Incentivizing and Coordinating Exploration

Part I: Alex Slivkins (Microsoft Research NYC)
Part II: Robert Kleinberg (Cornell)

Tutorial at ACM EC 2017

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  (medical trials)

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.

Exploration and incentives

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

1. Under-exploration: some actions remain unexplored, or get explored at a less-than-optimal rate

   Ex: best action may remain unexplored if it seems worse initially

2. Selection bias: both chosen action and observed outcome may depend on agent properties => not typical population

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

   a) rarely see some sub-population => learn slowly, at best
   b) data is unreliable at face value

Motivation: exploration in markets

- Markets under uncertainty

  large scale acquisitions, e.g.: start-ups, real-estate, art

  how much is this worth? how much would others bid?

  matching markets, e.g.: college admissions, job markets, …

  do I want this job? do I stand a chance?
  how good is this candidate? Are we likely to get him/her?

- Costly exploration: money and/or opportunity cost

  E.g.: hire a building inspector, interview a candidate

Misaligned incentives: one agent's info may be useful to others, but he lacks incentives to explore and/or reveal the info

Our scope: incentivizing exploration

- Agents choose among information-revealing actions:
  one agent's action may reveal info that is useful to others
- Principal wishes to incentivize/coordinate exploration:
  interacts with agents, but cannot force them;
  sends signals (e.g., recommendations) and/or pays money
- Principal and/or agents can *learn* over time

Recent work in CS, economics and operations research

Assumes: principal has the ***power to commit*** to a particular algorithm
(so that agents believe he is actually using this algorithm)

Distinctions inside our scope

- Who learns, the principal or the agents?
  Are monetary transfers allowed?

  Part I: recommender systems: principal learns, no payments
  Part II: recommender systems: principal learns, payments allowed
  exploration in markets: agents learn, but not the principal

- Bayesian or frequentist ML

  Part I: regret-minimization, no time-discounting
  Part II: Bayesian time-discounted rewards

- can agents observe other agents' actions/outcomes?
  does one agent's reward depend on other agents' actions?
  reward distribution, bounded/light-tailed vs heavy-tailed?

Just outside our scope

Other work on "exploration and incentives"

- Decentralized exploration without a principal
  ex: Bolton & Harris`99, Keller, Rady, Cripps`05
- Info-revealing actions are not controlled by agents

  dynamic auctions (ex: Athey & Segal`13, Bergemann & Valimaki`10)
  ad auctions with unknown CTRs (ex: Babaioff, Kleinberg, Slivkins`10)
  incentivize good reviews (ex: Ghosh and Hummel `13)
- Dynamic pricing: aggregated info is not new to agents
  Ex: Kleinberg & Leighton`03, Besbes & Zeevi'09, Wang, Deng, Ye`14,
  Badanidiyuru, Kleinberg, Slivkins`13

Related work -- bigger picture

- Our model w/o incentives: explore-exploit tradeoff

  Huge literature in ML, OR, Statistics, Economics … since 1933

- Single round of our model: designing policies for revealing info to agents (to incentivize them to act in a certain way)

  Bayesian Persuasion (Kamenica & Gentzkow`11)
  Information design (Bergemann & Morris`13)

- Growing literature on "ML meets Economics"

  ML methods in Econometrics
  Sample complexity in auction design
  Learning in repeated games
  Mechanisms to crowdsource labels for supervised ML
  ML models to predict human behavior in games

✓ Motivation & scope

Part I: Incentivizing exploration without payments

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
Dine in this restaurant
Vacation in this resort
Buy this product
Drive this route
See this 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

Our approach: use ***information asymmetry***
(algorithm knows more than each agent)
to create ***intrinsic incentives***

Recommendation systems

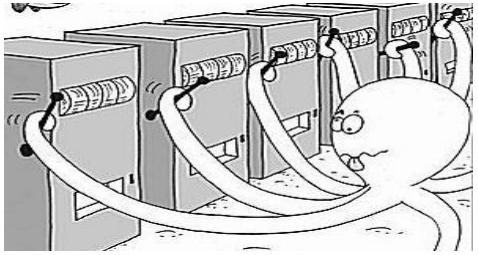
movie NETFLIX
restaurant yelp
this resort tripadvisor
oduct amazon.com
Drive this route waze
See this doctor suggestadoctor

Basic model

- K actions; T rounds
- In each round, a new agent arrives:
  - 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)

"actions" = "arms"

If agents follow recommendations ⇒ "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 vs. exploitation

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

- Revealing full history to all agents does not work
  (algorithm only exploits; ex: gets stuck on "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

✓ Motivation and scope

Part I: incentivizing exploration via information asymmetry

✓ basic model: BIC bandits

❑ results for BIC bandits

❑ algorithms and key ideas

❑ extensions

❑ discussion and open questions

How to measure performance?

μ_a expected reward of arm a after the prior is realized

For the first t rounds:

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

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

Results: optimal regret

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

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

For given $(\mu_1, \ldots, \mu_K)$: Δ is the *gap* between best and 2nd-best arm.
Optimal for given Δ.

optimal regret in the worst case

Depends on prior $\mathcal{P}$.
"Price" for BIC.

Conceptually: *exploration schedule is **adaptive** to previous observations*

Resolve BIC bandit exploration for constant #arms

Results: detail-free algorithm

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

Extra perks:

- Algorithm does not need to know N_0 and $\mu_{\min}$ exactly
- Agents can have different beliefs, if they believe that:

Results: 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 $c_{\mathcal{P}}$:
  Suppose $\mathcal{A}$ outputs a *prediction* ϕ_t in each round t.
  Then $\mathcal{A}'$ outputs a prediction ϕ'_t distributed as $\phi_{\lfloor t/c_{\mathcal{P}} \rfloor}$.

e.g., predicted best arm

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*)

- ✓ Motivation and scope

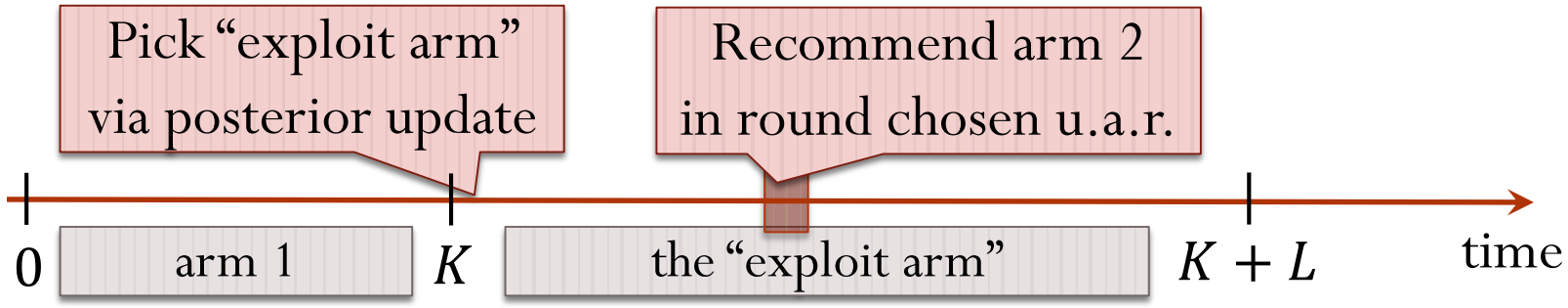
Part I: incentivizing exploration via information asymmetry

- ✓ basic model: BIC bandits
- ✓ results for BIC bandits
- ❑ algorithms and key ideas
- ❑ extensions
- ❑ discussion and open questions

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

How to sample the other arm?

Hide *exploration* in a large pool of *exploitation*

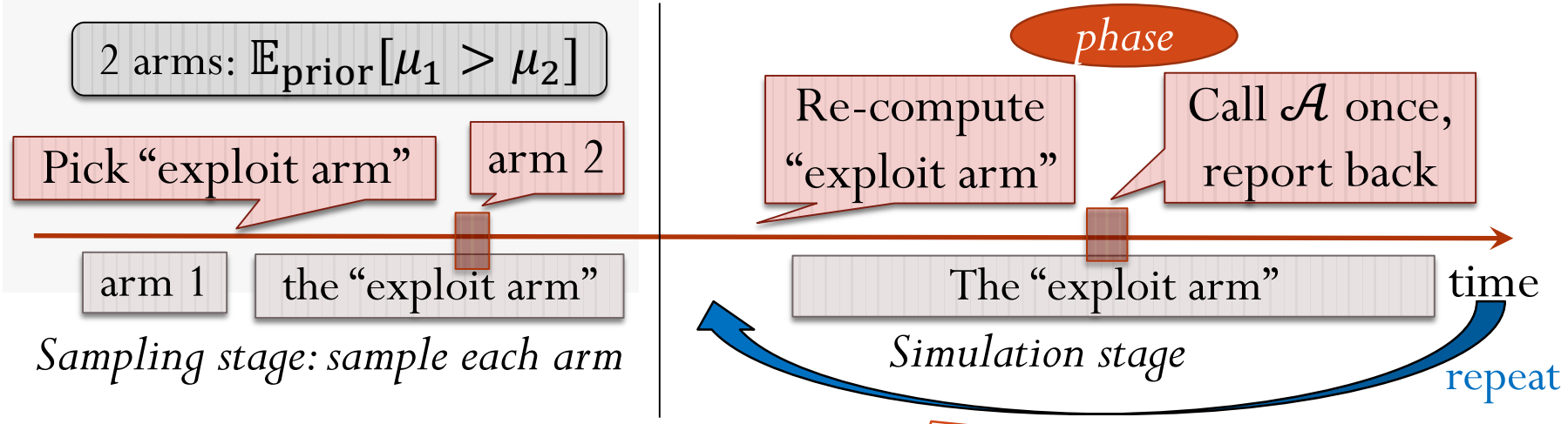


Enough samples of arm 1 ⇒ arm 2 could be the exploit arm!

Agent with rec=arm 2 for exploration does not know it!

Exploration prob. low enough ⇒ follow recommendation.

Black-box reduction from algorithm $\mathcal{A}$



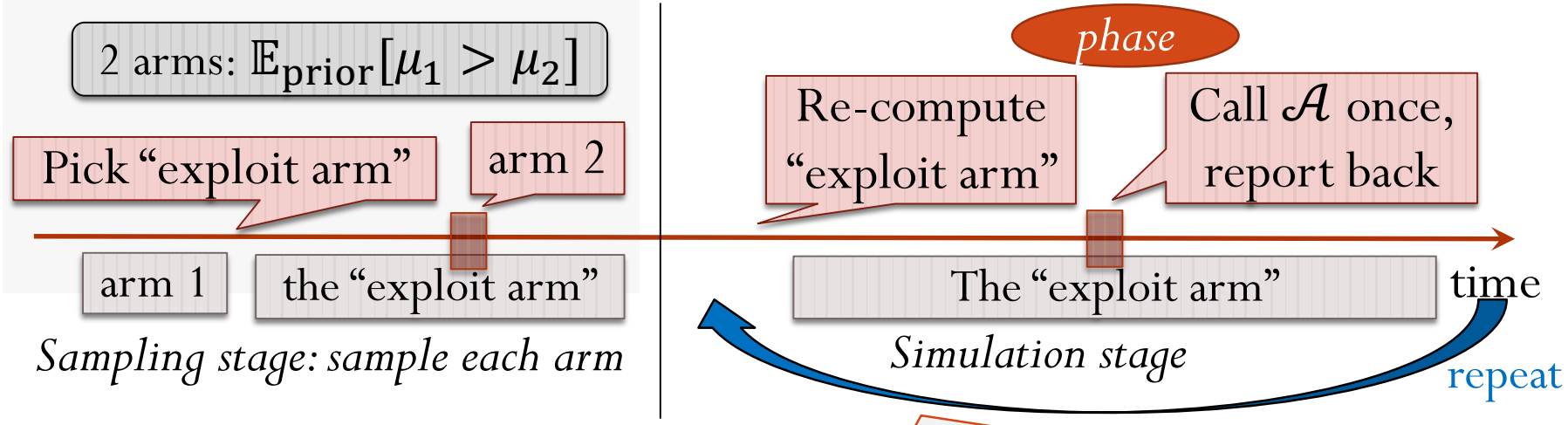
Enough initial samples ⇒ any arm could be the exploit arm!

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

"Algorithm" prob. low enough ⇒ follow recommendation.

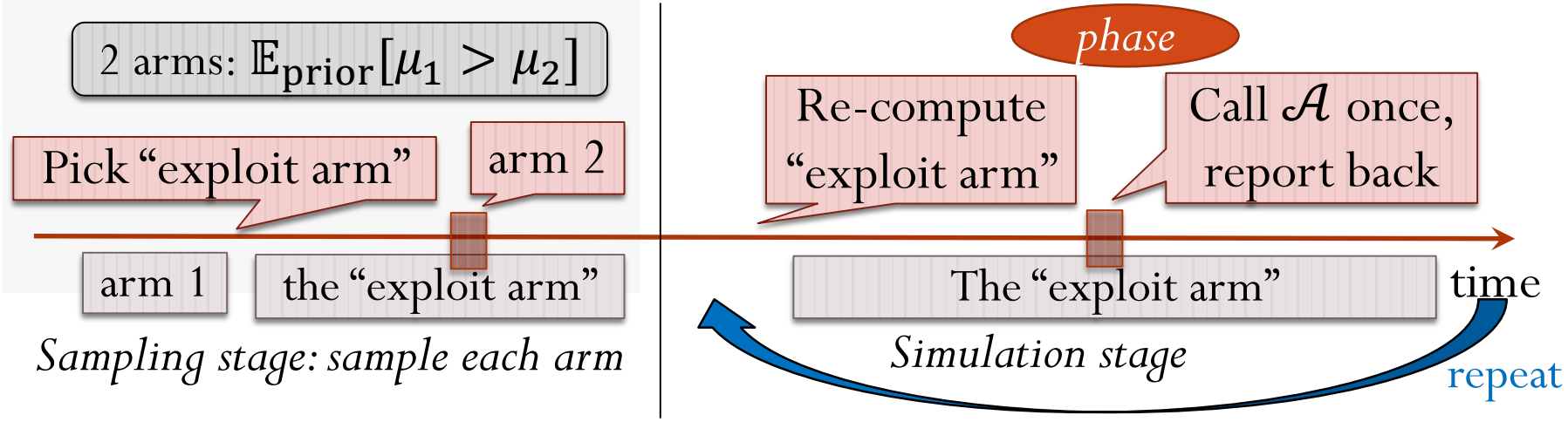
Performance: $\mathbb{E}_{\text{prior}}$[reward] of exploit arm ≥ that of $\mathcal{A}$

Black-box reduction from algorithm $\mathcal{A}$



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

Black-box reduction from algorithm $\mathcal{A}$



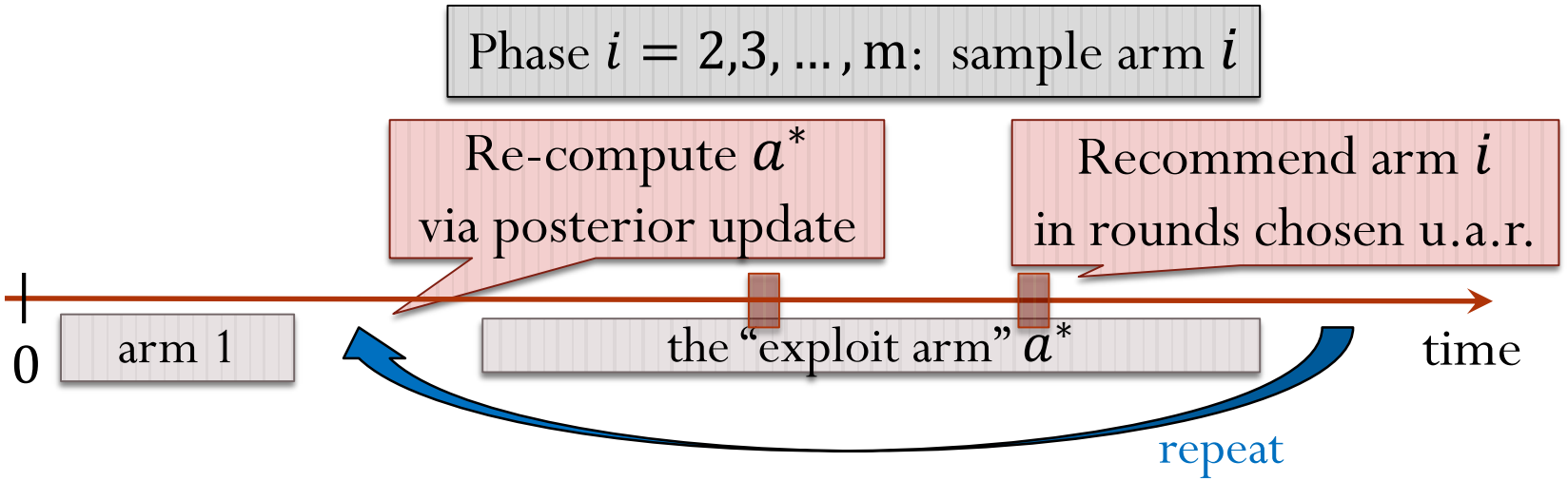
How low should explore prob. be to convince the agents?

Sufficient phase length should not grow over time!

Analysis of incentives should not depend on algorithm $\mathcal{A}$.

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

Sampling stage for many arms



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

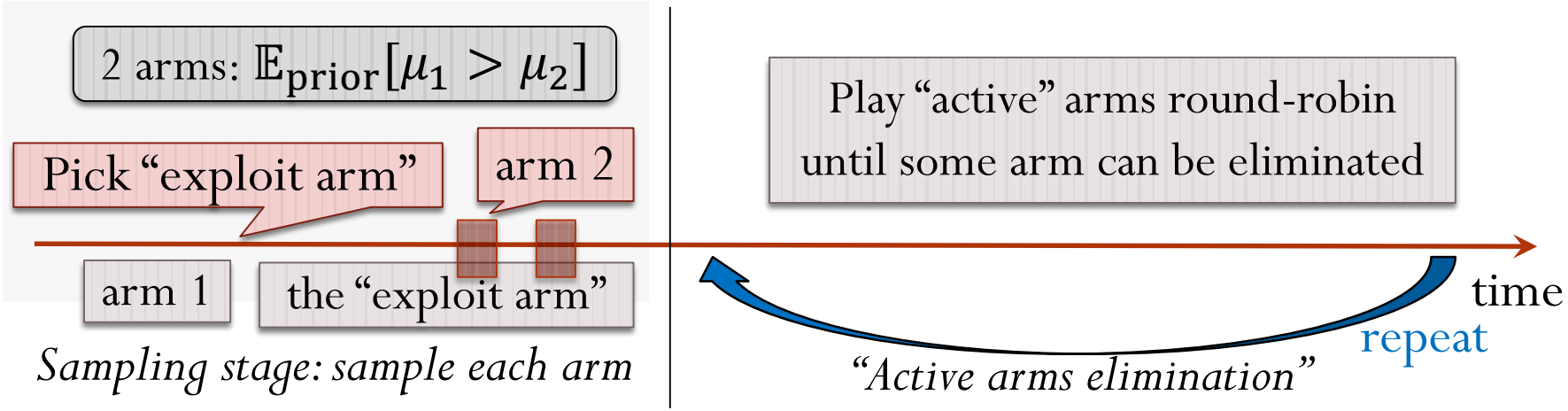
sample each arms $j < i$ enough times

Exploration prob. low enough ⇒ follow recommendation.

The detail-free algorithm

Detail-free ⇒ cannot use Bayesian update

Ex-post regret ⇒ best posterior arm may not suffice



Define "exploit arm" & "elimination condition" via sample averages.

For BIC, connect sample averages to Bayesian posteriors (tricky!).

Enough initial samples ⇒ "Active arms elimination" is BIC

Assumptions on the prior

- Hopeless for some priors
  e.g., if μ_1 and $\mu_1 - \mu_2$ are independent.

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

- 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]*

Suffices for the detail-free algorithm

- ✓ Motivation and scope

Part I: incentivizing exploration via information asymmetry

- ✓ basic model: BIC bandits
- ✓ results for BIC bandits
- ✓ algorithms and key ideas
- ❑ extension: auxiliary feedback
- ❑ extension: agents can affect one another
- ❑ discussion and open questions

Extension: auxiliary feedback

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

Combinatorial semi-bandits: arms $S \subset U$, observe reward for each $e \in S$.
Feedback graphs: observe rewards for chosen arm *and* all adjacent arms

Setup & result

Contextual Bayesian regret

$$R_{\Pi}(t) = \mathbb{E}_{\text{prior}}[W(t;\pi^*) - W(t;\mathcal{A})]$$

total reward

Policy π: {contexts} → {arms}

Fixed set of policies Π

π^*: best policy in Π

Bayesian incentive-compatibility (BIC):

$$\mathbb{E}_{\text{prior}}\left[\mu_{x,a} - \mu_{x,b} \,\middle|\, x_t = x, \text{rec}_t = a\right] \geq 0$$

$\forall$time t, context x, arms a, b

Arms a, contexts x.

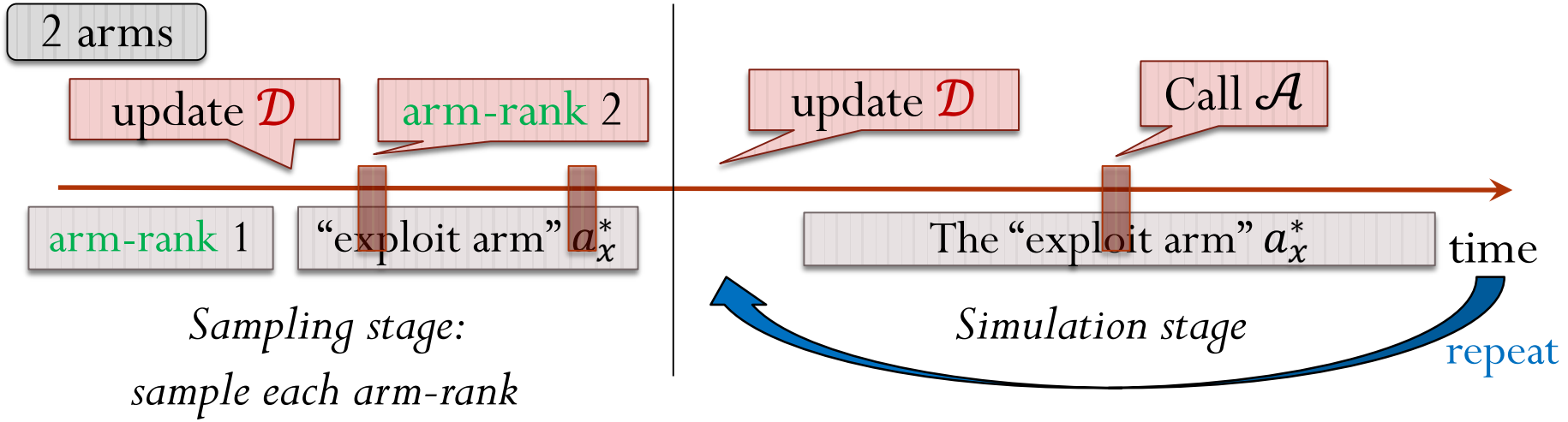
Expected reward $\mu_{x,a} \in [0,1]$.

Reduction: bandit algorithm $\mathcal{A}$ ⇝ BIC bandit algorithm $\mathcal{A}'$ with similar Bayesian regret & prediction quality

Unlike algorithms, our reduction does not depend on:
policy set Π, what is extra feedback, or what is predicted

Algorithm

- Defn: *arm-rank* i is a policy which maps each context x to i-th best arm given x, according to the prior.
- Key idea: *recommend arm-ranks instead of arms*.
- Dataset $\mathcal{D}$ of samples: (context, arm, reward, feedback).
  Exploit arm a_x^*: best posterior arm for context x given $\mathcal{D}$



- ✓ Motivation and scope

Part I: incentivizing exploration via information asymmetry

- ✓ basic model: BIC bandits
- ✓ results for BIC bandits
- ✓ algorithms and key ideas
- ✓ extension: auxiliary feedback
- ❑ extension: agents can affect one another
- ❑ discussion and open questions

Extension: agents affect one another

Agents affect each other's utilities (even without the principal)

- Drivers choose routes, congestion affects all
  Principal recommend routes
- Event ticket resellers choose prices in a shared market
  Principal recommends prices
- People choose an experience to share with others
  Principal coordinates to make it happen

Principal learns over time, needs to incentivize exploration

New aspect: agents play a game against one another

BIC bandit game

Principal's utility not restricted to welfare

In each round, a fresh batch of agents plays a game

- same game in every round, possibly with noisy payoffs
- algorithm recommends an action to each agent; observes utilities of all agents and the principal
- reward matrix is unknown, but there is a common prior

Single-round solution concept: *Bayes correlated equilibrium (BCE)*

- policy: observations → distribution over action profiles
- given policy and prior over observations, each agent prefers to follow (realized) recommended action

Single round: Bayesian Persuasion game (Kamenica & Gentzkow`11)

BIC algorithm: BCE in each round

What would be a natural benchmark?

BIC bandits $\Rightarrow$ best fixed action

single round $\Rightarrow$ best BIC policy *given prior over past observations*

OPT: best single-round BIC policy given all "learnable" info

Explorable action profile: can be chosen by some BIC algorithm
(then utilities for this action profile can, in principle, be learned)

Subtlety: action profiles may be explorable, but not immediately
… for some (but not all) realizations of the prior

utility vectors of all explorable action profiles

Which action profiles are "explorable" & how to explore them?
What is OPT and how to converge on it?

Results & techniques

BIC algorithm explores all explorable action profiles, matches OPT

deterministic utilities ⇒ constant regret w.r.t. OPT

IID utilities ⇒ O(log T) regret w.r.t. OPT

some small print

optimal up to prior-dependent constants

polynomial-time under generic input (prior as a big table)

Monotonicity-in-information for single-round game

if principal has more "relevant" info, things can only get better

- what utility can be obtained via a BIC policy ⇒ can't beat OPT
  which action profiles are immediately explorable

Subtlety: more "irrelevant" info does not help.

✓ Motivation and scope

Part I: incentivizing exploration via information asymmetry

✓ basic model: BIC bandits

✓ results for BIC bandits

✓ algorithms and key ideas

✓ extension: auxiliary feedback

✓ extension: agents can affect one another

❑ discussion and open questions

Auxiliary signals

Reviews, scores, …

Algorithm could send aux signals along with the recommendation

Is algorithm *required* to send some aux signals?

no

not sending any is w.l.o.g. if principal knows the prior
… and it is cleaner that way (and this is what we do)
however, aux signals may help for detail-free algorithms

yes

may hurt exploration, e.g., revealing full stats does not work!
may help to reveal more than required
what *must* and *can* be revealed may depend on application

??

Connection to Systems

- System with many settings/parameters (hidden or exposed)
  your laptop, smartphone, or facebook feed
- Optimal settings unclear => need for ***exploration***
  Settings are often hidden, exploration done covertly
- Alternative: expose the settings, let users decide

  explore via incentive-compatible recommendations
- The version without incentives is understood in theory,
  but (sort of) open in practice, need to really solve *that* first.

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
  - OK not to give the patients any data from the trial itself
  - extension to contexts and extra feedback very appropriate!

Open questions

theory→practice

Relaxed economic assumptions

Incorporate auxiliary signals

Agents with different, partially known beliefs
perhaps elicit some info from agents?

(Small) deviations from rationality

Long-lived agents

Improve ML & algorithms

Optimal dependence on the prior?

Better dependence on #actions?

(Large) action spaces with known structure?

Use exploration that happens anyway?

BIC bandit game with succinct representation?

Bring BIC exploration closer to theory of medical trials

Credits

- Original paper: Kremer, Mansour, Perry. Implementing the wisdom of the crowd. EC`14. J. of Political Economy, 2014.
  Deterministic rewards, two actions: optimal BIC mechanism.
  IID rewards, two actions: BIC mechanism with $T^{2/3}$ regret.
- This tutorial (part I):
  Mansour, Slivkins, Syrgkanis. Bayesian incentive-compatible bandit exploration. EC`15. Working paper (2017).
  BIC Bayesian Games: Mansour, Slivkins, Syrgkanis, Wu. Bayesian exploration: Incentivizing exploration in Bayesian games. EC'16. Working paper (2016).
- This tutorial (part II):
  Frazier, Kempe, Kleinberg, Kleinberg. Incentivizing exploration. EC'14 (Best Paper).
  Kleinberg, Waggoner, Weyl. Descending price optimally coordinates search. EC'16. Working paper (2016).

Other work on BIC exploration

Bimpikis et al. Crowdsourcing exploration. Management Science (2017).
Time-discounted rewards (2 actions, Bernoulli rewards).
Derives [slow] optimal BIC mechanism; fast heuristic based on same ideas.
Known reward for one action => BIC mechanism achieves first-best.

Che & Hörner. Optimal design for social learning. Working paper, 2013 –
Continuum of customers, continuous info flow (2 actions, 2 rewards).
Derives optimal BIC policy for a technically different model.

Bahar et al. Economic recommendation systems. EC`16.
Also observe friends' recommendations in a known social network.
(deterministic rewards, two actions, limited #high-degree nodes)

Recent working papers (2017)

Schmit & Riquelme, Human Interaction with Recommendation Systems: On Bias and Exploration.

"Free exploration" due to customer diversity suffices.

Each user knows her "idiosyncratic bias", reports "unbiased" feedback.

Algorithm reports estimated "common utility" for each action.

Mansour, Slivkins, Wu, Competing bandits: learning under competition.

Two exploration algorithms (e.g., search engines) compete for users.

Users give revenue and information: without users, you don't learn!

Kannan et al., Fairness Incentives for Myopic Agents (EC`17).

Incentivizing *fair* exploration via payments.

(ex: agents are lenders on a lending platform, actions are loan recipients)