Incentivizing Bandit Exploration: Recommendations as Instruments

Daniel Ngo Logan Stapleton Nicole Immorlica

Vasilis Syrgkanis Zhiwei Steven Wu

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 incentivize them to explore potentially sub-optimal arms. Under this setting, we provide a novel recommendation mechanism that views the planner's recommendations as a form of instrumental variables (IV) that only affect 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 mean rewards of the actions and enables the social learning process to minimize regret over the long term.

1 Introduction

In many online recommendation systems, including 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 the 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 incentives, a long line of work started by Kremer et al. [2013], Mansour et al. [2015] has studied mechanisms that incentivize users to explore by leveraging the information asymmetry between the recommendation system and users [Mansour et al., 2016, Immorlica et al., 2019, Sellke and Slivkins, 2020]. These papers consider a simple 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 alternatives (or called actions or arms) and receive a reward for their choice. Upon the arrival of each agent, the social planner provides a recommendation (one of the alternatives) that influences the agent's selection. The problem is to design a recommendation policy that incentivizes the agents to perform exploration and in the long run maximize the cumulative rewards of all the agents.

Prior work on "incentivizing exploration" typically approaches the problem by enforcing *Bayesian incentive-compatibility* (BIC) for every agent—that is, it is in each agent's interest to follow the planner's recommendation even if such action is inferior according to their prior belief. To achieve

Preprint. Under review.

BIC, these mechanisms typically need to assume a strong assumption that all of the agents and the planner share a *common prior* over the rewards of actions. However, in reality, agents tend to be inherently heterogeneous in terms of their beliefs and perceived rewards. In particular, some agents can have a stronger bias that favors one action over others and different agents. Moreover, even if the actions have the same effects on all of the 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.

To explicitly capture the underlying heterogeneity across agents, we study a new model in which each agent has unobserved private type that determines both their prior preferences across the actions and their action-independent confounding shift in the rewards. Since the heteroneneity of the agents confounds the observed rewards, we draw techniques from instrumental variable regression to estimate the reward effects across the actions. In particular, we take a novel view that the planner's recommendations can serve as a form of instrumental variables (IV) that only affect 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 recommendation. By running IV regression, our mechanism obtains a reliable estimate of reward effect of each action and in turn incentivize the agents' selections to converge on the optimal action.

A major advantage of our IV-based mechanism is that we no longer need to enforce BIC for all agents. Our algorithm can estimate the action effects accurately, as long as the recommendations induce variability in the agents' selections. An important special case is when a fraction of the types are more willing to explore an action that a-priori inferior. Furthermore, as soon as the mechanism obtains a sufficiently accurate estimate on the action effects, the policy can then be BIC for the remaining types. We demonstrate such partially BIC policy can significantly improve the regret bounds of a fully BIC policy obtained by extending the algorithm of Mansour et al. [2015] to our setting.

Organization. In this workshop version, we will present our formal model (Sec.2) and illustrate key ideas of our mechanism in a simple setting of two private types and two arms (Sec. 3). We defer the fully general results and details to the full version.

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. There is a set $\mathcal{X}$ of k possible actions for each agent to choose from. In each round t, an entirely new agent indexed by t arrives with their *private type* u_t drawn independently from a distribution $\mathcal{U}$. Each agent t receives an action recommendation $z_t \in \mathcal{X}$ from the planner and then selects an action X_t in $\mathcal{X}$ and receives a reward $y_t \in \mathbb{R}$.

Reward Given the choice of $X_t \in \mathcal{X}$, the reward y_t for each agent t is given by

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

where θ^a denotes the *exogenous mean rewards* for each action $a \in \mathcal{X}$, $g(u_t)$ is a *confounding term* that depends on the agent's private type u_t, and ε_t denotes the random 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 treatment reward $\theta^{X_t} \in [0, 1]$ and the confounding term $|g(u_t)| \leq \Upsilon$ for some constant Υ. Both $g(u_t)$ and ε_t are unobserved. The social planner's objective is to maximize the total reward of all agents over all T rounds. The agents are self-interested individuals whose objective is to maximize their own expected reward.

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 $\pi_t \colon (\mathcal{X} \times \mathcal{X} \times \mathbb{R})^{t-1} \to \Delta(\mathcal{X})$ 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 choice 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 mean reward vector θ, the agents' private types, and their reward noise. In round t, the action

X_t chosen is given by a selection function $f(u_t, z_t, t) = X_t$. Each agent chooses an action X_t that maximizes their Bayesian expected reward conditional on the recommendation z_t. We can define how action X_t is chosen by the selection function f as such:

$$X_t = f(u_t, z_t, t) := \operatorname*{argmax}_a \mathop{\mathbb{E}}_{\theta^a \sim \mathcal{P}_t} [\theta^a \mid z_t, t] \tag{2}$$

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

Regret We consider the following notions of regret that measure the ability of the algorithm to maximize the expected total reward of all agents. For the following definitions, let $\theta^{a^*} = \max_a \theta^a$. The regret of the mechanism is

$$R_\theta(T) = T\theta^{a^*} - \mathbb{E}\left[\sum_{t=1}^{T} \theta_t^{X_t} \big| \theta\right]. \tag{3}$$

The regret accumulated by agents of type π is then defined as

$$R_\theta^\pi(T) = \mathbb{E}\left[\sum_{t=1}^{T} \left(\theta^{a^*} - \theta_t^{X_t}\right) \mathbb{1}[u_t = \pi] \big| \theta\right]. \tag{4}$$

2.1 Recommendations as Instruments

We model the reward y_t and action X_t in the presence of confounding variables that depend on the type u_t. In a standard bandits setting, without confounding variables, one would find θ^{a^*} by taking average 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 does not work. Instead, we will use *instrumental variable (IV) regression* to estimate the exogenous mean reward vector 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 estimate of $\hat{\theta}_n$. To derive the estimator, we will rewrite the selection and reward functions in (1) and (2) as a linear relationship between X_i, y_i and z_i. We denote each action a in $\mathcal{X}$ with a standard basis vector $\mathbf{e}_a \in \{0,1\}^k$ with 1 at the a-th coordinate. Then

$$y_i = \langle \theta, X_i \rangle + g(u_i) + \varepsilon_i \tag{5}$$

$$X_i = \Gamma^\intercal z_i + \eta_i \tag{6}$$

where $\eta_i = \mathbb{E}_{\mathcal{U}}[X \mid z_i] - X_i$ and Γ is $k \times k$ *compliance matrix* such that each ab-th entry given by:

$$(\Gamma)_{ab} = \mathop{\mathbb{P}}_{\mathcal{U}}[X_t = \mathbf{e}_b | z_t = \mathbf{e}_a] \tag{7}$$

Two-Stage IV estimator. We estimate θ via a generalization of *Two-Stage Least Squares regression*. We first form an empirical estimate of Γ, denoted $\hat{\Gamma}_n$, wherein the ab-th entry is given by:

$$\left(\hat{\Gamma}_n\right)_{ab} = \frac{1}{n_a} \sum_{i=1}^{n} \mathbb{1}[X_i = \mathbf{e}_b \wedge z_i = \mathbf{e}_a] \tag{8}$$

where n_a denotes the number of times that arm a is recommended: $n_a = \sum_{i=1}^n \mathbb{1}[z_i = \mathbf{e}_a]$.

In the second stage, let $\beta = \hat{\Gamma}_n \theta$ and $\beta^* = (\beta, \mathbb{E}[g(u)])$ be the augmented vector created by adding a coordinate of $g(u_t)$ to β and let $\tilde{z}_i = (z_i, 1)$ be the augmented vector created by adding a coordinate of 1 to z_i. Note that we have $\mathbb{E}[y_i \mid z_i] = \langle \beta^*, \tilde{z}_i \rangle$, and so we can form a ridge regression estimate of β^*, denoted $\tilde{\beta}$:

$$\tilde{\beta} = \left(\sum_{i=1}^{n} \tilde{z}_i \tilde{z}_i^\intercal + \lambda \mathbf{I}\right)^{-1} \left(\sum_{i=1}^{n} \tilde{z}_i y_i\right)$$

Let $\hat{\beta}_n$ be such that $\tilde{\beta} = (\hat{\beta}_n, \hat{\mathbb{E}}[g(u)])$ where $\hat{\mathbb{E}}[g(u)]$ is the estimate of $\mathbb{E}[g(u)]$ and $\hat{\beta}_n$ is the estimate of β. Finally, our IV estimate is given by

$$\hat{\theta}_n = (\hat{\Gamma}_n)^{-1}\hat{\beta}_n.$$

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

Theorem 2.1. *Given n observations $\{(z_i, X_i, y_i)\}_{i=1}^n$ generated from (5) and (6), where each u_i is drawn independently from a private type distribution $\mathcal{U}$. Let $n_{\min} := \min_j n_j = \min_j \sum_{i=1}^n \mathbb{1}[z_i = \mathbf{e}_j]$ be the minimum number of times any arm is recommended and $\sigma_{\min}\{\hat{\Gamma}_n\}$ be the minimum singular value of the empirical compliance matrix $\hat{\Gamma}_n$. Suppose $n_{\min} > 0$, then with probability at least $1-\delta$:*

$$\left\|\hat{\theta}_n - \theta\right\|_2 \leq \frac{\sqrt{\lambda(k+\Upsilon^2)} + 2\sqrt{2\log\left(\frac{2}{\delta}\right) + (k+1)\log\left(1 + \frac{2n}{\lambda(k+1)}\right)}}{\sigma_{\min}\{\hat{\Gamma}_n\}\sqrt{n_{\min}}}.$$

Remark 2.2 (Lower bound on $\sigma_{\min}(\hat{\Gamma}_n)$)**.** *Suppose among the n observations $\hat{p}$ is the proportion of agents that follow our recommendations. Then with probability at least $1-\delta$, the minimum singular value of the empirical compliance matrix satisfies $\sigma_{\min}(\hat{\Gamma}_n) \geq \hat{p}^2/2$.*

3 Warmup: two arms and two types

In this section, we provide a concrete example to demonstrate how our algorithm works. We consider a simple setting which entails only two arms and two types. Each type has a different Bayesian prior over the rewards of these two arms, but both types agree on which arm is better than the other.

At a high level, our algorithm follows the structure of the BIC mechanism of Mansour et al. [2015] with two key modifications. In particular, our algorithm is no longer required to be BIC for all agents. As a result, it can obtain a better regret bound whenever one of the types is more compliant. Also, our mechanism estimates the action effects θ with IV regression. Our algorithm runs in two stages: first, the *sampling stage* and then the *racing stage*. The sampling stage sub-algorithm collects L_1 samples of each arm. These L_1 samples are then input as a parameter into the racing stage algorithm, which recommends each arm one-by-one until one arm "wins." We design the algorithm to incentivize one type to be compliant throughout. We are able to do so because we employ an instrumental variable (IV) regression in order to estimate the true treatment effect of each arm. Because we require only partial incentive-compatibility, our algorithm achieves better regret than a fully BIC algorithm similar to that of the detail-free algorithm in Mansour et al. [2015] would.

This section is organized as follows: Section 3.1 lays out the background assumptions for this two-arm two-type setting. Section 3.2 outlines assumptions needed for the sampling stage algorithm, demonstrates said algorithm in Algorithm 1, and proves that it is BIC for one type in Lemma 3.2. Section 3.3 also notes assumptions needed for the racing stage, demonstrates the racing stage algorithm (Algorithm 2), and proves incentive-compatibility first for one type, then later for both. Finally, Section 3.4 states the overall expected regret and the type-specific regrets for our algorithm in Lemmas 3.5, 3.6, and 3.7. In Remark 3.9, we demonstrate the advantage of our algorithm by comparing the type-specific regret of our algorithm to that of a fully BIC algorithm.

3.1 Setting Assumptions

Before we present the algorithm, we present our background assumptions.

Assumption 3.1. *The entire population of agents has two types, i.e. type 1 and type 2. The proportions of agents of either type in the population are equal. Each type i has a Bayesian prior, namely $\mathcal{P}_i$, over the distribution of the mean reward vector θ. For each type i, the prior mean reward of arm a is given by $\mu_i^a := \mathbb{E}_{\mathcal{P}_i}[\theta^a]$. We assume that the arms are ordered according to their prior mean rewards and that this ordering is shared across types, i.e. $\mu_i^1 > \mu_i^2$ for either type $i \in \{1, 2\}$.*

If agents were left to their own devices, we would only see samples of arm 1, since all agents prefer arm 1 to arm 2. Thus, we need to incentivize agents to choose arm 2 via our recommendations.

3.2 Sampling Stage Algorithm for Two Arms and Two Types

Following the two-arm fully BIC sampling stage algorithm in Mansour et al. [2015], Algorithm 1 capitalizes on an information asymmetry between the agents and the social planner in order to incentivize the agents to comply to the its recommendations. However, while the fully BIC sampling stage is BIC for all agents at all rounds, Algorithm 1 is designed to be BIC for agents of a given type i.[1] In our *partially BIC* algorithm, we implement Algorithm 1 to be BIC for agents of type 2, because these agents are easier to convince to be BIC under our recommendation policy than are agents of type 1. In Section 3.4, we demonstrate how implementing a partially BIC algorithm which is BIC for only agents of a single type allows for improvements in regret over a fully-BIC algorithm which is BIC for agents of all types.

In the first phase of Algorithm 1, we recommend only arm 1: this recommendation confers no information about the history and each agent chooses the better arm according to their prior, i.e. arm 1. Thus, we observe the rewards from L_1 samples of arm 1. For a given type i, there is a non-zero chance that arm 1 performs so poorly that the prior of arm 2 looks better than the posterior of arm 1 conditioned on these L_1 samples. For type i, define this event as

$$\xi_i = \left\{ \frac{1}{L_1} \sum_{t=1}^{L_1} y_t^1 + \frac{1}{2} + \Upsilon + \sigma_\varepsilon \sqrt{\frac{2\log(1/\delta)}{L_1}} < \mu_i^2 \right\}, \tag{9}$$

where δ is a sampling failure probability to be determined.

Starting in the second phase of the sampling stage, we recommend either an arm to exploit or an arm to explore. If event ξ_i occurs, the algorithm always recommends arm 2 to *exploit*. The algorithm also *explores* by recommending arm 2 to L_1 agents (out of ρL_1 in total, where $1/\rho$ represents the exploration probability). Thus, when any agent receives a recommendation for arm 2, there is a chance that it is because ξ_i occurred and there is a chance that it is because the algorithm chose them to explore on. If the exploration probability is small enough (i.e. ρ is large enough), then agents of type i are always inclined to follow a recommendation for arm 2.

We order the two types such that agents of type 1 require a smaller exploration probability than agents of type 2. As such, we implement Algorithm 1 to be BIC for only agents of type 2.

We use a set of assumptions to limit the prior to ensure that our algorithm has a chance to recommend the weaker arm and that our estimates are close to the true treatment effects. Hence, we restrict the priors to be independent across arms, with bounded treatment effects and confounder.

Algorithm 1: The sampling stage for two arms with two heterogeneous types

```
Input: parameters $\rho, L_1 \in \mathbb{N}$
In the first $L_1$ rounds, let the agents pick their preferred arm (arm 1). The sample average of reward
  $y^1$ after pulling arm 1 $L_1$ times is $\bar{y}^1_{L_1}$.
if $\mu_2^2 > \bar{y}^1_{L_1} + \frac{1}{2} + \Upsilon + \sigma_\varepsilon\sqrt{\frac{2\log(1/\delta)}{L_1}}$ then
    $a^* = 2$
else
    $a^* = 1$
end
From the set $P$ of the next $L_1.\rho$ 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 2.
```

We prove that the algorithm 1 is BIC for agents of type i as long as the number of phases ρ is larger than some prior-dependent constant. The intuition is when an agent of type i is recommended arm 2, they do not know whether this is due to exploration or exploitation. However, with large enough number of phases ρ, the exploration probability $1/\rho$ is low enough that the expected gain from exploiting exceed the expected loss from exploring and the agent would take arm 2.

[1] Mansour et al. [2015] does not explicitly consider agents of heterogeneous types. However, we consider an extension of their algorithm in our setting as an algorithm that is BIC for agents of all types at all rounds.

Lemma 3.2 (Two Arm Sampling Stage BIC for Type i)**.** *Algorithm 1 with parameters* (ρ, L_1) *completes in* $L_1\rho + L_1$ *rounds. Algorithm 1 is BIC for all agents of type* i *if we hold Assumption 3.1 above and the parameters satisfy:*

$$\rho \geq 1 + \frac{4(\mu_2^1 - \mu_2^2)}{\mathbb{P}_{\mathcal{P}_i}[\xi_i]}. \tag{10}$$

3.3 Racing Stage Algorithm for Two Arms and Two Types

From the sampling stage, we get a set of observations for $L_1\rho + L_1$ rounds. Because, as shown in Lemma 3.2, agents of type 2 will always follow our recommendation, this allows us to induce variability among the action choices of type 2 agents and get samples of arm 2. This allows the empirical compliance matrix $\hat{\Gamma}$ to have all positive singular values. Thus, we can use our recommendations as instruments and regress over them (using IV regression) to estimate the exogenous mean rewards $\hat{\theta}^1$ and $\hat{\theta}^2$.

In the racing stage (Algorithm 2), each arm will be recommended one-by-one until one of them can be determined to be optimal with high confidence. After that, the racing stage has ended the "winner" arm is recommended for the rest of the time horizon. Thus, when an agent gets a recommendation for arm 2, they are not sure whether they are receiving that recommendation because they are still in the racing stage or because it has ended and arm 2 has won. We will leverage this uncertainty to incentivize them to follow the recommendations.

By collecting more samples, we are able to form good enough estimates $\hat{\theta}^1$ and $\hat{\theta}^2$ and increase the likelihood that the racing stage has ended when an agent receives a recommendation for arm 2. Thus, we collect a sufficient number of samples in the sampling stage in order to make the racing stage algorithm BIC for agents of type 2. By similar reasoning as the recommendation in the sampling stage, we can treat the recommendation in the racing stage as an instrument because agents of type 2. We use all L_1 samples from the sampling stage and any additional samples in the racing stage in order to estimate the mean rewards $\hat{\theta}^1$ and $\hat{\theta}^2$.

Similar to the sampling stage, we also rely on agents of type 2 to take our arm 2 recommendation in the beginning of the racing stage. Hence, to account for the probability of agent of type 2 getting an arm 2 recommendation, we divide the racing stage into phases of $2h$ rounds each, where each arm is recommended h times in a phase.

After each phase, we check if the empirical gap $|\hat{\theta}^1 - \hat{\theta}^2|$ is larger than some threshold. If that happens, the arm with the higher estimate is determined with high probability to be the best arm, and all agents are only recommended this arm from now on. To ensure that the winner of the racing stage is indeed the best arm with high probability, we need to pick a threshold large enough, similar to the termination condition in the Active Arm Elimination algorithm from Even-Dar et al. [2006].

We also divide the racing stage into two smaller sub-algorithms. The first part of the racing stage is partially BIC, where we guarantee that only agents of type 2 will comply with our recommendations. After collecting sufficient samples of arm 2, if the algorithm has not found the "winner" yet, we can proceed to the second part of the racing stage. At this point, we have collected enough samples to also convince agents of type 1 to follow our recommendations. By having two consecutive racing stages, we can potentially reduce the regret of our algorithm and achieve a fully BIC algorithm, where every agent follows our recommendations.

We prove that the Algorithm 2 is BIC for all agents of type 2 as long as we have collected sufficient samples of arm 2 in the sampling stage and our estimate of θ is close to the true treatment effect. Due to information asymmetry, when an agent is recommended arm 2, they do not know whether the two arms are still racing each other or if the best arm has been found. However, with a small enough threshold, the expected gain from exploiting the best arm exceeds the expected loss from exploring a random arm and the agents of type 2 would follow our recommendation.

Algorithm 2: The racing stage for two arms with two heterogeneous types

```
Input: parameters $L_1 \in \mathbb{N}$; time horizon $T$; probability $\delta$, number of recommendations $h$ for a
    single arm in a single phase
Input: IV estimates of mean reward $\theta^1$ and $\theta^2$, denoted $\hat{\theta}^1_{L_1}$ and $\hat{\theta}^2_{L_1}$
Let $\hat{\theta}^1_{L_1}$ and $\hat{\theta}^2_{L_1}$ be the IV estimate of the mean reward $\theta^1$ and $\theta^2$ after $L_1\rho$ rounds of Algorithm 1;
Split remainder into consecutive phases of $2h$ rounds each, starting from phase $q = L_1$;
while $L_1 < \frac{589824(\log(3T/\delta)+3\log(2T(2+\Upsilon^2)))}{\tau^2(\mathbb{P}_{\mathcal{P}_1}[G\geq\tau]^2)}$ do
    while $|\hat{\theta}^1_q - \hat{\theta}^2_q| \leq 192\sqrt{\frac{\log(3T/\delta)+3\log(2T(2+\Upsilon^2))}{q}}$ do
        The next $2h$ agents are recommended both arms sequentially;
        For each arm $a \in \{1,2\}$, let $y^a_t$ be one sample reward of that arm in this phase;
        Let $\hat{\theta}^a_q$ is the IV estimate of the reward of each arm $a \in \{1,2\}$ for the racing stage up to and
          including phase $q$;
        $q = q + 1$;
    end
    For all remaining agents recommend $a^* = \mathrm{argmax}_{a\in\{1,2\}}\,\hat{\theta}^a_q$ and end the algorithm.
end
while $|\hat{\theta}^1_q - \hat{\theta}^2_q| \leq 192\sqrt{\frac{\log(3T/\delta)+3\log(2T(2+\Upsilon^2))}{q}}$ do
    The next $2h$ agents are recommended both arms sequentially;
    For each arm $a \in \{1,2\}$, let $y^a_q$ be one sample reward of that arm in this phase;
    Let $\hat{\theta}^a_q$ is the IV estimate of the reward of each arm $a \in \{1,2\}$ for the racing stage up to and
      including phase $q$;
    $q = q + 1$;
end
For all remaining agents recommend $a^* = \mathrm{argmax}_{a\in\{1,2\}}\,\hat{\theta}^a_q$
```

Lemma 3.3 (Type 2 BIC)**.** *Let $G := \theta^2 - \theta^1$. Fix an absolute constant $\tau \in (0,1)$ and let*

$$\delta_\tau = \frac{1}{4}\frac{\tau\,\mathbb{P}_{\mathcal{P}_2}[G\geq\tau]}{\tau\,\mathbb{P}_{\mathcal{P}_2}[G\geq\tau]+1}.$$

Algorithm 2 is BIC for all agents of type 2 if the parameters satisfy $\delta \leq \delta_\tau$ and

$$h \geq -\frac{\log\left(\frac{3\tau\,\mathbb{P}_{\mathcal{P}_2}[G\geq\tau]+4}{4\tau\,\mathbb{P}_{\mathcal{P}_2}[G\geq\tau]+4}\right)}{\log(2)}$$

and

$$L_1 \geq \frac{589824(\log(3T/\delta)+3\log(2T(2+\Upsilon^2)))}{\tau^2(\mathbb{P}_{\mathcal{P}_2}[G\geq\tau]^2)}.$$

During the first racing stage, we collect more samples of arm 2 through agents of type 2. After getting sufficient samples of arm 2, we can also convince agents of type 1 to follow our recommendations for arm 2. The following Lemma 3.4, which follows the same structure as that of Lemma 3.3, proves that the second racing stage algorithm is BIC for both agents of type 1 and type 2.

Lemma 3.4 (Type 1 BIC)**.** *Assume two arms. Let $G := \theta^2 - \theta^1$. Fix an absolute constant $\tau \in (0,1)$ and let*

$$\delta_\tau = \frac{1}{2}\frac{\tau\,\mathbb{P}_{\mathcal{P}_1}[G\geq\tau]}{\tau\,\mathbb{P}_{\mathcal{P}_1}[G\geq\tau]+1}.$$

Algorithm 2 is BIC for type 1 in the second part of the racing stage if the parameters satisfy $\delta \leq \delta_\tau$ and

$$L_1 \geq \frac{589824(\log(3T/\delta)+3\log(2T(2+\Upsilon^2)))}{\tau^2(\mathbb{P}_{\mathcal{P}_1}[G\geq\tau]^2)}.$$

3.4 Regret Analysis

With the number of samples of each arm collected in the sampling stage L_1, the number of sampling stage phases ρ and the accuracy guarantee of our estimate δ, our sampling stage and racing stage algorithms achieve sub-linear regret for a particular instance of the treatment effect vector 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 vector θ.

Lemma 3.5 (Regret)**.** *Algorithm 1 and 2 with parameters $L_1, \rho \in \mathbb{N}$ and $\delta \in (0, 1)$ achieves ex-post regret:*

$$R(T) \leq L_1(\rho + 1) + O(\sqrt{T \log(T/\delta)}) \tag{11}$$

where $\Delta = |\theta^1 - \theta^2|$, L_1 is the duration of one phase in the initial sampling stage and $\frac{1}{\rho}$ is the exploration probability in the sampling stage.

With the number of samples of each arm collected in the sampling stage L_1, the number of sampling stage phases ρ, and the accuracy guarantee of our estimate δ, our sampling stage and racing stage algorithms achieve sub-linear Bayesian regret over the randomness in the priors of the agents. Lemma 3.6 provides a basic performance guarantee of our algorithm.

Lemma 3.6 (Expected Regret)**.** *Algorithm with parameters $L_1.\rho \in \mathbb{N}$ and $\delta \in (0, 1)$ achieves expected ex-post regret.*

$$\mathbb{E}[R(T)] = O\left(\sqrt{T \log(T)}\right) \tag{12}$$

where T is the time horizon.

From the regret analyzed in Lemma 3.5, we can derive the type-specific regret of each type of agent. Since our algorithm does not guarantee that all agents follow our recommendations in the sampling stage and the first racing stage, the type-specific regret is divided into two cases, depending on the best arm overall. In addition to the regret in Lemma 3.5, this regret also depends on the proportion of each type of agents in the population, which consider to be equal between the two types. Lemma 3.7 below demonstrates our type-specific regrets guarantee.

Lemma 3.7 (Type-specific regret of a partially BIC algorithm)**.** *The following type-specific expected regret bounds in Table 1 hold.*

Best arm/Type	Type-specific Regret
Arm 1: Type 1	$O(\log(T/\delta))$
Arm 1: Type 2	$O\left(\frac{\log(T/\delta)}{\tau^2\,\mathbb{P}_{\mathcal{P}_2}[G \geq \tau]^2}\right) + O(\log(T/\delta))$
Arm 2: Type 1	$O\left(\frac{\log(T/\delta)}{\tau^2\,\mathbb{P}_{\mathcal{P}_2}[G \geq \tau]^2}\right) + O(\sqrt{T \log(T/\delta)})$
Arm 2: Type 2	$O\left(\frac{\log(T/\delta)}{\tau^2\,\mathbb{P}_{\mathcal{P}_2}[G \geq \tau]^2}\right) + O(\sqrt{T \log(T/\delta)})$

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

Definition 3.8. [Fully BIC Algorithm] A *fully BIC* algorithm is BIC for all agents at all rounds, similar to the detail-free algorithm in Mansour et al. [2015]. We implement Sub-algorithm (1) to be BIC for agents of type 1 (which will also be BIC for agents of type 2). Then, we run Sub-algorithm (2) starting from the second racing stage, wherein recommendations are BIC for all agents.

Remark 3.9 (Regret Comparison Between Fully BIC and Partially BIC Algorithms)**.** *The type-specific regret for the fully BIC algorithm (Definition 3.8) is the same as the type-specific regret for our partially BIC algorithm (see Table 1) wherein we replace the prior-dependent constant term $\tau^2\,\mathbb{P}_{\mathcal{P}_2}[G \geq \tau]^2$ for type 2 with the prior-dependent constant term $\tau^2\,\mathbb{P}_{\mathcal{P}_1}[G \geq \tau]^2$ for type 1, which can be arbitrarily large. In the full version of our paper, we work out a case wherein the priors over the treatment reward vector θ for each type are Gaussian and the fully BIC algorithm (Definition 3.8) achieves linear regret $\Omega(T)$, but our algorithm achieves sub-linear regret $O(\sqrt{T \log(T)})$.*

References

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.

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.

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.

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

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.

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