Incentivizing Exploration and Compliance without Money

Vasilis Syrgkanis

Stanford University

Joint with: Yishay Mansour, Aleksandrs Slivkins, Steven Wu, Daniel Ngo, Logan Stapleton

Exploration vs exploitation in recommendation systems

Goal. Recommend option of high value to user

Observation. Information about options comes from prior user experiences

- Users are both producers and consumers of information

For overall welfare optimization: balance exploration vs exploitation

- Explore many options to gather information about alternatives
- Exploit the current information by recommending the seemingly best option

Motivating applications: Waze - user based navigation

- Real time navigation recommendations
- Based on user inputs
  - Cellular/GPS
- Recommendation dilemma:
  - Need to try alternate routes to estimate time
  - Actually, done in practice

Motivating applications: User based recommendation systems

- Recommendation web sites
- Example: TripAdvisor
- User based reviews
- Popularity Index
  - Proprietary algo.
  - Self-reinforcement
- Can be used to induce exploration

Exploration problem

- Prior bias of users leads to lack of exploration
- Can miss good options that a priori seem inferior
- System needs to incentivize exploration
- This talk: incentivizing exploration through information asymmetry

Modelling Goals

- Repeated interaction between a planner and multiple agents
- Each agent picks one among a set of available options
  - Routes in a network, hotels, restaurants
- Agents arrive, pick an action and report feedback to planner
- Agents are strategic: maximize reward conditional on information
- Planner wants to learn best alternative and maximize overall welfare of agents

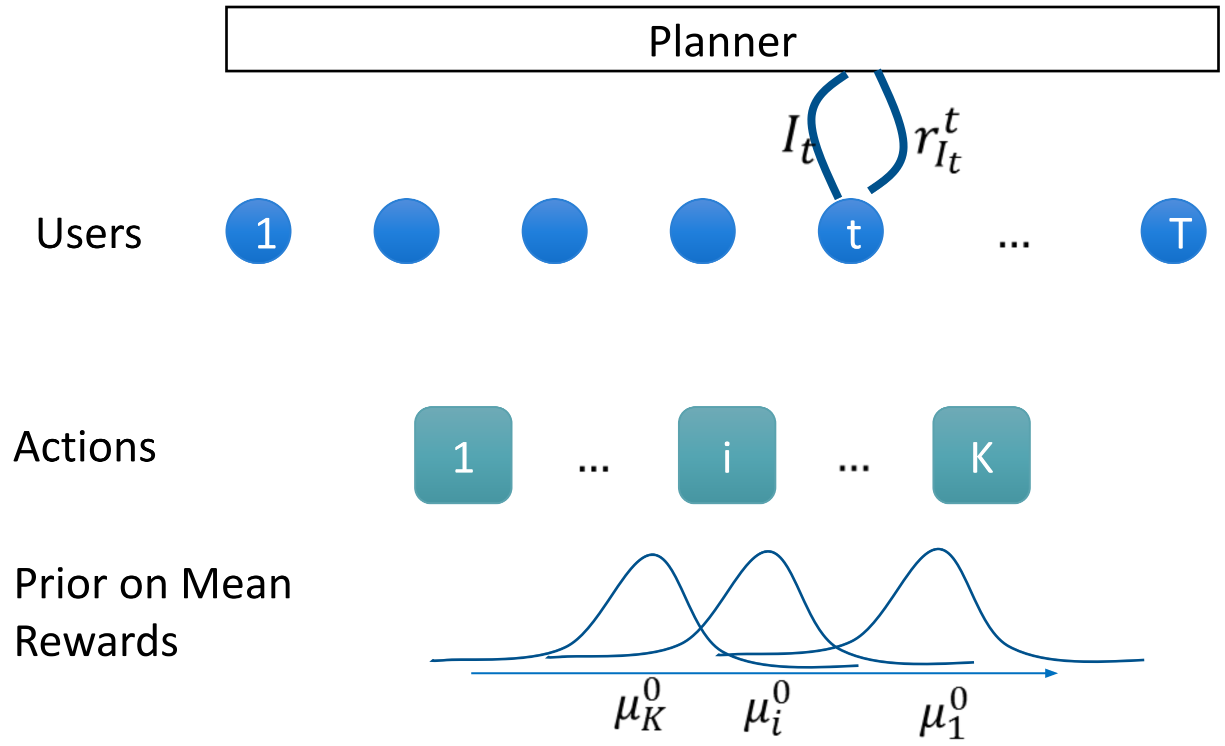
Research Questions

- Planner limitations
  - No monetary transfers
  - Controls information flow between agents
- Can the planner induce exploration?
  - Learn best alternative
- What is the rate of learning?
  - Impact of agent incentives on learning rate
- Extensions (briefly mention)
  - Multiple agents arrive at a time with interconnected payoffs (game)
  - Planner has arbitrary objective function
  - Observed and unobserved heterogeneity across agents

Main model

Bayesian incentive compatible bandit exploration

Bayesian Incentive-Compatible Bandit Model



- T users arrive sequentially
- Each can take one of K actions
- Each action has a mean reward of $\mu_i \in [0,1]$
- Common prior belief on each μ_i*
- Realized reward $r_i \in [0,1]$: stochastic i.i.d. with mean μ_i
- At each time-step planner recommends an action I_t
- Users report realized reward

*We will impose some assumptions on the priors

Planner's performance measure

- Asymptotic ex-post regret (think of $T \to \infty$)

$$\text{Regret}(\mu_1, \dots, \mu_K) = \underbrace{T \cdot \max_i \mu_i}_{\text{Welfare of always best action}} - \underbrace{\sum_{t=1}^{T} E[\mu_{I_t}]}_{\text{Expected welfare of recommendation algorithm}}$$

Weaker performance measure of Bayesian Regret

$$\text{Bayesian} - \text{Regret} = E_{\mu_1, \dots, \mu_K \sim \text{Prior}}[\text{Regret}(\mu_1, \dots, \mu_K)]$$

Remark. Regret vs Bayesian optimal policy

- Best fixed action benchmark is upper bound to Bayesian optimal
- Vanishing regret algorithm achieves average welfare close to Bayesian optimal as $T \to \infty$
- Interpreted as large market optimality
- Ex-post regret is prior-free (i.e. robustness to inaccuracies on prior)

- So far equivalent to Stochastic i.i.d. Multi-armed Bandit Model
  - Well studied in Econ, OR, CS, since 1933
  - Thompson sampling, Gittins index, [Lai-Robbins'85], UCB [Auer et al'92]

$\sqrt{T}$ regret achievable

Agents are strategic

Incentive Compatibility (IC). Playing recommended action has expected utility as high as any other action

$$\forall i:\ E[\mu_i | I_t = i] \geq E\left[\mu_{i'}\ \middle| I_t = i\right]$$

e.g. first user can only take action 1

If users observe everything will only take the posterior better action given previous rewards – cannot guarantee exploration

How to incentivize: Information Asymmetry

Users do not observe rewards or recommendations of previous users

Unaware whether rewards of previous steps have made a priori better arms worse than a priori worse arms

Information flow from prior users is at the hand of planner

Information is revealed only through recommended action and knowledge of planner policy

Main question

- Is $\sqrt{T}$ regret achievable under the incentive compatibility constraint?

Preview of main results: Bayesian Regret

- Black-box reduction: any bandit algorithm to an incentive compatible one (prior-dependent constant blow up in Bayesian regret)
- Implies $O(\sqrt{T})$ Bayesian regret IC algorithms
- T steps of any algorithm can be simulated in an incentive compatible manner in $c\,T$ time steps
  - Average expected reward as high as that of the algorithm

Enables modular design of IC recommendation systems

Preview of main results: Ex-post Regret

- $O(\sqrt{T})$ ex-post regret
- $O(\log(T))$ for instances with large "gap" in the means
  - Difference of best arm and suboptimal arms lower bounded by a constant

Detail-free algorithm (doesn't need to know full prior, but only an upper bound on a single parameter of the prior)

Preview of main results: extensions

- Observed agent heterogeneity
  - Recommendation takes observed features into account
  - Compete with best policy from target class of policies that map features to actions
- Unobserved heterogeneity with confounding
  - Recommendation can be viewed as "instruments"
  - Non-incentive compatible method that uses "compliers" and IV regression
- Multiple agents arrive simultaneously
  - Payoffs depend on all players actions (e.g. routing game)
  - Policy sends private signals to each player
  - Policy is a mapping from information to distribution over action profiles
  - Incentive compatibility ⇔ Bayes correlated equilibrium [Bergemann-Morris]
  - Which actions are explorable?
  - Computationally efficient policy which performs at least as good as Bayesian optimal policy after a few number of rounds

Some related work

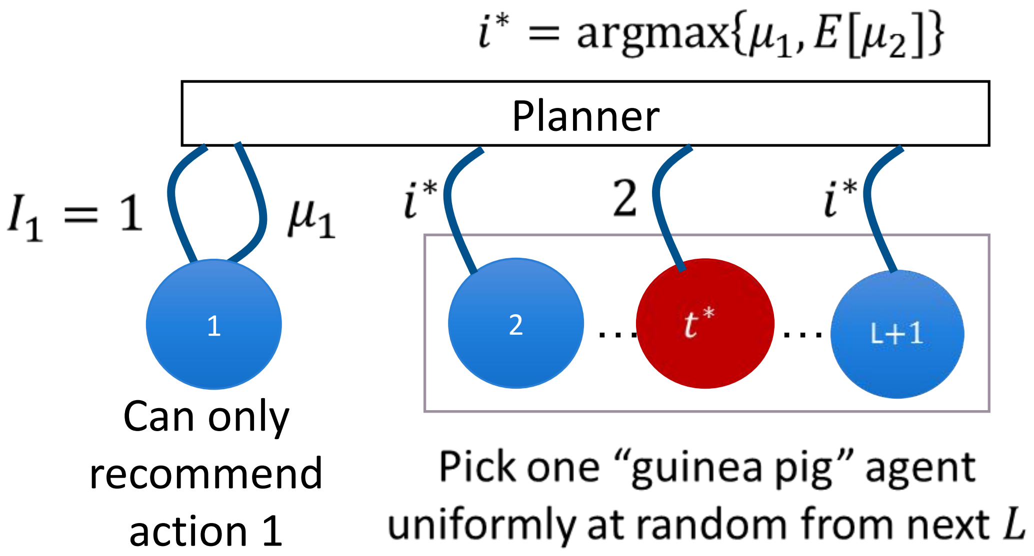
- Kremer, Mansour, Perry [2014]: Same model, two arms, primarily Bayesian optimal for non-stochastic rewards, $T^{2/3}$ for stochastic
- Che and Horner [2013]: continuous time stochastic model, two arms, binary reward, Bayesian optimal
- Papanastasiou, Bimpikis, Savva [2015]: discounted reward, heuristic for Bayesian optimal
- Frazier et al. [2014]: Monetary transfers allowed, users observe past actions, payments vs. information asymmetry

- Bayesian Persuasion: Kamenica, Gentzkow [2011]
- Herding and Information Cascades: Bikhchandani-Hirshleifer-Welch [1992], Banerjee [1992]

Main Ideas

Two actions, deterministic rewards

- Action 1: $\mu_1 \sim U[1/3,1]$, $\mu_1^0 = E[\mu_1] = 2/3$
- Action 2: $\mu_2 \sim U[0,1]$, $\mu_2^0 = E[\mu_2] = 1/2$
- Without planner everyone picks arm 1
- How to incentivize players to play action 2?
- Assume deterministic rewards: $r_i = \mu_i$
- Hide exploration in a pool of exploitation



Why should a player t follow recommendation:

$$\underbrace{E[\mu_1 - \mu_2 | I_t = 2] \Pr[I_t = 2]}_{\text{Gains from switching to 1}} = \underbrace{\frac{1}{L}(\mu_1^0 - \mu_2^0)}_{\text{Gains if you are unlucky guinea pig}} + \underbrace{\left(1 - \frac{1}{L}\right) E[\mu_1 - \mu_2^0 | \mu_1 < \mu_2^0] \Pr[\mu_1 < \mu_2^0]}_{\text{"Gains" if you are not and action 1 is worse than 1/2}} \leq 0$$

(+) **(-)**

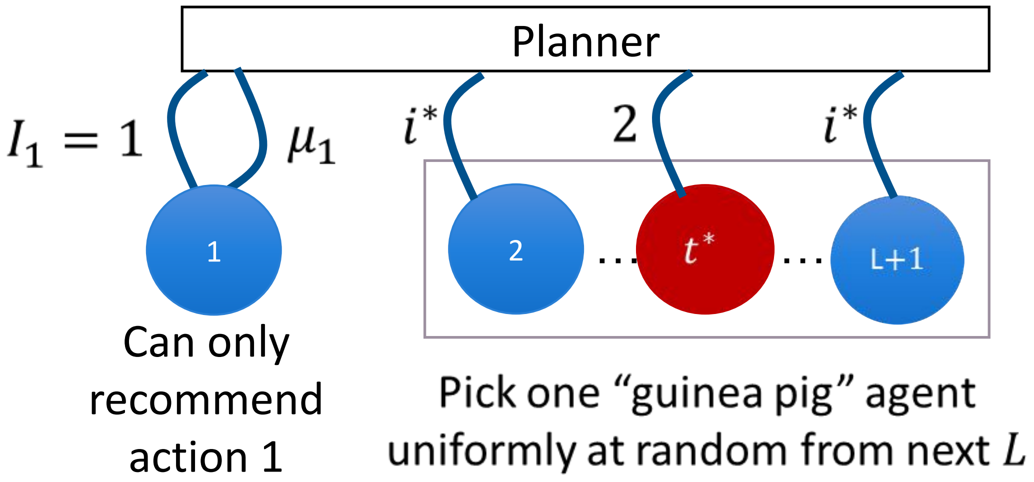
(Holds for $L >= 12$)

Two actions, deterministic rewards

- Action 1: $\mu_1 \sim U[1/3,1]$, $\mu_1^0 = E[\mu_1] = 2/3$
- Action 2: $\mu_2 \sim U[0,1]$, $\mu_2^0 = E[\mu_2] = 1/2$
- After $L+1$ rounds know both μ_1, μ_2
- Play best of two actions from then on

- Requires (necessary) assumption: Action 1 can be inferior after seeing its realization

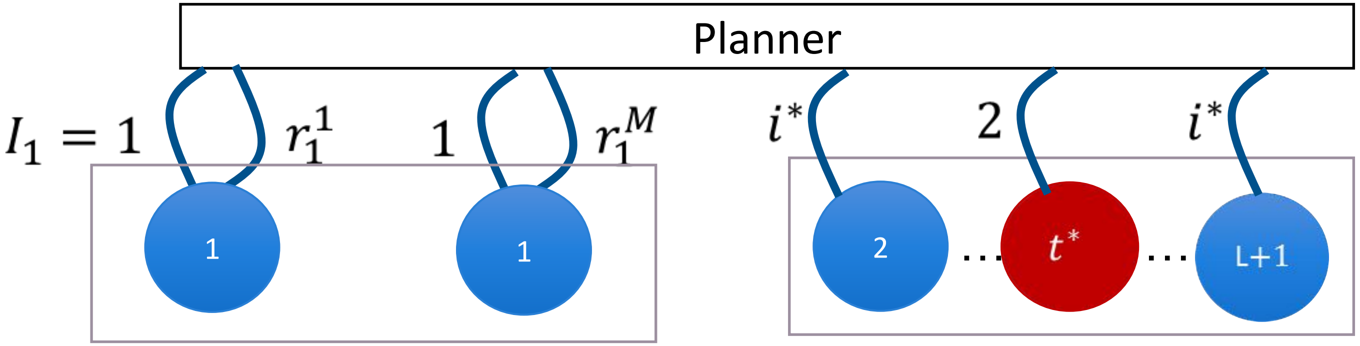
$$\Pr\left[\mu_1 < \mu_2^0\right] > 0$$



Two actions, stochastic rewards

- Rewards i.i.d. $r_i \sim D$, $E[r_i] = \mu_i$
- Requires slightly more complex assumption: Arm 1 posterior worse after seeing M signals

$$\Pr\left[E\left[\mu_1 \middle| r_1^1, \dots, r_1^M\right] < \mu_2^0\right] > 0$$



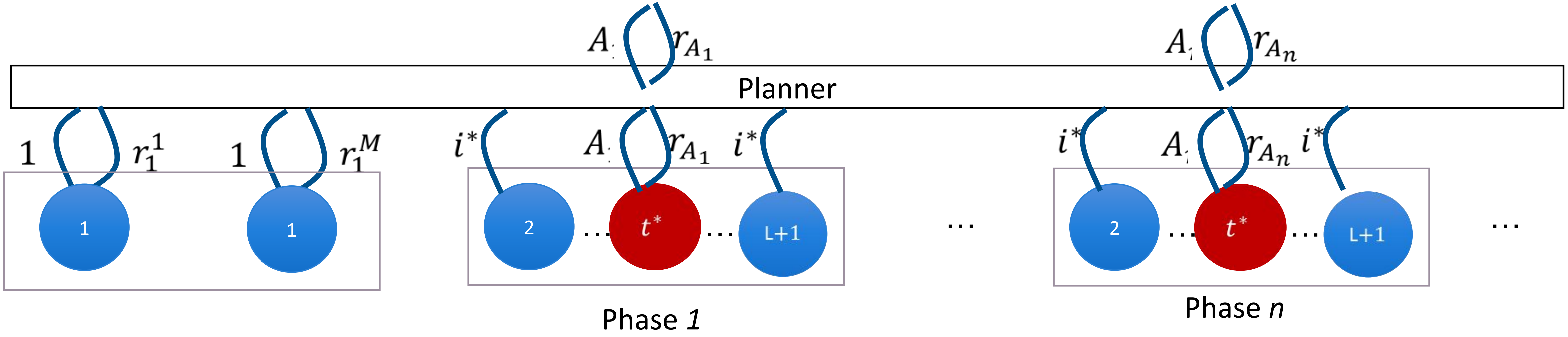
Collect M samples of arm 1
Enough to form a quite accurate posterior:
$E\left[\mu_1 \middle| r_1^1, \dots, r_1^M\right]$

Pick one "guinea pig" agent uniformly at random from next L

$$i^* = \operatorname{argmax}\left\{E\left[\mu_1 \middle| r_1^1, \dots, r_1^M\right], \mu_2^0\right\}$$

Two actions: black box reduction

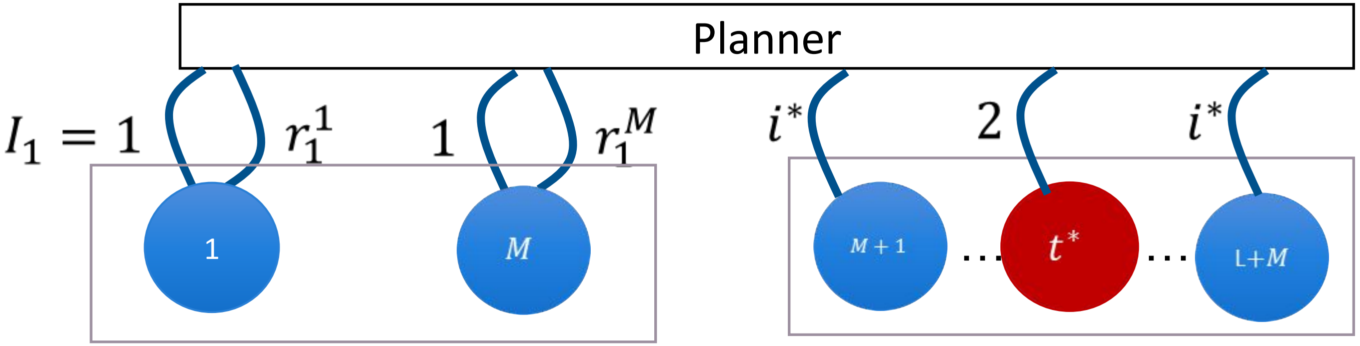
- Suppose we are given a multi-armed bandit algorithm A
- We can simulate this algorithm in an incentive compatible manner



- Expected reward of exploit users in phase n at least as good as algorithm's reward at phase n
- Expected welfare at least: $L \cdot \mathrm{Reward}_A\left(\frac{T}{L}\right) \Rightarrow$ Bayesian-Regret at most: $L \cdot \mathrm{Regret}_A\left(\frac{T}{L}\right)$
- If A is $\sqrt{T}$ algorithm $\Rightarrow \sqrt{L \cdot T}$ IC algorithm

Two actions, ex-post regret

- Instead of using posterior best, use sample means
- Make arm 2 the "exploit" action only if sample average of 1 is below μ_2^0 by a margin
- Chernoff bound analysis implies incentive compatibility



Collect M samples of arm 1
Use sample average as proxy of μ_1

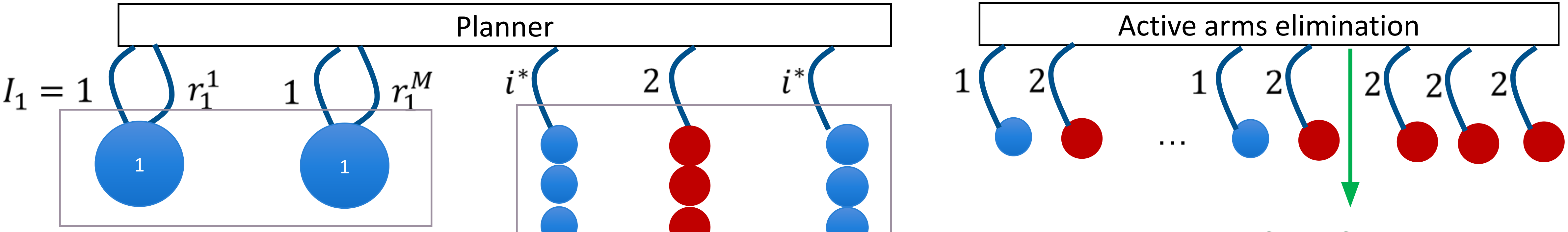
$$\hat{\mu}_1 = \frac{1}{M}\sum_t r_1^t$$

Pick one "guinea pig" agent
uniformly at random from next L

$$i^* = \begin{cases} 2, & \hat{\mu}_1 < \mu_2^0 - c_m \\ 1, & o.w. \end{cases}$$

Two actions, ex-post regret

- Similarly can get M samples of action 2
- Then do "active arms elimination"
  - Recommend actions in round robin
  - Until one sample average is above the other by margin $c_n \approx \frac{1}{\sqrt{n}}$



$\hat{\boldsymbol{\mu}}_2 > \hat{\boldsymbol{\mu}}_1 + \boldsymbol{c}_n$

Collect M samples of arm 1
Use sample average as proxy of μ_1

$$\hat{\mu}_1 = \frac{1}{M}\sum_t r_1^t$$

Pick M "guinea pigs" agent
uniformly at random from next L
$* M$

$$i^* = \begin{cases} 2, & \hat{\mu}_1 < \mu_2^0 - c_m \\ 1, & o.w. \end{cases}$$

Unobserved Heterogeneity and Confounding

Unobserved Heterogeneity

- Two actions: $x_t \in \{0,1\}$ (control, treatment)
- Agents are of two types $u_t \in \{0,1\}$
- Type is unobserved and affects baseline reward, $\left|g_t^{u_t}\right| \leq \sigma_g$

$$r_t = \theta x_t + g_t^{u_t}$$

- θ is the “effect” of the treatment
- Type affects prior bias on treatment effect $\theta \sim P^{u_t}$
- Type 1 prefers treatment $\mu_1 = E_{P^1}[\theta] > 0$
- Type 2 prefers baseline $\mu_0 = E_{P^0}[\theta] < 0$

Confounding Bias

- Suppose we calculate difference in means from treatment and control populations

$$E[\bar{y}_1 - \bar{y}_0] = \theta + E\left[g^1 - g^0\right]$$

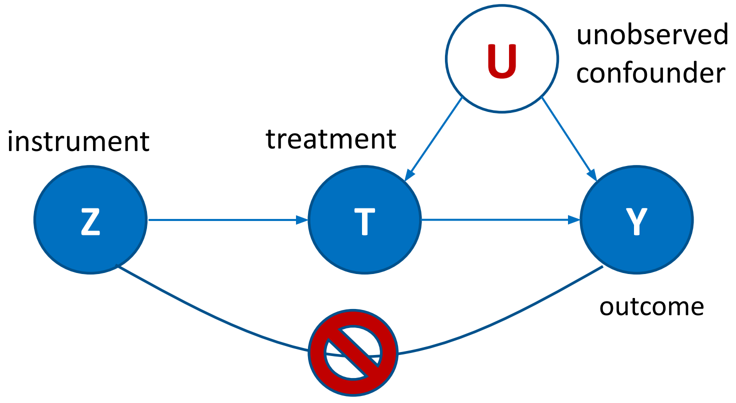
- Effect is heavily biased due to the fact that treatment take-up is correlated with baseline reward
- Example. Recommendation system for salespeople
  - Recommend a customer to go after
  - Commonplace misconception (belief) in the field, that customers that bring high revenue are customers we'll make a big difference
  - Prior on effect size positively correlated with baseline revenue

Recommendations as Instruments

- We don’t really need an incentive compatible mechanism
- Suppose that we give a recommendation z_t that is followed with positive probability
- Since recommendation is independent of unobserved type “confounder”, it can be viewed as what is known as an “instrument”
- Any variable that affects the taken treatment, but does not affect the outcome other than through the treatment

Instrumental Variables

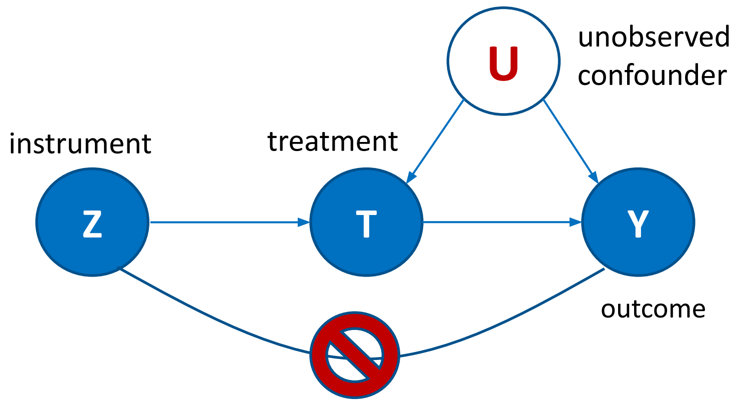
Instrumental Variable: any random variable **Z** that affects the treatment (log-price) **T** but does not affect the outcome (log-demand) **Y** other than through the treatment [Wright'28, Bowden-Turkington'90, Angrist-Krueger'91, Imbens-Angrist'94]



Instruments are widely used

- **Policy.** Judge leniency => Effects of incarceration
- **Healthcare.** Ambulance company assignment => Hospital quality
- **Digital experimentation.** Recommendation A/B test => Effects of user induced actions

Identification of Causal Effects via Instruments



Phillip Wright's idea (1928): the first causal path diagram analysis

- We can estimate effect of Z on y via a regression

$$\gamma = \frac{\mathbb{E}\left[\left(Z - \bar{Z}\right)\left(y - \bar{y}\right)\right]}{\mathbb{E}\left[\left(Z - \bar{Z}\right)^2\right]}$$

- We can estimate the effect of Z on T via a regression

$$\delta = \frac{\mathbb{E}\left[\left(\mathrm{Z} - \bar{\mathrm{Z}}\right)\left(\mathrm{X} - \bar{X}\right)\right]}{\mathbb{E}\left[\left(\mathrm{Z} - \bar{\mathrm{Z}}\right)^2\right]}$$

- The effect of Z on Y (γ) is the product of the effect of Z on T (δ) multiplied by the effect of T on y (θ)

$$\theta = \frac{\gamma}{\delta} = \frac{\mathbb{E}\left[\left(Z - \bar{Z}\right)\left(y - \bar{y}\right)\right]}{\mathbb{E}\left(\mathrm{Z} - \bar{\mathrm{Z}}\right)\left(\mathrm{X} - \bar{X}\right)}$$

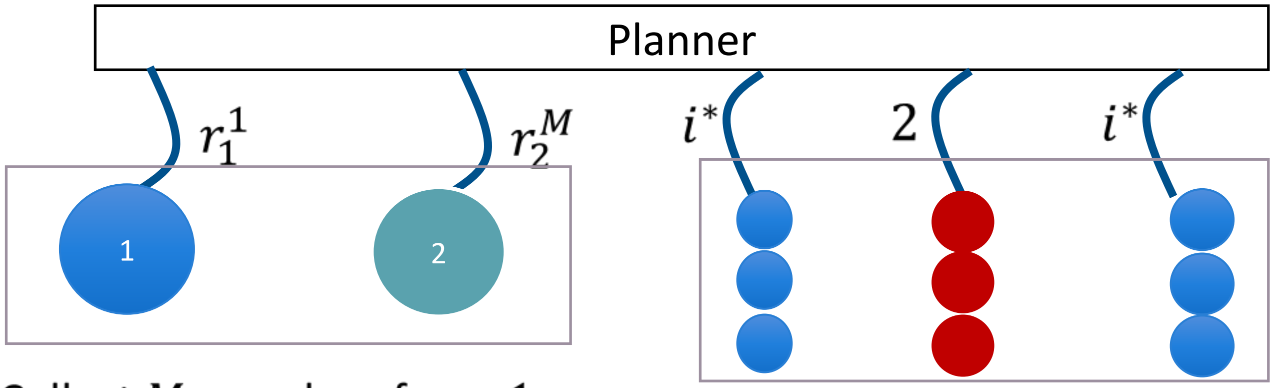
- In finite samples, replace expectations with empirical averages

$$\hat{\theta} = \frac{\mathbb{E}_n\left[\left(Z - \bar{Z}\right)\left(y - \bar{y}\right)\right]}{\mathbb{E}_n\left(\mathrm{Z} - \bar{\mathrm{Z}}\right)\left(\mathrm{X} - \bar{X}\right)}$$

Instrument Strength/Compliance Level

- If planner had no private information, then no matter what recommendation they send, taken treatment would be solely driven by private type
- Instrument strength would be 0:
$$\mathbb{E}\left(\mathrm{Z}-\bar{\mathrm{Z}}\right)\left(\mathrm{X}-\bar{X}\right)=0$$
- Finite sample result: w.p. $1-\delta$
$$\left|\hat{\theta}-\theta_0\right| \leq \frac{2\sigma_g\sqrt{2n\log\frac{2}{\delta}}}{\sum_i (x_i-\bar{x})(z_i-\bar{z})}$$

Online Instrumental Variable Regression

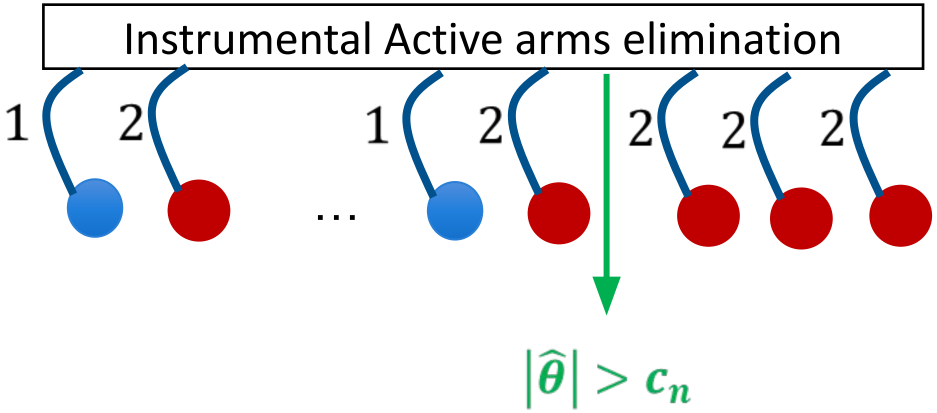


Collect M samples of arm 1
Use sample average as proxy of μ_1

$$\hat{\mu}_1 = \frac{1}{M}\sum_t r_1^t$$

Pick M "guinea pigs" agent uniformly at random from next $L * M$

$$i^* = \begin{cases} 1, & \hat{\mu}_1 > \hat{\mu}_0 + c_m \\ 0, & o.w. \end{cases}$$



Continuously run IV regression after every step of "both-arm collection"

- Can show $O\left(\sqrt{T\log(T)}\right)$ regret
- Constants much smaller than BIC exploration. We do not need to incentivize all agents to take all actions

Summary

- Black−box reduction: any bandit algorithm to an incentive compatible one (prior−dependent constant blow up in Bayesian regret)
- Enables modular design of IC recommendation systems
- O(√T) and O(log(T)) instance-based ex-post regret
- Detail-free algorithm (doesn't need to know full prior, only upper bound on a single parameter of the prior)
- Extensions: game theoretic setting, observed and unobserved heterogeneity

Take home message: Via control of information flow, incentivizing exploration is feasible. Can identify optimal option.

Thank you

Bayesian incentive compatible bandit exploration, *Conference on Economics and Computation, 2015*

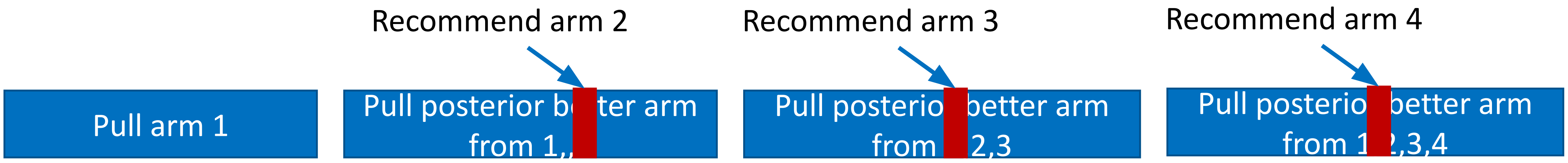
Bayesian exploration: incentivizing exploration in Bayesian games, *Conference on Economics and Computation, 2016*

User-Heterogeneity: Contextual Bandit Extension

-

Key idea: many arms

- Need to first sample actions $1, \ldots, i$ to convince to play $i + 1$
- Do a contest:



- Many technical difficulties to perform contest with sample means for detail-free
- Use of sample averages with a confidence bound not as straight-forward
- Not trivial to define exploit arm as a function of sample means