Incentivized Exploration

Alex Slivkins (Microsoft Research NYC)
https://www.microsoft.com/en-us/research/people/slivkins/

Collaborators: Xinyan Hu, Nicole Immorlica, Yishay Mansour, Jieming Mao, Daniel Ngo, Mark Sellke, Max Simchowitz, Vasilis Syrgkanis, Steven Wu

Survey: "Exploration & Persuasion" (2021) http://slivkins.com/work/ExplPers.pdf
Textbook chapter: Ch 11 in my bandits book (https://arxiv.org/abs/1904.07272)

Simons Workshop on "Quantifying Uncertainty", Sept 2022

Abstract: Incentivized Exploration

In a wide range of scenarios, individual decision-makers ("agents")
consume information revealed by the previous decisions,
and produce information that may help in the future decisions.
Each agent would individually prefer to "exploit" (optimize the current reward),
but would prefer the previous agents to "explore" (try out various alternatives).
A social planner, by means of carefully designed information disclosure,
can incentivize the agents to balance exploration and exploitation
in order to maximize social welfare.
We overview the current state of this problem space,
and highlight some recent developments.

Incentivized exploration

Incentivize self-interested agents
to *explore* when they prefer to *exploit*

Exploration & Incentives

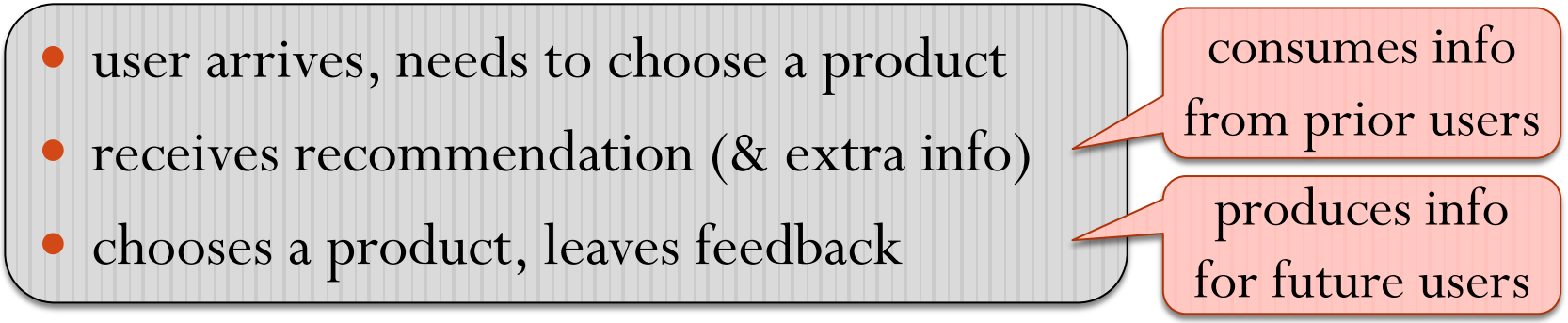
Bandits & agents:
bandits with Bayesian persuasion

agents choose actions (our model)
agents choose bids
(learning in repeated auctions)
agents only affect rewards
(dynamic pricing / assortment)
agents choose between algorithms
(platforms compete for users)

Agents: users in 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



For common good, users should balance explore & exploit

e.g., coordinate via system's recommendations.

Misaligned incentives: self-interested users (*agents*) prefer to exploit

- some actions may be explored at sub-optimal rate
- best action may remain unexplored if it seems worse initially

Our model

(Bayesian) GREEDY algorithm

default: full history

- T rounds, K actions ("arms").
  In each round t:
  new agent arrives, observes *something* (msg_t),
  chooses an arm, and reports her reward $\in [0,1]$
- IID rewards: reward of arm a drawn from some D_a with mean μ_a
  Distributions D_a fixed but unknown; common Bayesian prior
- Objective: social welfare (= cumulative reward)
- Agents' rational choice: $\text{argmax}_{\text{arms } a}\ \text{E}[\mu_a | \text{msg}_t]$

What goes wrong with GREEDY?

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

- Two arms, $G := E[\mu_1 - \mu_2] > 0$ — "prior gap"
- Round 1: arm 1 is chosen
- Deterministic rewards: μ_1 is observed — learning failure
  If $\mu_1 > E[\mu_2]$ then arm 2 is never chosen
- Randomized rewards: — learning failure
  Thm: Pr[arm 2 is never chosen] $\geq G$
  Cor: Bayesian Regret is (at least) linear in time — Sellke & Slivkins (2019)
  if the prior is independent across arms & each arm has positive density on [0,1]

How to incentivize?

How to incentivize agents to take actions that seem suboptimal (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

selection bias

not always feasible

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

Our model

- T rounds, K actions ("arms").
  In each round t:
  new agent arrives, observes *something* (msg_t),
  chooses an arm, and reports her reward $\in [0,1]$
- IID rewards: reward of arm a drawn from some D_a with mean μ_a
  Distributions D_a fixed but unknown; common Bayesian prior
- Objective: social welfare (= cumulative reward)
- Agents' rational choice: $\text{argmax}_{\text{arms } a}\ \text{E}[\mu_a | \text{msg}_t]$

chosen by algorithm

default: ~~full history~~

w.l.o.g. msg_t is a suggested arm, &
algorithm is *Bayesian Incentive-Compatible* (BIC):

$\text{E}[\mu_a - \mu_b | \text{alg}, \ \text{msg}_t = a] \geq 0 \quad \forall t, \text{arms } a, b$

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 & Simchowitz (2021)

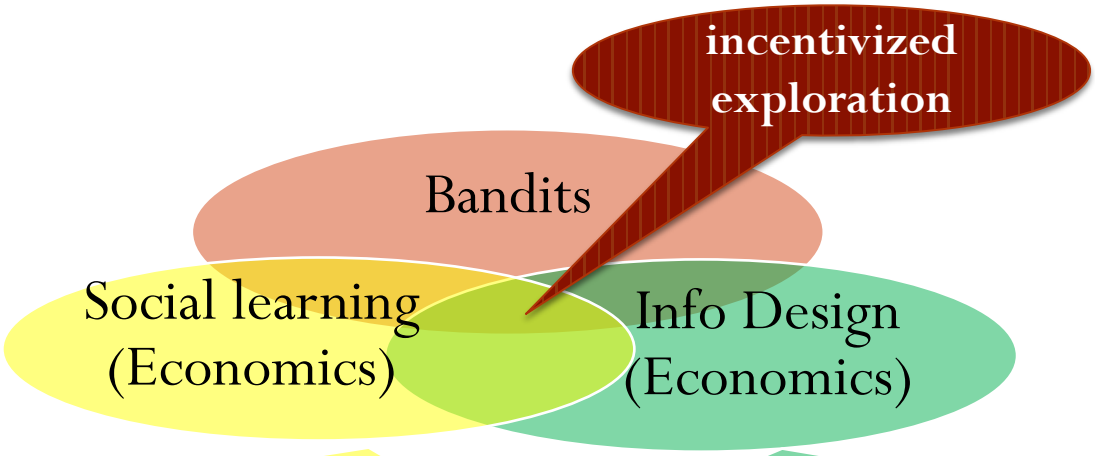
Li & Slivkins (2022)

Hu, Ngo, Slivkins & Wu (2022)

Home community:
economics & computation
(ACM EC)

Outline

- ✓ Intro
- ❑ Related work
- ❑ Basic stuff
- ❑ Advanced stuff
- ❑ Open questions



Agents learn in shared environment:
no principal, agents are on their own!

Most related:
"strategic experimentation"
long-lived agents observe each other, free-ride on exploration by others

Us: social learning with a mediator

one round: Bayesian Persuasion (BP)

- Sender (S) gets private signal, chooses a message for Receiver (R)
- Then, R chooses an action
- Focus: optimal 1-round policy for S

Us: sender= algorithm, signal= history

- Objective: explore & exploit
- Myopic: S's utility is $1_{\text{arm 2 is chosen}}$
- Long-term obj: can't model via BP

(More) related work

Greedy explores well assuming structure

see my survey for refs

- e.g., linearity of rewards & heterogeneity of agents
- structure substitutes for explicit exploration

Incentivized Exploration with money

Starting from Frazier et al. (2014)

- all agents observe full history
- incentives via money, not info asymmetry (us: vice versa)

Online Bayesian Persuasion

Castiglioni et al. (2020), Zu et al. (2021)

- algorithm's signal IID in each round (us: algorithm's history)
- w/o incentive constraints, problem is vacuous (us: bandits)

Outline

- ✓ Intro
- ✓ Related work
- ❑ Basic stuff
- ❑ Advanced stuff
- ❑ Open questions

2 arms: $E[\mu_1] > E[\mu_2]$

Basic conditions on the prior

- Hopeless in general: e.g., if μ_1 and $\mu_1 - \mu_2$ are independent
- Sufficient condition (as we prove):

  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*")

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

- 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]*

Basic Techniques

Hidden Exploration in lots of exploitation

- In each round, {*go wild*} w/prob p, else exploit
  $p > 0$ is a "constant": depends only on the prior
- general reduction: any algorithm -> BIC algorithm

Successive Elimination: iterate active arms, eliminate if you can

- modification: larger confidence bounds for incentives
- [pro] does not need to know the prior:
  approx. knowing two parameters suffices

- Thomson Sampling (with the actual prior)

Successive Elimination & Thompson Sampling require independent priors

Initialization for each algorithm: N samples from each arm
(N is prior-dependent constant), collected by a version of Hidden Exploration

Basic results

K arms, T rounds

c_P : prior-dependent constant

Optimal Bayesian regret $\tilde{O}(c_P \sqrt{T})$

Hidden Exploration

for constant #arms (but exponential in K)

Independent priors:

- optimal frequentist regret $\tilde{O}(c_P \min(\sqrt{T},\ 1/\text{Gap}))$

  Successive Elimination

  for constant #arms (but exponential in K)
- Optimal Bayesian regret $\tilde{O}(\sqrt{KT})$

  Thomson Sampling

  with $c_P \cdot K$ initial samples of each arm

Outline

- ✓ Intro
- ✓ Related work
- ✓ Basic stuff
- ❑ Advanced stuff
- ❑ Open questions

Advanced Questions

In the basic model

- dependence on #arms and the prior
- pure exploration: explore all explorable arms

ec21/OpRe r&r

Extend the ML model

- auxiliary feedback: e.g., contextual bandits

ec15/OpRe'20

- large, structured problems, *e.g.*, episodic RL

w.p. 2021; NeurIPS'22

Extend the Econ model

- heterogenous agents (public or private types)

WebConf'19

- multiple agents in each round

ec16/OpRe'22

- relax rationality assumptions

ec20

In the basic model

Price of Incentives

Loss in performance compared to bandits

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": $1/\min_{\mathcal{P}\in\mathcal{F}} \mathrm{Var}(\mathcal{P})$

Algorithm

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

Sellke & Slivkins (2021)

Extend the learning model

Incentivized Episodic RL

- In each episode: new agent arrives,
  observes a message (chosen by our algorithm),
  selfishly chooses a policy for the entire episode
- pure exploration: visit all reachable (stage, state, action) triples
- Issues: huge #policies, highly correlated rewards
- New technique: Hidden Hallucination
  - Essentially: "message" is algorithm's history
  - in each episode, show "true" history (exploit) w/prob $1 - p$,
    else, show "hallucinated" history which promotes exploration
    by making all explored stuff look bad

First paper combining RL & mechanism design

Simchowitz & Slivkins (2021)

[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"

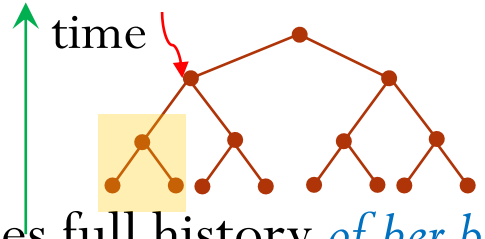
How to ensure predictable user behavior?

Immorlica, Mao, Slivkins, Wu (2020)

Selective disclosure

Users want full history.
Let's give them the next best thing

- Principal (only) chooses partial order (DAG) on rounds

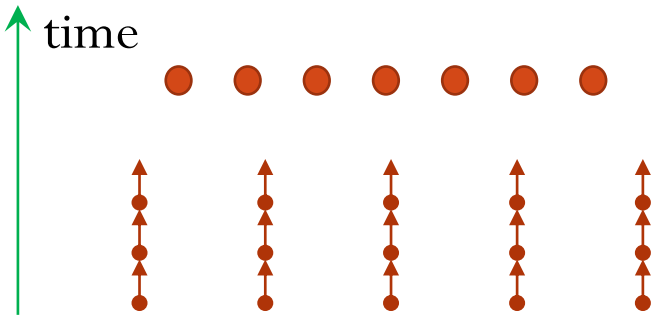


the relevant sub-algorithm

- Each user sees full history *of her branch*
  - *no cheating*: can't subsample all rounds that make arm 2 look good
  - *no need to second-guess* why this agent chose this arm
- Economics foundation: assumptions only on users that see full history
  - HardMax or SoftMax? anything consistent with confidence intervals
- Each agent is "locally greedy", and yet it works! (for some DAGs)
  - simple construction => implements explore-first, $T^{2/3}$ regret
  - tricky construction => implements adaptive exploration, $\sqrt{T}$ regret

Design the partial order

Each agent is "locally greedy", and yet it works!



Level 2: each agent sees all L1-agents

Level 1

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

Two "levels": implements non-adaptive exploration

Adaptive exploration

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

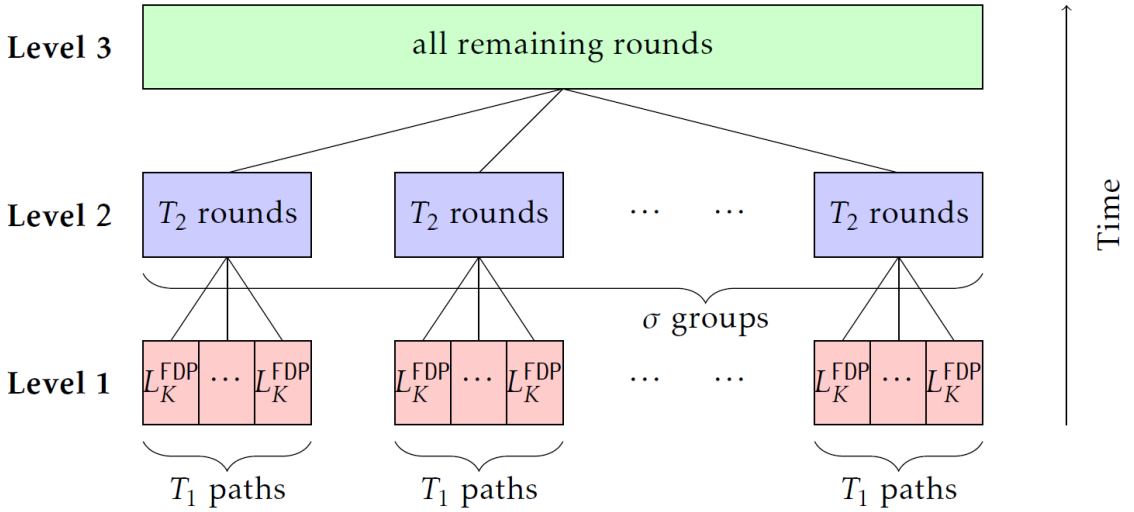


Figure 2: Info-graph for the three-level policy. Each red box in level 1 corresponds to T_1 full-disclosure paths of length L_K^{FDP} each.

Adaptive exploration

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

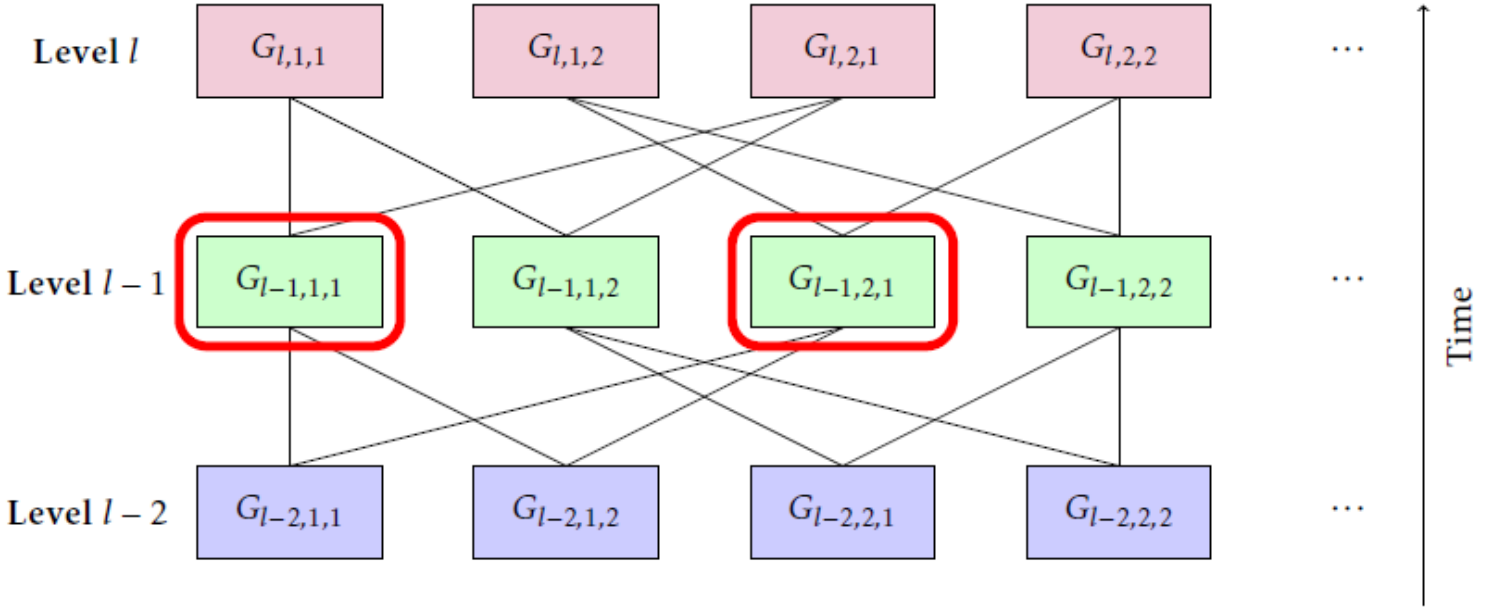


Figure 3: Interlacing connections between levels for the L-level policy.

Open questions

All directions very open, despite substantial prior work on some

- Relaxed economic assumptions: do we have the "right" ones?
  make the constructions simpler/ more general / more robust
- Partially known priors:
  what if the prior is not fully known initially?
- Long-lived agents:
  what if each agent is present for multiple rounds?
- Inevitable observations:
  what if some aspects of history are always observed by the agents
- Heterogenous agents: regret bounds?
  Use diversity to help BIC exploration?