Bubeck et al.'11.
- Bandits with agents/incentives:

dynamic pricing (E.g.: Kleinberg & Leighton'03, Besbes & Zeevi'09)
ad auctions with unknown CTRs (E.g.: Babaioff et al.'09,'10','13)
dynamic auctions (E.g.: Athey & Segal'13, Bergemann & Valimaki'10)

Related work: BIC exploration in Econ

- Kremer, Mansour, Perry (2014): same model, two arms. Bayesian-optimal algorithm for deterministic rewards, very suboptimal performance for IID rewards
- Frazier, Kempe, J.Kleinberg & R.Kleinberg (2014): payments allowed, agents observe past actions
- Connections to some high-profile work in Economics
 - Bayesian Persuasion (Kamenica & Gentzkow: Econometrica'11)
 - Strategic Experimentation (Bolton & Harris: Econometrica'99, Keller, Rady & Cripps: Econometrica'05)

Outline

- ✓ Basic model & motivation
- Main result & key ideas
- Other results
- Discussion and open questions

How to measure performance?

For the first t rounds:

μ_a expected reward of arm a
after the prior is realized

- Expected total reward of the algorithm $W(t)$
- **Ex-post regret** $R_{\text{ex}}(t) = t \cdot (\max \mu_a) - W(t)$
- **Bayesian regret** $R(t) = \mathbb{E}_{\text{prior}}[R_{\text{ex}}(t)]$

Can **BIC** bandit algorithms attain optimal regret?

Main result: black-box reduction

Given arbitrary bandit algorithm \mathcal{A} ,
produce **BIC bandit algorithm** \mathcal{A}' with similar performance:

- Bayesian regret increases only by constant factor $C_{\mathcal{P}}$
(which depends only on the prior \mathcal{P}).
- Learning rate decreases by factor $C_{\mathcal{P}}$: e.g., **predicted best arm**

Suppose \mathcal{A} outputs a *prediction* ϕ_t in each round t .

Then \mathcal{A}' outputs a prediction ϕ'_t distributed as $\phi_{\lfloor t/C_{\mathcal{P}} \rfloor}$.

Modular design: use existing \mathcal{A} , inject BIC

can incorporate auxiliary info (e.g., prior);
exploration preferences (e.g., arms to favor)

predict beyond
the *best arm*
(e.g., *worst arm*)

Two arms: $\mathbb{E}_{\text{prior}}[\mu_1 > \mu_2]$

How to sample the other arm?

Hide exploration in a large pool of exploitation

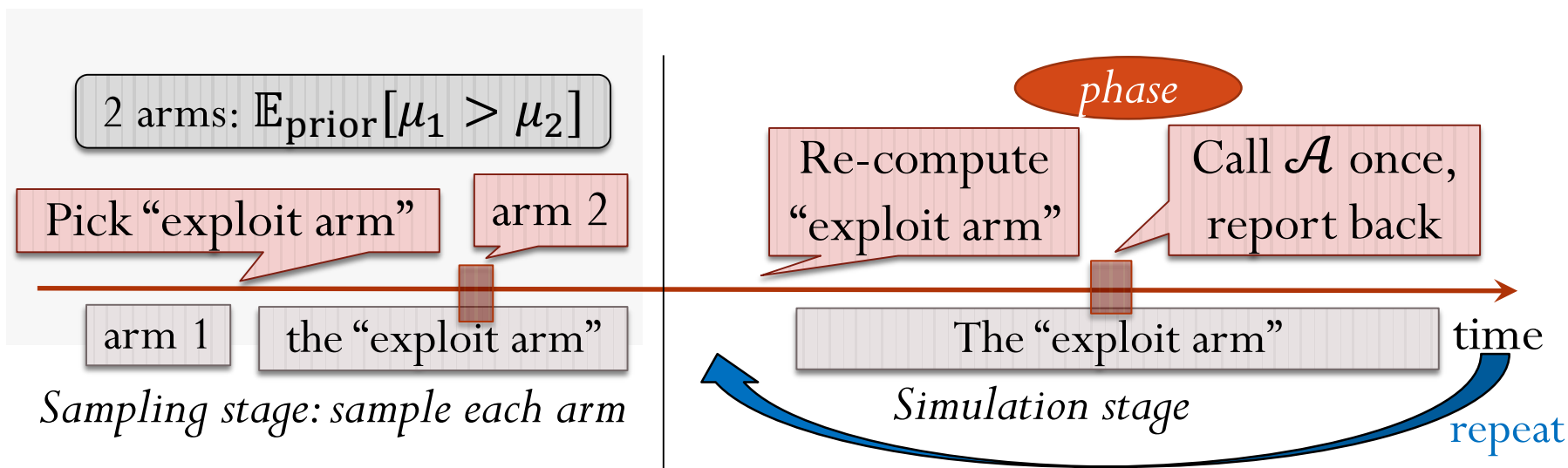


Enough samples of arm 1 \Rightarrow arm 2 could be the exploit arm

Agent recommended arm 2 *for exploration* does not know it!

Exploration prob. low enough \Rightarrow follow recommendation.

Black-box reduction from algorithm \mathcal{A}



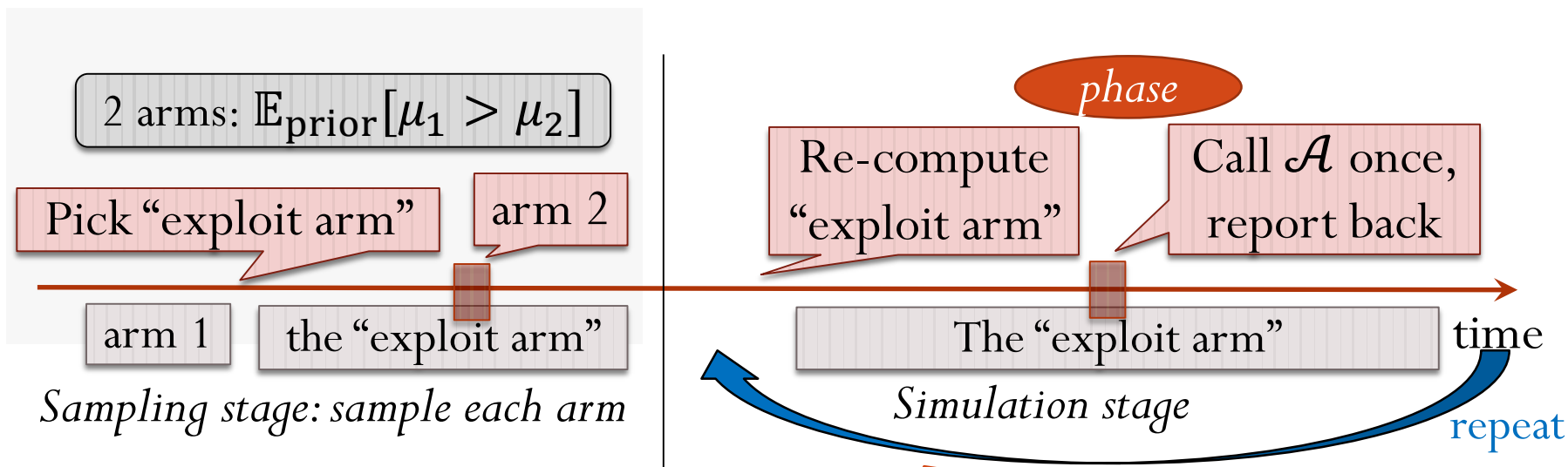
Enough initial samples \Rightarrow any arm could be the exploit arm!

Agent does not know: exploitation or algorithm \mathcal{A} ?

"Algorithm" prob. low enough \Rightarrow follow recommendation.

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

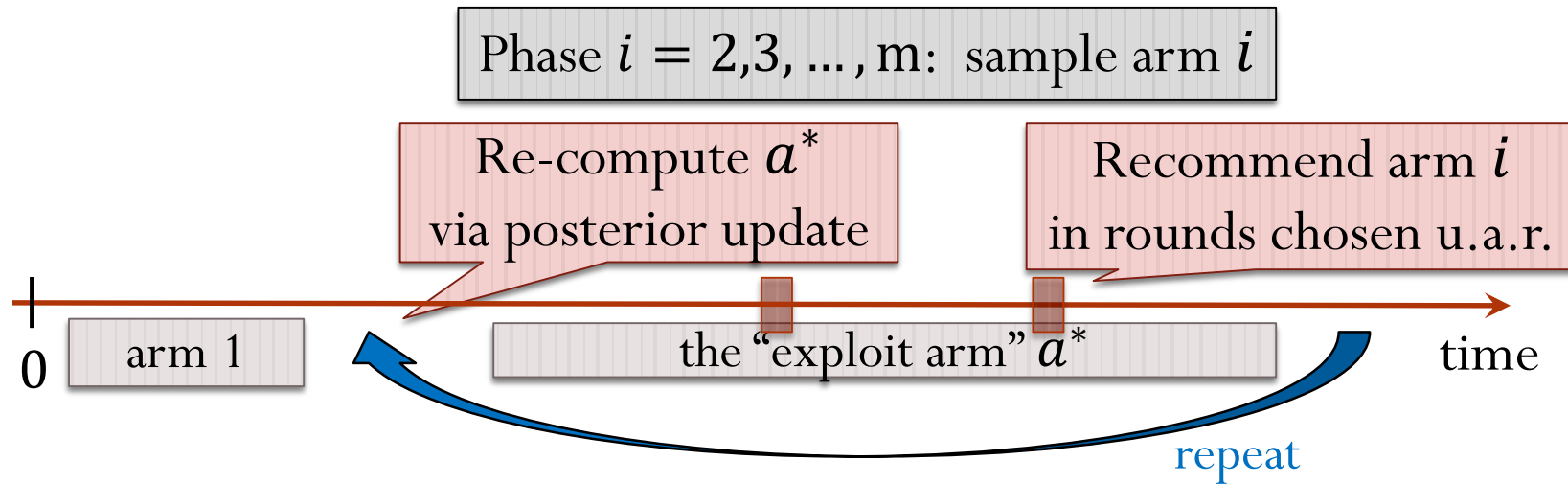
Black-box reduction from algorithm \mathcal{A}



If algorithm \mathcal{A} outputs a **prediction** ϕ_t in each round the new algorithm outputs the same prediction in all of next phase. Prediction in round t is distributed as $\phi_{\lfloor t/L \rfloor}$, $L = \text{phase length}$.

$$\mathbb{E}_{\text{prior}}[\mu_1 > \dots > \mu_m]$$

Sampling stage for many arms



Need to make sure that arm i could be the exploit arm!

sample each arms $j < i$ enough times

Exploration prob. low enough \Rightarrow follow recommendation.

Assumptions on the prior

- Hopeless for some priors

2 arms: $\mathbb{E}_{\text{prior}}[\mu_1 > \mu_2]$

e.g., if μ_1 and $\mu_1 - \mu_2$ are independent.

- Assumption for two arms: for k large enough,

$$\mathbb{P}(\mathbb{E}[\mu_2 - \mu_1 \mid k \text{ samples of arm 1}] > 0) > 0.$$

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

- Necessary for BIC algorithms (to sample arm 2).

Sufficient for black-box reduction!

- Similar condition for black-box reduction with > 2 arms

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

Outline

- ✓ Basic model & motivation
- ✓ Main result & key ideas
- Other results
- Discussion and open questions

Optimal “ex-post regret”: for each realization of the prior

BIC algorithm with optimal ex-post regret for constant #arms:

$$R_{\text{ex}}(T) = O\left(\min\left(\frac{\log T}{\Delta}, \sqrt{T \log T}\right)\right) + c_{\mathcal{P}} \log T$$

gap between best & 2nd-best arm.
Optimal for given Δ .

optimal in the
worst case

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

Our algorithm is *detail-free*: requires little info about the prior

- $N > N_0$, where N_0 is a constant that depends on the prior
- $\hat{\mu}$: approx. *min prior mean reward*

$$\mu_{\min} = \min_{\text{arms } i} \mathbb{E}_{\text{prior}}[\mu_i]$$

Agents can have different beliefs, if they believe that:

Black-box reduction with contexts

Our black-box reduction “works” in a very general setting

For each round t , algorithm **observes context x_t** , then:

- recommends an arm, and (possibly) makes a prediction
- agent chooses an arm, reports her reward & **extra feedback**

Distribution of reward & **feedback** depend on arm & **context**

e.g., customer profile @Amazon

e.g., detailed restaurant reviews

- allows (limited) agent heterogeneity
- incorporates three major lines of work on *bandits*:
with contexts, with extra feedback, and with predictions

BIC bandit games

In each round, a **fresh batch of agents plays a game** (possibly noisy payoffs, same game in every round)

- algorithm recommends an action to each agent
E.g., driving directions on Waze
- ... chooses a **distribution over action profiles**
- solution concept: **Bayesian correlated equilibrium (BCE)**

Which action profiles are “explorable” by a BIC algorithm?

How to explore all of them?

What is the best **BCE** achievable with all explorable info?

How to converge on this **BCE**?

Outline

- ✓ Basic model & motivation
- ✓ Main result & key ideas
- ✓ Other results
- Discussion and open questions

Auxiliary signals

For each agent, algorithm recommends an arm & **sends aux. signal**

- If algorithm can control whether to send the aux. signal
 - *not sending* is w.l.o.g. if the prior is fully observed & used
 - aux. signal may help for detail-free algorithms
 - cleaner without aux. signals (and we don't use them)
- If algorithm *is required* to send some aux. signals
 - complicated – e.g., revealing full stats does not work!
 - may help to reveal more info than required
 - what *must* and *can* be revealed may depend on application



Connection to medical trials

- Basic design: new drug vs. placebo (blind, randomized)
 - “advanced” designs studied & used (adaptive, >2 arms, contexts)
 - medical trials is one of original motivations for bandits
- Patients’ incentives: why participate & take less known drug?
Major obstacle, esp. for wide-spread diseases & cheap drugs.
 - Medical trial as a BIC recommendation algorithm
 - OK not to give the patients any data from the trial
 - extension to contexts and extra feedback very appropriate!

How to *really* convince the patients / model their incentives?

Connection to Systems

- System with many settings/parameters (hidden or exposed) your laptop, smartphone, or facebook feed
- Optimal settings unclear => need for *exploration*
 - often: settings are hidden, exploration done covertly
- Alternative: *expose the settings, let users decide*
 - *explore via incentive-compatible recommendations* (e.g., the defaults that users can override)

Open questions

Optimal dependence on the prior?

Better dependence on #actions?

Action spaces with known structure?

Use exploration that happens anyway?

ML

Fully detail-free algorithms?

Elicit some info from agents?
(ensure they do not lie)

BIC bandit game with
succinct game representation:
better regret, running time?