Incentivizing and Coordinating Exploration

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

Tutorial at ITA/ALT 2020
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

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

- **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 => learn slowly, at best
- data is unreliable at face value
Motivation: 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 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: incentivized exploration

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

Recent work in CS, economics and operations research

Part I: incentives via signals/information
Part II: incentives via money
Zoom out: Exploration & incentives

- agents choose actions => our scope
- agents choose bids => 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: Mansour, Slivkins, Wu ‘18, Aridor, Slivkins, Wu ‘19)
Incentivize exploration without payments

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

“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?

*External* incentives:

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

"External" incentives:

- prone to selection bias;
- not always feasible

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

Recommendation systems:

Dine in this restaurant
Vacation in this resort
Buy this product
Drive this route
See this doctor
Basic model: BIC bandits

- T rounds, K actions (“arms”). In each round $t$:
  - new agent arrives, algorithm recommends an arm;
  - agent 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 ($= \text{cumulative reward}$)

Agents follow recommendations $\Rightarrow$ “multi-armed bandits”

*Bayesian Incentive-Compatible (BIC)* if

$$E[\text{reward}(a) - \text{reward}(b) | \text{rec}_t = a] \geq 0 \quad \forall t, \text{arms} \ a, b$$
How much info to reveal?

What if the algorithm could send any “message” to each agent

Reveal full history ⇒
“greedy algorithm”: “choose arm that looks best right now”
Very bad in the worst case, in many examples.
⇒ algorithm should reveal less than it knows

Revelation principle: arbitrary messages give no extra power.
WLOG, message = recommended arm.
Existing work: BIC bandits

Kremer, Mansour, Perry (2013)
Che & Horner (2013)

Mansour, Syrgkanis, Slivkins (2015)
Papanastasiou, Bimpikis, Savva (2015)
Mansour, Syrgkanis, Slivkins, Wu (2016)
Bahar, Smorodinsky, Tennenholtz (2016)
Schmit & Riquelme (2018)

Immorlica, Mao, Slivkins, Wu (2019)

Immovrlica, Mao, Slivkins, Wu (2018-2020)
Bahar, Smorodinsky, Tennenholtz (2019)
Cohen & Mansour (2019)
Sellke & Slivkins (2020)
“Zoom out”

Bandits

Social learning (Economics)

BIC bandits

Info Design (Economics)
✓ Motivation and scope
Part I: incentivizing exploration via information asymmetry
✓ basic model: BIC bandits
❑ Some fundamental results
❑ Further directions
❑ Relaxing rationality assumptions
❑ Open questions
Regret of BIC bandit algorithms

\[ \text{Regret}(T) = T \cdot (\max \mu_a) - \mathbb{E}[\text{REW}(T)] \]

Can BIC bandit algorithms attain optimal regret?

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

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

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

Gap between best & 2nd-best arm

Optimal regret for given Gap.

Constant \# arms

optimal regret in the worst case

Mansour, Slivkins, Syrgkanis (2015)
Black-box reduction from algorithm $\mathcal{A}$

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

- 2 arms: $\mathbb{E}_{\text{prior}}[\mu_1 > \mu_2]$
- Call $\mathcal{A}$ once, report back
- Re-compute “exploit arm”
- The “exploit arm”

Simulation stage

---

Enough initial samples $\Rightarrow$ any arm could be the exploit arm!
Agent does not know if it is 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 $\mathcal{A}$

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

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

During each phase:

- **Arm 1**: The "exploit arm"
- **Re-compute “exploit arm”**: Re-compute the exploitation
- **Call $\mathcal{A}$ once, report back**: Call the algorithm $\mathcal{A}$ once and report back.

**Simulation stage**

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 $\mathcal{A}$.  

17
Beyond Bayesian regret

- “Exploit arm” computed via Bayesian update
  only good in expectation over the prior \( \Rightarrow \) only Bayesian regret
- For regret bounds that hold for each realization of the prior, different algorithm,
  use sample averages rather than posterior mean rewards
- Extra perk: algorithm is “detail-free”
  no need to know the prior exactly, two parameters suffice
  - Different agents can have different beliefs,
    as long as they are “consistent” with these two parameters
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 $\mu_1$ and $\mu_1 - \mu_2$ are independent.

- Assumption for two arms: for $k$ large enough,
  \[ P(\mathbb{E}[\mu_2 - \mu_1 | k \text{ samples of arm 1}] > 0) > 0. \]

  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
- Some fundamental results

Further directions
- Relaxing rationality assumptions
- Open questions
Further directions

- Black-box reduction $\rightarrow$ contextual bandits & aux feedback
- Bayesian-optimal mechanisms (for special cases)
  Kremer, Mansour, Perry `13, Che & Horner `13
  Papanastasiou, Bimpikis, Savva `15, Cohen & Mansour `19
- Explore all “explorable” arms (some arms aren’t)
  (Mansour, Syrgkanis, S., Wu `16, Immorlica, Mao, S., Wu `19)
- Heterogenous agents (Schmit & Riquelme `18, Immorlica, Mao, S., Wu `19)
- Multiple agents playing a game (Mansour, Syrgkanis, S., Wu `16)
- Inevitable revelation: some history observed no matter what
  Bahar, Smorodinsky, Tennenholtz (2015, 2019)
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
✓ Motivation and scope

Part I: incentivizing exploration via information asymmetry
✓ basic model: BIC bandits
✓ Some fundamental results
✓ Further directions

❑ Relaxing rationality assumptions
❑ Open questions
[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 (2018-2020)
Unbiased histories

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

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/round 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.
✓ Motivation and scope
Part I: incentivizing exploration via information asymmetry
✓ basic model: BIC bandits
✓ Some fundamental results
✓ Further directions
✓ Relaxing rationality assumptions
☐ Open questions
Unbiased histories

- Same regret bounds with a simpler construction?
- How to make the construction more robust? (“anything like this works” instead of “it needs to be just so”)
- Do we have the “right” behavioral assumptions? And what can we possibly mean by “right assumptions”?
- Does this approach generalize? E.g., to contextual bandits / heterogenous agent types.
Back to basics

- Large $K = \#\text{arms}$: so far, regret $\exp(K)$ vs. $\sqrt{K}$ for bandits
- Optimal dependence on the prior? (In which parameters?)
- Do we really need specialized algorithms?

Partial answers: Sellke & Slivkins (2020)

- **Thompson Sampling is BIC** with enough initial data
  - Caveats: independent priors; TS needs exact prior; inst.-dep. regret unclear

- Reduces the problem to sample complexity:
  - How many samples suffice? How many rounds to collect them?

- BIC algorithm to collect samples: $\text{UB} \leq \text{LB}^{O(1)}$
  - Beta-Bernoulli priors: $C_{\text{prior}} \cdot \text{poly}(K)$ rounds suffices for TS
  - Caveat: computationally inefficient
More open questions

- [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?

All directions very open, despite substantial prior work on some
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