Incentivizing Bandit Exploration: Recommendations as Instruments

Daniel Ngo
University of Minnesota
ngo00054@umn.edu

Logan Stapleton
University of Minnesota
stapl158@umn.edu

Vasilis Syrgkanis
Microsoft Research
vasy@microsoft.com

Zhiwei Steven Wu
Carnegie Mellon University
zstevenwu@cmu.edu

Abstract

We study a multi-armed bandit learning setting where a social planner incentivizes a set of heterogeneous agents to efficiently explore the set of available arms. At each round, an agent arrives with their unobserved private type that determines both their prior preferences across the actions as well as their action-independent confounding shift in the rewards. The planner provides the agent with an arm recommendation that may alter their belief and incentivizes them to explore potentially sub-optimal arms. Under this setting, we provide a novel recommendation mechanism that views the planner's recommendations as forms of instrumental variables (IV) that only affect an agents' arm selection, but not the observed rewards. We construct such IVs by carefully mapping the history –the interactions between the planner and the previous agents– to a random arm recommendation. Despite the unobserved confounding shift in the rewards, the resulting IV regression provides reliable estimates on the reward effects of the actions and enables the social learning process to minimize regret over the long term. Compared to the existing approaches in the literature of incentivizing exploration, our IV-based mechanism also mitigates selection bias and the negative externality that one group of agents may have on others.

1 Introduction

In many online recommendation systems, e.g. those used by Netflix, Amazon, Yelp, Stubhub, and Waze, users are both consumers and producers of information. Users make their selections based on recommendations from the system, and the system collects data about the users' experiences to provide better-quality recommendations for future users. To ensure the quality of its recommendations, the system typically needs to balance between *exploration*—selecting potentially suboptimal options for the sake of acquiring new information—and *exploitation*—selecting the best option given the available information. However, there is an inherent tension between exploration and users' incentives: since each user is primarily concerned with their short-term utility, their incentives naturally favor exploitation.

To resolve this tension between exploration and exploitation, a long line of work started by [11; 12] has studied mechanisms that incentivize users to explore by leveraging the information asymmetry between the recommendation system and users [13; 8; 12; 14]. These papers consider a multi-armed bandit model where the recommendation system is a *social planner* who interacts with a sequence of self-interested *agents*. The agents arrive one-by-one to choose from a given set of actions (or "arms"

Causal Discovery & Causality-Inspired Machine Learning Workshop at Neural Information Processing Systems, 2020.

or "alternatives") and receive a reward for their choice. Upon the arrival of each agent, the social planner provides a recommendation (of one of the actions) that influences the agent's selection. The problem is to design a recommendation policy that incentivizes the agents to explore and in the long run maximize the cumulative rewards of all the agents, even when the agents want to exploit.

Prior work on *incentivizing exploration* typically approaches the problem by enforcing *Bayesian incentive-compatibility* (BIC) for every agent—that is, the planner recommends actions which it is in each agent's interest to follow, even if such action is inferior according to the agent's prior belief. To achieve BIC, a recommendation mechanism typically needs to make strong assumptions that 1) all of the agents share a *common prior* over the rewards of actions and 2) all agents will *comply* with each recommendation. However, in reality, agents tend to be heterogeneous in terms of their beliefs and perceived rewards; as such, some may comply with a recommendation and some may not. In particular, some agents can have a stronger *bias* that favors one action over others. Moreover, even if the actions have the same effect or reward for all agents, their realized or observed utilities can be different. For example, different patients taking the same drug may report levels of pain and different drivers taking the same route may have different commute time.

Example 1.1. *(Discriminatory lending) Consider the following scenario where some agents have biases that favor one action over others concerns: a bank that lends in a discriminatory fashion towards small businesses, as opposed to big ones. Suppose a bank is testing out a new sales campaign where a loan officer can offer a business client a low-interest loan as an intervention. If the loan officer offers the loan, the bank incurs some cost based on: 1) how much money the business takes out; and 2) if the borrower defaults on their loan. Businesses generate* baseline revenue *that depends on their types (big or small) and is independent of the intervention (i.e. offering the loan or not). The loan officer knows their clients better than the bank does: so, the bank does not know the clients' private types, but the loan officer does. The loan officer knows their clients' types, but the bank does not (say, because the loan officer knows their clients better than corporate does). Big businesses already have more investments with the bank, so they generate more baseline revenue than small businesses regardless of whether or not they're offered a low-interest loan. Whether the loan officer offers the loan depends on the business type and based on personal sales goals that are separate from the bank's goals. The officers are biased against small businesses, since they expect them to be likely to default (which may be unfounded or not). Thus, the* intervention choice *is correlated with the* baseline revenue *and causes a confounding effect. This confounding means that the bank, who only observes the intervention choice and the total revenue and not the baseline revenue (nor the client type), cannot estimate the effect of the low-interest loan campaign by standard methods (e.g. OLS). Thus, they must use an* instrument *—here, the bank's recommendation— to accurately judge whether giving a low-interest loan causes higher profits than offering no loan and, similarly, to estimate the effect of this intervention on profits. By giving recommendations to induce variability, the bank can incentivize loan officers to lend to small businesses. Thus, the bank's recommendation policy mitigates* selection bias*, preventing discrimination against small businesses in this setting.*

1.1 Related Work

Our work applies instrument variable (IV) regression [1; 2; 7] in a multi-armed bandits setting [3; 10; 15]. We use IV regression because it yields consistent causal effect estimates in presence of confounding, unlike OLS or ANOVA. The confounding effects of our model are a byproduct of the more general setting of the model considered in [12]. Particularly, our model and algorithm accommodates for heterogeneous priors, noncompliance, and binary treatment effect estimation; those in [12] do not. Both [9] and [4] consider confounding in multi-armed bandits settings. However, [4] does not consider incentive-compatibility nor do they employ IV regression. [9] uses IV regression in a Bayesian bandits setting, but they consider an entirely different "instrument-armed bandit" problem that does not consider incentivizing agents.

1.2 Results

We consider a Bayesian Incentive-Compatible bandits model that is more general than the benchmark [12] insofar as it considers heterogeneous priors, noncompliance, confounding, and binary treatment effect estimation. We provide an algorithm (see sub-algorithms 1 and 2) that employs IV regression in a novel way in order to yield sub-linear regret (section 4.2) and a confidence interval on the true treatment effect (theorems 3.2, 4.2 and 4.4). As an ancillary result, we provide a bound on the

difference between a true treatment effect and its finite-sample IV estimate theorem 2.1. In this workshop paper, we showcase our results in the simplified binary treatment setting with only two types of agents to build intuition. Our results also apply to general settings with arbitrarily many treatments and types of agents: however, this is reserved for our full paper.

2 Model

We study a sequential game between a *social planner* and a sequence of *agents* over T rounds, where T is known to the social planner.[1] There are two actions to choose from: either the control or the treatment. In each round t, an entirely new agent indexed by t arrives with their *private type* u_t drawn independently from a distribution $\mathcal{U}$ defined over the set U of all privates types. Each agent t receives an action recommendation $z_t \in \{0, 1\}$ from the planner and then selects an action $x_t \in \{0, 1\}$ and receives a reward $y_t \in \mathbb{R}$.

Reward Given the choice of $x_t \in \{0, 1\}$, the observed outcome y_t for each agent t is given by

$$y_t = \theta x_t + g(u_t) + \varepsilon_t \tag{1}$$

where $\theta \in [-1, 1]$ denotes the unknown exogenous *treatment effect*, the term $g(u_t)$ denotes a *baseline reward* that depends on the agent's private type u_t, and the term ε_t denotes an independent reward noise with a subgaussian norm of σ_ε and conditional expectation $\mathbb{E}[\varepsilon_t \mid x_t, u_t] = 0$. We assume that for all rounds t, the confounding term $|g(u_t)| \leq \Upsilon$ for some constant Υ. Both $g(u_t)$ and ε_t are unobserved by the social planner. The social planner's objective is to maximize the total observed reward of all agents over all T rounds.

The utility for an agent t is different than the observed reward insofar as the agent incurs some cost for taking the treatment that is not accounted for in the observed reward. Let the utility for agent t be:

$$y_t^{\text{agent}} := y_t - \phi(u_t)x_t = \theta x_t - \phi(u_t)x_t + g(u_t) + \varepsilon_t \tag{2}$$

where $\phi(u_t)$ denotes the cost to an agent of type u_t for taking the treatment for cost function $\phi : U \to [-\nu, \nu]$. The social planner observes neither cost function ϕ nor agent reward y_t^{agent}.

Agents are self-interested individuals who aim to maximize their own expected reward y_t^{agent}, while the social planner aims to maximize the long-term cumulative rewards across all agents.

History and recommendation policy The interaction between the planner and the agent t is given by the tuple (z_t, x_t, y_t). For each t, let H_t denote the history from round 1 to t, that is the sequence of interactions between the social planner and first t agents: $((z_1, x_1, y_1), \ldots, (z_t, x_t, y_t))$. Before the game starts, the social planner commits to a recommendation policy $\pi = (\pi_t)_{t=1}^T$ where each $\pi_t \colon (\{0, 1\} \times \{0, 1\} \times \mathbb{R})^{t-1} \to \Delta(\{0, 1\})$ is a randomized mapping from the history H_{t-1} to a recommendation z_t. The policy π is fully known to the agents.

Beliefs, incentives, and action choices Each agent t knows their place t in the sequential game. As part of their private type u_t, each agent t has a *prior belief distribution* $\mathcal{P}_t$, which is a joint distribution over the treatment effect θ, the agents' private types, and their reward noise.

In round t, the action x_t chosen is given by a selection function f that takes the following form:

$$x_t = f(u_t, z_t, t) := \mathbb{1}\left[\mathop{\mathbb{E}}_{\mathcal{P}_t}[\theta \mid u_t, z_t, t] > \phi(u_t)\right], \tag{3}$$

where, as above, $\phi(u_t)$ denotes the cost to an agent of type u_t for taking the treatment, for some bounded function $\phi : U \to [-\nu, \nu]$.

We say that the recommendation is *Bayesian incentive compatible (BIC)* for agent t if $x_t = z_t$.

[1] In example 1.1, the *social planner* would be the bank and the *agent* would be the loan officer–business pair that arrives on a given day.

2.1 Recommendations as Instruments

We model the observed reward y_t and action x_t in the presence of confounding variables that depend on the type u_t. In a bandits setting without a confounder, we could find the treatment effect θ by taking the mean over sufficiently many observed rewards of each arm. However, because of the unknown confounding variable $g(u_t)$ that is correlated with x_t, such simple estimation is inconsistent. Instead, we use *instrumental variable (IV) regression* to estimate the exogenous treatment effect θ. Note that each recommendation z_t can be viewed an *instrumental variable* because: (1) z_t influences the selection x_t; and (2) z_t is independent from the endogenous error term $g(u_t)$. Criterion (2) follows because planner chooses z_t randomly independent of the type u_t.

Our mechanism periodically solves the following IV regression problem: given a set of n observations $\{(x_i, y_i, z_i)\}_{i=1}^n$, find an accurate estimator $\hat{\theta}_n$ of θ. To derive the estimator, we will rewrite the selection and reward functions in (1) and (3) as a linear relationship between x_i, y_i and z_i. Then

$$y_i = \theta x_i + g(u_i) + \varepsilon_i \tag{4}$$

$$\begin{aligned} x_i &= \gamma_1 z_i + \gamma_0(1 - z_i) + \eta_i \\ &= \gamma z_i + \gamma_0 + \eta_i \end{aligned} \tag{5}$$

where $\eta_i = \mathbb{E}_{\mathcal{U}}[x \mid z_i] - x_i$ and $\gamma_1 = \mathbb{P}_{\mathcal{U}}[x_i = 1 \mid z_i = 1]$ and $\gamma_0 = \mathbb{P}_{\mathcal{U}}[x_i = 1 \mid z_i = 0]$ make up the *compliance coefficient* γ, which is given by:

$$\gamma = \gamma_1 - \gamma_0 = \mathop{\mathbb{P}}_{\mathcal{U}}[x_i = 1 \mid z_i = 1] - \mathop{\mathbb{P}}_{\mathcal{U}}[x_i = 1 \mid z_i = 0] \tag{6}$$

Let the operator $\bar{\cdot}$ denote the mean, e.g. $\bar{y} := \frac{1}{n}\sum_{i=1}^n y_i$ and $\bar{g} := \frac{1}{n}\sum_{i=1}^n g(u_i)$. The mean reward

$$\bar{y} = \beta\bar{z} + \theta\gamma_0 + \theta\bar{\eta} + \bar{g} + \bar{\varepsilon}$$

Thus, the centered reward and treatment choice at round i are given as:

$$\begin{cases} y_i - \bar{y}_i = \beta(z_i - \bar{z}) + \theta(\eta_i - \bar{\eta}) + g(u_i) - \bar{g} + \varepsilon_i - \bar{\varepsilon} \\ x_i - \bar{x}_i = \gamma(z_i - \bar{z}) + \eta_i - \bar{\eta} \end{cases} \tag{7}$$

Instrumental Variable Estimator. Using these formulations of the centered reward $y_i - \bar{y}$ and treatment choice $x_i - \bar{x}$, we form the estimate $\hat{\theta}_n$ via *Instrumental Variable (IV) regression*. We first form empirical estimates $\hat{\gamma}_n$ and $\hat{\beta}_n$ by regressing the centered treatment choice $x_i - \bar{x}$ and the centered reward $y_i - \bar{y}$ over the centered recommendation $z_i - \bar{z}$, respectively. These estimates are formed over n samples $\{x_i, z_i, y_i\}_{i=1}^n$ as such:

$$\hat{\beta}_n := \frac{\sum_{i=1}^n (y_i - \bar{y})(z_i - \bar{z})}{\sum_{i=1}^n (z_i - \bar{z})^2} \qquad \text{and} \qquad \hat{\gamma}_n := \frac{\sum_{i=1}^n (x_i - \bar{x})(z_i - \bar{z})}{\sum_{i=1}^n (z_i - \bar{z})^2} \tag{8}$$

Second, we take their quotient as the predicted treatment effect $\hat{\theta}_n$, i.e.

$$\hat{\theta}_n = \frac{\hat{\beta}_n}{\hat{\gamma}_n} = \frac{\sum_{i=1}^n (y_i - \bar{y})(z_i - \bar{z})}{\sum_{i=1}^n (z_i - \bar{z})^2} \frac{\sum_{i=1}^n (z_i - \bar{z})^2}{\sum_{i=1}^n (x_i - \bar{x})(z_i - \bar{z})} = \frac{\sum_{i=1}^n (y_i - \bar{y})(z_i - \bar{z})}{\sum_{i=1}^n (x_i - \bar{x})(z_i - \bar{z})} \tag{9}$$

We provide a finite-sample *approximation bound* on $\hat{\theta}_n$, which may be of independent interest.

Theorem 2.1 (Treatment effect approximation bound)**.** *Given a sample set $(z_i, x_i, y_i)_n$, which contains n samples of instrument z, treatment x, and reward y, we bound the difference between the true treatment effect θ and the predicted treatment effect $\hat{\theta}_n$ derived via IV regression over $(z_i, x_i, y_i)_n$. Let Υ be a upper bound on the confounding term $g(u_i)$. For some confidence $\delta > 0$, with probability at least $1 - \delta$:*

$$\left|\hat{\theta}_n - \theta\right| \leq \frac{(2\sigma_\varepsilon + 4\Upsilon)\sqrt{2n\log(4/\delta)}}{\left|\sum_{i=1}^n (x_i - \bar{x})(z_i - \bar{z})\right|}.$$

Proof. See appendix A.1.2 for a proof. □

3 Setting Assumptions and Sampling Stage for Control-Treatment with Two Mixed-Preference Types

Assume that there are two types of agents. Let p_i for $i \in \{0, 1\}$ be the proportion of agents of type i in the population. The types are *mixed*, since agents of each type prefer different arms, i.e.

- agents of type 0 initially prefer the control (arm 0), i.e. $\mu_0^0 > \mu_0^1$; and
- agents of type 1 initially prefer the treatment (arm 1), i.e. $\mu_1^0 < \mu_1^1$.

Our algorithm splits into two stages. First, in the sampling stage (algorithm 1), we incentivize only type 0 agents to take all of our recommendations (see lemma 3.1). (Since they already prefer arm 0, the primary difficulty here is to incentivize them to take arm 1 when we recommend it.) We do so by recommending arm 0 as an "exploitation" arm most of the time and recommending arm 1 as an "exploration" arm with a low probability. Because we recommend the exploration arm with low probability, our instrument is weak and it takes a lot of rounds to get an IV estimate with a good approximation bound. So, once our approximation bound is good enough, we move to the second phase: the racing stage (algorithm 2). In the racing stage, we play Active Arms Elimination [6] and recommend arms 0 and 1 sequentially in equal proportion. In the first part of this stage, only type 0 agents are BIC, insofar as they always follow our recommendations. Up until this point, agents of type 1 will always choose arm 1, regardless of what we recommend. However, in the second part of the racing stage, our estimate $\hat{\theta}$ is good enough so that all agents become BIC and follow our recommendations. Once one arm 'wins,' we end the racing stage and recommend the winning arm for the remainder of the time horizon. At each stage, we get a confidence interval on the true treatment effect relative to our estimate $\hat{\theta}$ and approximation bound (given in theorem 2.1).

3.1 Sampling Stage for Control-Treatment with Two Mixed-Preference Types

In the first phase of algorithm 1, we let the agents pick their preferred arm. Hence, each agent chooses the better arm according to their prior: agents of types 0 will choose control (arm 0) while agents of type 1 will take treatment (arm 1). We observe the reward from these arm pulls. Assume that there are ℓ_0 samples of arm 0 and ℓ_1 samples of arm 1 observed in this phase. For a given type i, there is a non-zero chance that arm 0 performs so poorly that the posterior of arm 1 looks better than that of arm 0 conditioned on the observed samples. We define this event as

$$\xi_i = \left\{ \frac{1}{\ell_1} \sum_{t=1}^{\ell_1} y_t^1 - \frac{1}{\ell_0} \sum_{t=1}^{\ell_0} y_t^0 > 2\Upsilon + \sigma_\varepsilon \sqrt{\frac{2\log(1/\delta)}{\ell_0}} + \sigma_\varepsilon \sqrt{\frac{2\log(1/\delta)}{\ell_1}} + \frac{1}{2} + \nu \right\} \tag{10}$$

where $\delta > 0$ is a small failure probability and at round t, the term y_t^a denotes the observed reward of taking arm a at time t (or no reward if arm a is not observed at time t), i.e. $y_t^a = y_t \mathbb{1}[X_t = a]$.

Algorithm 1: Sampling stage for control-treatment with two mixed-preference types

Input: parameters $\rho, \ell_1 \in \mathbb{N}$ and $0 < \delta < \delta'$, where $\delta' := \frac{\mathbb{P}_{\mathcal{P}_0}[\xi_0]}{8}$
In the first $\ell_0 + \ell_1$ rounds, let the agents pick their preferred arm. The sample average of control reward y^0 after ℓ_0 pulls of arm 0 is denoted $\bar{y}_{\ell_0}^0$. Similarly, the sample average of treatment reward y^1 after ℓ_1 pulls of arm 1 is denoted $\bar{y}_{\ell_1}^1$. After that, do the following:

if $\bar{y}_{\ell_0}^0 - \bar{y}_{\ell_1}^1 > 2\Upsilon + \sigma_\varepsilon \sqrt{\frac{2\log(1/\delta)}{\ell_0}} + \sigma_\varepsilon \sqrt{\frac{2\log(1/\delta)}{\ell_1}} + \frac{1}{2} + \nu$ **then**
  $a^* = 0$
else
  $a^* = 1$
end
From the set P of the next $\ell_1.\rho$ agents, pick a set Q of ℓ_1 agents uniformly at random.
Every agent $t \in P - Q$ is recommended arm a^*.
Every agent $t \in Q$ is recommended arm 1.

Lemma 3.1 (Sampling Stage BIC for Type 0)**.** *Algorithm 1 with parameters (ρ, ℓ_0, ℓ_1) completes in $\ell_1\rho + \ell_0 + \ell_1$ rounds. Algorithm 1 is BIC for all agents of type 0 if we hold the assumption above*

and the parameters satisfy:

$$\rho \geq 1 + \frac{4(\mu_0^0 - \mu_0^1 + \nu\,\mathbb{P}_{\mathcal{P}_0}[\xi_0])}{\mathbb{P}_{\mathcal{P}_0}[\xi_0]}. \tag{11}$$

For type 0, the *fighting chance* event is given as

$$\xi_0 = \left\{ \frac{1}{\ell_1}\sum_{t=1}^{\ell_1} y_t^1 - \frac{1}{\ell_0}\sum_{t=1}^{\ell_0} y_t^0 > 2\Upsilon + \sigma_\varepsilon\sqrt{\frac{2\log(1/\delta)}{\ell_0}} + \sigma_\varepsilon\sqrt{\frac{2\log(1/\delta)}{\ell_1}} + \frac{1}{2} + \nu \right\}, \tag{12}$$

and happens with probability at least $1-\delta$ for confidence $\delta > 0$.

Proof. See appendix A.2.1 for a proof. □

Theorem 3.2 (Sampling Stage Treatment Effect Confidence Interval)**.** *With n total samples collected from algorithm 1 –run with exploration probability ρ large enough so that our recommendation BIC for type 0 agents (see lemma 3.1),– we form an estimate θ_n of the treatment effect θ. With probability at least $1-\delta$,*

$$\left|\hat{\theta}_n - \theta\right| \leq \frac{2\rho\,(\sigma_\varepsilon + 2\Upsilon)\,\sqrt{2n\log(5/\delta)}}{\left(np_0(1-1/\rho) - \sqrt{\frac{n(1-1/\rho)\log(5/\delta)}{2}}\right)}$$

Proof. See appendix A.2.2 for a proof. □

4 Racing Stage and Regret Analysis

We observe a set of $\rho\ell$ samples from the sampling stage. With large enough $\rho\ell$, we run the racing stage algorithm, which is a modification of the *Active Arms Elimination* algorithm [6] with added BIC constraints, i.e. that agents of type 0 are BIC in the first part of the algorithm, then all agents in the second part. Finally, once an arm is found to be sub-optimal, it will be dropped from the race and the other arms "wins". To accelerate the sample collecting process, we employ the doubling trick [5] to increase the phase length of the racing stage.

4.1 Racing Stage for Control-Treatment with Two Mixed-Preference Types

Lemma 4.1 (Racing Stage BIC for Type 0)**.** *Fix some constant $\tau \in (0,1)$. Remember that $\phi(u_t) \in [-\nu, \nu]$ for all u_t and $\mathbb{P}_{\mathcal{P}_0}[\theta > \tau]$ be the prior probability that $\theta > \tau$ for agents of type 0. Let the approximation bound s_L after the sampling stage be*

$$s_L := \frac{(2\sigma_\varepsilon + 4\Upsilon)\,\sqrt{2L\log(4/\delta)}}{\left|\sum_{i=1}^{L}(x_i - \bar{x})(z_i - \bar{z})\right|} \leq \frac{\tau\,\mathbb{P}_{\mathcal{P}_0}[\theta > \tau]}{4} - \frac{\nu}{2}. \tag{13}$$

Then, during the first phase of the racing stage, the recommendations from algorithm 2 are BIC for agents of type 0.

Proof. See appendix A.3.1 for a proof. □

Theorem 4.2 (First Part Racing Stage Treatment Effect Confidence Interval)**.** *With n total samples collected from the first part of algorithm 2 where the type 0 BIC criterion on the sampling stage approximation bound is met (see lemma 4.1), form an estimate θ_n of the treatment effect θ. With probability at least $1-\delta$,*

$$\left|\hat{\theta}_n - \theta\right| \leq \frac{8\,(\sigma_\varepsilon + 2\Upsilon)\,\sqrt{2n\log(5/\delta)}}{np_0 - \sqrt{n\log(5/\delta)}}$$

Proof. See appendix A.3.2 for a proof. □

Algorithm 2: The racing stage for control-treatment with two mixed-preference types

Input: samples $S_L := \{x_i, z_i, y_i\}_1^L$ sampling stage sans the first phase $L \in \mathbb{N}$; time horizon T; minimum phase length h; probability δ; and upper bound ν on cost function $\phi(u_t)$

Input: IV estimate of mean reward θ, denoted $\hat{\theta}_L$ and approximation bound $s_L := \frac{(2\sigma_\varepsilon + 4\Upsilon)\sqrt{2L\log(4/\delta)}}{\left|\sum_{i=1}^L (x_i - \bar{x})(z_i - \bar{z})\right|}$ defined over samples S_L from the sampling stage

Let $\hat{\theta}_0 := \hat{\theta}_L$ and $s_0 := s_L$ and $S_0 := S_L$;

Split the remainder into consecutive phases of $h2^q$ rounds each, starting $q = 1$;

while $\hat{\theta}_{q-1} \leq s_{q-1}$ **do**

The next $h2^q$ agents are recommended both arms sequentially;

For each arm $a \in \{0, 1\}$, let y_t^a be one sample reward of that arm in this phase;

Let the samples gathered from round q be denoted $S_q := \{x_i, z_i, y_i\}_{L+h2^{q-1}}^{L+h2^q}$

Define S_q^{BEST} as the samples from the phase (including the sampling stage samples S_L) with the best approximation bound, i.e. $S_q^{\text{BEST}} = \underset{S_r,\, 0 \leq r \leq q}{\operatorname{argmin}} \frac{2(\sigma_\varepsilon + 2\Upsilon)\sqrt{2|S_r|\log(4/\delta)}}{\left|\sum_{S_r}(x_i - \bar{x})(z_i - \bar{z})\right|}$;

Define $\hat{\theta}_q$ and s_q as the estimate and approximation bound defined over the best samples S_q^{BEST};

$q = q + 1$;

end

For all remaining agents recommend $a^* = \mathbb{1}\left[\hat{\theta}_q > 0\right]$

Lemma 4.3 (Racing Stage BIC for Type 1)**.** *Fix some constant $\tau \in (0, 1)$. Remember that $\phi(u_t) \in [-\nu, \nu]$ for all u_t and $\mathbb{P}_{\mathcal{P}_1}[\theta < -\tau]$ be the prior probability that $\theta < -\tau$ for agents of type 1. Let the approximation bound s_{L_1} after the first racing stage be*

$$s_{L_1} := \frac{(2\sigma_\varepsilon + 4\Upsilon)\sqrt{2L_1\log(4/\delta)}}{\left|\sum_{i=1}^{L_1}(x_i - \bar{x})(z_i - \bar{z})\right|} \leq \frac{\tau\,\mathbb{P}_{\mathcal{P}_1}[\theta < -\tau]}{4} - \frac{\nu}{2}. \tag{14}$$

Then, during the second part of the racing stage, the recommendations from algorithm 2 are BIC for agents of type 1 (as well as for type 0).

Proof. See appendix A.3.3 for a proof. □

Theorem 4.4 (Second Part Racing Stage Treatment Effect Confidence Interval)**.** *With n total samples collected from the second part of algorithm 2 where the BIC criterion for both types on the sampling stage approximation bound is met (see lemmas 4.1 and 4.3), form an estimate θ_n of the treatment effect θ. With probability at least $1 - \delta$,*

$$\left|\hat{\theta}_n - \theta\right| \leq \frac{8\,(\sigma_\varepsilon + 2\Upsilon)\sqrt{2n\log(5/\delta)}}{n}$$

Proof. See appendix A.3.4 for a proof. □

4.2 Regret Analysis

The goal of this algorithm is to maximize the cumulative reward of all agents. We measure the algorithm's performance through the definition of regret. We are interested in minimizing the ex-post regret of the algorithm, which is the regret specific to a particular treatment effect θ. Since the priors are not exactly known to the social planner, this ex-post regret is correct for any realization of the priors and treatment effect θ.

Definition 4.5 (Ex-post Regret)**.** The ex-post regret of the algorithm is

$$R_\theta(T) = T\theta a^* - \sum_{t=1}^{T} \theta x_t \tag{15}$$

The Bayesian expected regret of the algorithm is:

$$R_{\mathcal{P}}(T) = \mathbb{E}\left[T\theta a^* - \sum_{t=1}^{T} \theta x_t\right] = \mathbb{E}[R_\theta(T)] \tag{16}$$

where if $\theta > 0$, then $\theta^{a^*} = \theta$; and otherwise if $\theta \leq 0$, then $\theta^{a^*} = 0$.

Using these definitions, our entire algorithm (with the sampling stage and the racing stage) achieves sub-linear ex-post regret.

Lemma 4.6 (Regret)**.** *Algorithms 1 and 2 with parameters $\ell_1, \rho \in \mathbb{N}$ achieves ex-post regret*

$$R(T) \leq \ell_1 \rho + O(\sqrt{T \log T}) \tag{17}$$

where ℓ_1 is the sampling stage's phase length and $1/\rho$ is the exploration probability in the sampling stage.

Proof. See Appendix B.1 for a proof. □

With similar analysis, we also achieve sub-linear expected regret over the randomness in the priors of the agents. Lemma 4.7 provides a basic performance guarantee of our algorithm.

Lemma 4.7 (Expected Regret)**.** *Algorithms 1 and 2 with parameters $\ell_1, \rho \in \mathbb{N}$ achieves expected regret*

$$\mathbb{E}[R(T)] = O(\sqrt{T \log T}) \tag{18}$$

Proof. See Appendix B.2 for a proof. □

These results are analyzed via standard technique and are comparable to the regret of the classic multi-armed bandit problem, with some added constants factors for the BIC constraints [6]. Although the regret of our algorithm is the same as the fully incentive-compatible algorithm for both types, our algorithm can finish in a more timely manner and has smaller prior-dependent constants in the asymptotic bound.

From the analysis in lemma 4.6, we can derive the type-specific regret for each type of agent. Since our algorithm does not guarantee that all agents follow our recommendation, this regret analysis is divided into two cases, depending on the best arm overall. If we assume that proportions of each type in the population are equal, lemma 4.8 below demonstrates the type-specific regret guarantee of our algorithm.

Lemma 4.8 (Type-specific regret)**.** *Algorithms 1 and 2 with parameters $\ell_1, \rho \in \mathbb{N}$ achieves type-specific regrets in table 1.*

Table 1: Type-Specific Regret for Two Arms & Two Types with partially BIC Algorithm

Best arm / Type	Type-specific regret
Arm 0 / Type 0	$O\left(\frac{\log T}{(\tau\, \mathbb{P}_{\mathcal{P}_0}[\theta \geq \tau] - \nu)^2}\right) + O(\sqrt{T \log T})$
Arm 0 / Type 1	$O\left(\frac{\log T}{(\tau\, \mathbb{P}_{\mathcal{P}_0}[\theta \geq \tau] - \nu)^2}\right) + O\left(\frac{\log T}{(\tau\, \mathbb{P}_{\mathcal{P}_1}[\theta < -\tau] - \nu)^2}\right) + O(\sqrt{T \log T})$
Arm 1 / Type 0	$O\left(\frac{\log T}{(\tau\, \mathbb{P}_{\mathcal{P}_0}[\theta \geq \tau] - \nu)^2}\right) + O(\sqrt{T \log T})$
Arm 1 / Type 1	$O(\sqrt{T \log T})$

Proof. See Appendix B.3 for a proof. □

5 Full results

For simplicity, we showcased our how we apply our algorithm with recommendations as instrumental variables in a binary treatment setting with only two types of agents. However, our results apply to general settings with many treatments and many types. Please see our full paper for these fully general results.

References

[1] Joshua Angrist and Alan Krueger. Instrumental variables and the search for identification: From supply and demand to natural experiments. *Journal of Economic Perspectives*, 15(4):69–85, 2001.

[2] Joshua D. Angrist and Guido W. Imbens. Two-stage least squares estimation of average causal effects in models with variable treatment intensity. *Journal of the American Statistical Association*, 90(430):431–442, 1995. doi: 10.1080/01621459.1995.10476535. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1995.10476535.

[3] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47:235–256, 2002. doi: 10.1023/A:1013689704352.

[4] E. Bareinboim, A. Forney, and J. Pearl. Bandits with unobserved confounders: A causal approach. In *Advances in Neural Information Processing Systems,*, page 1342–1350, 2015.

[5] Lilian Besson and Emilie Kaufmann. What doubling tricks can and can't do for multi-armed bandits, 2018.

[6] Eyal Even-Dar, Shie Mannor, and Yishay Mansour. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of Machine Learning Research (JMLR)*, 7:1079–1105, 2006.

[7] Guido Imbens, Joshua Angrist, and Donald Rubin. Identification of causal effects using instrumental variables. *Journal of Econometrics*, 71(1-2):145–160, 1996.

[8] Nicole Immorlica, Jieming Mao, Aleksandrs Slivkins, and Zhiwei Steven Wu. Bayesian exploration with heterogeneous agents. In *The World Wide Web Conference*, WWW '19, page 751–761, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450366748. doi: 10.1145/3308558.3313649. URL https://doi.org/10.1145/3308558.3313649.

[9] Nathan Kallus. Instrument-armed bandits. *ArXiv*, abs/1705.07377, 2018.

[10] M. Katehakis and A. F. Veinott. The multi-armed bandit problem: Decomposition and computation. *Math. Oper. Res.*, 12:262–268, 1987.

[11] Ilan Kremer, Yishay Mansour, and Motty Perry. Implementing the "wisdom of the crowd". In Michael J. Kearns, R. Preston McAfee, and Éva Tardos, editors, *Proceedings of the fourteenth ACM Conference on Electronic Commerce, EC 2013, Philadelphia, PA, USA, June 16-20, 2013*, pages 605–606. ACM, 2013. doi: 10.1145/2492002.2482542. URL https://doi.org/10.1145/2492002.2482542.

[12] Yishay Mansour, Aleksandrs Slivkins, and Vasilis Syrgkanis. Bayesian incentive-compatible bandit exploration. *In 15th ACM Conf. on Economics and Computation (ACM EC)*, 2015.

[13] Yishay Mansour, Aleksandrs Slivkins, Vasilis Syrgkanis, and Zhiwei Steven Wu. Bayesian exploration: Incentivizing exploration in bayesian games. In Vincent Conitzer, Dirk Bergemann, and Yiling Chen, editors, *Proceedings of the 2016 ACM Conference on Economics and Computation, EC '16, Maastricht, The Netherlands, July 24-28, 2016*, page 661. ACM, 2016. doi: 10.1145/2940716.2940755. URL https://doi.org/10.1145/2940716.2940755.

[14] Mark Sellke and Aleksandrs Slivkins. Sample complexity of incentivized exploration. *CoRR*, abs/2002.00558, 2020. URL https://arxiv.org/abs/2002.00558.

[15] Aleksandrs Slivkins. *Introduction to Multi-Armed Bandits*. Foundations and Trends in Machine Learning, 2019.